Università degli Studi di Firenze

DET
Dipartimento di Elettronica
e Telecomunicazioni

Dipartimento di Fisica

3rd
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

December 10-12, 2003
Firenze, Italy

PROCEEDINGS

Firenze University Press

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

3rd INTERNATIONAL WORKSHOP

December 10-12, 2003
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2003

Models and analysis of vocal emissions for biomedical applications : 3rd international workshop: december 10-12, 2003 : Firenze, Italy / edited by Claudia Manfredi. – Firenze : Firenze university press, 2003.
http://digital.casalini.it/8884531543
Stampa a richiesta disponibile su http://epress.unifi.it

ISBN 88-8453-154-3 (online)
ISBN 88-8453-155-1 (print)
612.78 (ed. 20)
Voce - Patologia medica

Cover designed by: CdC, Firenze, Italy

Università degli Studi di Firenze
Firenze University Press
Borgo Albizi 28, 50122 Firenze, Italy
http://epress.unifi.it/

Printed in Italy

MAVEBA 2003

INTERNATIONAL PROGRAM COMMITTEE

S. Aguilera (ES)
S. Cano Ortiz (CU)
U. Eysholdt (D)
A. Izworski (PL)
C. Larson (USA)
C. Marchesi (I)
X. Pelorson (F)
O. Schindler (I)
J. Sundberg (SE)
U. Uergens (D)
A. Barney (UK)
R. Carlson (SE)
O. Fujimura (JP)
M. Kob (D)
F. Locchi (I)
W. Mende (D)
P. Perrier (F)
A. Schuck (BR)
J. Svec (CZ)
G.Valli (I)
D. Berckmans (BE)
M. Clements (USA)
H. Herzel (D)
A. Krot (BY)
J. Lucero (BR)
V. Misun (CZ)
R. Ritchings (UK)
R. Shiavi (USA)
R. Tadeusiewicz (PL)
K. Wermke (D)
P. Bruscaglioni (I)
J. Doorn (AR)
D.Howard (UK)
U. Laine (FI)
C. Manfredi (I)
C. Moore (UK)
S. Ruffo (I)
H. Shutte (NL)
I. Titze (USA)
W. Ziegler (D)

LOCAL ORGANISING COMMITTEE

C. Manfredi, Dept. of Electronics and Telecommunications, Faculty of Engineering - Conference Chair
P. Bruscaglioni, Dept. of Physics, Faculty of Maths, Physics and Natural Sciences - Conference Chair
F. Locchi, Dept. of Clinical Physiology, Faculty of Medicine.
C. Marchesi, Dept. of Systems and Computer Science, Faculty of Engineering.
S. Ruffo, Dept. of Energetics, Faculty of Engineering.
G.Valli, Dept. of Electronics and Telecommunications, Faculty of Engineering

SPONSORS

MPS - Banca Monte de' Paschi di Siena

ISCA - International Speech and Communication Association

AIIMB - Associazione Italiana Ingegneria Medica e Biologica

INFM – Istituto Nazionale per la Fisica della Materia

CONTENTS

Poster session

Plenary lecture

Mechanical models

Pathology classification

Special session on Singing voice

Devices

Voice/hearing impairment

Numerical models

Pathology detection

Voice analysis

FOREWARD

On behalf of the organising committee, we would like to welcome all the participants to the 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA 2003, held 10-12 December 2003, Firenze, Italy.
The workshop is organised every two years, and aims to stimulate contacts between specialists active in research and industrial developments, in the area of voice analysis for biomedical applications. The Workshop aims at offering the participants an interdisciplinary platform for presenting and discussing new knowledge in the field of models and analysis of speech signals and images, both as far as adults and children voices are concerned.
The scope of the Workshop includes all aspects of voice modelling and analysis, ranging from fundamental research to all kinds of biomedical applications and related established and advanced technologies.
Some of the relevant topics are: linear and non-linear analysis and modelling of the normal and pathological voice source, for parameter definition and quantification, analysis of pathological voices, for diagnostic and classification purposes, enhancing voice quality during rehabilitation and after surgery, development of vocal prostheses and devices for impaired. Moreover, protocols and reliable objective parameters are among the Workshop topics. Strong focus of interest is in understanding the relationship between speech and neurological dysfunction (e.g. epilepsy, autism, schizophrenia, stress etc.) and the interaction with hearing impairment. Finally, singing voice analysis is also considered, with emphasis on pitch control for training purposes.

This third edition of the Workshop has gained great interest from the international scientific community, with more than 60 papers received, all of high scientific level, covering the most relevant fields of research in voice analysis. Moreover, two plenary lectures and two special sessions exploit specific themes: infant cry, singing voice, music and medicine, prosody.

We would like to thank the members of the organising committee and all the reviewers, who gave freely of their time to assess the highly disparate work of the workshop, helping in improving the quality of the papers.

We have also benefited from the efforts of the administrative staff within our University, office for Research and International Relations, and the Department of Electronics and Telecommunications, that devoted a lot of time and efforts to make this workshop a successful one. Special thanks to our University Orchestra and Chorus for their generous participation.

Finally, our gratitude goes to the supporters and sponsors, who contribute much to the success of the MAVEBA workshop.

Dott. Claudia Manfredi
Conference Chair

Prof. Piero Bruscaglioni
Conference Chair

Voice disorders

3rd International Workshop MAVEBA 2003, 5-8

AUTOMATIC ANALYSIS OF VOCAL MANIFESTATIONS OF APPARENT MOOD OR AFFECT

E. P. Rosenfeld[1], D. W. Massaro[2], J. Bernstein[1]
[1]Ordinate Corporation, Menlo Park, California, USA
[2]Department of Psychology, University of California at Santa Cruz, USA

Abstract: **Skilled clinicians are able to integrate linguistic, paralinguistic, and non-linguistic cues in the assessment of mood disorders. This project identified duration- and amplitude-based aspects of the speech signal that can be measured automatically by computer and which provide paralinguistic information about the apparent affect of a speech sample. A group of 40 experimental subjects produced 1584 spoken renditions of sentences, in 3 conditions, uninstructed, depressive, or manic. An automatic speech recognition system extracted 10 paralinguistic parameter values from each of these spoken responses. Psychotherapists have a relatively uniform model of depressive and manic speech patterns, which shows up in distinct paralinguistic features of their speech when simulating these states. Several features are significantly different in the three simulated emotional states and these features can be detected automatically.**
Keywords: automatic, speech recognition, mood, affect.

I. INTRODUCTION

Skilled clinicians are able to integrate linguistic and non-linguistic cues in the assessment of mood disorders. This ability is part of what makes a skilled clinical interview the preferred method of assessment for mood disorders. Among all the non-linguistic aspects of a patient's behavior, non-linguistic aspects of speech may be the easiest to record and analyze. These paralinguistic aspects of the manner of speaking can be collected unobtrusively and analyzed objectively. Previous research has identified stable patterns of acoustic indicators of mood and emotion [1-15]. Among many reported patterns, sad or depressive speech tends to be quieter, slower, with longer pauses, lower in pitch and more monotonous than normal speech.

This research project [16] identified certain duration- or amplitude-based aspects of the speech signal that can be measured completely automatically by computer and which provide paralinguistic information about the apparent affect of a speech sample. Specifically, the project identified measurable physical differences in speech signals that can be used to estimate how depressed or elated a person would sound to a panel of experienced clinicians.

The purpose of the project was the development and evaluation of techniques that may contribute to the measurement of affective states like depression. This project is an empirical study preliminary to building an integrated computerized instrument for administering structured interviews to patients, via the telephone, that can provide non-obtrusive, objective data that may improve assessment accuracy and validity. The project created a corpus of elicited speech and developed an automatic analysis of the recordings. The experiment reported here accomplished two preliminary objectives:

A. Replicate the reported relations between timing and amplitude of speech and perceived affect, for example, by [9, 11], but using fully automatic means;
B. Find and verify additional temporal manifestations of affect in speech signals.

The project focused on answering three main questions:

1) Which measurable paralinguistic characteristics of speech (e.g. response latency, speech rate, amplitude) can be reliably related to the simulated mood of a speaker?
2) Which of these characteristics can be derived automatically from the acoustic signals of spoken responses to test questions?
3) Which observed measures of paralinguistic variables show significant differences across speakers, and which show significant differences only within speakers?

II. METHODOLOGY

The data collection procedure followed a single session experimental design, wherein each speaking subject took a seven-minute speaking test by telephone, three times in succession, under three different conditions: once without instruction, once instructed to speak as if severely depressed, and once instructed to speak as if they were extremely manic (the order of the second and third conditions was counterbalanced). The seven-minute speaking test is a "sample" form of the PhonePass SET-10, a language test developed by Ordinate Corporation in California [17] to measure spoken English proficiency.

The experiment compared acoustic variables extracted from the speech samples corresponding to the uninstructed (or "normal") renditions, to the same variables from the speech samples that the speaking subjects intended to be simulations of depressive and manic speech.

Analysis of data from this experiment was intended to determine whether or not there were observable differences in the speech samples according to the speakers' intentions.

Subjects: The Speaking Subjects comprised 40 psychotherapists who were all native speakers of English, between the ages of 30 and 71; mean age was 53 years old. Of the 40 speaking subjects, 23 were women and 17 were men. Each speaking subject spent approximately 35 minutes of time in the experiment.

Instrumentation: The PhonePass SET-10 Sample Form comprises a set of 32 items administered in a 7-minute telephone call. Each item presents a recorded prompt over the telephone that solicits a spoken response from the subject that is recorded via telephone. The items used in this experiment form part of a single test form that prompts a subject to speak 32 times. Items of five different types are presented to the examinee: first, six one-sentence readings, then eight elicited repetitions of sentences, then eight opposite words, then eight short-answer questions, and, finally two open questions – each allowing the subjects thirty seconds to deliver their response. Most items elicit one-sentence responses or one-word responses that are about 0.5 to 5 seconds in duration.

Assuming that the average response length is six words and an average word has four phonemes, with 26 spoken responses measured per subject per condition, the data set potentially contains about 624 dependent measures per subject condition and more than 1800 dependent measures per speaking subject.

Ten dependent variables were measured:

TST: total speaking time (milliseconds)
TPT: total pause time (milliseconds)
TUT: total utterance time (milliseconds)
ROS: rate of speech (phonemes/second)
ART: articulation rate (phonemes/second)
LAT: response latency (milliseconds)
MPD: mean pause duration (milliseconds)
SDP: segment duration probability (log probability)
PDP: pause duration probability (log probability)
MaxSA: maximum speech amplitude (signal value)

III. Results

The results are presented numerically in Tables 1 and 2. Table 1 presents the data grouped across subjects, each cell showing the mean and standard deviation of each sample of 480 responses (12 selected responses x 40 subjects) per condition as measured on each of the 10 paralinguistic acoustic parameters under study. Table 2 presents the data organized by within-subject, within-item differences when the Speaker Subjects responded to the same item with two different intended moods.

The data as presented in Table 1 represent a comparison of groups of Speaking Subjects according to their instructed intentions. Table 1 presents measures that describe the central tendency and dispersion of the paralinguistic parameters of the responses when these Speaking Subjects talked in three different moods, as these parameters were automatically estimated by the speech recognition and signal processing internal to the PhonePass system.

Table1:
Mean, s.d. of Parameters for Intended Mood (N=480)

Param	D (=Depressed)		N (=Normal)		M (=Manic)	
	mean	s.d.	mean	s.d.	mean	s.d.
TST	2891.67	985.83	2556.67	787.10	2234.92	925.30
TPT	178.60	360.61	40.29	192.50	124.67	580.90
TUT	3070.27	1148.26	2596.96	836.97	2359.58	1276.64
ROS	9.72	1.97	11.37	1.68	13.06	2.72
ART	10.15	1.73	11.50	1.62	13.33	2.39
LAT	1360.21	759.0	656.79	287.60	533.58	498.38
MPD	20.12	41.24	4.57	20.43	12.51	59.16
SDP	-5.23	0.39	-4.90	0.29	-5.05	0.27
PDP	-2.63	0.93	-2.32	0.79	-2.20	0.82
MaxSA	6.62	4.24	9.96	4.74	15.35	8.23

The columns in Table 1 are ordered D – N – M (Depressed, Normal, Manic) in the expectation that the parameter values will generally be increasing or decreasing in that order. That is, from the literature, one would expect the Normal value of most of these parameters to be between the Depressed and the Manic value. This presumed ordering was observed for seven of the ten paralinguistic parameters in this study.

Table 2: Within-Subject Within-Item Paired Differences

Param	D-N (N = 375)		M-N (N = 373)		M-D (N = 386)	
	mean	s.d.	mean	s.d.	mean	s.d.
TST	344.80	595.62	332.44	730.24	655.80	838.19
TPT	141.07	399.59	56.94	413.48	76.14	517.90
TUT	485.87	809.41	275.50	989.47	731.94	1163.39
ROS	-1.75	2.14	1.66	2.53	3.32	2.69
ART	1.43	1.85	1.79	2.22	3.15	2.26
LAT	672.96	697.75	136.59	465.24	806.27	739.19
MPD	15.38	44.77	6.34	54.71	8.88	67.12
SDP	-0.34	0.43	-0.15	0.37	0.19	0.46
PDP	-0.33	1.09	0.03	0.94	0.40	1.15
MaxSA	3.41	4.05	5.35	6.72	9.00	7.44

Table 2 presents the data in a way that is more relevant to the ultimate question: how well would one expect an automatic system to detect changes in a known speaker's paralinguistic parameters under the instructions of this experiment. Table 2 presents paired differences. Each normal item response by each subject is a control on the measures for that item in the other two conditions. This way of treating the data should eliminate expected inter-subject and inter-item variance, yielding smaller standard errors of the mean, while the mean differences are approximately equal to the differences in the means for the various moods. This expected reduction in variance should promote rejection of the null (no-difference) hypotheses.

To test the significance in the differences in the mean parameter values, as shown in Table 1, across the population of speaking subjects and across the various sets of 12 items measured per call, a t-test for two population means with variances unknown and unequal [18] was used. The results indicate that 29 of the 30 observed differences in means are significantly different from zero ($t > 1.96$, $p = 0.05$), and even under the stricter criterion corrected for 10 simultaneous variables ($t > 2.81$), 26 of the 30 t-tests are still significant ($p < 0.0025$). All four of the comparisons that fail the stricter significance test, TPT (M-D), MPD (M-N, M-D), and PDP (M-N), are based in part on the measures of pause time in the manic experimental condition.

A simple and conservative test of the statistical significance of the differences between intended normal, depressive and manic speaking on the 10 paralinguistic acoustic parameters is a sign test [19]. The sign test assumes related samples, considered in pairs where members of the pairs can be ranked. The sign test does not assume that the data under study carry more than ordinal information, and it does not assume a normal distribution. The differences in 28 of the 30 possible comparisons are statistically significant ($z > 1.96$, $p = 0.05$). Only the manic-normal differences for TPT and MPD fail to reject the null hypothesis of no difference. If a 10-variable correction is accepted, and the rejection region is divided by 10 so that $p < 0.0025$ is the criterion for significance, the boundary of significance for the statistic increases from 1.96 to 2.81.Under this stricter criterion and with a test that makes no assumptions about distribution shape, 28 of the 30 tests show the mean difference to be significantly different from zero. Note that differences with values of zero were not counted in the calculation of the sign test.

A convenient measure of discriminability is "d-prime" (written d′). The parameter d′ is a standardized difference between two means [20]. Table 3 displays the value of d′ for depressed speech when this condition is to be discriminated from normal speech. The d′ is a normalized standard score. A d′ value of 0.0 indicates that there is no information useful in discriminating the depressed speech samples from the background expectation of normal speech. Larger d′ values indicate greater discriminability in a parameter and greater usefulness for automatic categorization of signals.

Table 3: Values of d′ for depressed vs. normal speech within- and across-subject groupings

Parameter	d′ (population	d′ (person-item)
TST	0.376	0.819
TPT	0.478	0.499
TUT	0.477	0.849
ROS	0.902	1.155
ART	0.805	1.097
LAT	1.226	1.365
MPD	0.478	0.486
SDP	0.962	1.120
PDP	0.365	0.436
MaxSA	0.744	1.192

Table 4: Agreement of significant experimental results from literature reviews

Parameter	Significant Order	Scherer (1986) agrees	Murray & Arnott (1993) agree
TST	D > N > M	yes	yes
TPT	D > N > M	no info	no info
TUT	D > N > M	yes	yes
ROS	M > N > D	yes	yes
ART	M > N > D	yes	yes
LAT	D > N > M	no info	no info
MPD	D > N , D > M	no info	no info
SDP	N > M > D	no info	no info
PDP	M > N > D	no info	no info
MaxSA	M > N > D	yes	yes

IV. Discussion

The data are generally consistent with an alternative hypothesis that psychotherapists have a relatively uniform model of depressive and manic speech patterns that do show up in their simulations and agree with the patterns reported in the literature. Of the parameters (often vaguely specified in the literature) that seem to have an analog in the parameters of this experiment, the significant observed orders are uniformly in accordance with published literature reviews, as is shown in Table 4.

Many of the statistical tests show effects that are extremely unlikely under the null hypothesis, yet the single parameter d′ values are not particularly large, which indicates that a device that used any single one of these parameters to classify an unknown person could make a substantial number of errors. The d′ values are generally greater for the within-speaker comparisons, which supports the intuitive and expected result that a device or a person would do better using paralinguistic information to discriminate among the moods of a known person than to identify the moods of an unknown person. From a single voice recording by itself, a listener can presumably recognize a mood shift in a friend more reliably than that same listener could classify the mood of a stranger.

All ten of the paralinguistic acoustic variables that were studied had statistically significant association with one or the other of the two moods (depressed or manic) that were intentionally simulated by the psychotherapists who served as speaking subjects; eight out of ten parameters were significantly different in both moods from the uninstructed (normal) condition. Two variables failed the test of significance for the manic speech only in manic versus normal comparisons.

When analyzed within subject and within item, both simulated moods are significantly different from the uninstructed (normal) mood in nine of the ten parameters, instead of the eight of ten in the group comparison. The only failure of significance was in one manic versus nor-

mal comparison.

Certain conditions of this experiment limit the scope of the conclusions. The foremost limitation concerns the use of psychotherapists as subjects. The variety of initial speaking patterns and courses of change over time that is found in real clinical populations is simply not found in the speech data from people simulating moods. Likewise, there is no possibility to compare the speech data with concurrent scores on cognitive, emotional, physiological, or motor-performance tests. Thus, none of the hypotheses about the cognitive or psychomotoric nature of mood disorders as discussed by [7] or by [14] can be tested with this new data. Finally, an important limitation is that voice fundamental frequency (F0) was not measured and therefore not analyzed.

V. Conclusion

Psychotherapists can imitate (without any instruction or guidance) some of the vocal patterns of depressed and manic people in a way that is relatively consistent over the population of therapists and is also consistent with the paralinguistic changes reported in the literature on speech in emotion and mood disorders. For many traditional paralinguistic parameters, the ordering of {depressed, normal, manic} is monotonic increasing or decreasing. Generally, for the psychotherapists simulating mood or pathology, the depressed direction from normal is more reliably and distinguishably produced.

The differences in paralinguistic parameters between groups of people when speaking normally and when simulating moods are very significant, but these differences may be relatively difficult to use for mood identification from any single one of the duration- or amplitude-based parameters that were studied in this project.

If these results can be replicated with an appropriate clinical population, then this study provides a system and the core of an algorithm for rating the paralinguistic evidence of mood disorders by telephone, automatically, on demand. Note that to be useful or interesting, the system does not have to be highly accurate, it may suffice that the system performs as well as a skilled therapist, and only on that aspect of the therapist's judgment that relates to manner of speaking.

References

[1] M Alpert, A. Rosen, J. Welkowitz, C. Sobin & J. Borod, "Vocal acoustic correlates of flat affect in Schizophrenia". *British Journal of Psychiatry*. 154, 51-56, 1989.

[2] R. Banse & K. Scherer, "Acoustic profiles in vocal emotion expression". *Journal of Personality and Social Psychology*, 70 (3), 614-636, 1969.

[3] R. Cowie & R.R. Cornelius, "Describing the emotional states that are expressed in speech" Speech *Communication*, 40, (1-2),5-32, 2003

[4] H Ellgring, & K. Scherer, "Vocal indicators of mood change in depression." *Journal of Nonverbal Behavior*, 20 (2), 1996.

[5] G Fairbanks & L Hoagland,. "An experimental study of the durational characteristics of the voice during the expression of emotion". *Speech Monographs*, 8, 85-90. 1941.

[6] T Goldbeck F. Tolkmitt, & K Scherer, "Experimental studies on vocal affect communication". In K. Scherer (Ed.), *Facets of Emotion. Recent Research*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988, pp. 119-137.

[7] J Greden. & B. Carroll, "Decrease in speech pause times with treatment of endogenous depression". *Biological Psychiatry*, 15 (4), 575-587, 1980.

[8] S. Kuny. & H. Stassen "Speaking behavior and voice sound characteristics in depressive patients during recovery". *Journal of Psychiatric Research*, 27 (3), 289-307. 1993.

[9] I. Murray, & J. Arnott, "Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion" *Journal of the Acoustical Society of America* 93 (2), 1097-1108, 1993.

[10] J. Pittam, & K Scherer "Vocal expression and communication of emotion". In M. Lewis & J.M. Haviland (Eds),.*Handbook of Emotion,* New York: Guilford Press, 1993, pp. 185-198.

[11] K. Scherer). "Vocal affect expression: A review and a model for future research," *Psychological Bulletin*, 99 (2), 143-165, 1986.

[12] H. Stassen, G. Bomben, & E. Gunther, "Speech characteristics in depression", *Psychopathology*, 24, 88-105, 1991.

[13] E. Szabadi, C. Bradshaw & J. Besson, "Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression", *British Journal of Psychiatry*, 129, 592-597, 1976.

[14] D. Widlocher, & A. Ghozlan, "The measurement of retardation in depression". In I. Hindmarch & P. Stonier (Eds.), *Human Psychopharmacology: Measures and Methods,* Vol. 2. New York: John Wiley & Sons Ltd., 1989, pp. 1-22.

[15] C. Williams & K. Stevens, "Emotions and speech: some acoustical correlates. *Journal of the Acoustic Society of America,* 52, 1238-1250, 1972.

[16] E. P. Rosenfeld, "Automatic analysis of vocal manifestations of psychological states." Unpublished doctoral dissertation. Western Graduate School of Psychology, Palo Alto, California. 2000.

[17] Ordinate Corporation *Validation Summary for the SET-10 Test,* Menlo Park, CA: Ordinate Corporation., 2000.

[18] G. Kanji, *100 Statistical Tests*. London: SAGE Publications, 1993.

[19] S. Siegel, *Non-parametric Statistics*. New York: McGraw-Hill, 1956.

[20] C. Coombs, R. Dawes & A. Tversky, *Mathematical Psychology.* Englewood Cliffs, NJ: Prentice-Hall, 1970.

3rd International Workshop MAVEBA 2003, 9-12

AUTOMATIC DETECTION OF STRESS IN SPEECH

H. J. Fell[1], J. MacAuslan[2]
[1]College of Computer and Information Science, Northeastern University, MA, USA
[2]Speech Technology and Applied Research, Lexington, MA USA

Abstract: **We have developed software based on the Stevens landmark theory to extract features in utterances in and adjacent to voiced regions. We then apply two statistical methods, closest-match (CM) and principal components analysis (PCA), to these features to classify utterances according to their emotional content. Using a subset of samples from the Actual Stress portion of the SUSAS database as a reference set, we automatically classify the emotional state of other samples with 75% accuracy, using CM either alone or with PCA and CM together. The accuracy apparently does not depend strongly on measurement errors or other small details of the present data, giving confidence that the results will be applicable to other data.**

Keywords : **automatic detection, emotion, speech, stress**

I. INTRODUCTION

If computers are to interact with humans in a natural way, they will need a speech interface that recognizes emotional as well as linguistic content of speech. Scherer *et al* [1998] argue that modeling of speaker states and emotions can improve the quality of automatic speech recognition, speech synthesis, and speaker verification and that such emotion effects are relatively robust to changes in the phonetic context. Imagine your computer responding with sympathy when you are sad, explaining things more simply when you are frustrated, or speaking calmly to you when you are stressed.

Speech scientists have been able to identify a number of acoustic speech parameters that correlate with the speaker's emotional state. Johnstone & Scherer [6] report that analysis of glottal opening and closing characteristics proved useful in interpreting the emotion-dependent characteristics of the acoustic waveform. Quast [10] identifies a number of parameters that appear to carry crucial information, e.g. location of the sentence foci, intensity values, relation of the fundamental frequencies (F_0) at the focus and ends of the sentence, speech rate, and spectral histogram.

There have been few attempts and limited success at actually recognizing and classifying affect in speech. Roy and Pentland [11] used six acoustic measurements (F_0 mean and variance, Energy variance and derivative, open quotient, and spectral tilt) to classify spoken sentences as approving or disapproving. They achieved 65% to 85% classification accuracy for speaker dependent, text independent data. Their results suggest that energy and F_0 statistics may be effectively used for automatic affect classification. Stolcke et al. [14] used prosodic cues as part of a statistical approach to model dialogue acts in conversational speech. They achieved a 71% accuracy in labeling act-like units such as statement, question, agreement, disagreement, and apology. Dellaert et al. [1] applied several statistical pattern recognition techniques to classify utterances according to their emotional content. For the purposes of classification they used only pitch information extracted from the utterances. They also introduced a spline approximation of the pitch contour to extract features. Their best method resulted in a 20.5% error rate in classifying four emotions: happiness, sadness, anger, fear. Human performance at the same task resulted in an 18% error rate.

We have had success in applying landmark detection coupled with Principal Component Analysis in detecting significant differences in the vocalizations of typically-developing and at-risk infants [2, 3, 4] and in detecting fatigue in adult speech [8]. Here, they apply similar techniques to classifying stress in speech.

II. THE DATA

We are using the Actual Speech Under Stress portion of the SUSAS (Speech Under Simulated and Actual Stress) database [5]. A common highly confusable vocabulary set of 35 aircraft communication words make up the database. All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8kHz. We are using samples recorded under four conditions: neutral - Neutral Speech, medst - low Dual-Tracking task stress, hist - high Dual-Tracking task stress, and scream - Scream Machine Roller Coaster stress. We have restricted this study to the four male speakers: m1, m2, m4, who have a General USA Accent; and m3: who has a Southern USA Accent.

We formed a base of features for classification using only the first sample of each of the 35 words for each speaker in each emotional state whenever such samples were present. Table 1 shows the number of words for each speaker/emotional-state used to create the base.

Table 1: Number of words used to create the base for classification

	neutral	medst	hist	scream
m1	35	35	35	29
m2	34	35	35	29
m3	34	35	35	23
m4	35	35	35	23

We then created test cases for classification using the second sample of each of the 35 words for each speaker in each emotional state whenever such samples were present. Table 2 shows the number of words for each speaker/emotional-state test case.

Table 2: Number of words per sample for the 16 test cases

	neutral	medst	hist	scream
m1	35	35	35	15
m2	8	35	35	24
m3	34	35	34	15
m4	35	35	35	2

III. METHODOLOGY

We listened to many words in the SUSAS Actual Stress database before attempting to perform automatic classification. One subjective impression was that the vowels were longer, relative to word duration, in the medst and hist words than in the corresponding neutral words. Another impression was that the consonants were clipped, shorter and less structured than their neutral correspondents. To model these impressions, we needed to extract more than pitch information.

Using software that we have developed [2, 3, 4] based on the Stevens landmark detection theory [7, 13] for the recognition of phonetic features in speech, we extracted measurements on twenty-five features from the ~35-word sets of speech samples. These served to summarize the speaker, state, and sample.

From Syllables:
Timing:
mean duration, mean duration of voicing, mean voiced fraction (i.e. mean of voiced duration/total duration), maximum and mean voice onset time (VOT), maximum and mean offset time, mean rate (i.e. mean of 1/duration), mean voiced rate (i.e. mean of 1/voiced duration).
Pitch (F_0):
median and mean F_0, fraction of syllables in which the pitch rises (falls) during the first half (second half) of the syllable.
Structure:
mean, median, and maximum number of landmarks per syllable.

From Words:
Pitch:
root mean square standard deviation of F_0, relative range of pitch (see below), 10^{th}, 50^{th}, and 90^{th} percentile value of the relative range, 10^{th}, 50^{th}, and 90^{th} percentile value (over all the words) of the "central" F_0 value, i.e., the value in the middle of the word.

The relative range of pitch: defined as the maximum (over each word) of the 90^{th} percentile values of the pitch, minus the minimum of the 10^{th} percentile values, divided by the median value (over the word). Thus, it is a non-negative number, and typically less than 1. We divide by the median F_0 so that the results are not strongly skewed for irrelevant reasons, such as a generally lower F_0 for men than women.

For each state, we normalized the four speakers' data by comparing their values for each of these features to the mean and standard deviation σ of all four in that state. Specifically, we subtracted the mean and then divided by a certain variability measure. This measure consists of σ and an *a priori* estimate of measurement error, combined in RSS (root sum-of-squares) fashion. Thus, for example, the squared measure for an F_0-related feature consists of the sum of the observed four-subject value of that feature's variance σ^2 plus $(5\text{ Hz})^2$, because 5 Hz represents an estimate of the irreducible measurement uncertainty for F_0. Such irreducible measurement uncertainties depend primarily on the recording environment or computational details (for F_0, at least).

Observe that this normalization process yields feature values of zero mean and approximately unit variance for the base cases. As 25-element vectors, then, the normalized base-case summaries have norm (Euclidean length) $\sim 25^{1/2}$.

When comparing one speaker/state/sample summary to another, we simply evaluate the RSS of the vector of differences in feature values. By construction, this also produces values $\sim 25^{1/2}$ to $50^{1/2}$ when comparing two base cases, and we might anticipate similar or even smaller results when comparing two samples from the same speaker and state. In fact, this was routinely observed.

To identify a state from a test set of speaker/state/sample, we hypothesize a state, normalize the corresponding summary using the mean and variability parameters for that state, and compare to each of the base cases of the state. Across all speakers and states defining the base, 16 summaries in all, the lowest RSS difference identifies the closest-matching, or CM, state (and, in principle, speaker).

An important refinement is available. Of the 16 subject/state normalized feature vectors that define the base, some linear combinations may be redundant. Eliminating these would improve the robustness of the results, because the redundant components would otherwise tend to model inappropriately small details of the data, i.e.,

"noise". Principal Components Analysis (PCA: equivalently, singular value decomposition, SVD) determines the extent to which this occurs among the set of vectors. In this case, the first three PC's accounted for 99% of the total variance, suggesting both a high degree of linear dependence and a high degree of linear predictability.

IV. Software and Algorithms

Our landmark detector is based on Stevens' acoustic model of speech production [13]. Central to this theory are *landmarks*, points of abrupt spectral change in an utterance around which listeners extract information about the underlying distinctive features. They mark perceptual foci and articulatory targets. Our program detects three types of landmarks:

glottis (+g, -g): marks the time when the vocal folds start and stop vibrating;
sonorant (+s, -s): marks sonorant consonantal closures and releases;
burst (+b, -b): aspiration/frication ends due to stop closures.

Our analysis is based on a low-resolution spectrogram. The SUSAS signals are sampled at 8 kHz and analyzed into a small number, nominally 32, of separate, frequency intervals of ~256 Hz each. An 8 kHz rate provides information only up to 4 kHz, but this is sufficiently high to include at least 3-4 formants for an adult and to show the distinction between voicing and other speech sounds: fricatives, stop releases, bursts, etc. (See Fig. 1.)

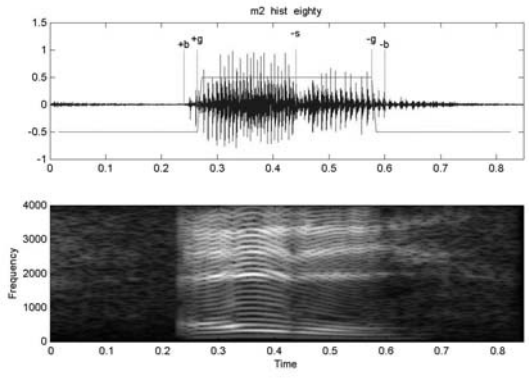


Figure 1: Waveform and Landmarks (top) and Spectrogram (bottom) of "eighty" as spoken by male 2 in high stress conditions.

To locate the landmarks, spectral intervals are grouped into six broad bands. An energy waveform is constructed in each of the six bands, the time derivative of the energy is computed, and peaks in the derivative are detected. These peaks thus represent times of abrupt spectral change in the six bands. Energy in bands 2 (1200 - 2500 Hz.) and 3 (1800 - 3500 Hz), e.g., provides evidence of voicing or, in some cases, of bursts. The distinction between these is readily made in the time domain (voicing persists much longer than bursts) as well as by appeal to information in the other spectral bands: voicing provides a power spectrum that decays with frequency approximately as $1/\text{frequency}^2$, whereas most other speech sounds have flatter spectra.

V. Results and Discussion

Our small study with sixteen test cases, as seen in Table 3, resulted in a 25% error rate.

Table 3: Results of the CM (closest-match) comparison. Boldfaced values represent correct identification of speaker state. *The listed states had nearly equally small distances.

	neutral	medst	hist	scream
m1	**neutral**	neutral	neutral	**scream**
m2	**neutral**	**medst**	neutral	**scream**
m3	**neutral**	hist* neutral	**hist**	**scream**
m4	**neutral**	**medst**	**hist**	**scream**

To test the stability of the results, we performed a Principal Components Analysis (PCA, or, equivalently, singular value decomposition, SVD [9]). This permitted us to discard several of the principal components (PCs) that described only noise-level variations in the data. Retaining eight of the original 16 PCs, accounting for 95% of the variance, produced only small variations in the results, and no overall degradation in accuracy.

Table 4: Results of the PCA/CM comparison. Boldfaced values represent correct identification of speaker state. *The listed states had nearly equally small distances.

	neutral	medst	hist	scream
m1	**neutral**	neutral	neutral **hist***	**scream**
m2	**neutral**	**medst**	neutral	**scream**
m3	**neutral**	hist* neutral	**hist**	**scream**
m4	**neutral** hist*	**medst**	**hist**	**scream**

Inspection of the Tables reveals that the classification has no errors for the neutral or scream states. Furthermore, most errors occurring in the other states are manifest as neutral, that is, the closest-match algorithm selects the "conservative" interpretation that the data represent no departure from the neutral state.

VII. Conclusion

We have shown that a simple knowledge-based analysis of American English speech and some measures of F_0 can classify a speaker's emotional state among four choices moderately well. We achieve 75% accuracy when comparing new data from a speaker that is already represented among the base cases. PCA indicates that this result does not depend sensitively on small details such as noise level. We are currently investigating the performance when the speaker is not so represented.

Acknowledgements

This work was supported in part by National Science Foundation grant SGER 0206940.

References

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," *Proc. ICSLP*, 1996.

[2] H.J. Fell, L.J. Ferrier, D. Sneider, and Z. Mooraj, "EVA, An early vocalization analyzer: an empirical validity study of computer categorization," *Assets '96*, pp. 57-61, 1996.

[3] H.J. Fell, J. MacAuslan, L.J. Ferrier, K. Chenausky, "Automatic Babble Recognition for Early Detection of Speech Related Disorders," *J. Behaviour & Inf. Tech.*, **18**, no. 1, pp. 56-63, 1999.

[4] H.J. Fell, J. MacAuslan, L.J. Ferrier, S.G. Worst, and K. Chenausky, "Vocalization Age as a Clinical Tool," *Electronic Proc. ICSLP (Int. Conf. on Speech & Language Processing)*, Denver, 2002.

[5] J.H.L. Hansen, "SUSAS -Speech Under Simulated and Actual Stress*," Robust Speech Processing Lab.*, http://www.ee.duke.edu/Research/Speech/, 1997

[6] T. Johnstone, and K.R. Scherer, "The effects of emotions on voice quality*", Proc. XIVth Int. Congress of Phonetic Sci*, 1999.

[7] S. Liu, "Landmark detection of distinctive feature-based speech recognition," *J. Acc. Soc. Amer.,* **96**, 5, Part 2, p. 3227, 1994.

[8] J. MacAuslan "Speech Analysis for Fatigue Assessment", US Air Force Final Report, 2002.

[9] Press, W., S. Teukolsky, W. Vetterling, & B. Flannery. (1992). Numerical Recipes in C, 59-70. New York: Cambridge University Press.

[10] H. Quast, "Robust Machine Perception of Nonverbal Speech," *http://ergo.ucsd.edu/~holcus/Speech.html*, 2000.

[11] D. Roy & A. Pentlland, "Automatic Spoken Affect Classification and Analysis," *Pro Second Int. Conf. Automatic Face & Gesture Recognition*, pp. 363—367, 1996.

[12] K.R. Scherer, T. Johnstone, and J.Sangsue, "L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole," *Actes des XXIIèmes Journées d'Etudes sur la Parole,* Martigny, 1998.

[13] K.N. Stevens, S. Manuel, S. Shattuck-Hufnegel, and S. Liu, "Implementation of a model for lexical access based on features," *Proc. Int'l. Conf. Spoken Language Processing*, Banff, Alberta, **1**, 499-502, 1992.

[14] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech", *Computational Linguistics* **26**(3), pp. 339-373, 2000.

3rd International Workshop MAVEBA 2003, 13-16

INVESTIGATION OF GLOTTAL FLOW SPECTRAL SLOPE AS POSSIBLE CUE FOR DEPRESSION AND NEAR-TERM SUICIDAL RISK

Asli Ozdas[1], Hasmila Omar[3], Richard G. Shiavi[2], Stephen E. Silverman[4], Marilyn K. Silverman[4], and D. Mitchell Wilkes[3]

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
[2] Department of Biomedical Engineering, Vanderbilt University, Nashville, Tennessee, USA
[3]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, Tennessee, USA
[4]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut, USA

Abstract: **When reviewing his clinical experience in treating suicidal patients, one of the authors observed that successful predictions of suicidality were often based on the patient's voice independent of content. In this study we investigated the discriminating power of an excitation-based speech parameter, the glottal flow spectrum. There were two sets of subjects, male and female. Each set consisted of 10 high-risk near-term suicidal patients, 10 major depressed patients, and 10 non-depressed control subjects. As a result of two sample statistical analyses, the slope of the glottal flow spectrum, was a significant discriminator in five of six comparisons ($p<0.05$). A maximum likelihood classifier, developed by combining the *a posteriori* probabilities of two features, yielded correct classification scores between 60 and 95%.**
Keywords: **Speech, glottal flow spectrum, suicide, depression, classification**

I. INTRODUCTION

Identification of individuals at imminent suicidal risk is often one of the most important judgments that clinicians must make. This task requires gathering and weighing of a variety of information and data from numerous sources by experienced clinicians [1]. These methods help in categorizing individual patients as "high risk", but they are not sufficient to determine if a patient is at imminent risk. Stephen and Marilyn Silverman describe suicidal speech as similar to depressed speech but exhibiting significant perceptual changes in its qualities when a patient becomes near-term suicidal. The exhibition of these qualities was often a decisive factor in alerting the clinicians to the need to take preventative action [2]. These clinical findings together with the literature on the clinical importance of a patient's voice in psychiatry led to the hypothesis that near-term suicidality may be associated with changes in speech production and articulation that differ from non-suicidal persons. Our own studies are showing this [3][4].

Many studies have been done using the fundamental frequency. However, the fundamental frequency provides information only about the duration of the glottal cycle. Besides fundamental frequency, glottal flow waveform was also reported to be altered as a result of excessive tension or lack of coordination in the laryngeal musculature under emotional stress [5]. Investigation of this phenomenon showed an increase in the amount of high frequency energy in the glottal pulses under emotional stress. In this paper, we explore the significance of the slope of the glottal flow spectrum (spectral tilt) as an indicator of near-term suicidal risk.

II. DATABASE FORMULATION

Glottal flow spectral analyses were performed on sets of audio recordings for males and females. Each set contained 10 near-term suicidal patients, 10 depressed patients, and 10 non-depressed control subjects collected from existing databases. All the patients used in this research were white Caucasians between the ages of 25 and 65. Because of the inability to record psychiatric speech in controlled settings, all of the speech samples were recorded during real-life situations (i.e., therapy sessions, suicide notes left on tapes, etc with various tape recorders at various recording environments). A high-risk, near-term suicidal patient was defined as one who has committed suicide or attempted suicide and failed within minutes to weeks from the time of their voice recordings. The audio recordings of the depressed and control groups were extracted from the database of an ongoing study in the Vanderbilt University Department of Psychiatry. The control group was comprised of depressed individuals who, after receiving cognitive therapy or pharmacotherapy, were judged to be no longer depressed and not in need of further treatment. The selected non-depressed control subjects met the following criteria: 1) a Hamilton rating scale (17 item version) for a depression score of 7 or less [6]; 2) a Beck depression score of 7 or less [7]. The depressed patients met the following criteria: 1) major depressive disorder as defined by the research diagnostic criteria [8]; 2) a Beck depression score of 20 or greater; 3) a Hamilton rating scale for depression score 14 or greater.

All of the selected audio recordings were digitized using a sixteen-bit analog to digital converter. The sampling rate was 10 KHz, with an anti-aliasing filter (i.e.,

5KHz low-pass) precisely matched to the sampling rate. The digitized speech waveforms were then imported into a MicroSound Editor where silence pauses exceeding 0.5 seconds were removed to obtain a record of continuous speech. Thirty seconds of continuous speech from each subject were stored for analyses.

III. METHODS

A. Glottal Spectral Slope Feature Extraction

Vocal tract effects were removed from the speech spectrum while estimating glottal flow spectrum. It was assumed that the frequency response of the vocal tract shapes the speech spectrum for different vowels and glottal flow spectrum stays the same for all vowels. Therefore, the glottal flow spectrum can be estimated if energy normalized frames from voiced speech spectra are averaged to remove the effects of vocal tract shaping. The averaged vocal tract response will have an all pass characteristics if a wide variety of vowel spectra are used, and the average energy normalized frames will yield the glottal flow spectrum. This approach provides a representation that reflects the properties of glottal flow waveform.

A1. Estimation of Glottal Spectrum

a) The patient speech is broken into segments containing 256 samples.
b) Voiced and unvoiced speech detection is performed on each segment. However, only voiced segments are retained for analysis. The method used is based on wavelets and developed by Ozdas [4].
c) The periodogram for each voiced segment is calculated using the discrete Fourier transform.
d) Each periodogram is normalized by its energy.
e) All normalized periodograms are then averaged to remove the effects of varying vocal tract response.
f) The average energy of all voiced segments is then used to scale the average normalized periodogram back to its original amplitude. This is the glottal flow spectrum estimate.

A2 Estimation of Glottal Spectral Slope

The spectral slope is calculated using a least squares line fit on a log-log scale is performed over 300-3000 Hz frequency band of the glottal flow spectrum. The slope given by the least square error approximation gives the glottal spectral slope for each patient. Fig. 1 shows an example of the estimation procedure.

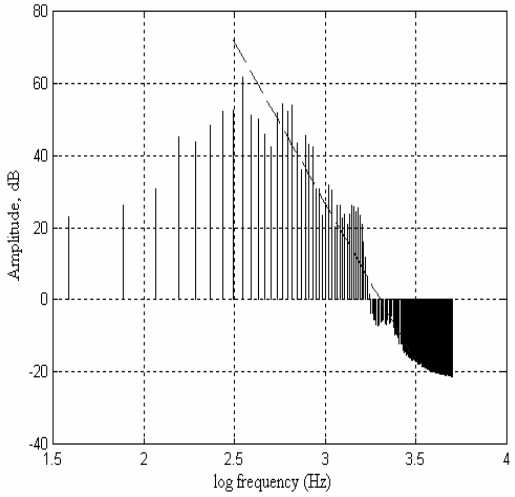


Fig. 1: Glottal Flow Spectrum and estimation of slope

B. Comparative Statistical Analyses and Classification

B1. Statistical Tests

Two-sample (i.e., control-depressed, control-suicidal, and depressed-suicidal) t-tests were performed separately on glottal spectral slope estimates to determine any statistically significant differences in means [9].

B2. Maximum Likelihood Classifier

In order to evaluate the discriminating power of the slope among groups, a Maximum Likelihood (ML) classifier was developed for each parameter. The ML classifier employs the Probability Density Function (PDF) of each class to make a decision as to which class PDF results in the closest match for a test data sample. The PDFs of the class distributions were assumed to be unimodal Gaussian and were generated by using the means and variances estimated from the training samples. Given the trained class model, classification of the test samples was accomplished according to Bayes' decision rule, where a test subject was assigned to the class for which it had the maximum *a posteriori* probability for its set of observations.

Ideally, this procedure is conducted by splitting the total data set into a training set and a test set. Because of the limited number of patients in this case, Lachenbruch's *holdout* procedure was employed [10]. This procedure is very useful for small data class sizes because it makes it possible to use the same subject for both training and testing rather than using only half of the data for each part.

IV. Results

A. Magnitude of Glottal Slopes

The estimated magnitudes of the slopes of the glottal flow spectra for each subject are given in Figs. 2 and 3.

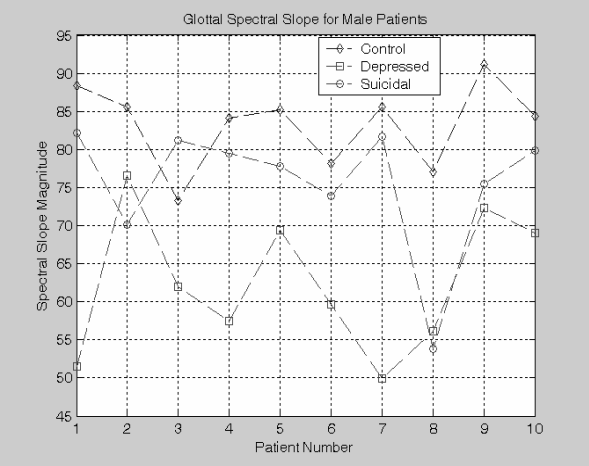


Fig. 2: Magnitude of Glottal Slopes for Males

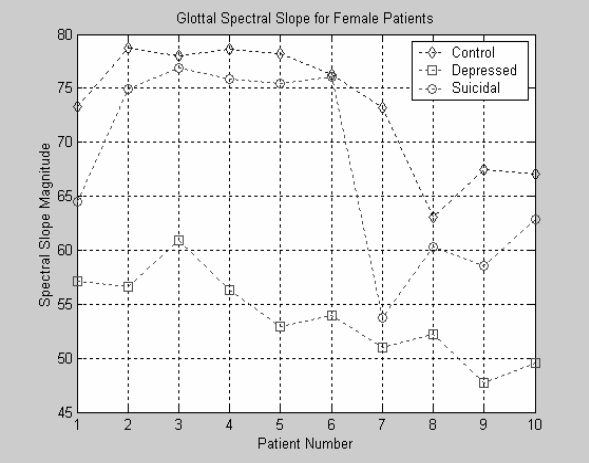


Fig. 3: Magnitude of Glottal Slopes for Females

Notice that the controls have the highest values and the depressed subjects the lowest values.

B1. Statistical Test

The p-values for pair-wise comparison of the means are shown in Table 1.

Table 1. P-Values for Mean Comparisons of Spectral Slope Estimates

P-Values	Female	Male
Control/Depressed	0.0000	0.0000
Depressed/Suicidal	0.0005	0.0035
Control/Suicidal	0.1156	0.0266

The means of the groups of spectral slope values are significantly different ($p < 0.05$) for five of the six comparisons. The only comparison that is not different is the one between control and suicidal females.

B2. Maximum Likelihood Classifier

The ML classification results for glottal flow spectral slope are presented in Table 2.

Table 2. ML Pairwise Classification Results (%) For Spectral Slope Estimates

% Classification	Female	Male
Control/Depressed	95	90
Depressed/Suicidal	85	75
Control/Suicidal	60	60

The ML classifier yielded overall classification scores between 60% and 95%. The highest between depressed and control classes and the lowest between suicide and control classes. This was consistent in the male and female populations.

V. Discussion and Conclusion

Analyses of glottal spectral slope measurements indicated that both near-term suicidal and depressed patients exhibit significantly higher energies in the upper frequency bands of the glottal flow spectrum compared to healthy controls. These shifts are significantly different among most of the comparisons. The spectral content of the glottal spectra is more similar between controls and suicidal subjects while those for depressed subjects have a broader bandwidth. In addition it is possible to use the spectral slope to classify subjects as belonging to one of three groups. Evidence for similar energy shifts in long-term energy spectra during depression and near-term suicidal states have been reported by various researchers [11]. Most of the studies that investigated this phenomenon have revealed that the speech of patients who suffer from major depressive illness contains more energy at higher frequency bands, which was shifted toward lower frequencies after treatment. Here, it is important to note that it is not possible to collect speech samples from suicidal persons shortly before their suicide attempts in a

systematic manner. Therefore expanding the database requires a considerable amount of time.

REFERENCES

[1] W.J., Fremouw, M. Perczel, and T.E. Ellis, *Suicide Risk: Assessment and Response Guidelines*, New York: Pergamon Press, 1990.

[2] S.E. Silverman, "Vocal parameters as predictors of near-term suicidal risk", U.S. Patent 5 148, 483, Sept. 1992.

[3] France, D.J., Shiavi, R.G., Silverman, S.E., Silverman, M.K., Wilkes, D.M., "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Transactions on Biomedical Engineering, vol. 47, no. 2, pp. 829-837, 2000

[4] Ozdas, A., "Analysis of Paralinguistic Properties of Speech for Near-Term Suicidal Risk Assessment", PhD. Thesis, Vanderbilt University, 2000.

[5] K.E. Cummings, M.A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech", *J. Acoust. Soc. Am.*, vol. 98, pp. 88-98, July 1995.

[6] M. Hamilton, "A rating scale for depression", *J. Neurosurg Pscychiatry,* vol. 23, pp. 56-62, 1960.

[7] A.T. Beck, C.H. Ward, M. Mendelson, J.E. Mock, J.K. Erbaugh, "An inventory for measuring depression", *Arch. Gen. Psych*., vol. 24, pp. 561-571, 1961.

[8] R.L. Spitzer, J. Endicott, E. Robins, "Research diagnostic criteria: rationale and reliability", *Arch. Gen. Psych.*, vol. 35, pp. 773-782, 1978.

[9] B. Rosner, *Fundamentals of Biostatistics*, Fifth edition, CA: Duxbury, Thomson learning, 2000.

[10] R. Johnson, D. Wichern, *Applied multivariate statistical analysis*, Third edition, NJ: Prentice Hall, Englewood Cliffs, 1992.

[11] P. Ostwald, P., *Soundmaking: The Acoustical Communication of Emotion*. Springfield, IL: Charles C Thomas, 1963.

Laryngectomy

3rd International Workshop MAVEBA 2003, 19-22

MODELLING OF THE LARYNGECTOMEE SUBSTITUTE VOICE

[1]J. Lohscheller, [2]M. Döllinger, [1]R. Schwarz, [1]U. Eysholdt and [1]U. Hoppe
[1]Department of Phoniatrics and Pediatric Audiology, University Erlangen-Nuremberg, Germany
[2]Laryngeal Dynamics Laboratory, University of California Los Angeles, USA

Abstract: **A bio-mechanical model is derived which describes the fundamental principles of laryngectomee substitute voice production. Within the model the substitute voice generator (PE segment) is modelled as an elastic tube which is set into vibrations by streaming air. The model bases on the well known two-mass-model by Ishizaka and Flanagan (1972) wich has been successfully used to describe regular phonation. The morphology of the PE segment is considered by several two-mass-models which are orbitally coupled with spring and damping elements. The main parameters which affect oscillation are vibrating masses, muscle tensions and lung pressure. Within the model, the time dependent minimum aperture serves as measure of PE segment deformations. The performance of the PE-Model is demonstrated by adapting the PE-Model to experimental PE segment vibrations which are extracted from high-speed sequences.**

Keywords: **Substitute Voice, High-Speed-Recording, Two-Mass-Model, PE-Model.**

I. INTRODUCTION

Therapy of cancer of the throat may require a surgical excision of the larynx which results into the loss of voice. During a so-called total laryngectomy, trachea and esophagus are separated in order to prevent uncontrolled mixing of breathing and swallowing [1]. Breathing is maintained by suturing the trachea into the frontal skin of the neck (tracheostoma). In order to achieve voice rehabilitation a substitute voice generating element has to undertake the task of the excised larynx. State of the art therapy is the insertion of a silicon shunt valve which reconnects the separated trachea and esophagus and establishes an unidirectional connection from the trachea to the esophagus [2]. When closing the tracheostoma during expiration, air passes through the voice prothesis into the esophagus. The airflow excites vibrations of soft tissue at the upper esophagus sphincter, i.e. pharyngeal-esophageal segment (PE segment). These tissue vibrations modulate the airstream which poses as substitute voice signal (tracheoesophageal voice production). The anatomy of the PE segment consists of a mucosal coated ring-shaped muscle structure. The aim of this work is to introduce a bio-mechanical model of the PE segment (PE-Model) which describes PE segment dynamics in order to gain insight into the voice generating process.

II. METHODOLOGY

A. *Analyzing PE Vibrations in High-Speed Sequences*

High-speed recordings are performed during phonation using an endoscope coupled with a digital high-speed camera which allows the observation of PE segment vibrations in real-time. The patients are instructed to articulate the vowel /a/ in a 'comfortable' way. The frame rate of the high-speed system is 3704 Hz while the resolution of the CCD-array is 128 x 64 pixel. Simultaneously, the acoustic signal is recorded. For two high-speed recordings the tissue vibrations are quantitatively analyzed during a time interval of 95 ms using an image processing algorithm [3, 4]. The size and shape of the pseudoglottis $a(t)$, which is determined by the algorithm, serves as measure for PE segment deformations.

B. *Model of the Pharyngeal Esophageal Segment*

The principle properties of voice production of laryngeal (vocal folds) and tracheoesophageal phonation (PE segment) are similar to each other. In both cases tissue vibrations are excited by aerodynamic forces which are caused by airflow. The aerodynamic forces can be described by the Bernoulli law while the myoelastic tissue vibrations follow bio-mechanics. Therefore, the here proposed model of substitute voice generation is derived from the model of vocal folds by Steinecke and Herzel [5] which bases on the Two-Mass-Model (2MM) developed by Ishizaka and Flanagan [6]. Though the 2MM contains a lot of simplifications concerning both the myoelastic and the aerodynamic part, it allows the description of the most important features of vocal fold dynamics. It has successfully been used to study vocal fold vibrations in voice production.

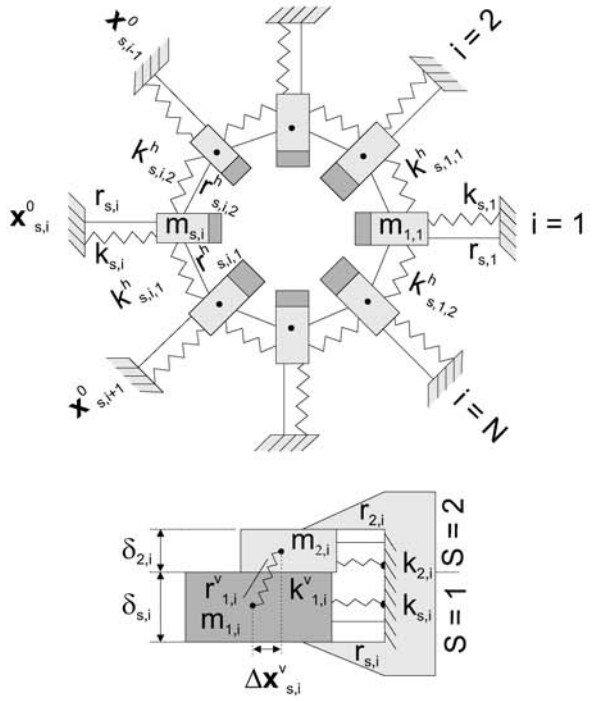


Figure 1: Top view and front view onto a PE-Model comprising eight masses per plane.

As first approximation the upper part of the esophagus is regarded as a flexible tube. The morphology is integrated into the PE-Model by placing several 2MM orbitally onto a horizontal circle. The center of the circle is regarded as point of origin of a cartesian coordinate system. Each 2MM is orientated to this point of origin. As the esophagus is a closed elastic tube the 2MMs are horizontally connected to each other. Therefore, adjacent 2MMs are coupled by additional spring and damping elements k^h, r^h. Fig. 1 shows the PE-Model with circular geometry, comprising eight masses per plane. Since the horizontal coupling extends the degree of freedom each mass is capable to move within the entire (x,y)-plane. The PE-Model is described by the following differential equation:

$$\begin{aligned} \mathbf{0} \;=\; & m_{s,i}\,\ddot{\mathbf{x}}_{s,i} + r_{s,i}\,\dot{\mathbf{x}}_{s,i} + k_{s,i}\,|\Delta\mathbf{x}_{s,i}|\,\mathbf{u}^0_{s,i} + \\ & k^v_{s,i}\,|\Delta\mathbf{x}^v_{s,i}|\,\mathbf{u}^v_{s,i} + r^v_{s,i}\,(\dot{\mathbf{x}}_{s,i} - \dot{\mathbf{x}}_{s+1,i}) + \\ & \mathbf{F}^D_{s,i} + \mathbf{F}^I_{s,i} + \mathbf{F}^H_{s,i}\,. \end{aligned} \tag{1}$$

The indices i denote the number of masses $m_{s,i}$ within the lower ($s = 1$) and upper ($s = 2$) plane. The differential equation contains tissue properties of the PE segment, i.e. masses $m_{s,i}$, stiffness $k_{s,i}$, $k^v_{s,i}$, $k^h_{s,i}$, and damping coefficients $r_{s,i}$, $r^v_{s,i}$, $r^h_{s,i}$. $\mathbf{x}_{s,i}$ denote the position of the masses $m_{s,i}$ within the cartesian coordinate system. Spring length variations in respect to the rest position of the masses $\mathbf{x}^r_{s,i}$ are expressed by the supplement Δ. The unit vectors $\mathbf{u}_{s,i}$ indicate the directions of spring and damping elements of mass $m_{s,i}$ and are illustrated in Fig. 2.

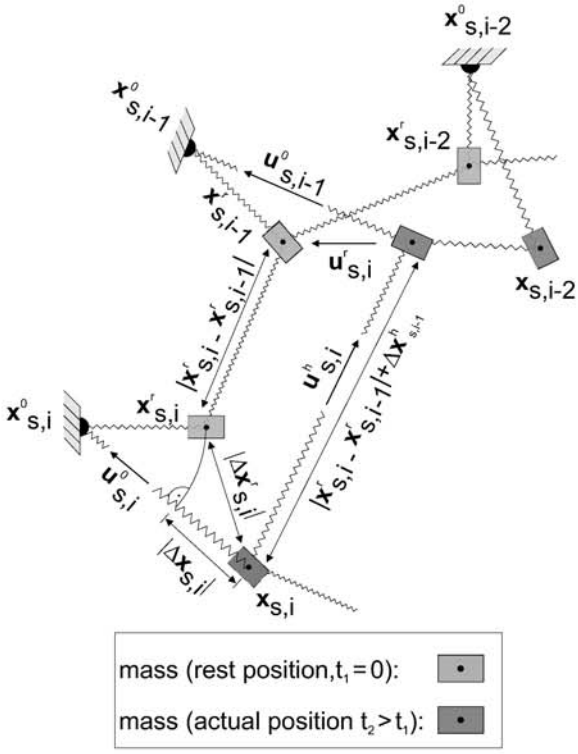


Figure 2: Graphical definition of mass positions $\mathbf{x}_{s,i}$, unit-vectors $\mathbf{u}_{s,i}$, and spring length variations $\Delta\mathbf{x}_{s,i}$. The damping elements are not illustrated.

The driving forces $\mathbf{F}^D_{s,i}$ result from pressure variations within the PE segment and depend on height of the lower plane δ_1 the area ratio of the minimal area a_{min} of both planes and the area of the lower plane a_1:

$$\mathbf{F}^D_{1,i} = P_L \cdot L_{1,i} \cdot \delta_1 \cdot \left(1 - \left(\frac{a_{min}}{a_1}\right)^2\right) \cdot \mathbf{u}^D_{1,i}\,. \tag{2}$$

The directions $\mathbf{u}^D_{1,i}$ of the driving forces $\mathbf{F}^D_{1,i}$ are defined in Fig. 3. The influence of colliding tissue is considered by additional spring constants $k^c_{s,i}$. Fig. 4 illustrates the collision force $F^c_{s,i,n}$ for a single impact. Collisions occur when a mass $m_{s,i}$ collides with a coupling spring of two adjacent masses $m_{s,j}$, $m_{s,j+1}$.

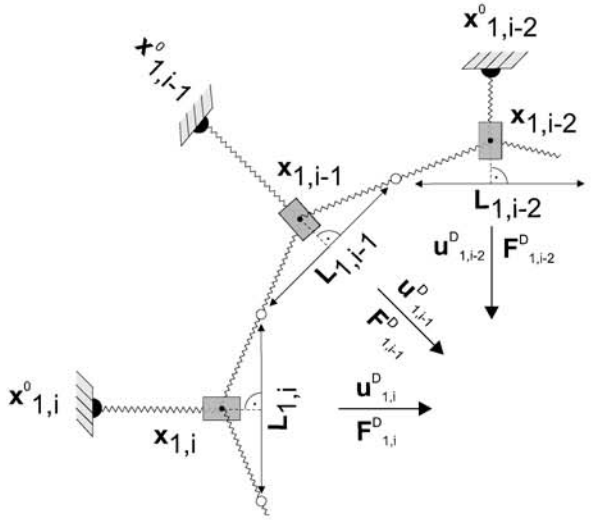


Figure 3: Graphical definition of the driving Force $\mathbf{F}^D_{1,i}$ and its corresponding unit vector $\mathbf{u}^D_{1,i}$. The damping elements are not illustrated.

At impact a collision spring $k^c_{s,i}$ acts along the dotted line in direction of $\mathbf{u}^0_{s,i}$ which results into the impact force

$$\mathbf{F}^I_{s,i} = k^c_{s,i,j} \sum_{j=1}^{N-2} \gamma_{s,i,j} \cdot \mathbf{u}^0_{s,i} \,. \qquad (3)$$

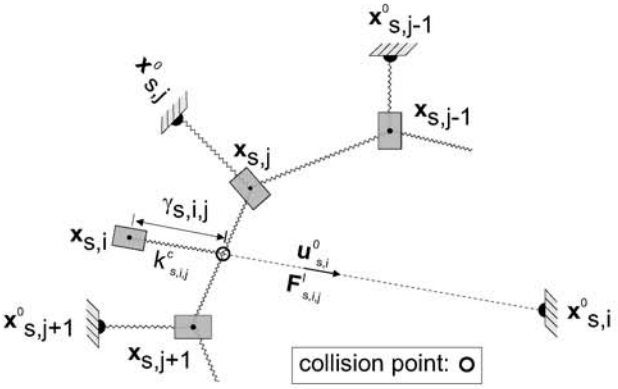


Figure 4: Graphical definition of a collision between the mass $m_{s,i}$ and the spring of the adjacent masses $m_{s,j}$ and $m_{s,j+1}$. The penetration depth $\gamma_{s,i,j}$ and the spring $k^c_{s,i}$ indicate the strength of the collision. The damping elements are not illustrated.

Finally, the horizontal coupling forces

$$\mathbf{F}^H_{s,i} = \sum_{n=1}^{2} r^h_{s,i,n}(\dot{\mathbf{x}}_{s,i} - \dot{\mathbf{x}}_{s,n}) + k^h_{s,i,n}|\Delta\mathbf{x}^h_{s,i,n}|\mathbf{u}^h_{s,i,n} \qquad (4)$$

are considered by additional coupling elements $k^h_{s,i,n}$ and $r^h_{s,i,n}$ ($n = 1, 2$ denotes the left and right coupling string and damping element of mass $m_{s,i}$).

III. RESULTS

A. *Adjustment to high-speed recordings*

The performance of the PE-Model is exemplarily demonstrated by modifying the parameters P_L and $k_{s,i}$ to match the PE-Model to observable pseudoglottis deformations $a(t)$. These pseudoglottis deformations are extracted from the high-speed sequences HS-I and HS-II which had been recorded during the examination of two different laryngectomees.

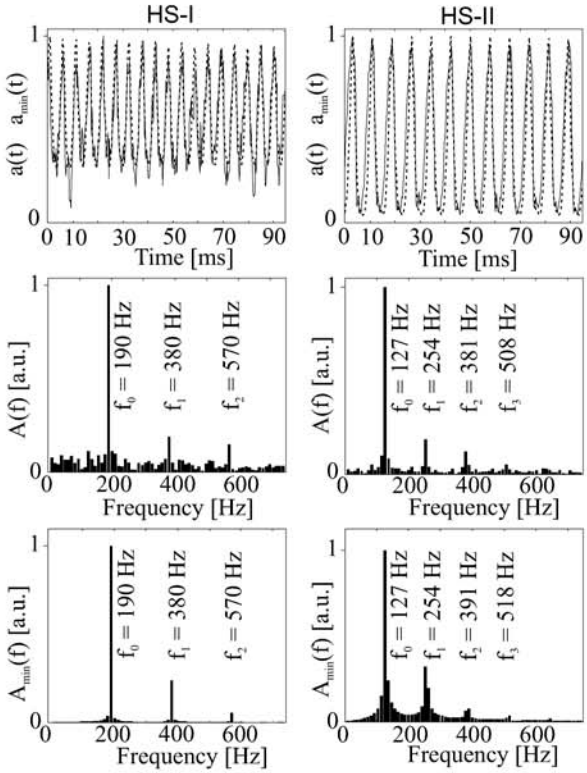


Figure 5: Top: The solid lines show the experimental PE deformations a(t) while the dotted lines show the simulation results $a_{min}(t)$. Middle: Amplitude spectra of experimental PE deformations $A(f)$. Bottom: Amplitude spectra of the modelled PE deformations $A_{min}(f)$.

The parameters of each 2MM within the PE-Model are initially derived by dividing the standard parameter set of Ishizaka and Flanagan [6] by the number of single 2MM used within the PE-Model.

Thus, the dynamic properties of the PE-Modell do not depend on the number of 2MM within the PE-Model. In both cases the horizontal coupling is defined as

$$\begin{aligned} k^h_{s,i,n} &:= (k_{s,i} + k_{s,i\pm1}) \cdot 0.0854 \\ r^h_{s,i,n} &:= (r_{s,i} + r_{s,i\pm1}) \cdot 0.025 \end{aligned} \quad (5)$$

while the springs and masses within different planes show the following relation

$$\begin{aligned} k_{s+1,i} &:= 0.1 \cdot k_{s,i} \\ m_{s+1,i} &:= 0.2 \cdot m_{s,i}. \end{aligned} \quad (6)$$

Finally, the lunge pressure P_L and the spring constants $k_{s,i}$ are manually modified. For HS-I the determined lunge pressure is $P_L = 42.5\,cm\,H_2O$ while the spring constants are $k_{1,i} = 0.0153$. For HS-II the lunge pressure is $P_L = 31.1\,cm\,H_2O$ while the spring constants are $k_{1,i} = 0.0077$. The adaptation results obtained with the two modified parameter sets are shown in Fig. 5. Within the upper two graphs the dotted lines show the simulated PE deformations represented by $a_{min}(t)$ while the solid lines show the experimental PE deformations a(t) extracted from the high-speed recordings during a time interval of 95 ms. The differences between the curves are hardly visible, since the simulated PE vibrations match very precisely the experimentally extracted PE deformations. The amplitude spectra $A(f)$ of both experimental and simulated PE dynamics are shown. The constant components in Fourier-Space are eliminated. Within each spectra characteristic frequencies f_i can clearly be identified. Besides the fundamental frequencies of 190 Hz and 127 Hz the PE-Model simulates successfully the characteristic frequencies f_i.

IV. DISCUSSION

This paper describes a bio-mechanical model which allows the simulation of the substitute voice generating process. Within the PE-Model the two dimensional morphology of the PE segment is considered by coupling orbitally multiple harmonic oscillators with additional spring and damping elements. The PE-Model can successfully be used to model the fundamental characteristics of PE segment vibrations. This is demonstrated by adapting the PE-Model manually to experimental PE segment vibrations which had been extracted from to different high-speed recordings. The simulation results showed identical vibratory characteristics as experimental PE segment vibrations. In summary, this work is the first approach to model PE segment dynamics in order to gain insight into the substitute voice generating process. In a further project a fully automatic adaptation and optimization of the PE-Model to match experimental PE segment vibrations is intended [7] in order to derive physiological parameters of the PE segment. Furthermore, the PE-Model shall be used to investigate the correlation between PE-dynamics and substitute voice quality.

V. ACKNOWLEDGEMENT

The work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 603 (SFB 603, sub-project B5) and DFG AZ EY 15/11.

REFERENCES

[1] J. Kirchner, J. Scatliff, F. Dey, and D. Shedd, "The pharynx after laryngectomy: changes in its structure and function," *Laryngoscope*, vol. 73, pp. 18–33, 1963.

[2] E. Blom, "Tracheoesophageal voice restoration: Origin - evolution - state-of-the-art," *Folia. Phoniatr. Logo.*, vol. 52, pp. 14–23, 2000.

[3] J. Lohscheller, M. Döllinger, M. Schuster, U. Eysholdt, and U. Hoppe, "The laryngectomee substitute voice: Image processing of endoscopic recordings by fusion with acoustic signals," *Meth. Inf. Med.*, vol. 42, no. 3, pp. 277–281, 2003.

[4] J. Lohscheller, M. Schuster, U. Eysholdt, and U. Hoppe, "Investigation of the tracheoesophageal voice generating element by means of active contour models," in *Advances in Quantitative Laryngology, Voice and Speech Research*, 2003.

[5] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal fold model," *J. Acoust. Soc. Am.*, vol. 97, pp. 1874 – 1884, 1995.

[6] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Techn. J.*, vol. 51, pp. 1233–1268, 1972.

[7] M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, and U. Eysholdt, "Vibration parameter extraction from endoscopic image series of the vocal folds," *IEEE T Biomed Eng*, vol. 49, pp. 773–781, 2002.

3rd International Workshop MAVEBA 2003, 23-26

A STUDY OF A DIRECT SPEECH TRANSFORM METHOD ON LARYNGECTOMEE SPEECH

Koji Murakami, Kenji Araki, Makoto Hiroshige[1], and Koji Tochinai[2]
[1]Graduate School of engineering, Hokkaido University, Japan
[2]Graduate School of business administration, Hokkai-gakuen University, Japan

Abstract: **This paper proposes and evaluates a new direct speech transform method with waveforms from laryngectomee speech to normal speech. Most conventional speech recognition systems and speech processing systems are not able to treat laryngectomee speech with satisfactory results. One of the major causes is difficulty preparing corpora. It is very hard to record a large amount of clear and intelligible utterance data because the acoustical quality depends strongly on the individual status of such people. Our proposed method focuses on the acoustic characteristics of speech waveform of laryngectomee people and transforms such characteristics directly into normal speech. The proposed method is able to deal with esophageal and alaryngeal speech in the same algorithm. The method is realized by learning transform rules that have acoustic correspondences between laryngectomee and normal speech. Results of several fundamental experiments indicate a promising performance for real transform.**

Keywords : **Esophageal speech, Alaryngeal speech, Speech transform, Transform rule, Acoustic characteristics of speech**

I INTRODUCTION

Speech is a perfect medium and the most common for human-to-human information exchange because it is able to be used without hands or other tools, being a fundamental contributor to ergonomic multi-modality. Much research have been developed to realize such advantages for human-machine interaction. Many applications are produced and they are recently contributing to human life.

On the other hand, many people who are unable to use their larynxes are not able to benefit from such advances in technology although such assistance is expected. Both esophageal and alaryngeal speech, which laryngectomee people practice to enable conversation, are understandable and enable adequate communication. However, conventional speech processing systems are not able to accept them as inputs because almost all current systems deal with only normal speech. Many intelligible utterances spoken by normal people have to be prepared as learning data to construct useful acoustic models for the systems. It is easy to find a lot of corpora valuable in both quality and quantity in many languages. However, there are not many resources of laryngectomee or other disordered speech because it is very difficult to sample a number of intelligible and clear utterances. One of the major causes is dependence on individual status of speech Thus it is not easy to obtain a high acoustic quality of corpora.

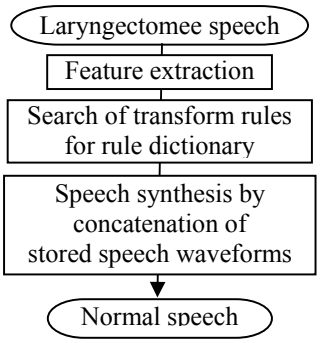


Fig.1 Processing of the proposed method

We focus on laryngectomee speech waveforms themselves to transform them into normal speech. Many studies have attempted to transform laryngectomee speech to normal speech, for example: re-synthesizing the fundamental frequency or formant of normal speech[1], or by utilizing a codebook[2]. We propose a radically different speech transform approach which handles only acoustic characteristics. Fig.1 shows the processing stages of our method. The proposed method is realized by dealing with only the correspondence in acoustic characteristics of speech waveforms. Our basic conception is based on our belief that laryngectomee utterances contain acoustic characteristics although these are inarticulate and quite different from normal speech waveforms. Thus acoustic common and different parts extracted by comparing with two utterances within the same speech side have correspondences of meaning between two different types of speech. We generate transform rules and register them in a translation dictionary. The rules also have the location information of acquired parts for speech synthesis on time-domain. Deciding the correspondence of meaning between two speech sides is the unique condition necessary to realize our method.

In a transform phase, when an unknown utterance of laryngectomee speech is applied to be transformed, the system compares this sentence with the acoustic information of all rules within the speech side. Then several matched rules are utilized and referred to their

corresponding parts of the normal speech side. Finally, we obtain roughly synthesized normal speech utterance by simply concatenating several suitable parts of rules in the normal speech side according to the information of location.

The boundaries of word, syllable, or phoneme are not important for our method because we acquire only acoustic common and different parts as transform knowledge by comparing speech utterances.

We evaluate effectiveness of the transform rules through fundamental experiments and offer discussion on behaviors of the system.

II. LARYNGECTOMEE SPEECH

Laryngectomee people try to acquire esophageal or alaryngeal speech as second speech to enable them to once again communicate effectively in society. The characteristics of these types of speech are explained in this section.

2.1 Esophageal speech

Characteristics of esophageal speech mainly depend on difference of sound source mechanism. Several remarkable features are as follows: lower fundamental frequency than normal speech, including a lot of noise and lower volume[3]. Moreover, differences on prosody and spectral characteristics of speech are also reported[4].

2.2 Aralyngeal speech

Aralyngeal speech has an unnatural quality and is significantly less intelligible than normal speech. The utterances spoken using artificial larynx, are not able to contain any accent and intonation despite the speaker's intention. The cause is that this device is only able to vibrate fixed impulse source. Therefore, it is impossible to express their emotion or intention with speech.

2.3 Speech recognition for laryngectomee speech

We need to reveal the actual performance of usual speech recognition for laryngectomee speech. We utilized Julius[5] as a speech recognition tool. The acoustic and language models in the system were constructed by the learning of normal speech utterances. Table 1 explains the result of recognition performance. It is very clear that the system is not able to treat laryngectomee speech without rebuilding the acoustic model of many esophageal or alaryngeal speech utterances.

Table 1 Results of speech recognition.

Type of Speech	Number of Utterances	Accuracy of correct words[%]
Normal Speech	80	65.82%
Alaryngeal Speech	119	29.61%
Esophageal Speech	107	24.32%

III. SPEECH PROCESSING

3.1 Speech data and spectral characteristics

Various acoustic parameters specific to disordered speech have been developed and applied to many studies[6]. One such study has succeeded to show acoustic differences by a clustering method using these values between normal and disordered female voices[7]. However, we have focused on results of comparison experiments using only spectral analysis [4].

We recorded utterance data with 16bit and 48kHz sampling rate, and downsampled to 16kHz. These data were spoken by three people whose speech is normal, esophageal and alaryngeal, respectively. Table 2 shows parameters adopted for speech processing, and Table 3 shows these speaker's characteristics. In this report, LPC Cepstrum coefficients were chosen as spectral parameter, because we focused on frequency characteristics of speech and could obtain better results than other representations of speech characteristics[8].

Table 2 Parameters for speech processing.

Size of analysis frame	30msec
Frame cycle	15msec
Speech window	Hamming Window
AR Order	14
Cepstrum Order	20

Table 3 Information of speakers.

Type of Speech	Age/Gender	Speaker's feature
Normal Speech	24/male	Student
Alaryngeal Speech	70/male	Operation in 1990
Esophageal Speech	65/male	Operation in 1994

3.2 Searching for the start point of parts between utterances

When speech samples were being compared, we had to consider how to normalize the elasticity on time-domain. We meditated upon suitable methods that would be able to give a result similar to dynamic programming[9] to execute time-domain normalization. We adopted a method to investigate the difference between two characteristic vectors of speech samples for determining common and different acoustic parts. We also adopted the Least-Squares Distance Method for the calculation of the similarity between these vectors.

Two sequences of characteristic vectors named "test vector" and "reference vector" are prepared. The "test vector" is picked out from the test speech by a window that has definite length. At the time, the "reference vector" is also prepared from the reference speech. A

distance value is calculated by comparing the present “test vector” and a portion of the “reference vector". Then, we repeat the calculation between the current “test vector” and all portions of the “reference vector" that are picked out and shifted in each moment with constant interval on time-domain. When a portion of the “reference vector” reaches the end of the whole reference vector, a sequence of distance values is obtained as a result. The procedure of comparing two vectors is shown in Fig.2. Next, the new “test vector” is picked out by the constant interval, then the calculation mentioned above is repeated until the end of the “test vector”. Finally, we can get several distance curves results between two speech samples.

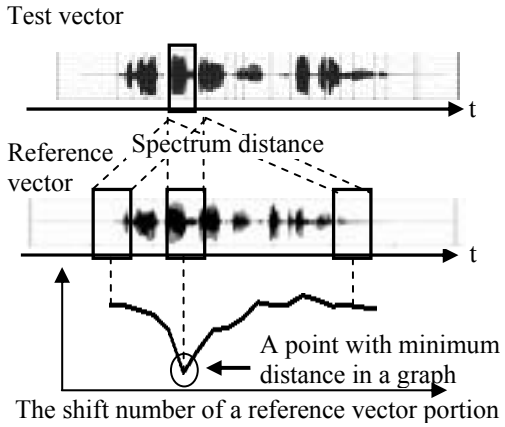


Fig.2 Comparison of vector sequences.

Fig.3 shows an example of the difference between two utterances. This applied speech sample is spoken by the same normal speaker and the contents of the utterances are the same. The horizontal axis shows the shift number of reference vector on time-domain and the vertical axis shows the shift number of test vector, i.e., the portion of test speech. In the figures, a curve in the lowest location has been drawn by comparing the head of the test speech and whole reference speech. If a distance value in a distance curve is obviously lower than other distance values, it means that the two vectors have much acoustic similarity.

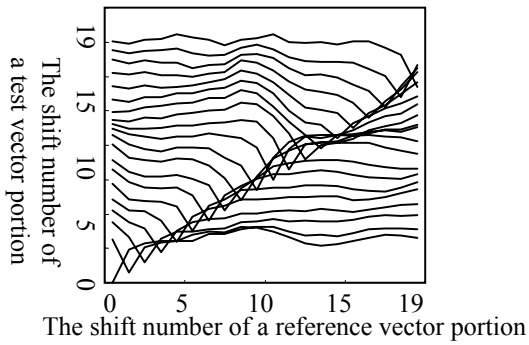


Fig.3 Difference of utterances:”airmail.”

As shown in Fig.3, when the test and reference speech have the same content, the minimum distance values are found sequentially in distance curves. According to these results, if there is a position of the obviously smallest distance point in a distance curve, that point should be regarded as a frame in the “common part” by evaluating the point by a decision method in our previous research[8]. Moreover, if these points sequentially appear among several distance curves, they will be considered a common part. At the time, there is a possibility that the part corresponds to several semantic segments, longer than a phoneme and a syllable.

IV. Generation and Application of Transform Rules

4.1 Acquisition of transform rules

Acquired common and different parts are applied to determine the rule elements needed to generate translation rules. At the time, there are three cases of sentence structure as the “rule types”. If two compared utterances were almost matching or did not match at all, several common or different parts are acquired, respectively. And the other case is that these utterances have both parts at the same time. Combining sets of common parts of both normal and laryngectomee speech become elements of the transform rules for rule generation. The set of common parts extracted from the laryngectomee speech, which have a correspondence of meaning with a set of common parts in normal speech, are kept. The sets of different parts become elements of the transform rules as well.

Finally, these transform rules are generated by completing all elements as below. It is very important that the rules are acquired if the types of sentences in both speech sides are the same. When the types are different, it is impossible to obtain the transform rules and register them in the rule dictionary because we are not able to decide the correspondence between two speech sides uniquely. Information that a transform rule has are as follows:

- rule types as mentioned above
- index number of an utterance in both speech sides
- sets of start and end point of each common and different parts

4.2 Transform and speech synthesis

When an unknown utterance of a laryngectomee person is applied to be transformed, acoustic information of acquired parts in the transform rules are compared in turn with the unknown speech, and several matched rules become the candidates to transform. The inputted utterance should be reproduced by a combination of

several candidates of rules. Then, the corresponding parts of the normal speech in candidate rules are referred to obtain transformed speech. Although the final synthesized normal speech may be produced roughly, speech can directly be concatenated by several suitable parts of rules in the normal speech side using the location information on time domain in the rules.

Table 4 Condition for experiments.

Frame length of test vector	120msec
Frame rate of both vectors	60msec
Margin of time delay	±180ms, ±120ms

V. Rule acquisition experiments

All data in experiments are achieved through several speech processes as explained in 3.1. We applied 80 utterances of each speaker. The system is prepared with the same parameters throughout the experiments between both esophageal or alaryngeal and normal speech to evaluate the generality of the system. The conditions shown in Table 4 are also adopted in these experiments. The rule dictionary has no rule or initial information at the beginning of learning.

We evaluate that the system could obtain a number of useful transform rules created by only the calculation of acoustic similarity. Moreover, location of parts on time-domain is also evaluated because this characteristic expresses the accuracy of correspondence of parts to those in another speech side. We allow a margin for parts appearing in time domain, ±180ms and ±120ms to consider for individual uttering differences. When corresponding parts between two speech sides in a rule appear in appropriate location on time-domain with suitable length, the rule included these parts is regarded as a correct rule because the correspondences are able to be decided uniquely. Table 5 shows a number of acquired rules and those that have appropriate correspondence.

Table 5 Comparison of correspondences of acquired rules.

Speech Data	Num. of Data	Num. of acquired rules	±180ms	±120ms
Alaryngeal ↔ normal	80	2,284	1,665 [73.9%]	1,315 [57.6%]
Esophageal ↔ normal	80	1,378	1,055 [76.6%]	846 [61.4%]

V. Discussion

Many appropriate rules are obtained in both experiments through the same parameters. The results shows common and different parts appear approximately close location on time-domain independent of speech type. They also indicate that calculation of acoustic similarity is able to be a criterion to partition laryngectomee utterances although these are not clear and intelligible and are not able to be dealt with in conventional speech recognition. Therefore, these rules indicate promising possibilities for speech transform. The number of appropriate rules from esophageal speech is lower than from alaryngeal speech. Noises accrued from injecting volumes of air into the esophagus are one of the major causes.

We need to increase the number of speech utterances to obtain more suitable transform rules, and it is also necessary to consider the contents of utterances for more effective rule acquisition and application.

VI. Conclusion and future works

In this paper, we have described the proposed method and have evaluated rule acquisition without being parameter tuning specific for esophageal or alaryngeal speech. We have confirmed that appropriate acoustic information is able to be extracted by calculation of acoustic similarity and that rules have been generated.

We will have to implement transform experiments with a large amount of data, and confirm the synthesized speech in normal speech by listening.

References

[1] Wen Ding and N. Higuchi, "A voice conversion method based on complex RBF network," In *Proc. the 1997 autumn meeting of ASJ(Japanese)*, pp.335-336, 1997

[2] Oyton Turk and Levent M.Arslan, "Subband based Voice Convesation," In *Proc. ICSLP2002*, pp.289-292, 2002

[3] K. Matsui and E. Noguchi, "Enhancement of esophageal speech," In *Proc. the 1996 autumn meeting of the ASJ(Japanese*), pp.423-424, 1996

[4] Jinlin Lu, Y. Doi, S. Nakamura and K. Shikano, "Acoustical Characteristics of Vowels of Esophageal Speech ," *SP96-126*, pp.233-240, 1997

[5] Lee Akinobu, T. Kawahara and K. Shikano, "Julius – a Open Source Real-Time Large Vocabulary Recognition Engine," Proc. EUROSPEECH '01, pp.1691-1693, 2001

[6] Y. Katoh, "Acoustic characteristics of speech in voice disorders," In *Proc. the 2000 spring meeting of the ASJ(Japanese)*, pp.309-310, 2002

[7] Daniel Callan, Ray D. Kent, Nelson Roy and Stephen M. Tasko, "Self-organizing Map for the Classification of Normal and Disordered Female Voices," *Journal of Speech, Language, and Hearing Research*, Vol.43, pp.355-366, April, 1999

[8] K. Murakami, M. Hiroshige, K. Araki and K. Tochinai, "Evaluation of rule acquisition for a new speech translation method with waveforms using inductive learning," In *Proc. ACL-02 Workshop on Speech-to-Speech Translation*, pp.45-52, 2002

[9] H.F.Silverman and D.P.Morgan, "The application of dynamic programming to connected speech recognition,", In *IEEE, ASSP Magazine,* pp.6-25, 1990

3rd International Workshop MAVEBA 2003, 27-30

EXTERNAL EXCITATION OF THE VOCAL TRACT AFTER LARYNGECTOMY

V. Misun
Department of solids bodies, Brno University of Technology, Brno, Czech Republic

Abstract: **The vocal tract, along with the vocal folds, is the organ generating the human voice. The vocal folds alone generate what is called source voice which differs depending on whether a person wants to speak in a loud voice or in a whisper. The patients after laryngectomy are not able to use the source voice for voice generation because their vocal folds are surgically removed. Than it is necessary to use other artificial possibility for source voice generation. The paper deals with the external excitation of the vocal tract, that is without the vocal folds engaged – after totally laryngectomy.**
Keywords : **External excitation, laryngectomy, voice**

I. INTRODUCTION

The vocal tract can also be excited by an external source independent of the vocal folds´ activity. The possibility of the external excitation of the vocal tract appears to be the supply of the compressed air through a jet placed in the sinus nasal – Fig.1.

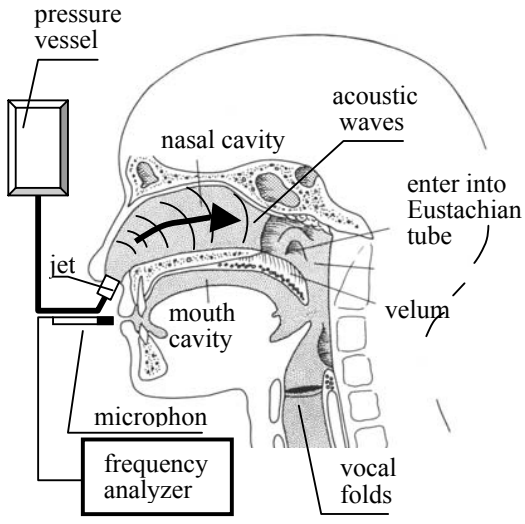


Fig.1 Diagram of external excitation of the vocal tract

The jet [3], [4] has specific geometric parameters, thus it generates noise with the required flat shape of the spectrum within the broadest possible frequency range. This requirement follows from the need to excite at least three formants of each vowel. The flowing air generates the consonants based on the settings of the different parts of the vocal tract.

The source voice for speaking in a whisper is generated by the jet as described in this paper, the jet is replacing the voice prostheses as a result.

II. METHODOLOGY

The diagram of the external vocal tract excitation for speaking in a whisper is presented in Fig.1. The excitation is reached by the compressed air supplied by the jet; in this case the jet is placed in the nostril at the sinus nasal beginning – Fig.1.

The compressed air expands on the jet outlet and generates noise with a continuous spectrum. The acoustic waves generated are transported through the nostrils to the guttural cavity where they excite formants of the vowels studied.

Through a valve the compressed air is let into the jet where the noise is generated with the continuous spectrum. The acoustic waves thus generated pass through the nostril as far as the vocal tract along with the flowing air. These acoustic waves in turn excite the individual formants of the vowel concerned. The flowing air generates the consonants based on the settings of the different parts of the vocal tract.

At this point it is necessary to meet both the conditions (acoustic waves creation and the air flow) so that both vowels and consonants can be generated with the convenient intensity.

The method of external vocal tract excitation can be modified in different way:

- e excitation by means of the external compressed air source (vessel) – Fig.1
- the supply of the compressed air using a hose from the lungs of another individual
- supply of the compressed air from the stoma of the patient himself.

Now we need to emphasize that:

- the consonants must be excited by the flowing air from the rear part of the mouth cavity
- the vowels can be excited by the acoustic waves in any position of the vocal tract; the most convenient is the position of the maximum amplitude of the acoustic mode of the vocal tract cavity. This is on the rear side of the vocal folds to be removed.

III. RESULTS

We will present three cases of the voice generation in a patient after totally laryngectomy. We will compare the vowel spectra and consonant ones to be generated after laryngectomy, further by using the electrolaryng and finally generated by means of the external vocal tract excitation to be defined above.

The voice spectra have been measured in front of the mouth cavity.

Results of the voice spectra :

a) the voice generation after laryngectomy

The patient was trying to speak without any additional aids. In Fig.2 there are vowel spectra presented to be generated step by step in the following order : a, e, i, o, u and during 14 seconds. In Fig.3 there are presented a consonants in the following order : s, ch, f, r, s.

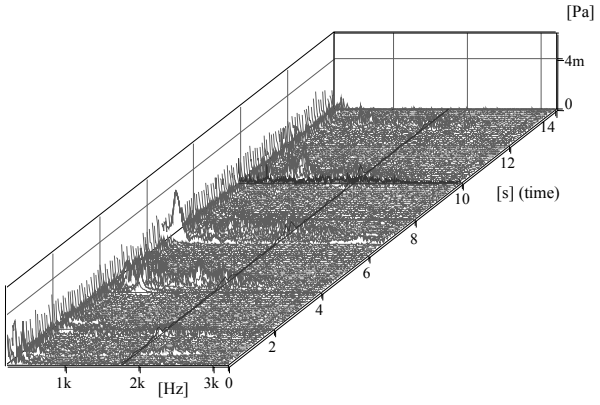


Fig.2 Spectra of vowels generated after laryngectomy

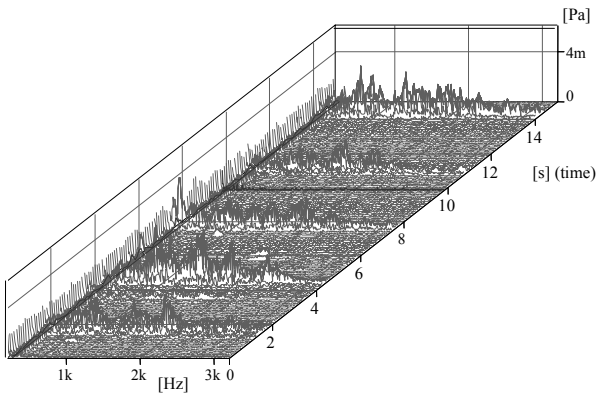


Fig.3 Spectra of consonants generated after laryngectomy

It is clear from these spectra that quality of the voice to be generated is insufficient. The communication of the patient with other people is not satisfactory since the source voice cannot be generated in this case.

b) the voice generation by means of electrolaryng

The spectra of both the vowels and consonants are defined more correctly and more accurately in the following case. The formants of individual vowels are defined more distinctively – Fig.4. In the same way the spectra of consonants are defined more accurately – Fig.5.

More satisfactory spectra are produced as a result of the periodical compression of the vocal tract walls, which is a necessary condition for the vowel formants excitation. In the same way the mouth walls motion causes the air motion in the mouth cavity, which is a condition for the consonants excitation.

c) the voice generation by using the external vocal tract excitation

The individual vowel and consonant spectra excited by an external source voice – see Fig.1, are presented in Fig.6 and others.

From these spectra it is possible to define individual vowel formants correctly and exactly.

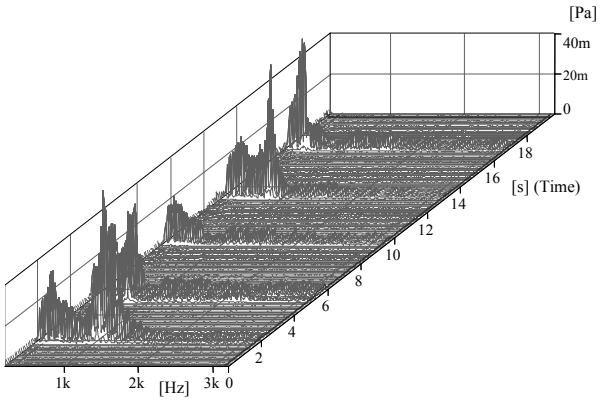


Fig.4 Spectra of vowels: a, e, i, o, u

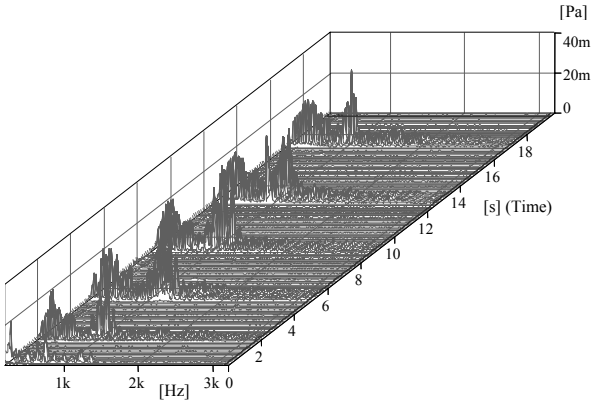


Fig.5 Spectra of consonants: s, ch, f, r, s

The accurate determination of the formants of individual vowel is easy due to the continuous shapes of the spectra.

These spectra correspond to the spectra of the vowels to be generated aloud and not generated in a whisper. It is due to the rear vocal tract section which is closed after laryngectomy.

IV. DISCUSION

The vowel spectra to be excited by an external source voice have the similar shapes as those generated by the people with healthy vocal folds.

The spectra of the noise generated by the jet stimulate the excitation of the vowel formants while the flowing air is a condition for generating the individual consonants. Therefore this method enables the excitation both the vowels and the consonants.

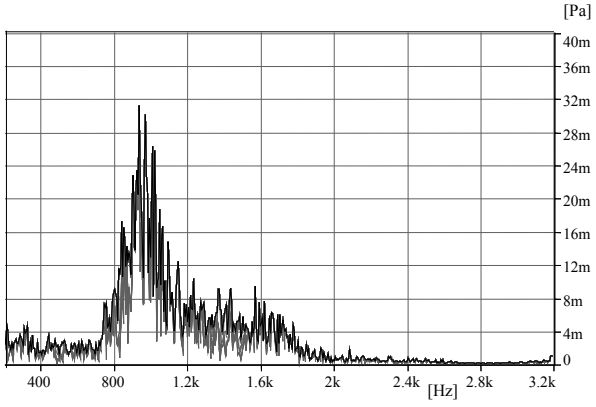


Fig.6 Spectrum of vowel „a“ excited by an external source

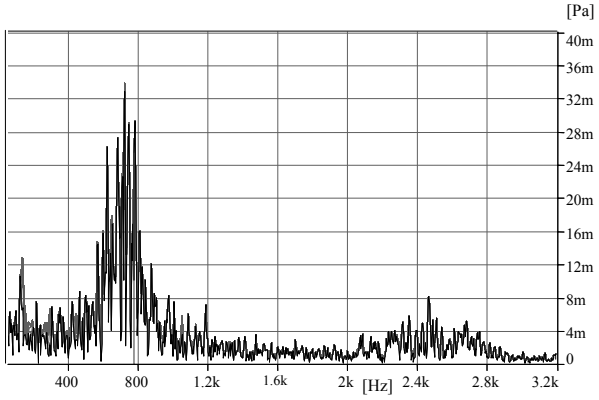


Fig.7 Spectrum of vowel „e“ excited by an external source

It is necessary to remember that the spectra in Fig.6 and the other figures are continuous because the vocal folds prosthesis (jet) have been used for voice generation in a whisper. So that this jet after compressed air expansion was generating the continuous excitation spectrum of the vocal tract.

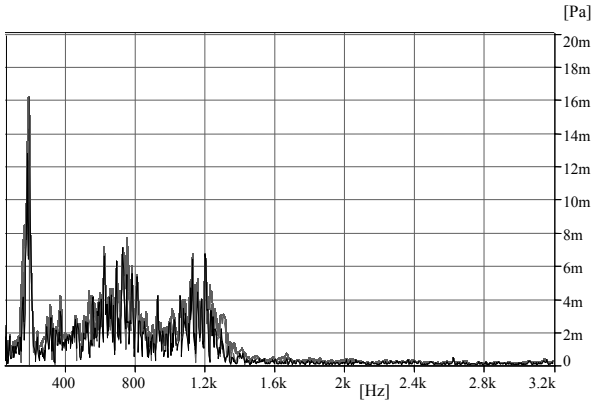


Fig.8 Spectrum of vowel „o“ excited by an external source

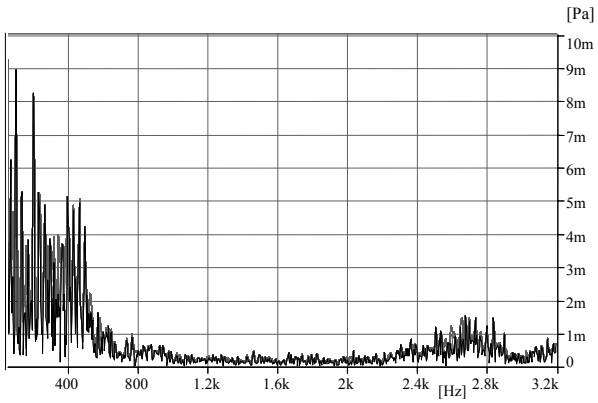


Fig.9 Spectrum of vowel „i“ excited by an external source

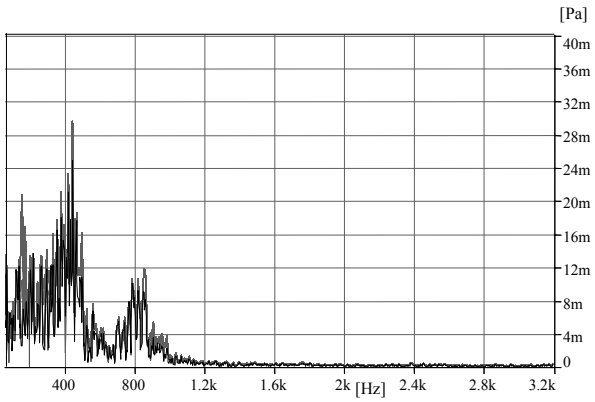


Fig.10 Spectrum of vowel „u“ excited by an external source

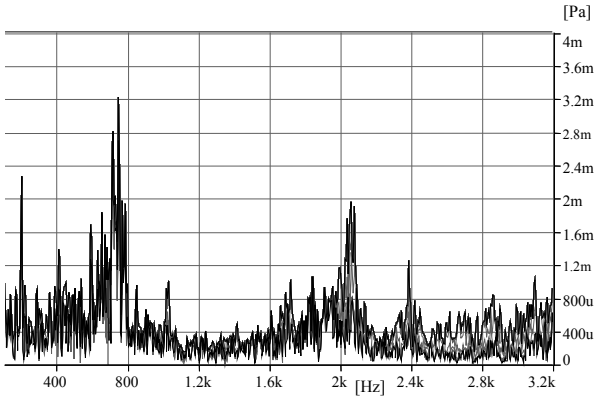


Fig.11 Spectrum of consonant „ch“ excited by an external source

If the vocal folds prosthesis for speaking in loud are used for external vocal tract excitation than the individual vowels have a discrete spectra with corresponding harmonic components structure.

But it is necessary to say at this point that the continuous spectra in Fig.6 and other correspond to the speaking in loud. It is due to the vowel formants position on the frequency axis and which is the same when the healthy vocal folds would be closed. In our case the vocal folds are removed, so that the setting up of the vocal tract is the same in both cases.

This method may be used both by the people with the healthy vocal folds and by the patients who had their vocal folds surgically removed (after totally laryngectomy).

It is apparently rather difficult to acquire this method and apply the principle of the vocal tract external excitation. Still we can state that having gained some experience with it, people are able to communicate satisfactorily using this method with those around them.

V. Conclusion

The paper describes an experimentally verified method of the external excitation of the vocal tract. The external source is the compressed-air supply in this case, coming from an outer source, e.g. from a pressure vessel etc.

The air is supplied by means of a jet which is placed in the nostril at the end of the sinus nasal. This situation does not disconcert or restrict the patient in any way. However the jet must have appropriate properties particularly concerning the flat spectrum shape required, generated by the air leaving the jet and expanding. The spectrum must be generated within the frequency range which is defined by the range of at least three formants of all the vowels.

The method described may primarily provide a means of communication for the patients who had their vocal folds taken out in the total laryngectomy, unless the patient uses a different aid for generating the voice, such as „vocal folds substitutes“ for generating the guttural voice or electrolarynx.

This method is also useful for defining, developing and verifying the functionality of the different vocal folds substitutes, without any surgical intervention needed.

Acknowledgement: The work associated with the preparation of this contribution was supported by the GAČR grant project no. 106/98/K019 and the CEZ Research Plan: 2600001.

References

[1] I.R.Titze, “*Principles of voice production”*, Prentice Hall, Englewood Cliffs, New Jersey 1994

[2] J. M. Pickett, “*The Acoustics of Speech Communication*, Allyn&Bacon, USA, 1998

[3] V. Misun, P. Svancara, P. Janovsky, “Vowel Formants Determination by External Vocal Tract Excitation“, Proc. Eurospeech 2003 – Switzerland, September 1-4, 2003, Geneva, Switzerland, pp. 4

[4] V. Misun, “External excitation of the vocal tract through the sinus nasal“, Proc. wc2003 – World Congress on Medical Physics and Biomedical Engineering, Sydney, Australia, August 24-29, 2003, pp. 4

[5] V. Misun, “The influence of the glottal aperture on the resonance properties of the vocal tract”, Proc. of the 5th International Workshop Advances in Quantitative Laryngoscopy, Voice and Speech Research, Groningen, Holand, 2001

[5] V. Misun, “Modelling of the Vocal Folds Function”, Proc. of EMBEC’02 (Europen Medical&Biological Engineering Conference), Vienna, (4 – 9) December, 2002, pp. 242-243.

[6] V. Misun, “Vocal Folds Function Modelling by means of the Principle of Compressive Air Bubble”, Proc. of ICVPB (Inter. Conf. on Voice Physiology & Biomechanics), Denver, Colorado, USA, (14 – 16) September, 2002, pp.1-4

[7] V. Misun, „The principle of pressure air bubble within the fonation“,. *Proc.Conf. Interaction of dynamic systems*, Praque, 2001, (in Czech)

[8] V. Misun, K. Prikryl, „The modeling of the vocal folds function“. *Proc. 4rd International acoustic conference*, Kocovce, pp. 48-53, 2001, (in Czech)

[9] K. Prikryl, V. Misun, „Phonation simulation of 2D vocal folds model using the FEM“, *Proc. of 3rd Inter.Conf. Mechanics, Robotics and Biomechanics*, Trest, 2001, (in Czech).

[10] J.A. Seikel, D.W. King, D.G. Drumright, “*Anatomy and Physiology for Speech, Language and Hearing”,* Sing.Publish.Group, San Diego, 2000

Special session on

Infantry cry analysis

3rd International Workshop MAVEBA 2003, 33-34

TIME VARIATIONS OF THE FUNDAMENTAL FREQUENCY (MELODY) AND RESONANCE FREQUENCIES IN INFANT'S CRYING – KEY PARAMETERS FOR PRE-SPEECH DEVELOPMENT

Werner Mende[1] & Kathleen Wermke[2]

[1]Berlin-Brandenburg Academy of Science, Berlin, Germany, mende@bbaw.de
[2]Center for pre-speech development & developmental disorders, Department of Orthodontics, Julius-Maximilians-University Würzburg, Germany

The ability to perceive and to produce the time varying fundamental frequency (melody) is an extremely important component of auditory information and a fundamental aspect of language. The fundamental frequency is an essential parameter of prosody.

In adult language perception, prosody can guide the syntactic analysis of spoken sentences [1]. Concerning infant language perception it was shown that young infants recognize utterances in their language based on prosodic cues before they become sensitive to its segmental characteristics [see review in 2]. Speakers use the F0 modulation to stress particular elements in an utterance or to indicate the beginning or end of a syntactic phrase.

Recently, Drayna et al. [3] demonstrated in a twin study the influence of genes on the ability to recognize correct pitch and melodies. They could show that the perception of pitch is highly heritable. Research examining patients with brain damage has indicated that melodic information may be processed primarily by a cortical system in the right hemisphere. A close link between the processing of melodies and the processing of language has been demonstrated in a recent study by Maess et al. [4] who found that music processing involves a neural network normally seen to be active during language processing. This finding strongly supports a direct relationship between the processing of language and music from a functional and neuroanatomical view.

The importance of F0 and related parameters is also well described for infant's and children's sound production. The importance is not only given by research results in the framework of "cry-diagnosis", but also by findings within the field of pre-speech development and language acquisition [5-10]. Moreover, the interaction between laryngeal (melody) and pharyngeal (resonance frequencies) activity is one of the key aspects for pre-speech research [e.g. 11]. Tuning processes between the cry melody and resonance frequencies are preparatory activities for an intentional articulation in speech.

BRIDGING FROM CRYING OVER BABBLING TO SPEECH BY USING A MATHEMATICAL MODEL

Melody (fundamental frequency (F0) as a function of time) is one of the essential features of prosody. We would like to outline, that different resolutions in respect of time and frequency as well as in the degree of smoothness are possible for melody analysis and sketch our approach to reduce the melody curves to smooth, simple arcs or plateaus with only one maximum (mono-modal melodies). The smoothing is strong enough to have none or only one inflection point on each of the increasing/ decreasing flanks of the melody. In this sense, we use the term "shape" of the melody.

Cry and babbling melody analysis are only qualitatively investigated so fare, because a suitable modelling-approach was lacking. Analysis of single F0 values, measured at only a few marked points of the utterances, common in pre-speech research, are not sufficient for melody analysis. The diversity of melody curves of infant's utterances is reducible to a manageable extent by the use of a minimal-parametric model. Our theoretical model has the form of double power-law with a non-linear kernel $Y^{\gamma}(1-Y)^{\delta}$. The first exponent $\gamma=\alpha*\beta$ expresses concavity/ convexity of the increasing flank. The exponent $\delta=(1-\alpha)*\beta$ expresses the concavity/ convexity of the decreasing flank. Both are power laws. This model allows separating the asymmetry of the arcs from kurtosis properties. The model is minimal-parametric in the sense, that only two parameters (γ, δ) are scale-independent and it characterizes the shape of the melodies completely. The model describes the majority of produced melodies accurately and with amazing ease.

The application of the model to melodies allows reliable comparisons and evaluations of intra- and inter-individual differences. Therefore, the application of the proposed melody-shape-model to cries, babbling and speech sounds could considerably improve comparative studies on melodies by starting from a quantitative representation of the melody. It will be possible to define normal values and to measure objectively deviations from the norm.

A future significance of the melody analysis in combination with the study of resonance properties of the infant's vocal tract we see in the following fields: Firstly, this approach seems to be promising for the field of "Cry diagnosis" with respect to the possibility to develop quantitative scores. Analogous scores for normative values of melody features of early babbling and speech sounds would be very helpful within the framework of studies of specific language disorders. The application of this or similar models can also help to develop early diagnosis tools under the presupposition that infants at-risk for the development of specific language disorders are different with respect to prosodic features including melody features, rhythmical characteristics (time shrinking and expansion of melody) and intensity-melody interaction at very early ages.

REFERENCES

[1] K. Steinhauer, K. Alter and A.D. Friederici, „Brain potentials indicate immediate use of prosodic cues in natural speech processing," Nat. Neurosci. 2, pp. 191-196, 1999.

[2] P.W. Jusczyk, The discovery of spoken language. MIT Press, Cambridge, 1997.

[3] D. Drayna, A. Manichaikul, M. de Lange, H. Snieder and T. Spector, " Genetic correlates of musical pitch recognition in humans." Science 291, pp. 1969-1972, 2001.

[4] B. Maess, S. Koelsch, T.C. Gunter and A.D. Friederici, „ Musical syntax is processed oin Broca's area: an MEG study," *Nat. Neurosci.* 4, pp. 540-545, 2001.

[5] P. Lieberman, "The acquisition of intonation by infants: Physiology and neural control," in *Intonation in Discourse*, C. Johns-Lewis, Ed. San Diego: College-Hill Press, 1986, pp. 239-257.

[6] B. Boysson-Bardies, *How language comes to children; from birth to two years.* Cambridge, MA/London: A Bradford Book, MIT Press, 1999.

[7] J. L. Locke, *The child's path to spoken language*. Cambridge, MA/London: Harvard University Press, 1995.

[8] W. Mende, K. Wermke, S. Schindler, K. Wilzopolski, and S. Hoeck, "Variability of the cry melody and the melody spectrum as indicators for certain CNS disorders," *Early Child Develop. Care,* 65, pp. 95-107, 1990.

[9] K. Wermke and W. Mende, "Ontogenetic development of infant cry- and non-cry vocalizations as early stages of speech abilities," in *Proceedings of the 3rd congress of the ICPLA, 9.-11.8.93, Helsinki/ Finland*, R. Aulanko and A.M. Korpijaakko-Huuhka, Eds. Helsinki: University Press, 1994, pp. 181-189.

[10] K. Wermke, W. Mende, H. Borschberg, and R. Ruppert, "Voice characteristics of prespeech vocalizations of twins during the first year of life," in *Pathologies of Speech & language: Contributions of Clinical Phonetics & Linguistics*, New-Orleans, LA: ICPLA, pp. 1-8, 1996.

[11] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," *Medical Engineering & Physics* 24, pp. 501-514, 2002.

3rd International Workshop MAVEBA 2003, 35-38

RESONANCE DEVELOPMENT AND FORMANT TUNING PHENOMENA IN INFANT'S CRYING

Claudia Manfredi[1], Werner Mende[2], Pierro Bruscaglioni[3], Kathleen Wermke[4]

[1]Dept. of Electronics and Telecommunications, Faculty of Engineering, University of Firenze, Italy
[2]Berlin-Brandenburg Academy of Science, Berlin, Germany
[3]Dept. of Physics, Faculty of Mathematics, Physics and Nat. Science, University of Firenze, Italy
[4]Center for pre-speech development & developmental disorders, Department of Orthodontics, Julius-Maximilians-University Würzburg, Germany

***Abstract*: The tracking of resonance frequencies and the analysis of their interaction with the fundamental frequency (F0) allows a description of (pre-) articulatory activity in very young infants. Subjects are six healthy infants. Spontaneous cries were recorded weekly from the 4th until the 20th week. For resonance frequency estimation a spectral parametric technique was applied, which was based on autoregressive models whose order is adaptively estimated on subsequent signal frames [1]. Cry melodies exhibiting different degrees of complexity (e.g. single-arc-melodies, multiple-arc-melodies) were selected for analysis. We found that resonance (formant) tuning occurs much earlier than expected. Here we demonstrate the early occurrence of a tuning between resonance frequencies and the cry melody in infants from 8 weeks onward. A more intense tuning between the melody and the lower resonance frequencies was found beginning about the 2nd / 3rd month. This tuning is interpreted as an early articulatory activity in infant's crying. In a broader perspective it is attributed to a language-related behaviour preparing formant tuning in speech. Medical applications are seen for infants with disturbances of the vocal tract transfer function, e.g. infants with cleft-lip-palate.**

***Keywords*: cry melody, vocal tract resonance, formant analysis, pre-speech development**

I. INTRODUCTION

In a preceding paper [2] we have outlined both, the high control capacity of mechanisms underlying laryngeal sound production in infants, and the interaction between laryngeal (melody) and pharyngeal (resonance frequencies) activity. The results of this earlier study provide good reasons to consider in more detail the resonance properties of the infant's vocal tract during the earliest phases of pre-speech development.

The hypothesis that cry melody patterns might be direct precursors of melodic features of speech is not new [3-8]. Meanwhile there is good evidence that the development of certain cries (mitigated cries) serves as a preparatory activity for language acquisition [9-11]. Tuning processes between the cry melody and resonance frequencies need a certain training - period before they are at disposal for intentional use, e.g. imitating surrounding speech sounds at the babbling age. Starting about the fourth month of life a rapid expansion of non-cry vocalizations (marginal babbling) occurs, including many vowel-like sounds and near-syllables [12, 13]. So, we should expect that intentional articulatory activity is developed well before this age.

II. METHODOLOGY

SUBJECTS: We investigated six healthy, term-born German infants. All infants were without clinical history of pre- and postnatal illness and free of clinical signs of developmental or hearing disorders.

DATA ACQUISITION: Spontaneous cries of all six infants were recorded in weekly intervals from the 4th to 20th week. Cries were recorded in home environment by trained persons using a SONY-DAT-recorder (TCD-D100). The sampling frequency was 48 kHz and the amplitude resolution was 16 Bit.

DATA ANALYSIS: A set of 100 harmonic cries with a high signal-to-noise ratio was selected for analysis out of a total amount of 2000 recorded cries. Cry analysis was performed in a first step by an evaluation of broad-band and narrow-band spectrograms made with a CSL-4300-Model (Kay Elemetrics Corp., NJ/ USA). In-depth data processing was performed by means of a software tool developed on a PC under Matlab 5 environment at the Dept. of Electronics and Telecommunications, Faculty of Engineering, University of Firenze, Italy.

Fundamental frequency F0 is estimated by means of a robust two-step procedure [14]. As for formant estimation, the parametric AutoRegressive (AR) approach is applied. This method is particularly suited for newborn infant cries, which are characterised by higher resonance frequencies than those of adults. Many criteria have been defined for finding the best model order p, including both the estimated variance σ^2 and the model

complexity p in one statistics. The DME (Dynamic Mean Evaluation) model order selection criterion is applied here to the decreasing sequence of variance values, on subsequent data frames of varying length [1]. For comparisons, also fixed model orders were tested. We got better formant tracking than traditional approaches [1, 15].

In the figures presented here resonance frequencies estimated with the AR-method are shown, as we found this method producing the most coherent resonance tracks. Note that the resonance frequencies in infant cries (roughly seen as spectrographic amplitude enhancements) are in most cases not yet identical to formants of speech sounds. We call these resonances "R1", "R2" and "R3", because it is not yet known how they are related to the vowel formants in later speech.

In order to visualize the interaction between melody and time varying resonance tracks we made a special diagram. This diagram contains a background pattern with the melody and the corresponding harmonics (F0 ≡ first harmonic) together with the resonance tracks. This representation is well-suited to assess relations between resonance frequencies and harmonics of the melody up to the 7th harmonic of the melody.

III. RESULTS

Here we present typical examples of melodies and the corresponding spectral resonance functions during crying for the age period 8 – 14 weeks. The selected examples demonstrate also developmental changes of tuning processes. In the oral presentation we will present developmental sequences of the mentioned tuning processes for all infants. We will show both, the lack of such tuning in crying of the youngest infants and the step by step development of melody–resonance–tuning in older infants. We found between three and four main resonance frequencies up to 10 kHz within the age range under investigation.

During the first weeks of life, the resonance frequencies (particularly R1) were relatively constant without movements over the central part of the cry. At the age of about 8 weeks already a partial tuning between the first resonance frequency (R1) and the melody is observable.

We could observe relatively longer periods of a strong resonance, where the resonance tracks take a course closely following a certain harmonic of the melody. In contrast to this, there were relatively fast transitions of resonance frequencies from one harmonic to the other. This fact allows us to conclude that there exists a longer time period of coupling of the resonance movement and the melody, which can be interpreted as the action of a neuro-physiological tuning mechanism between cry melody and resonance frequencies.

Fig. 1 displays the first two resonance tracks (R1, R2) together with the cry melody and its harmonics. At the maximum of the first melody arc (at about 0.85 sec), R1 and R2 show a strong resonance peak around the 5th harmonic.

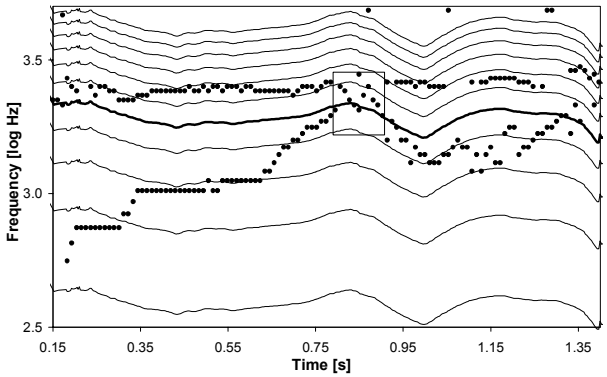


Fig. 1: First two resonance frequencies (black points) of a mitigated cry from a healthy infant at the age of 8 weeks. They are displayed together with the cry melody (lowest line) and its harmonics. R1 and R2 show a conspicuous convergence toward a punctuated resonance (rectangle) at the 5th harmonic of F0 (bold line).

R1 is moving step-wise from a resonance near the second harmonic at the beginning of the cry to the third harmonic. For about 30 ms R1 is fairly well-tuned with the third harmonic and is then moving toward the 5th harmonic at the maximum of the first melody arc. R2 is relatively constant, exhibiting a resonance tuning at the 7th harmonic for about 200 ms (0.4 – 0.6 sec); then R2 moves to the 6th harmonic. Note that R2 seems to support the R1-melody-tuning by a short down shift to the 5th harmonic exactly at the time point of the maximum of the first melody arc. The coordinated action of R1 and R2 seems to stabilize the melody at its apex and produces a punctuated resonance. This is interpreted as a pre-articulatory training process, which indicates an active neuro-physiologically controlled tuning. However, the infant did not repeat the tuning in the second melody arc. Although at the age of 10 weeks already more complex melodies (multiple-arc-melodies) are produced, we selected for reasons of comparability again cries consisting of a double-arc melody (Fig. 2a, b). In contrast to the punctuated resonance in Fig. 1 (8th week), the resonance track "R1" is coupled to the 6th harmonic over the whole cry (lasting resonance). Only in the transition region the melody-arc-tuning is lost, but immediately with the beginning of the second melody-arc the resonance tuning at the 6th harmonic occurs again. R2 shows an independent course in relation to the melody in both cries (Fig. 2a, b). Both cries of the infant from the same day exhibit the same tuning phenomenon between the melody and R1. The regular recurrence of tuning

events supports the assumption that the observed tuning is not by chance, but a controlled behaviour.

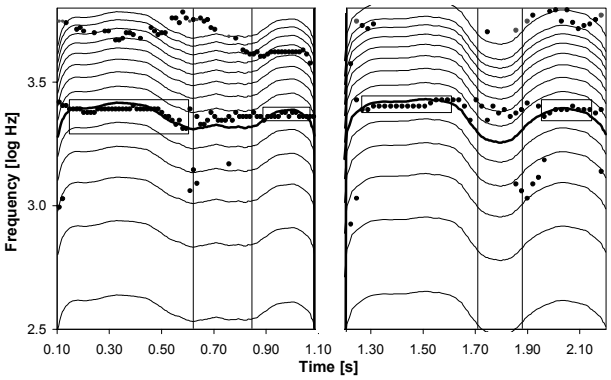


Fig. 2a, b: Resonance frequencies (black points) of two double-arc cries (a, b) from a healthy infant at the age of 10 weeks. In both cries the resonance frequency "R1" is coupled to the 6th harmonic of F0 (bold line). In the transition regions (within vertical lines) between both melody-arcs the tuning is lost (tuning is indicated by a rectangle).

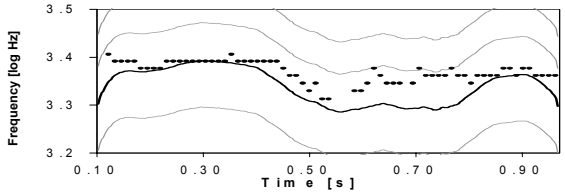


Fig. 3: Enlargement of the region 3.2–3.5 log Hz of Fig. 2a (source spectrum corrected).

In Figure 3, the frequency region between 3.2–3.5 log Hz is zoomed in, in order to demonstrate the good tuning. Note that in Figure 3 a rather precise coincidence of melody and R1 in the strongest resonance regions results when the necessary frequency correction is made taking into account the slope of the laryngeal source spectrum. We did that after we observed that in the strong resonance regions (i.e., a close coupling of melody and resonance) the resonance frequency is mostly situated a certain ratio (approximately 13%) down a harmonic of the melody (Fig. 1, 2a, b, 4). We believe that this small but significant discrepancy of frequency is due to the spectral amplitude slope of the harmonics of the laryngeal source signal.

In Fig. 4b, a selected example of a cry from an older infant (14th week) demonstrates a well-developed tuning between the first two resonance frequencies with harmonics of the melody (Fig. 4b).

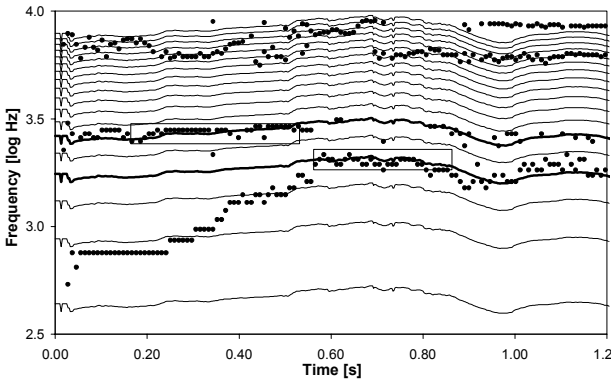


Fig. 4: Results of the resonance frequency tracking of a mitigated cry from a healthy infant at the age of 14 weeks. The first two resonance tracks (black points) are displayed together with the cry melody (lowest line) and its harmonics. This example demonstrates a well-developed tuning (rectangles) between R1 and the 4th harmonic and R2 and the 6th harmonic (bold lines). Tuning of higher resonances with the melody is hardly possible, because the harmonics are too dense at higher frequencies.

IV. Discussion

The results of previous studies [e.g. 16, 17] of resonance frequencies of infant cries are difficult to compare to the present results, because former studies provide only averaged values for resonance frequencies (formants) in pre-speech utterances. In contrast, the present study provides time functions of the resonance frequencies and investigates the interaction between these and harmonics of the melody. Our approach (tested in a preliminary study with twins [2]) allows investigating an (pre-) articulatory activity at a very early age and enables us to characterize developmental processes directed toward language acquisition.

We could confirm our former results concerning developmental changes, but the more fine-grained analysis applied here (coupling analysis of resonance frequency and melody) allows discovering the tuning process in infant's crying even earlier in life. In the preceding study [2] we found a coupling of the lower resonance frequencies to the melody or its harmonics in infant cries beginning at the age of 15 – 17 weeks of life. In the present study we observed tuning processes between the resonance frequencies and the melody, at least during short parts of the cry, much earlier (8th week). At the age of 8 weeks we observed already coordinated actions of R1 and R2 during short parts of the melody. This behaviour seems to stabilize the melody at its apex. Later stages of development are mainly characterized by longer well-tuned times between the first two resonance frequencies and the melody. Already two weeks later the resonance frequency "R1" is coupled to a harmonic over the whole cry. Only in the transition

region between two melody-arcs the tuning is lost. At the age of 14 weeks, a well-developed tuning between the first two resonance frequencies with the melody regularly occurs and reflects a neuro-physiological maturation.

V. Conclusion

The regularly occurring tuning as observed is probably a result of four factors: Firstly, the preceding "training" at earlier weeks. Secondly, the anatomical restructuring of the supra-laryngeal vocal tract at about 3 months [4], thirdly, a better control of sub-glottal air pressure at about 3 months of life [21], and fourthly, a co-ordination between laryngeal and pharyngeal activity. At this age also more voluntary phonation occurs and the infant has to co-ordinate and exercise sub-glottal air pressure and laryngeal - pharyngeal control. The "training"- idea, suggested by Philip Lieberman already in 1986, is confirmed by our results of the present study. These tuning processes have undoubtedly a preparatory function for intentional articulatory activities at later ages. The analysis presented here supports strongly the assumption of a continuous development from the first infant's sound productions to speech. So, this findings further support that cry development is an integral part within pre-speech development.

References

[1] A. Fort and C. Manfredi, "Acoustic analysis of newborn infant cry signals," Medical Engineering & Physics, 20, pp. 432-442, 1998.

[2] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," *Medical Engineering & Physics* 24, pp. 501-514, 2002.

[3] M. M. Lewis, Infant Speech: A study of the beginnings of language. Harcourt Brace, New York, 1936. Reprinted: Arno Press, New York, 1975.

[4] P. Lieberman, "The physiology of cry and speech in relation to linguistic behavior," in Infant crying. Theoretical and research perspectives, vol. 2, B.M. Lester and C.F.Z. Boukydis, Eds. New York: Plenum Press, 1985, pp. 29-57.

[5] K. Wermke and W. Mende, "Ontogenetic development of infant cry- and non-cry vocalizations as early stages of speech abilities," in *Proceedings of the 3rd congress of the ICPLA, 9.-11.8.93, Helsinki/ Finland*, R. Aulanko and A.M. Korpijaakko-Huuhka, Eds. Helsinki: University Press, 1994, pp. 181-189.

[6] W. Mende, K. Wermke, S. Schindler, K. Wilzopolski, and S. Hoeck, "Variability of the cry melody and the melody spectrum as indicators for certain CNS disorders," *Early Child Develop. Care,* 65, pp. 95-107, 1990.

[7] J. L. Locke, *The child's path to spoken language*. Cambridge, MA/London: Harvard University Press, 1995.

[8] M. M. Vihman, *Phonological Development. The Origins of Language in the Child*. Oxford/Malden: Blackwell Publishers, 1996.

[9] K. Wermke, W. Mende, H. Borschberg, and R. Ruppert, "Voice characteristics of prespeech vocalizations of twins during the first year of life," in *Pathologies of Speech & language: Contributions of Clinical Phonetics & Linguistics*, New-Orleans, LA: ICPLA, pp. 1-8, 1996.

[10] K. Wermke, Ch. Hauser, G. Komposch, and A. Stellzig, "Spectral analysis of prespeech sounds (spontaneous cries) in infants with unilateral cleft lip and palate (UCLP): a pilot study," *Cleft Palate Craniofac J.*, 39 (3), pp. 285-294, 2002.

[11] K. Wermke, "*Untersuchung der Melodieentwicklung im Säuglingsschrei von monozygoten Zwillingen in den ersten 5 Lebensmonaten,*" PhD Thesis, Humboldt-University of Berlin/ Germany, 2002

[12] B. Boysson-Bardies, *How language comes to children; from birth to two years.* Cambridge, MA/London: A Bradford Book, MIT Press, 1999.

[13] D. K. Oller, *The emergence of the speech capacity*. Mahwah/London: Lawrence Erlbaum Associates Publishers, 2000.

[14] C. Manfredi, "Adaptive noise energy estimation in pathological speech signals," *IEEE Transactions on Biomedical Engineering*, vol. 47, n.11, p.1538-1542, 2000.

[15] A. Fort, A. Ismaelli, C. Manfredi, and P. Bruscaglioni, "Parametric and non-parametric estimation of speech formants: application to infant cry," *Med. Eng. Phys.*, 18 (8), pp. 677-691, 1996.

[16] H.L. Golub and M.J. Corwin, "Infant cry: a clue to diagnosis," *Pediatrics*, 69, pp. 197-201, 1980.

[17] M.P. Robb, Y. Chen, and H.R. Gilbert, „Developmental Aspects of Formant Frequencies and Bandwidth in infants and toddlers," *Folia Phoniatr Logop.*, 49, pp. 88-95, 1997.

[18] P. Lieberman, "The acquisition of intonation by infants: Physiology and neural control," in *Intonation in Discourse*, C. Johns-Lewis, Ed. San Diego: College-Hill Press, 1986, pp. 239-257.

3rd International Workshop MAVEBA 2003, 39-42

USING EARLY VOCALIZATION ANALYSIS FOR VISUAL FEEDBACK

H. J. Fell[1], J. MacAuslan[2], C. J. Cress[3], L. J. Ferrier[4]
[1]College of Computer and Information Science, Northeastern University, MA, USA
[2]Speech Technology and Applied Research, Lexington, MA USA
[3]Communication Disorders Department, University of Nebraska – Lincoln, NE, USA
[4]Department of Speech Language Pathology and Audiology, Northeastern University, MA, USA

Abstract: **The Early Vocalization System (EVA) applies the Stevens landmark theory to infant vocalizations (babbles). The landmarks are grouped to identify syllable-like productions in these vocalizations. The visiBabble system processes vocalizations in real-time. It responds to the infant's syllable-like productions with brightly colored animations and records the landmark analysis. The system reinforces the production of syllabic utterances that are associated with later language and cognitive development. We report here on the development of the visiBabble prototype and our initial field-testing.**

Keywords : **acoustic analysis, babbles, landmarks**

I. INTRODUCTION

Communication skills are vital to educational and vocational success. Cerebral palsy, developmental apraxia (DAS), neurological insult/injury (e.g. head injury, encephalitis, meningitis), oral/motor dysfunction, cognitive impairments, tracheotomy, and deafness can all cause a child to be at risk for being non-speaking. A child having any of these or other syndromes may not be able to produce a sound when he or she wants to, may produce a limited range of sounds (often vowels and 1-2 consonants), or may not have learned to associate his or her sounds with meaningful referents [2]. During an intervention to promote speech-like vocalizations, non-speaking children tended to have difficulty initiating sounds and participating in vocal imitation play. They produced atypical sounds such as elongated vowels, distorted consonants, and non-speech sounds.

Because of the atypical sound production of infants in this population [8], traditional intervention strategies to prompt or respond to infant vocalizations may not be sufficient to promote change. Children at risk for being nonspeaking may produce a higher percentage of vowel-like sounds (*vocants*) and consonant-like sounds (*closants*) during later development than would be expected for typically developing children. Without strategies to detect and respond appropriately to these sound approximations, listeners may not be able to tailor their activities and responses appropriately to children's sound productions.

There is considerable research to support the position that infant vocalizations are effective predictors of later articulation and language abilities [7, 10, 12]. These studies have been carried out on normally developing children and on children with a variety of early diagnosed problems. These research studies emphasize the importance of early speech intervention for children at risk for being non-speaking. They also point out the difficulty of providing sufficient speech practice and feedback for children with such atypical speech patterns through traditional forms of intervention and interaction.

Closants and oral-cavity openings can be detected in the sound waveform from acoustic evidence of discontinuities in the spectrum of sound. These discontinuities have been called landmarks by some researchers of adult speech [9, 13]. Landmarks that result from the creation or release of a narrow constriction or closure along the vocal tract are also found in pre-linguistic vocalizations. We can hypothesize that the development of the ability to produce sounds exhibiting landmarks is a necessary skill underlying the production of syllables.

Vocants appear early in the vocalizations of infants and are characterized by slowly time-varying spectral patterns. These sounds result from movements of the tongue body, the jaw, and the lips, and are usually produced with the vocal folds positioned to vibrate. A variety of vowel-like sounds appear as the infant learns to control the positioning of these articulators. [1].

As babbling develops, the infant begins to coordinate control of the vocal folds and the velopharyngeal opening with control of the tongue blade and the lips, and the true consonants appear. In the landmark model, the larynx and the velum are considered secondary articulators, and they are "bound" to control by the primary articulators, in that implementation of the laryngeal and nasal features depends, in some ways, on the implementation of the primary articulator. This landmark model has proved useful in various applications concerning adult speech and has been successfully applied to analysis of infant vocalizations [3, 4, 5]. This analysis has, in turn, been used to formulate a "vocalization age" that clinically distinguishes between typically developing infants and infants at risk for later speech difficulties [6]. A vocalization age is a normative age-equivalence estimate

of the range of speech sounds (landmark sequences) expected for typically developing children.

The visiBabble system processes vocalizations in real-time. It responds to the child's syllable-like productions with brightly colored animations and records the landmark analysis. The system reinforces the production of syllabic utterances that are associated with later language and cognitive development. As a child interacts with visiBabble, the program collects and analyzes the infant's utterances so that it can be used by a child as a toy/trainer or as a clinical or research implement.

II. METHODOLOGY

A. The visiBabble System

The visiBabble system includes a modern notebook computer (Dell Inspiron, 2.4 GHz Pentium 4 running Windows XP), a microphone, a 15" flat-panel display, and software, which carries out the following functions:

- Landmark detection – detects landmarks in a child's vocalizations in real-time.
- Graphic feedback -- provides real-time visual response to sound input;
- Data collection – records each session and saves the result as a wav file, collects data on the types and duration of vocalizations produced;
- Experimental formats -- allows the system to run and data to be collected in single-case study formats.

B. Finding Landmarks

Our landmark detector is based on Stevens' acoustic model of speech production [13]. Central to this theory are landmarks, points in an utterance around which listeners extract information about the underlying distinctive features. They mark perceptual foci and articulatory targets. The program detects three types of landmarks:

glottis: marks the time when the vocal folds start (+g) and stop (-g) vibrating;

sonorant: marks sonorant consonantal closures (-s) and releases (+s) (e.g., voiced closants);

burst: designates stop/affricate bursts (+b) and points where aspiration/frication ends (-b) due to stop closure.

The visiBabble system can track simple aspects of the acoustic signal in real time, based on a low-resolution spectrogram. That is, the signal is sampled at 16 kHz and analyzed into a small number, nominally 64, of separate, frequency intervals of ~256 Hz each. A 16 kHz rate provides information up to 8 kHz, sufficiently high to include at least 3-4 formants for an infant and to show the distinction between voicing and other speech sounds: fricatives, stop releases, bursts, etc. (These parameters are suitable for using the FFT and impose no delay of their own beyond 4 ms, i.e., 1/256-th of one second.) The visiBabble system uses only one-half of these intervals because the others differ only in phase.

The spectral intervals are grouped into six broad bands. An energy waveform is constructed in each of the six bands, the time derivative of the energy is computed, and peaks in the derivative are detected. These peaks represent times of abrupt spectral change in the six bands. Energy in bands 2 (1200 - 2500 Hz.) and 3 (1800 - 3500 Hz), e.g., provides evidence of voicing or, in some cases, of bursts. The distinction between these is readily made in the time domain (voicing persists much longer than bursts) as well as by appeal to information in the other spectral bands: voicing provides a power spectrum that decays with frequency approximately as $1/\text{frequency}^2$, whereas most other speech sounds have flatter spectra.

For the poorly formed or unstable closants and vocants typical of infants, wide frequency bands are well suited to recognition: Higher frequency resolution would require averaging over bands anyway. It would require spending more time computing and – worse – more time sampling the signal for the initially higher resolution.

C. Graphic Feedback

The visiBabble prototype responds to the child's utterances with five different brightly colored animations that cycle to avoid habituation: (a train, a bird, a frog and two cartoon creatures that move across the screen). It responds to the start of each syllable it detects by advancing the current animation one step.

It determines that a syllable has started either by voicing onset or by a voiced closant that occurs at least 100 ms after start of the previous syllable. Admittedly, a syllable might start with a burst before the voicing onset but, to avoid responding to noise, visiBabble waits for the onset of voicing. The system responds in no more than 0.1 second of the corresponding acoustic event.

C. Data Collection

As visiBabble runs, it makes a digital recording of the session in wav format. It also saves a record of the times and types of landmarks it found during the session. A second program uses this landmark data to produce a syllable and utterance summary as shown in Table 1.

D. Experimental Formats

Single case study designs [11] are particularly suited to our preliminary tests of visiBabble since they provide the freedom to conduct a study on a small heterogeneous group of subjects. The prototype program can be run in a variety of "formats":

1) Baseline (recording, no graphic display);

2) Response (graphic display is always present, while recording);

3) A-B-A (no display, display on, no display). The length of A or B phases can be changed.

Data is collected during all phases of all formats to allow a comparison of behavior during the baseline and active phases. The analyses of landmarks and syllables are conducted and recorded separately for the B phase and two A phases.

E. Field Testing

As part of the software development, a prototype of the system, visiSyl 1.2, was beta-tested by a typically-developing one-year-old and is currently being evaluated in trials with four at-risk children, ranging in age from 28 months to 7.5 years, and three premature but typically developing infants with ages, corrected for prematurity, from 8 to 11 months. The system will be iteratively modified in response to the results of this field-testing.

Preliminary questions on the use of the visiBabble include:

1) What features of infant vocalization can the system respond to in real time?

2) What graphic feedback do infants find appealing?

3) What changes have to be made in the graphic feedback to avoid habituation?

4) Do the infants show increased babbling during the treatment (B) phases?

5) Do infants adjust the amplitude of their utterances in response to the visual reinforcement?

6) Do infants adjust the pitch of their utterances in response to visual reinforcement?

7) Do infants increase the variety of syllable types and complexity of their utterances?

8) Is there any change in the distribution of utterances as an infant matures?

9) Do parents perceive changes in their infants' vocalizations in response to the visiBabble program?

The ABA design allows direct comparisons of the child's productions (items 4 to 8) with and without the system's visual feedback. Both the rate and the variety of syllables may be tested for the stimulating effect of the system by several techniques.

III. Results

Our beta-testing with a typically developing one-year old showed that our system was responsive to a child of that age. On days when he wasn't cranky, as reported by his parents, he showed an interest in the visual response screens. These sessions were run by the child's parents in a particularly noisy environment. Noise from the heating system, a vacuum cleaner, parents talking, and the computer itself were often louder than the child and clearly affected the output. The child was also very interested in the buttons on the display.

As a result of these sessions, we now ask that the computer be placed behind the microphone and that observers, if they must speak, do so as quietly as possible and also behind the microphone. We have also placed black tape over the display buttons.

Our current tests are being run by trained speech pathology students. The system rarely responds to noise and whispering that can be heard in the background. The exception to this is when such sounds overlap with the child's utterances. The results of a sample session are shown in Table 1. Landmarks that were clearly caused by noise or adults were removed before the syllable analysis.

The subject of this session was a 6 year old male child with cerebral palsy and cortical visual impairments (but who focuses intently on book pictures and loves TV). He is a symbolic communicator with signs and word approximations, limited range of vowel and consonant sounds (about 4 consonants in repertoire).

IV. Further Development

There are several features we plan to add to the visiBabble system. We have observed that some young infants are not always interested in our visual feedback. They may not be focused on the part of the screen where the bird is flying or the frog is hopping. We will add feedback that occupies more of the screen, e.g. fireworks or large faces that wink or smile. We may add sound or tactile feedback to the responses.

Though our prototype system just responds to the detected start of syllables, it is also capable of responding to other aspects of the child's vocalizations, e.g. variation in pitch or energy, the duration of syllables or utterances, or the complexity of syllables in terms of landmark structure. We plan further tests with infants and children on these aspects of the system. We envision a system where a speech pathologist, for example, might choose to work with a child on producing longer utterances and set the visiBabble system accordingly.

For research purposes, we plan to add to the information saved by the visiBabble system. We currently save a digital audio recording of each session and the landmark analysis as it was computed in real time. From this, we are able to compute the syllables that visiBabble found and hence responded too. In future systems, we will likewise record which response was displayed so that we might determine whether certain responses are particularly effective. We will also save the pitch information as it was computed during the session. Our summary program will then be augmented to classify syllables according to pitch contours as well as landmark content.

We hope to see visiBabble become a product that is useful as a clinical and research tool for work with at-risk infants or older non-speaking children. We also intend to produce a version that can be used as a training toy for these infants and children.

Table 1: Sample Summary of Data Collected During a 10 minute A-B-A visiBabble Session
A1 - 2.5 minutes with no display; B - 5 minutes with responsive display; A2 - 2.5 minutes with no display

Syllables	**Entire Session**		**A1**		**B**		**A2**	
Type	number	average duration	number	average duration	number	average duration	number	average duration
+g-g	7	0.164			6	0.167	1	0.144
+g-s	1	0.048			1	0.048		
+s-g	1	0.120					1	0.120
+s-s	3	0.048			3	0.048		
+b+g-s	1	0.016			1	0.016		
+g+s-g	5	0.199			5	0.199		
+g+s-s	3	0.109			3	0.109		
+g-s-g	3	0.131			2	0.165	1	0.064
+s-s-g	4	1.211			4	1.211		
+g+s-g-b	1	0.112			1	0.112		
+g+s-s-g	3	0.230			2	0.265	1	0.161
+g-s-g-b	1	0.707			1	0.707		
+s-s-g-b	1	0.273			1	0.273		
+s+s	2	3.962			2	3.962		
+g+s+s	1	3.318					1	3.318
+g+s-s-s-g	2	0.591			1	0.490	1	0.691
+g+s-s+s-s-g	2	0.972			2	0.972		
Totals	**41**	**0.590**	**0**	**NaN**	**35**	**0.562**	**6**	**0.750**
Average Number of Landmarks per Syllable		3.049		NaN		3.029		3.167
Utterance Summary:								
	number	avdur	number	avdur	number	avdur	number	avdur
	33	0.756	0	NaN	28	0.728	5	0.911
Average Number of Syllables per Utterance:		1.242		NaN		1.250		1.200

ACKNOWLEDGEMENT

This work was supported in part by National Institutes of Health grant R41 DC005534.

REFERENCES

[1] C. Bickley, "Acoustic evidence for phonological development of vowels in young children," *MIT Speech Communication Working Papers **IV***, 111-124, 1984.

[2] C.J. Cress and L. Ball, "Strategies for promoting vocal development in young children relying on AAC: Three case illustrations," *Proc. Rehab. Eng. & Assist. Tech. Soc. North Am.*, RESNA Press, pp.44-46, 1998.

[3] H.J. Fell, L.J. Ferrier, D. Sneider, and Z. Mooraj, "EVA, An early vocalization analyzer: an empirical validity study of computer categorization," *Assets '96*, pp. 57-61, 1996.

[4] H.J. Fell, J. MacAuslan, L.J. Ferrier, Chenausky, "Automatic Babble Recognition for Early Detection of Speech Related Disorders," *Assets* '98, pp. , 1998.

[5] HJ Fell, LJ Ferrier, C. Espy-Wilson, S.G. Worst, E.A. Craft, K. Chenausky, J. MacAuslan, G. Hennessey, "Automatic Analysis of Infant Babbling in EVA, the Early Vocalization Analyzer", *ASHA Proc.*, 2000.

[6] H.J. Fell, J. MacAuslan, L.J. Ferrier, S.G. Worst, and K. Chenausky, "Vocalization Age as a Clinical Tool," *Proc. ICSLP (Int. Conf. on Speech & Language Processing)*, Denver, September 2002.

[7] T.S. Jensen, B. Boggild-Andersen, J. Schmidt, J. Ankerhus, and E. Hansen, E. "Perinatal risk factors and first-year vocalizations: Influence on preschool language and motor performance ," *Develop. Med. & Child Neur.*, **30**, pp. 153-161, 1988.

[8] K. Levin, "Babbling in infants with cerebral palsy" *Clin. Ling. and Phonetics*, **13** (4), pp. 249-267, 1999.

[9] S. Liu, "Landmark detection of distinctive feature-based speech recognition," *JASA,* **96**, 5, Part 2, p. 3227, 1994.

[10] J.L. Locke, "Babbling and early speech: Continuity and individual differences," *First Language*, **9**, pp. 191-206, 1989.

[11] L.V. McReynolds, K.P. Kearns, *Single Subject Experimental Designs in Communication Disorders*, Baltimore: University Park Press, 1983.

[12] P. Menyuk, J. Liebergott, M. Shultz, M. Chesnick, and L.J. Ferrier, "Patterns of Early Language Development in Premature and Full Term Infants," *JSHR* **34**, p. 1**,** 1991.

[13] K.N. Stevens, S. Manuel, S. Shattuck-Hufnegel, and S. Liu, "Implementation of a model for lexical access based on features," *Proc. ICSLP*, Banff, Alberta, **1**, 499-502, 1992.

3rd International Workshop MAVEBA 2003, 43-46

CRY FEATURES AS A MEASURE OF PAIN INTENSITY IN NEWBORNS

R. Sisto[1], C. V. Bellieni[2], D.M. Cordelli[2], G. Buonocore[2]
[1] Department of Occupational Health, ISPESL, Monteporzio Catone (Rome), Italy
[2] Department of Pediatrics, Obstetrics and Reproductive Medicine, University of Siena, Italy

Abstract: **Acoustical characteristics of the cry of 57 newborns during heel-prick were correlated to pain intensity, as evaluated according to the DAN index. A time-frequency analysis of the acoustic waveform showed that the fundamental frequency and the rms normalized pressure level are both correlated to DAN score. Moreover, a typical "siren cry" pattern was observed in more than 60% of the subjects with DAN score ≥9 and in none of those with DAN score ≤8. This observation and the rapid increase of the fundamental frequency above DAN=8 suggest that this DAN score represents a threshold level. Above this level, the acoustic features of the cry change significantly, conveying a message of unbearable pain and danger.**
Keywords : **cry, neonate, pain**

I. INTRODUCTION

Crying is simultaneously a sign, symptom and signal [1]. It is the infant's earliest form of communication, but the significance and meaning of neonatal crying are still unclear. It does not actually seem to differ in quality for hunger, pain and fussiness [2] as it appears not to be unitary and isomorphic with respect to discrete causes: it is a graded signal [3-5]. Gradations of crying may help a listener to whittle down the range of possible causes, usually with the help of contextual information, [3,5-8]. In the last few years some pain scales have been developed to discriminate the level of pain a newborn is suffering [9-14] but they have rarely been used in sound spectral analysis of crying [15]. Pain, has different levels, from zero to a maximum, and babies' behavior varies accordingly. The aim of this study was to investigate to what extent crying features vary with the level of pain, or in other words, to assess cry characteristics of different pain levels expressed by a validated pain scale.

II. METHODOLOGY

Subjects

This report is based on analysis of a cohort of 57 healthy term newborns, already analyzed in a previous study [16], who underwent heel-prick for neonatal screening. Selection criteria were: Apgar score at least 9 at 5min; gestational age 38-41 weeks; age more than 48h; more than 2h since last meal. A video of about one minute was made for each neonate to record behavior and cry. A composite measure of neonatal pain, ranging from 0 to 10 (Douleur Aiguë du Nouveau-né - DAN - scale) [17], based on facial expression and behaviour was attributed to the babies by the same double-blinded scorer. Although sucking and oral sugar were effective analgesic methods, SS was found to have even greater analgesic power. Siena University Ethical Board approved the present study. Informed consent was obtained from the parents of babies enrolled.

Procedure

The digital acoustic signal was extracted from the original .AVI file using Goldwave software, and the waveforms of cries visualized. The data were converted to ASCII format and analyzed with special software developed in Labview (National Instruments) for cry analysis. The acoustic signals were sampled at 44.1kHz corresponding to a Nyquist frequency of 22.05kHz. A digitized 25s file (2^{20} samples) was extracted from each record, starting immediately after the heel-prick.

The cry signals were further analyzed by short time Fourier transform (STFT) in the time and frequency domains. The length of the elementary time interval to be Fourier-transformed fixes the time and frequency resolutions, which are inversely proportional to each other and the same for all frequencies in the spectrum.

The 25s files were divided into 1024 (2^{10}) time intervals, each of 23.22ms. The power spectrum of the signal was computed for each interval to give a time sequence of 1024 spectra for each neonate, with a time resolution of 23.22ms and a frequency resolution of 43Hz. To avoid introducing spurious spectral features caused by cutting the waveform, a Hanning window was applied to each interval. The time evolutions of these spectra were visualized as time-frequency intensity plots, which were used for preliminary heuristic analysis. The acoustic pressure signal of each crying sequence was normalized to its maximum amplitude, and evaluated over the whole 25s interval. In this way, problems arising from absolute signal amplitude evaluation, which is a function of the microphone-to-neonate distance, were avoided. The root-mean-square (rms) value of normalized acoustic pressure was calculated for each waveform.

The mean square of pressure is directly proportional to the average power of the wave. In the present study, rms pressure normalized to its maximum is not a measure of absolute cry intensity, but rather a measure of the constancy of emission: in other words, it measures the

fraction of the observation time during which the signal amplitude was near its maximum.

With the STFT technique, a shorter time sequence, roughly matching the first burst of crying after heel prick, was analyzed for each neonate. A file of duration 1.5s (2^{16} samples) was extracted from each neonate's recording. The 1.5s files were divided into 64 time intervals of 1024 samples each, to obtain a time sequence of 64 spectra for each neonate. The average of those spectra was calculated.

When the average spectrum showed peaks with a quasi-periodic structure, the lowest frequency peak was identified as the fundamental excitation frequency (F_0). F_0 is the base frequency of harmonic vibration of the vocal cords. It is usually heard as the pitch of the cry [6, 18].

Statistical analysis of cry features in relation to DAN score

The rms normalized pressure of the cry signal and first cry F_0 of each neonate was related to his/her DAN index by linear regression analysis. Data corresponding to a DAN≤3 were not considered in this analysis, because when the DAN index is very low the neonate is rather quiet, and the recording is often dominated by background noise.

Cry spectra were analyzed visually for peculiar characteristics. A chi-square non-parametric Pearson test was applied to quantify the frequency of occurrence of a particular feature (siren cry, see later) in groups with different DAN score. The "siren" pattern was defined by visual inspection as a pattern in which the fundamental frequency and its multiple frequencies were modulated periodically for a continuous time interval of at least 10s.

First cry F_0 was compared by means of a standard Student's t-test (significance criterion $p<0.05$) between the populations of neonates with DAN≥9 and ≤8.

III. Results

Simple visual inspection of the time-frequency intensity plots obtained by STFT showed major qualitative differences between neonates with high and low DAN scores. A characteristic feature of the high-DAN group was the regularity and reproducibility of the amplitude pattern on a slow time-scale, on the order of 1s (siren-cry, see Fig.1). The time-frequency intensity patterns of this siren cry showed periodic modulation of the fundamental frequency F_0 and its multiple frequencies (Fig.1b) and the average power spectrum had a quasi-periodic peak structure. The "siren" pattern was not recognized in any cry of the 36 babies with DAN≤8 whereas it was recognized in 13 of the 21 babies with DAN≥9 ($p<0.001$).

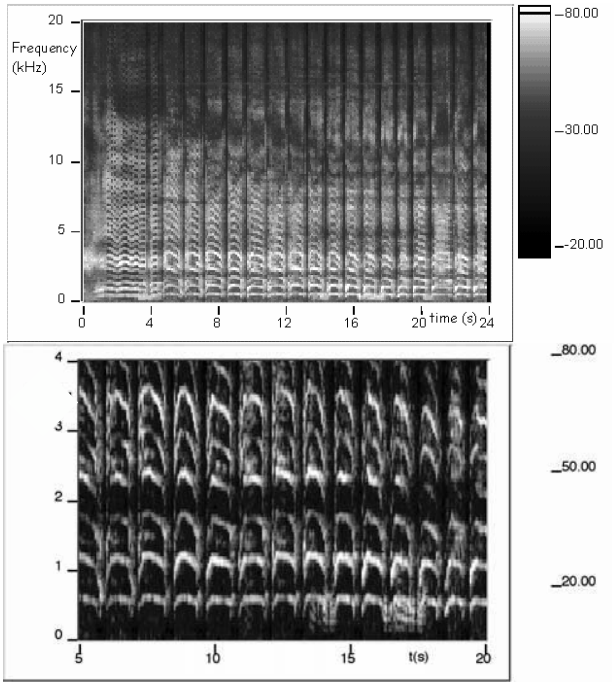


Figure 1: Top: time-frequency crying intensity plot for a neonate with DAN = 10; bottom: low-frequency zoom.

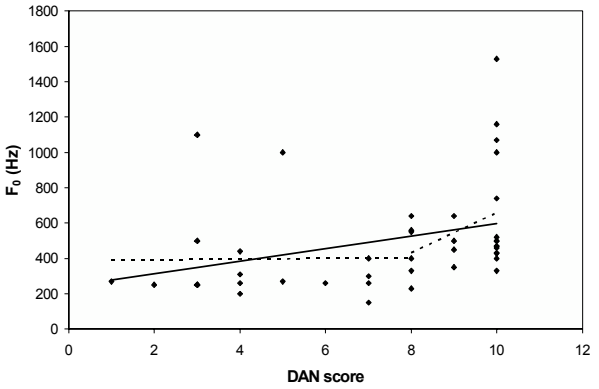


Figure 2: Fundamental frequency F_0 versus neonate DAN index. The solid line is a linear regression to all the data; the dotted line consists of two regressions for DAN≤8 and DAN≥8.

Mean F_0 of these two groups (DAN ≤8 and ≥9) were compared and a statistically significant difference in F_0 between the two groups was found. The fundamental frequency showed a shift to higher frequencies in

neonates with higher DAN. The result of this analysis is a mean F_0 of 630Hz, with a standard deviation of 330Hz for high-DAN neonates (DAN≥9), and 400±240Hz for the other group. The difference between the two groups was statistically significant (p=0.016).

First cry F_0 showed a significant correlation with DAN score (r=0.33, p<0.05). This correlation is not due to a monotonic increase of F_0 with DAN score, but rather to the sharp increase of F_0 above a DAN score of 8, as shown in Fig.2, where two distinct regression lines, relative to the data subsets with DAN≤8 and DAN≥8 are plotted along with the regression line relative to the whole data set.

The rms normalized pressure over a recording time of 25s showed a significant correlation with DAN score (r=0.86, p<0.01) (Fig.3).

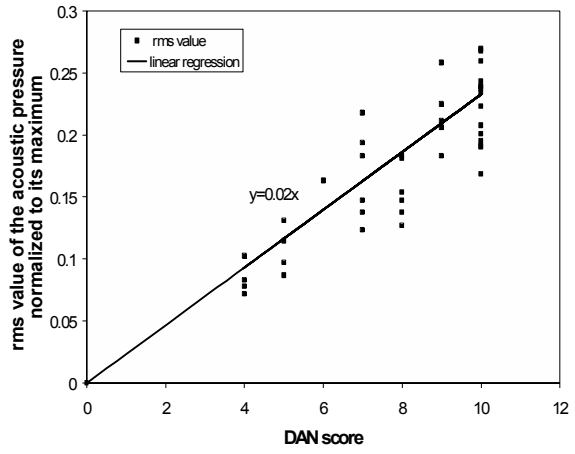


Figure 3: Rms normalized sound pressure during a 25s cry sequence, plotted against the neonate DAN score.

IV. Discussion

It is important to correlate crying due to pain with a validated pain scale. The results of this paper show that pain intensity (DAN) was correlated with normalized rms sound pressure. In other words, the stationary character of the overall cry intensity increased with increasing pain. The most interesting finding in the present study was the regularity and stereotyped pattern of cries with a DAN score greater than or equal to 9. Above this threshold, the features and meaning of crying changed. For DAN ≤8, crying was less regular in the modulation of the fundamental frequency and moan-like. When DAN was greater or equal to 9, a stereotyped cry was produced, the regularity of which suggests a call for attention and help. The spectrogram shown in Fig.1 is typical of high-DAN cries: after a few seconds of intense irregular and continuous sound, a periodic pattern starts, made up of repeated cries of almost the same duration (of the order of 1s) and spectral composition, separated by very short quieter intervals. Each cry shows symmetric modulation of the fundamental frequency. We called this pattern "siren cry". Obsessive repetition of the same sound signal seems an effective way of alerting the listener. Internal modulation makes each single cry more noticeable, and immediate repetition communicates a sense of alarm. Similar repeated sound patterns, in the same frequency range, are commonly used in different human cultures for communicating alarm. It is interesting that all cries with DAN score lower than 9 lack the periodic pattern shown in Fig.1.

First-cry F_0 showed a statistically significant difference between newborns with DAN score ≤8 and ≥9. This indicates that when pain exceeds a DAN score of 8, even the first cry is at a higher pitch. The abrupt change of the slope of the relation between F_0 and DAN score, shown in Fig. 2, also suggests the existence of a threshold at a DAN score of about 8, where the cry behavior changes qualitatively.

These features are easily recognized by listeners: first cries at a higher pitch, followed by the siren pattern, with a sound level constantly near its maximum indicate pain exceeding a DAN score of 8.

V. Conclusion

We have demonstrated that the spectral features of crying in term newborns communicate the level of stress and suffering, as measured by a validated scale, the DAN index. Our results help to recognize the threshold at which a response from bystanders becomes compulsive, and may effectively help caregivers to discriminate the threshold of unbearable pain in neonatal crying.

References

[1] Barr RG, Hopkins B, Green JA, "Crying as a sign, a symptom and a signal: evolving concepts of crying behavior," in *Crying as a sign, a symptom and a signal.* Barr RG, Hopkins B, Green JA, Eds. Cambridge University Press, 2000, pp. 1-7.

[2] Fuller BF, "Acoustic discrimination of three types of infant cries," *Nurs Res* pp. 156-160, 1991.

[3] Gustafson GE, Wood RM, Green JA, "Can we hear the causes of infants' crying?" in *Crying as a sign, a symptom and a signal.* Barr RG, Hopkins B, Green JA, Eds. Cambridge University Press, 2000, pp. 8-22.

[4] Porter FL, Miller RH, Marshall RE, "Neonatal pain cries: effect of circumcision on acoustic features and perceived urgency," *Child Dev,* vol.57, pp. 790-802, 1986.
[5] Wood RM, Gustafson GE, "Infant crying and adults' anticipated caregiving responses: acoustic and contextual influences," *Child Dev*, vol. 72(5), pp. 1287-1300, 2001.
[6] Corwin MJ, Lester BM, Golub HL, "The infant cry: what can it tell us?" *Curr Probl Pediatr*, vol. 26, pp. 325-334, 1996.
[7] Zeskind PS, Marshall TR, "The relation between variations in pitch and maternal perceptions of infant crying," *Child Dev* , vol. 59, pp. 193-196, 1988.
[8] Lester B, Boukydis Z, Garcia-Coll CT, Peucker M, McGrath MM, Vohr BR, Brem F, Oh W, "Developmental outcome as a function of the goodness of fit between the infant's cry characteristics and the mother's perception of her infant's cry," *Pediatrics* , vol. 95(4), pp. 516-521, 1995.
[9] Stevens B, Johnston CC, Petryshen P, Taddio A, "Premature infant pain profile: developmental and initial validation," *Clin J Pain*, vol. 12, pp. 13-22, 1996.
[10] Grunau RVE, Oberlander TF, Holsti L, "Bedside application of the neonatal facial coding system in pain assessment of premature neonates," *Pain*, vol. 76, pp. 277-286, 1998.
[11] Lawrence J, Alcock D, McGrath P, Kay J, MacNurray SB, Dulberg C: The development of a tool to assess neonatal pain. *Neonatal Netw,* vol. 12, pp. 59-66, 1993.
[12] Carbajal R, Paupe A, Hoenn E, Lenclen R, Olivier-Martin M, "DAN: une échelle comportamentale d'évaluation de la douleur aigue du nouveau-né, " *Arch Pediatr*, vol. 4, pp. 623-628, 1997.
[13] Krechel SW, Bildner J, "CRIES: a new neonatal postoperative pain measurement score: initial testing of validity and reliability, " *Paediatr Anaesth*, vol. 5, pp. 53-61, 1995.
[14] Sparshott M, "The development of a clinical distress scale for ventilated newborn infants: identification of pain based on validated behavioural scores," *J Neonatal Nurs* vol. 2, pp. 5-11, 1996.
[15] Craig KD, Gilbert-Mac Leod CA, Lilley CM, "Crying as an indicator of pain in infants," in *Crying as a sign, a symptom and a signal*. Barr RG, Hopkins B, Green JA, Eds. Cambridge University Press, 2000, pp. 23-40.
[16] Bellieni CV, Bagnoli F, Perrone S, Nenci A, Cordelli DM, Fusi M, Ceccarelli S, Buonocore G, "The effect of multi-sensory stimulation on analgesia in term neonates: a randomized controlled trial," *Pediatr Res*, vol. 51(4), pp. 460-463, 2002.
[17] Carbajal R, Chauvet X, Couderc S, Olivier-Martin M, "Randomised trial of analgesic effects of sucrose, glucose, and pacifiers in term neonates," *BMJ* vol. 319, pp. 1393-1397, 1999.
[18] Fort A., Manfredi C, "Acoustic analysis of newborn infant cry signals," *Med Eng Phys*, vol. 20, pp. 432-442, 1998.

3rd International Workshop MAVEBA 2003, 47-50

VOCAL IDENTITY - DIFFERENCES AND SIMILARITIES BETWEEN CHILDREN FROM CROATIA AND FINLAND

Natalija Bolfan-Stosic[1], Anneli Yliherva[2], Graham Welch[3]

[1]Dept. of Logopedics, Laboratory for Hearing and Speech Acoustics
Faculty of Education and Rehabilitation University of Zagreb, Croatia
[2]Dept. of Pediatrics and Adolescents, University Hospital of Oulu and Dept. of Saami, Finnish and Logopedics, University of Oulu, Finland
[3]Institute of Education, School of Arts and Humanities,
University of London, United Kingdom

Abstract: **The purpose of the present study was to find out if children between 8 and 10 years of age, from Croatia and Finland, are (i) able to identify appropriate voices from non-appropriate voices and (ii) are abusive in their voices. The third (iii) aim was to compare girls' and boys' vocal identity to each other. A structured questionnaire (Bolfan-Stosic, 2000) was used to investigate the children's voice habits. Results indicated that participant children did not differ with regard to country of origin. However differences appeared in relation to gender. The Croatian and Finnish girls (n=24) were better in identification of voice quality and vocal abuse compared to the Croatian and Finnish boys (n=16). It is suggested that future studies should continue to consider cultural environment in children's identification and understanding of own voice status.**

Keywords: **vocal identity, vocal abuse, pitch, loudness, school age, culture**

I. INTRODUCTION

A child's awareness of his/her own voice status and the identification of excitement when abusing their own voice is still very little studied. Emotions can change child's vocal timbre, loudness and pitch, such as in possible stress situations as non-regular family settings, school or in some other similar environments. Voice is a mirror of human emotions, especially in the very formative and sensitive period of life – childhood. Vocal abuses in children are mostly exampled as screaming, crying, speaking too loud or too quietly, speaking too fast or too slow, and speaking on the rest etc. Abuse happens usually at home, at school or in the playground and could have an influence on habitual voice characteristics [1]. The researchers of child voice conclude that voice habits established at an early age may persist into adulthood [2]. For example, hoarseness is common among school-aged children and may cause severe organic changes in the vocal fold [3]. In a study published in Finland [4] school-aged children with voice disorders had also been given more general remedial education than those children with healthy voices. It seems important, therefore, to track children's voice habits and to teach them how to identify when they are abusing their voices.

II. METHODOLOGY

In the present study one section of a Vocal Identity Questionnaire [5] was employed to study participant children's ability to identify **nice** and **bad** voices. Pictures were used to help the children to understand the instructions. The children's vocal abuse at home was also studied to find out if the child screams or yells a lot at home. The questionnaire was translated into Finnish language by one of the authors (AY) and then back into English to make sure the questions remain unchanged.

The following variables were used in the data analyses:

- ID = identification of differences between **nice** and **bad** voice
- LOUD = identification of differences between three levels of **voice loudness**
- PITCH = identification of differences between three levels of **voice pitch**
- VOCABUSE = **screaming** or **yelling** at home

Instructions:

- First the voice teacher sang the vowel **a** nicely whilst pointing to the white flower, and then badly, this time pointing to the

- black flower. The teacher then asks the child to listen to how the teacher sings and choose one picture that matches the sound. The same procedure is followed for the next two tasks where each child must recognize bad or nice vocal loudness and pitch whilst pointing to one of the bells and trees (there are three sizes of bells and trees corresponding to three levels of voice: loud-normal-silent and high-normal-low).
- The children evaluated themselves individually concerning whether they liked to scream or yell at home. The question is extracted from the Questionnaire as a variable of vocal abuse at home.

Data coding variables and statistical analysis:

ID, LOUD and PITCH: false answer = 1, correct answer = 2

VOCABUSE: positive answer/yes = 1, negative answer/no = 2

The statistical analysis was made by Statistics for Windows (version 6.0). The differences between girls and boys in the whole sample of Croatian and Finnish school-age children were analyzed using t-tests and Analysis of Variance and correlation between girls and boys were computed by the use of Correlation matrices.

III. RESULTS

The results (Table 1) indicate, that there are statistically significant differences in ID (i.e. identification of differences between **nice** and **bad** voice) between girls and boys. The means also demonstrate differences in the LOUD variable as well, but these are not significant. Usually girls were better than boys (Fig.1).

Table 1. Differences between the participant girls (N1=24) and boys (N2=16) from Croatia and Finland aged from 8 to 10 years in their identification of **nice** and **bad** voices (ID), voice loudness (LOUD) and pitch (PITCH)

	Mean1	Mean2	SD1	SD2	N1	N2	t-value	df	p
ID	1.80	1.42	.41	.35	24	16	2.41	38	.0210
LOUD	1.80	1.50	.44	.50	24	16	1.64	38	.1097
PITCH	1.63	1.50	.50	.83	24	16	.77	38	.4463

ID = identification of differences between **nice** and **bad** voice, LOUD = identification of differences between three levels of voice loudness, PITCH = identification of differences between three levels of voice pitch, N 1 = girls from Croatia and Finland aged from 8 to 10, N 2 = boys from Croatia and Finland aged from 8 to 10

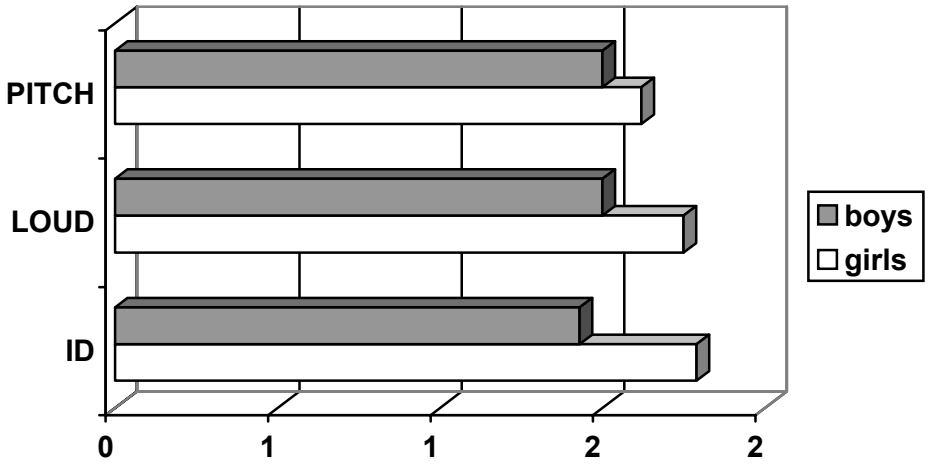


Fig. 1. Differences between the girls and boys aged from 8 to 10 years in ID, LOUD and PITCH according to the means.

In Table 2 there is a significant correlation shown between boys' groups evaluations of vocal abuse and identification of **nice** and **bad** voices and also in three levels of voice loudness. In a group of 24 girls we did not find any corresponding significant correlations. But we found statistically significant differences between children from two countries in three of the four variables which indicate identification of differences of the voice quality.

Table 2. Correlations between Croatian and Finnish boys

	ID	LOUD	PITCH	VOCABUSE
ID	1.00	.38	-.13	.68
LOUD	.38	1.00	.25	.77
PITCH	-.13	.25	1.00	.26
VOCABUSE	.68	.77	.26	1.00

Significant at level < .05

In Table 3 the differences between children of different ages by country are presented. The Finnish children from the oldest group (10 y) were best in ID (identification of **nice** and **bad** voices) compared to the other three groups. Both groups from Finland were better than the Croatian groups in ID and the younger group (8-9 y) from Finland were better in all three levels of LOUD than other three groups. The oldest Croatian children (10 y) were best in PITCH compared to other groups. Generally, the older children performed better in almost all the tasks (ID, LOUD, PITCH, VOCABUSE) than younger children, regardless of their country of origin.

Table 3. Differences in ID, LOUD and PITCH between children of different ages from Croatia (G1, G2) and Finland (G3, G4).

	ID	ID	LOUD	LOUD	PITCH	PITCH			
	Mean	SD	Mean	SD	Mean	SD	df	F	p-value
G1	1.50	.53	1.30	.48	1.30	.48	3	3.60	.0025
G2	1.40	.52	1.80	.42	2.00	.00	3	3.60	.0029
G3	1.70	.48	2.00	.00	1.40	.52	3	5.61	.0053
G4	2.00	.00	1.50	.53	1.60	.52	3	5.00	.0053

Significant at level < .05

G1: children from Croatia age 8; G2: children from Croatia age 10; G3: children from Finland age from 8 to 9; G4: children from Finland age 10

IV. Discussion

The significant difference between participant girls and boys were in their identification of nice and bad voice. Even though the girls performed better in their identification of nice and bad voices, they tend to have more longitudinal voice disorders according to the comprehensive Finnish study of school-age children [4]. On the other hand, there is also other evidence that girls' maturation is faster than boys' [6] which could be why they have better understanding and ability to find differences in different voice parameters is discern differences in cultural and economical background. Differences between the girls from Finland and Croatia and the boys from these countries were more significant than the cultural differences even though Finland and Croatia are very different countries both in cultural and economical background.

The vocal identity develops by the age. The youngest children (8 y) may not be aware of their voice status and they abuse their voices at home or at school. Based on clinical experience, parents usually do not pay attention to their children's voices even if the voices are very hoarse. One reason is lack of knowledge. There are very few studies of children's voice disorders and how to treat them. There is no some kind of "warning" program of vocal hygiene to alarm parents and society on different types of vocal abuse. On the contrary, voice is for the children more abstract than speech. It is easier for them to perceive for example singing, rate of speech (too fast or slow) or misarticulations.

Children learn speech sounds by listening their mother tongue around them. They develop vocal identity, too. They have different voice model around and they observe, listen and imitate other people's voices. If phoneticians ask themselves if children learn to discriminate all sounds very early, the voice researches could ask if children establish awareness of voice status or vocal identity before voice maturation.

We found statistically significant differences between children from two countries in three of the four variables which indicate identification of differences of the voice quality. Although differences between children from these two countries were not so big as we expected. Could we find explanation of these outcomes in possible higher voice awareness of children from Finland? One possible explanation could derive from differences in schooling. If school provides children with musical education from early age, including focused development of voice and the development of communication skills, there may be greater awareness of appropriate vocal behaviours.

In the future it could be interesting to study voice habits from the linguistic point of view. According to the language typological categorization the Croatian language belongs to the Indo-European languages and especially to the South-Slavic group, whereas the Finnish language to the Uralic group [7]. The Croatian and Finnish language belongs to different typological categories. The former is a typical flexion language with many prefixes, suffixes and word internal changes, and the latter an agglutinative language with many morphemes attached to word stem [7]. That difference might also be reflected in voice habits. For example the Finnish language is very monotonous with little changes in pitch, and maybe that is why the Finnish speaking children performed worse than the Croatian children in discriminating PITCH. These language bound features may also impact on vocal identity abilities.

The results of the present study did not offer explicit answer concerning vocal identity differences or similarities between the children from two cultures. Instead it supported the universal phenomenon that girls' maturation is faster than that of boys' even in vocal identity.

V. CONCLUSION

According to the results we conclude that:
1) Girls were better in ID (identification of nice and bad voices) than boys regardless of the cultural and economical background, which means that girls' maturation is faster than boys' in vocal identity.
2) The Finnish children identify better nice and bad voices compared to the Croatian children.
3) The younger group (8-9 y) from Finland identified the best LOUD (loud, normal or silent voice).
5) The oldest Croatian children (10 y) were better in discriminating PITCH (high, normal or low voice) compared to other groups.
6) Generally, the older children performed better in almost all the tasks (ID, LOUD, PITCH, VOCABUSE) regardless from which country they come from. In practice the younger children have more problems in identifying different voice parameters.

REFERENCES

[1] N. Bolfan-Štosic, V. Tokic, S. Jelčić-Jakšić, Some differences of voice quality of children from different social environments. *Proceedings of the 24th IALP* , Amsterdam, Netherlands, October, pp. 157-160, 1998.
[2] P. White, "Long-term average spectrum (LTAS) analysis of sex- and gender-related differences in children's voices, " *Log Phon Vocol,* vol 26, pp. 97-101, 2001.
[3] E. Sederholm, A. McAllister, "Group therapy for dysphonic children," in *Child Voice* , White, Ed. Stockholm: KTH Voice Research Centre, pp. 143-147, 2000.
[4] R. Koivusaari, "School-aged children's voice disorders: persistence of symptoms and their connections with psychomotor performance". A doctoral dissertation. *Acta Univ. Oul. B 29*. Oulu: Oulu University Press, 1998.
[5] N. Bolfan-Stosic, Vocal Identity Questionnaire. An unpublished questionnaire of vocal habits for school-aged children, 2000.
[6] G. Welch, "The developing voice" In L. Thuman & G. Welch (eds). *Bodymind and Voice: Foundations for Voice Education* (pp. 704-717). Iowa: National Centre for Voice and Speech, 2000.
[7] A. Joki, *"Languages of the world"* (in Finnish: *Maailman kielet*), Juva: WSOY, 1984.

3rd International Workshop MAVEBA 2003, 51-54

PHYSIOLOGY OF VOCAL PRODUCTION IN THE NEWBORN

Nicollas R[1], Ouaknine M[1], Giovanni A[1], Berger J[2], To JP[2], Dumoulin D[2],Triglia JM[1]

[1] Laboratoire d'audiophonologie clinique de l'Université de la Méditerranée. Fédération ORL. CHU Timone. 264 Rue Saint Pierre. 13385 Marseille cedex 5. France
[2] Institut de mécanique de Marseille (mécanique des fluides). Université de la Méditerranée. Technopôle de Château-Gombert, 60 rue Joliot Curie. 13453 Marseille cedex 13.

Abstract: Vocal folds of newborns are histologically different from children and adults. Reinke's space is not clearly individualized. As shown by Titze, this structure is absolutely needed for vocal fold vibration [1]. The hypothesis for vocal production in newborn is that the air column generates itself the acoustic turbulences (vortex) from which the sound merges. Some other possible vibrators within the mammalian production system include the vocal tract [2].
Acoustic analysis of excised larynx of 38 weeks-time dead human foetus was performed. An acoustic analysis and a phase portrait were calculated on each recorded sample. A newborn cry was also recorded with the same DAT. Anatomical measurements were performed and a virtual model (Gambit®) was designed to modelize turbulences with vocal folds in phonatory position (Fluent® 6.0). All data were correlated with those obtained by Laser Doppler Velocimetry.
The fundamental frequency of the sound produced by a fixed larynx was higher than those produced by fresh sample or newborn. Phase portraits are very different in each sample. High-frequency whirlwinds were modelized upon each vocal fold.
Preliminary results suggest that newborn phonation is a vortex effect coupled with a vibration of supraglottic structures.
***Keywords*:newborn,phonation, vocal folds, aerodynamic, modelization.**

I. INTRODUCTION

As shown by Titze, Reinke's space is absolutely needed for vocal fold vibration [1]. So, the hypothesis for vocal production in newborn is that the air column generates itself the acoustic turbulences (vortex) from which the sound merges. Some other possible vibrators within the mammalian production system include the vocal tract (arytenoid cartilages, ventricular folds and epiglottis) [2]. All those potential vibrators can be coupled by mechanical or aerodynamic forces. So, there is no surprise that newborn cry generates nonlinear phenomena [2;3]. The aim of this work was to correlate acoustic and aerodynamic data, using excised larynx of 38 weeks-time dead human foetus, and try to understand the physiology of newborn phonation.

II. MATERIAL AND METHODS

Three excised larynx of 38 weeks-time dead human foetus were used for the study. They were chosen after discussion with pathologist who confirm that the cause of death was not a malformation of respiratory tract. One of the larynx was included in formaldehyde during 6 months to avoid mucosal vibration. The other two were "fresh" larynx and supraglottic structures were resected in one of them. The three organs were hang on the experimental bench of the "Audio-phonology clinical laboratory", a Portex tracheal tube 2.5 mm was inserted in the trachea and a continue 4 l/mn airflow (corresponding to a physiological cry flow) was delivered. A sound could be generated with whole larynx (and not with the resected sample) and was recorded using a DAT (Aiwa DAT HDS1, microphone Sennheiser MKH 20 P48). An acoustic analysis and a phase portrait were calculated on each recorded sample. Airflow was "marked" with incense and a Laser (TSI® Model IFA 600, lengthwave 686nm _red_) was used to perform Doppler Velocimetry. A 135.5 mm focus was used allowing a 1.84 micrometers fringespacing. Laservec® software performed calulations. Laser position was modified with an electronical control system with an accuracy of 1 micrometer.
A newborn cry was also recorded with the same DAT.
On the other hand, anatomical measurements were performed. A virtual geometrical model using Gambit® software was made. It was then exported to Fluent® 6.0 software to modelize turbulences with vocal folds in phonatory position.

III. RESULTS

Acoustic datas are summed up with those graphs : first the temporal signal, second the Fourier transformation, and third the phase portraits. Temporal signal shows a "double source" signal on newborn cry. Fourier transformation is more intersting : it shows a fundamental frequency at 2650 Hz for the fixed larynx, a fundamental frequency at 725 Hz for fresh larynx. In the newborn cry sample, we can see a main frequency at 1000 Hz with its harmonics and another one around 2500 Hz which can be the result of another source. portraits shows a "noise" aspect for fixed larynx, a "neurological" aspect (very similar to those we found in Parkinson

disease voices) for excised fresh larynx and a deterministic chaotic attractor for newborn cry.
Laser Doppler Velocimetry shows that speed (in blue) and turbulence intensity (in pink) profiles are really different in larynx with (left) and without (right) supraglottic structures.
Modelization of whirlwinds was validated by Laser Doppler Velocimetry. So, we extracted calculated data to determine whirlwind frequency. This last was closely correlated to frequency found in excised fixed larynx.

IV. Discussion

All those results seem to show the importance of supraglottic structures in newborn phonation as shown by the impossibility to produce a sound when supraglottic structures are resected. Vocal folds generate a turbulence which probably induce supraglottic vibration. Vorticity by itself doesn't explain voice generation but remains contributing to the voice signal as shown by the acoustic data.
Laser Doppler Velocimetry and virtual model calculated by Fluent ® software shows higher speeds at the anterior part of larynx. Objective measures performed confirm the validity of the virtual model.
Further works are in process to try to establish the role of each supraglottic structure in newborn phonation. Others aerodynamic studies and models are still in process.

References

[1] Titze IR, Talkin DT : A theoretical study of the effects of various laryngeal configurations of the acoustics of phonation. *J Acoust Soc Am* 1979 ; 66 : 60-74.

[2] Fitch WT, Neubauer J, Herzel H : Calls out of chaos : the adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal behav* 2002 ; 63 : 404-418.

[3] Mende W, Herzel H, Wermke K : Bifurcations and chaos in newborn infant cries. *Physics Letters A* 1990 ; 145 : 418-424.

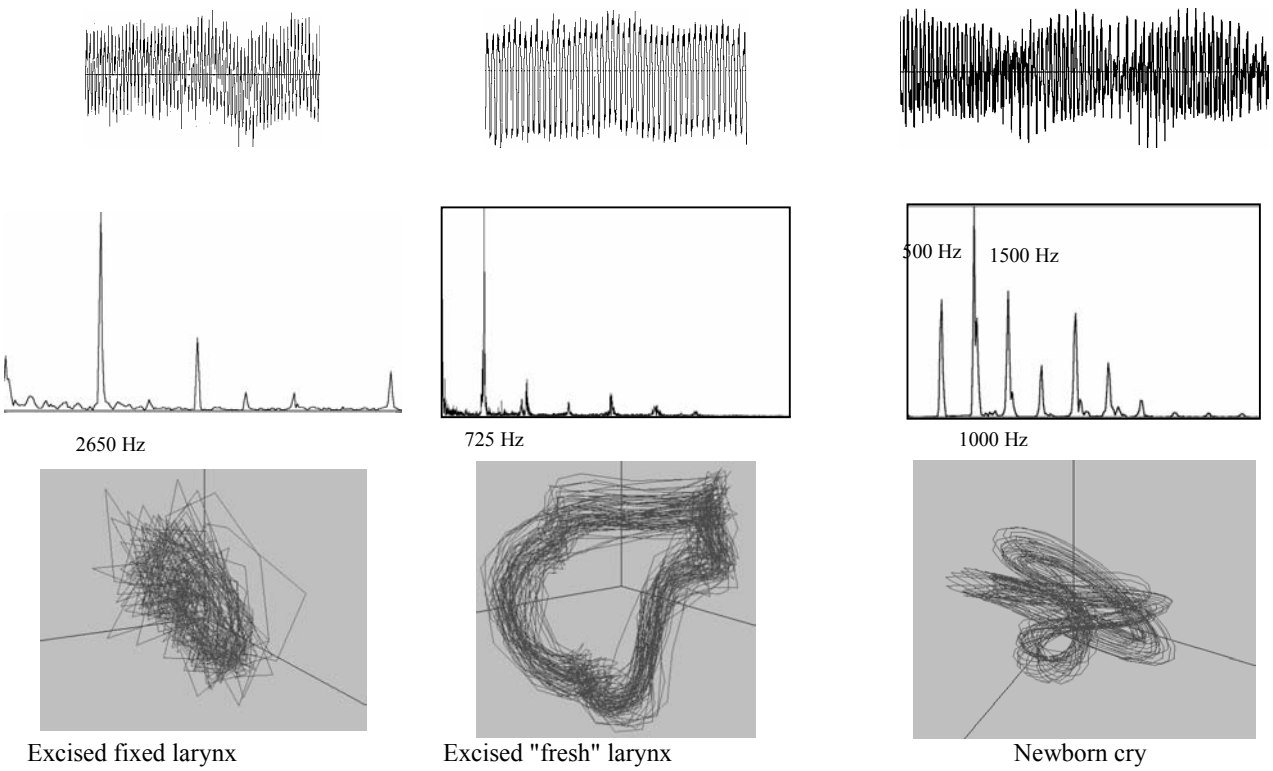


Figure 1 : Temporal signals, Fourier transformations and phase portraits represented in a 3 dimension space (3 axis) of recorded samples

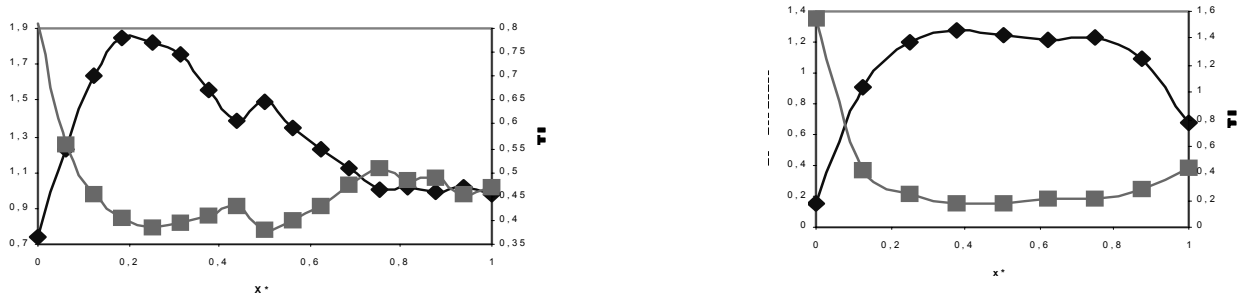


Figure 2 : Laser Doppler Velocimetry of larynx with (left) and without (right) supraglottic structures

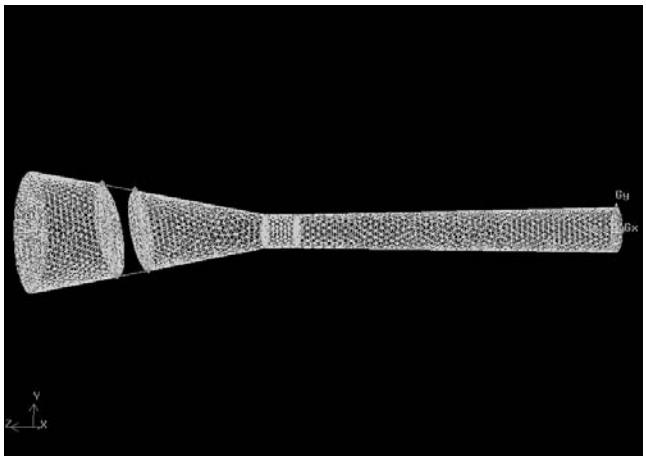
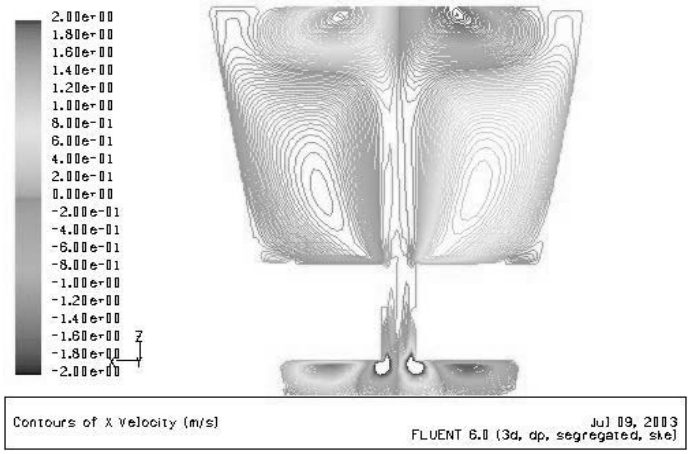


Figure 3 : Geometrical modelization of trachea and larynx with Gambit® software



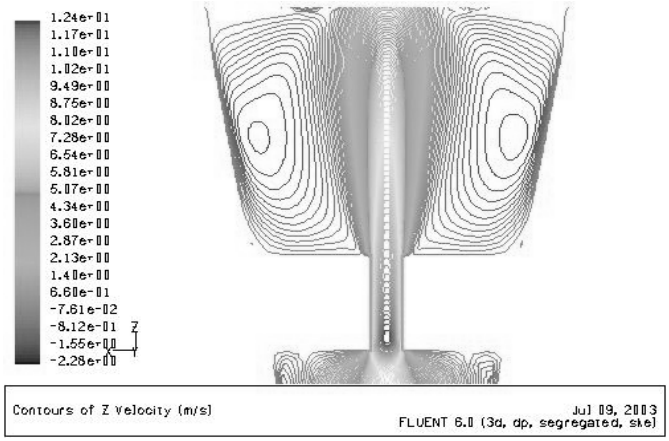


Figure 4 : Whirlwinds modelized with Fluent® software (velocity fields in X and Z axis)

Noise estimation/denoising

3rd International Workshop MAVEBA 2003, 57-60

A TWO-CHANNEL SPEECH DENOISING METHOD COMBINING WAVEPACKETS AND FREQUENCY COHERENCE

L. Helaoui, S. Ben Jebara, A. Benazza-Benyahia

Département MASC, Ecole Supérieure des Communications de Tunis
Cité Technologique des Communications de Tunis,
Route de Raoued 3.5 Km 2083 Ariana, Tunisie
sofia.benjebara@supcom.rnu.tn

Abstract: **In this paper, we are interested in multichannel speech denoising in the context of mobile communications. The conventional method exploits the "similarity" between the avalaible observations in the sense of the coherence function, measured in the Fourier domain. In this work, we alleviate the limitations of this approach by assessing the coherence function in the Modulated Lapped Transform (MLT) domain. Indeed, the MLT allows to take into account the local statistics of the underlying speech signal. Experimental simulations indicate the outperformance of the proposed method w.r.t. the conventional method: some distortions are reduced and the intelligibility is enhanced.**

I. INTRODUCTION

In order to guarantee a satisfactory quality and security, mobile communications systems offer the functionality of hands-free telephone. To this respect, several microphones (and, loudspeakers) are placed inside the car and, hence, several measurements of the speech signal are available [1]. However, these signals are very often, corrupted by noise (reverberation effects, echo, motor and wheels noise, ...). Therefore, it is required to denoise the multichannel observations. The most conventional denoising method consists in exploiting the inter-channel "similarities". Very often, the retained similarity measure is the Coherence Function (CF) [2], computed in the frequency domain [3]. Recently, an improved noise reduction procedure has been achieved thanks to a coherence function calculated in the Wavelet Transform (WT) domain [4]. In this work, we are interested in investigating a more efficient denoising method in the WT domain. To this regard, our contribution is two-fold. Firstly, by analyzing phoneme by phoneme the reconstructed signal, we emphasize the inadequacies of the conventional frequency approach. Secondly, we propose to compute a time-localized CF in order to take into account the phonemes nature. More precisely, the time axis is split thanks to the Modulated Lapped Transform (MLT) [6, 7]. This paper is organized as follows. In Section 2, we describe the background of our work. In Section 3, we present the proposed denoising method. In Section 4, some experimental results are given and some conclusions are drawn.

II. BACKGROUND

2.1. Observation model

Two microphones record the speech signal s uttered by the speaker during N_0 instants. As a result of background noise, the observation x_m at time n, registred at the m-th sensor is:

$$x_m(n) = s(n) + b_m(n), \quad n = 1, \ldots, N_0,\ m = 1, 2 \quad (1)$$

where b_m is the realization of the noise at microphone m, assumed to be a zero-mean, wide-sense stationary process, decorrelated from s. Furthermore, if the microphones are separated by a large distance, the processes b_m are assumed to be mutually decorrelated. It is worth noting that all the observations contain the *same* clean component. It is obvious that such simplified observation model holds only after a suitable delay-compensation [8].

2.2. Principle of two-channel denoising

A denoising procedure consists in estimating s from the available registrations. The nonstationarity of the observations is handled by a frame-by-frame approach. More precisely, the whole record set is split into overlapping frames each one being of size N and in the sequel, we will denote $x_m(l, n)$ the n-th sample of the l-th frame ($n = 0, \ldots, N-1$). Generally, the first step of a multichannel denoising method consists in passing from the temporal space to a suitable transform domain (of variable $\mathbf{u}$). As a

consequence, the temporal observations are mapped into coefficients $X_m(l,\mathbf{u})$ through a reversible linear transform $\mathcal{T}$:

$$\{X_m(l,\mathbf{u})\}_{\mathbf{u}} \stackrel{\triangle}{=} \mathcal{T}[\{x_m(l,n)\}_{n=0}^{N-1}], \quad m=1,2. \quad (2)$$

Obviously, the transform $\mathcal{T}$ is expected to generate coefficients whose processing is more tractable than the direct processing of the temporal samples. Usually, the transform $\mathcal{T}$ is taken as the Short Term Fourier Transform (STFT) and $\mathbf{u}$ is the frequency variable. In [4], the wavelet transform was also chosen as a $\mathcal{T}$ transform and, hence, $\mathbf{u}$ is a time-scale variable. The principle of multichannel denoising consists in exploiting the similarities between x_1 and x_2. The temporal correlation is discarded because it is not implicitly carry frequency information since it is not possible to discriminate between signals with different frequency components. This is the reason why the coherence function C has been introduced:

$$C(l,\mathbf{u}) \stackrel{\triangle}{=} \frac{\Gamma_{X_1X_2}(l,\mathbf{u})}{\sqrt{\Gamma_{X_1X_1}(l,\mathbf{u})\Gamma_{X_2X_2}(l,\mathbf{u})}}, \quad (3)$$

where $\Gamma_{X_1X_1}$, $\Gamma_{X_2X_2}$ and, $\Gamma_{X_1X_2}$ denote respectively the auto- and cross- power spectral densities of X_1 and X_2. At this level, it is worth noting that a common update of the spectral densities is made in a recursive way:

$$\begin{aligned}\Gamma_{X_mX_{m'}}(l,\mathbf{u}) &= \lambda\Gamma_{X_mX_{m'}}(l-1,\mathbf{u}) \\ &\quad +(1-\lambda)X_m(l,\mathbf{u})X_{m'}(l,\mathbf{u})^*\end{aligned} \quad (4)$$

where $(m,m') \in \{1,2\}$ and λ is a forgetting factor, empirically set by the user. If $|C(l,\mathbf{u})|$ is close to 0, the l-th frame of coefficients consists of noise and it has to be suppressed. More precisely, the average of the observation coefficients is filtered as follows:

$$\hat{S}(l,\mathbf{u}) = |C(l,\mathbf{u})| \left(\frac{X_1(l,\mathbf{u}) + X_2(l,\mathbf{u})}{2} \right). \quad (5)$$

Then, the inverse transform $\mathcal{T}^{-1}$ is applied to the processed coefficients $\hat{S}$ in order to derive the estimate $\hat{s}$ in the time domain.

III. Proposed method

3.1. Motivation

As mentioned previously, the STFT is usually chosen as the $\mathcal{T}$ transform. However, this conventional method presents some limitations. Our first contribution is to stress these limitations by analyzing the temporal variations of the denoised signal. In our experiments, the sentence "this day, the chicken leg is a real dish," uttered by a male speaker, sampled at 8 kHz is considered ($N_0 = 32768$). The two observations are obtained by adding artificial noise at each channel. In each frame ($N = 256$), the registred signal is weighted by a Hamming window (overlapped at 66.67%). In spite of high values of the Signal-to-Noise Ratio (SNR) obtained with very noisy environments, a residual noise, commonly known as "musical noise", remains in $\hat{s}$ during the (informal) perceptual tests. More precisely, in [9], it has been noted that the amount and the quality of noise reduction both depend on the phoneme type: voyels are correcly enhanced but the voiced occlusive phonemes, the fricative phonemes of short duration and low magnitude are poorly reproduced. During pause intervals, a modified background noise persists and it hampers the listener. The objective of this work is to avoid as much as possible, such artifacts. We aim at processing differently the observation according to the phoneme type. To this purpose, the transform $\mathcal{T}$ should provide coefficients that are well localized in time and frequency in order to take into acount the phoneme instants and the spectral features. The Modulated Lapped Transform (MLT) performs firstly an appropriate temporal segmentation then a local Fourier analysis and it is an appealing tool since it offers a flexibility in the choice of the intervals according to the signal structure. For example, long (resp. small) intervals could be envisaged for the stable (transition) parts of a phoneme.

3.2. Proposed modification

In each channel, the whole registred sequence (of size N_0) is subdivided into subblocks of length K. A MLT is applied to each subblock k as a transform $\mathcal{T}$. The variable of the transform domain consists of the block index k and the frequency variable f: $\mathbf{u} = (k,f)$. It amounts to a wavelet transform with the basis function $p_k(f,n)$ given by:

$$p_k(f,n) \stackrel{\triangle}{=} \sqrt{\frac{2}{K}} w_K(n) \cos[\frac{\pi}{K}(f+\frac{1}{2})(n+\frac{K+1}{2})], \quad (6)$$

where w_K is the analysis window and the time index varies from $n = 0,\ldots,2K-1$ and the frequency index f varies from 0 to $K-1$. The window $w_K(n)$ is set so as to ensure a maximum DC concentration:

$$w_K(n) = -\sin[(n+\frac{1}{2})\frac{\pi}{2K}]. \quad (7)$$

In the practice, the data x_m are preprocessed : the overlapping parts if the window w_K are folded back into the interval then the standard fast discrete cosine IV algorithm is used to calculate the expansion. More precisely, the resulting coefficients X_m ($m = 1,2$) can be expressed as follows:

$$X_m(k,f) = \sum_{n=0}^{2K-1} p_k(f,n) x_m(k,n). \quad (8)$$

Once the coefficients X_m computed, the CF is considered for each block k of size K. Although a MLT provides a complete representation of the analyzed signal, it is not time-invariant. As a consequence, the estimation error could be sensitive to the positions of the discontinuities in the signal and, the reproduced signal could exhibit Gibbs phenomena. Therefore, it is recommended to apply a Translation Invariant (TI) MLT [10]. It consists in applying the proposed denoising procedure to the shifted observation, for any feasible shift. Then, all the resulting estimates are averaged over all the shifts.

IV. Experimental results

Performance criteria should be defined for a fair comparison between several denoising methods. Ideally, the quality of the reproduced signal $\hat{s}$ should be assessed through psycho-acoustical tests. In the practice, these perceptual tests have a prohivitive cost because of the required specialized equipements. As a consequence, several objective criteria have been defined [5]. In this paper, we will consider the gain G (in dB) in term of SNR. However, in very noisy environments, high values of G could eventually be achieved but, at the cost of a great alteration of the spectral features of s. Thus, it is recommended to use distances that control the frequency content of the estimate $\hat{s}$ as the cepstral $d_{\rm cep}$, the Itakura $d_{\rm I}$ and, the Itakura-Saito $d_{\rm IS}$ distances.
In our simulations, we have used two test signals. The first set (denoted set I) corresponds to the sentence described in Subsection 3.1, which was artificially corrupted and Monte-Carlo simulations were performed. The second set (denoted set II) is a real speech sequence recorded by two microphones distant from 5 cm placed in a Volvo car, moving at 90 Km/h. The noise level is very high w.r.t. the clean signal.

4.1. Local analysis

Table 1 indicates that the global gain G increases with the size block K. This fact corroborates the relevance of the time-frequency localized analysis performed by the MLT. Besides, the improvement is considerable during the silence parts (which form 40% of the whole record) since a gain of around 9 dB is achieved whereas the average value of the global gain is 3.5 dB. A finer analysis of Table 1 indicates that the spectral alteration is more limited for different values of K according to the considered phoneme. As a result, we will retain the value $K = 128$ since it provides a good tradeoff between the precision and the preservation of the spectral content.

4.2. Global performances

Figure 1 provides the evolution of G *versus* the initial SNR (average observation SNR) for the test set I ($K = 128$). It clearly shows that an IT MLT is more advantageous than a time-variant one whatever are the initial values of the SNR. It can be noted that our method outperforms significantlty the classical method. The latter becomes inappropiate for weakly noisy environments on the contrary of the proposed approach. However, slight improvements are obtained in term of spectral alteration as shown in Figure 2.
Concerning real sequences (test set II), spectrograms are considered in Figures 3-5. It is worth pointing out that the denoised signal by the classical method is affected by a musical noise (especially in the silence intervals). The amount of such noise is reduced by the MLT-based approach. Another compelling of the proposed technique is the dramatical enhancement of speech components.

4.3. Conclusion

In this paper, we have proposed a denoising method based on a coherence function computed beteween the coefficients of the modulated lapped transform. The outperformance of this method in term of gain and residual musical noise are very encouraging. Further investigations should consist in selecting the wavelet basis according to an appropriate criterion.

References

[1] Y. Kaneda, J. Ohga, "Adaptive microphone-array system for noise reduction," IEEE Trans. Acoust. Speech, Signal Processing, vol. 34, no. 6, pp. 1391-1400, December 1986.

[2] G.C. Carter, "Coherence and time-delay estimation," *Proc. of the IEEE*, vol. 75, pp. 236-255, February 1987.

[3] R. Le Bouquin, G. Faucon, "Using the coherence function for noise reduction," *IEEE Proceedings*, vol. 139, no. 3, pp. 276-280, June 1992.

[4] D. Mahmoudi, A. Drygajlo, "Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement," *Proc. of ICASSP'98*, Seattle, WA, USA, pp. 385-388, 1998.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, USA, 1978.

[6] H.S. Malvar, "The LOT: A link between block transform coding and multirate filter banks," *Proc. IEEE Int. Symp. Circ. and Syst.*, pp. 835-838, Espoo, Finland, June 1988.

[7] H.S. Malvar, D.H. Staelin, "The LOT: transform coding without blocking effects," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 37, no. 4, pp. 553-559, April 1989.

[8] S. Fisher, K.U. Simmer, "Beamforming microphone arrays for speech acquisition in noise environment," *Speech Communication*, vol. 20, no. 3-4, pp. 215-227, December 1996.

[9] L. Helaoui, *Débruitage de la parole bi-capteur : apport de la technique de cohérence dans le domaine des ondelettes de Malvar*, Master Thesis (in French), Ecole Supérieure des Communications de Tunis, June 2003.

[10] R. Coifman, D. Donoho, "Translation-invariant denoising," in A. Antoniadis Editor, *Wavelets and Statistics*, Lecture Notes in Statistics, Springer-Verlag, pp. 125-150, 1995.

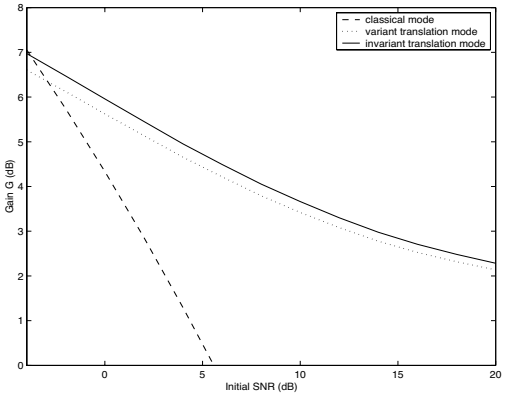


Fig. 1. Test set I: resulting gain G (dB) *vs* the initial SNR for all considered methods (block size $K = 128$).

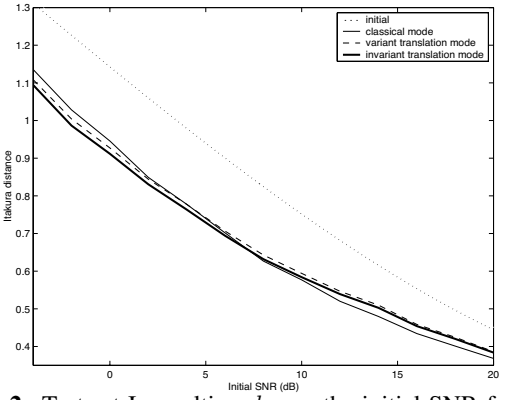


Fig. 2. Test set I: resulting d_{IT} *vs* the initial SNR for all considered methods (block size $K = 128$).

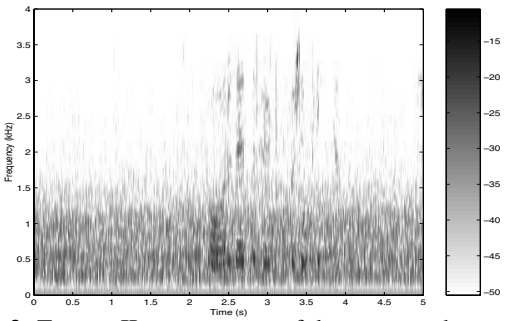


Fig. 3. Test set II: spectrogram of the average observation $\frac{x1+x2}{2}$.

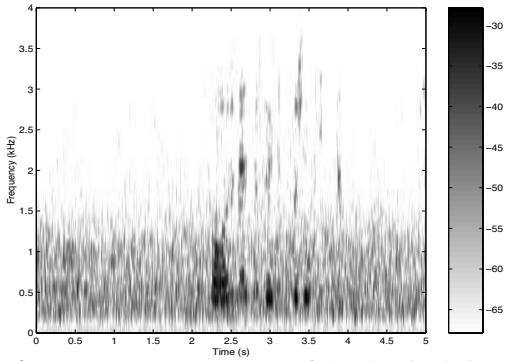


Fig. 4. Test set II: spectrogram of the denoised signal by the conventional method.

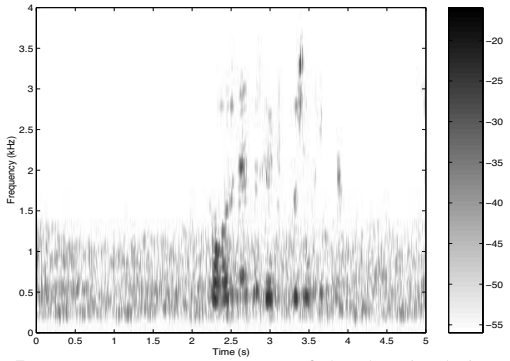


Fig. 5. Test set II: spectrogram of the denoised signal by the IT MLT based method ($K = 128$).

Table 1. Test set I: performances of the IT proposed method with the block size K for $SNR_{\mathrm{init},1} = 8$dB and $SNR_{\mathrm{init},2} = 6$dB.

	K	512	256	128	64	32
whole set	G	2.82	3.30	3.69	3.79	**3.88**
	d_{cep}	10.49	9.91	9.45	**9.30**	9.34
	d_I	0.61	0.60	0.58	**0.57**	0.59
"e"	G	0.45	1.15	**1.56**	1.42	1.25
	d_{cep}	0.080	0.069	0.065	0.062	**0.061**
	d_I	0.227	0.183	**0.131**	0.143	0.192
"th"	G	6.1	7.89	8.52	**8.65**	8.63
	d_{cep}	4.03	3.318	**3.195**	3.217	3.37
	d_I	1.873	1.997	2.155	**2.053**	2.126
"g"	G	2.766	3.03	3.09	3.37	**3.88**
	d_{cep}	4.149	3.876	3.6	3.44	**3.202**
	d_I	1.723	1.532	1.639	1.594	**1.471**
silence	G	8.46	9.11	9.43	9.56	**9.6**
	d_{cep}	32.98	31.5	**30.72**	30.87	31.16
	d_I	0.377	**0.39**	0.4123	0.44	0.53

3rd International Workshop MAVEBA 2003, 61-64

Wavelet-Based Noise Estimation Techniques for speech enhancement

E. Jafer, A. E. Mahdi
[1]Department of Computer and Electronic Engineering, University of Limerick, Limerick, Ireland

Abstract: **In this paper, we describe the implementation of three noise estimation algorithms using two different wavelet decomposition methods: Second-generation and Perceptual wavelet packet transform. The three-presented algorithms are: (a) smoothing based adaptive noise estimation, (b) quantile based noise estimation and (c) minimum variance tracking-based noise estimation These algorithms, which do not need a speech activity detector nor signal statistics learning histograms, are based on estimating the noise power from the noisy speech itself. The performance of presented algorithms has been evaluated and compared for different noise types and levels. A new robust noise estimation technique utilizing a combination of the quantile-based and smoothing based algorithms has been proposed. Reported results demonstrate how these algorithms are capable to track different noise types adequately but with varying degree of accuracy.**

I. INTRODUCTION

Reliable noise estimation remains a challenging problem in different speech enhancement systems. Accurate instantaneous noise power estimation is crucial for the success and robustness of any single-channel speech enhancement system. Over the last few years, various noise estimation techniques have been proposed and their performance evaluated. These include techniques that are based on tracking the minima of the noise power [1,2], and those, which utilize a quantile computation algorithm [3-4]. Although efficient, all these techniques involve relatively high computational complexity.

Three different noise estimation algorithms are considered in this paper: (a) an adaptive technique with a smoothing parameter that depends on the estimated subband signal-to-noise ratio (SNR) [5]; (2) a one-pass quantile-based technique [6]; and (3) a technique that is based on tracking the minimum variance of the subband noisy signal [7], are considered. First, we describe the implementation of these three algorithms using two signal representation schemes that provide different resolutions. The first is based on the application of second-generation wavelet transform (SGWT) [8,9], and the second is based on critical-band motivated perceptual wavelet packet decomposition (PWPD) [10]. We then propose a new and robust wavelet-based noise estimation technique that is based on combining the best features of algorithms (1) and (2). This is then followed by performance evaluation of all the above four-noise estimation techniques using a variety of speech signals distorted by different types of noise. The evaluation has been affected by using an objective assessment measure based on the average relative error between the true and estimated noise.

II. WAVELET-BASED SPEECH SIGNALS DECOMPOSITION

A. Second Generation Wavelet (SGWT)

Wavelet functions are traditionally defined as the dyadic translates and dilates of one single mother wavelet function. Such wavelet decomposition requires a regular mesh and unbounded domain. Therefore, such decomposition works well for infinite or periodic signals, but special adaptations of the basis functions near the boundaries are required in order to handle non-periodic boundary conditions, which are often encountered in natural speech. The second generation wavelet transform (SGWT) have been introduced to provide such adaptations as well as maintaining other powerful properties offered by classical WT such as time-frequency localization, multi-resolution and fast implementation [8]. The basic idea behind the second-generation wavelet is to first split a signal $x(n)$ into an even set, $\{x(n): n \text{ even}\}$, and an odd set, $\{x(n): n \text{ odd}\}$, by predicting the odd signal from the even part. What is missed by the prediction is called the detail. The even samples are then adjusted to serve the coarse version of the original signal. The adjustment is needed to maintain the same average for the fine and coarse versions of the same signal. The above process can be summarized as follows (see Figure 1): 1) Split data: even and odd, 2) Predict odd using even: detail = odd − P (even) and 3) Update even using detail: Coarse=even + U (detail).
The inverse transform can be easily constructed by "rewiring" the forward transform. As illustrated in Fig.1, the process of computing a prediction and recording the detail is called a lifting step. In general, the lifting scheme speeds up the implementation as compared to the case of classical WT. All operations within one lifting step can be done in parallel while the only the sequential part is the order of the lifting operations, resulting in an adaptive wavelet transform .

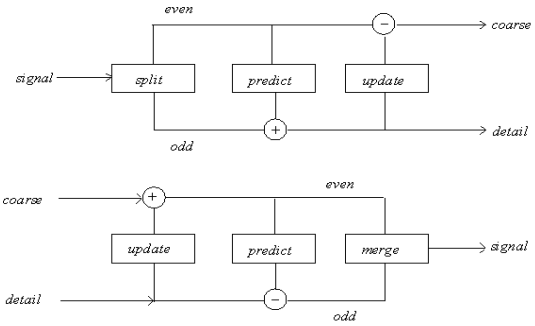


Figure 1. *Representation of the forward and inverse SGWTs.*

B. Perceptual wavelet transform

As widely used in perceptual auditory modelling, we utilise a wavelet packet decomposition (PWPD) scheme which designed to represent the critical bands of a given speech signal. The scheme, which was first proposed by Black and Zeytinoglu [9], is based on an efficient 6-stage tree structure decomposition using 16-tap FIR filters derived from the Daubechies wavelet function, and provides for an exact invertible decomposition. For speech signals sampled at 8 kHz, this decomposition results in 18 critical bands.

III. DESCRIPTION OF THE NOISE ESTIMATION ALGORITHMS

A brief description of the three different noise estimation algorithms and their wavelet-based implementation are given in this section. In what follows we assume that $y(n)$ represents a band limited and sampled noisy speech signal, consisting of a clean speech signal $s(n)$ and a noise signal $w(n)$, such that $y(n) = s(n) + w(n)$. The noisy speech is first decomposed into a appropriate number of bandpass signals, $y_i(n)$, where i denotes the subband index, using either the SGWT or the PWPD, then framed using an appropriate sliding window. Also, $\hat{\sigma}^2_{w_i} = E\{w_i^2\}$ will be used to denote the estimated noise power (or noise variance) at frame p.

A. Adaptive smoothing-based noise estimation

In this technique, the noise and speech are assumed to be independent signals and that the noise power changes slowly. This adaptive noise estimation technique is based on the use of a smoothing parameter that is controlled by the estimated subband *posteriori* SNR [5]. The subband noisy signal power (or variance), $\sigma^2_{y_i}(p) = E\{y_i^2(n)\}$, is estimated on a frame-by-frame basis using [5]:

$$\hat{\sigma}^2_{y_i}(p) = \frac{1}{N}\sum_{n=0}^{N-1} y_i(pN+n) \qquad (1)$$

where $\hat{\sigma}^2_{y_i}(p)$ is the estimated noise power calculated at frame p, and N is the size of the frame. Similarly, the subband noise power is estimated using:

$$\hat{\sigma}^2_{w_i}(p) = \alpha_i(p)\hat{\sigma}^2_{w_i}(p-1) + (1-\alpha_i(p))\hat{\sigma}^2_{y_i}(p) \qquad (2)$$

where $\hat{\sigma}^2_{w_i}(p)$ is the estimate of subband noise power at frame p using a smoothing parameter, $\alpha_i(p)$, which is computed for each frame p using the following formula:

$$\alpha_i(p) = 1 - \min\left\{1, \left(\frac{\hat{\sigma}^2_{y_i}(p)}{\overline{\sigma}^2_{w_i}(p-1)}\right)^{-Q}\right\} \qquad (3)$$

Here Q is an integer and $\overline{\sigma}^2_{w_i}(p-1)$ is the average of the noise estimates of the previous 5 to 10 frames, such that

$$\overline{\sigma}^2_{w_i}(p-1) = 1/10\sum_{k=1}^{10}\hat{\sigma}^2_{w_i}(p-k)$$

B. *Quantile-based noise estimation*

A one-pass noise estimation algorithm is considered here. After decomposition, each subband noisy signal is framed into frames of length L_{frame}. Let $L_{win} > L_{frame}$ be the length of a finite window observation of $y_i(n)$, ranging from 200ms to 2000ms. The method involves first sorting the previous set of data over the last M frames $\left\{y_i^p(n), n = 0, \cdots, L_{win}-1\right\}$ in an ascending order of their values according to the requirement of the quantile-based approach [6]. The noise power in the ith subband of the pth frame, $\hat{\sigma}_{w_i}$, is then estimated as:

$$\hat{\sigma}_{w_i}{}^2 = \beta \sum_{j=0}^{\mathrm{int}(q.L_{win})} (y_i^p(j))^2 \Big/ L_{win} \qquad (5)$$

Where β is an appropriate scaling factor and $q = 0.2$. Here, L_{frame} and L_{win} are chosen to be equal to 32 ms and 512 ms, respectively, with the frames overlapped by 50%.

C. Minimum variance tracking-based noise estimation

The variance in this algorithm can be estimated on a frame-by-frame basis, since both the noisy signal and the noise are considered to be stationary over a short period of time, such that. The noisy signal variance, $\sigma^2_{y_i}$, for each band is calculated as:

$$\sigma^2_{y_i}(p) = \alpha_i\,\sigma^2_{y_i}(p-1) + (1-\alpha_i)\sigma^2_{y_i,new}(p) \qquad (6)$$

where $\sigma^2_{y_i,new}(p) = \frac{1}{N}\sum_{k=0}^{N-1} y_i{}^2(pN+k)$ (7)

is the most recent approximation of the noisy signal variance using the new data at frame *p*. The parameter α_i is a smoothing factor chosen as $0.45 \le \alpha_i \le 0.95$.

The noise estimate $\sigma^2_{w_i}(p)$ is updated such that

$$\sigma^2_{w_i}(p) = \alpha_i\, \sigma^2_{w_i}(p-1) + (1-\alpha_i)\sigma^2_{w_i,new}(p) \quad (8)$$

where $\sigma^2_{w_i,new}$ is the minimum value of $\sigma^2_{y_i}(p)$ in the neighboring frames, i.e. if

$\sigma^2_{y_i}(p-1) < \sigma^2_{y_i}(p)$ & $\sigma^2_{y_i}(p-1) < \sigma^2_{y_i}(p-2)$ & $\sigma^2_{y_i}(p-1) < 2\sigma^2_{w_i}(p-1)$, then

$$\sigma^2_{w_i,new}(p) = \sigma^2_{y_i}(p-1) \quad (9)$$

otherwise $\sigma^2_{w_i,new}(p) = \sigma^2_{w_i}(p-1)$.

C. New noise estimation technique

Based on the modification of the quantile-based method and the addition of a smoothing parameter presented in Section 3.1, a new adaptive noise estimation technique is proposed here. The new technique proceeds as follows. The noise power in the *i*th subband of the *p*th frame, $\hat{\sigma}^2_{wi}$, is estimated as in the standard quantile-based method (Eq.5). This estimate of the noise power is considered here to be equivalent to the average of the noise estimates used in eq.3. Based on this, a smoothing factor, $\alpha_i(p)$, is then introduced such that:

$$\alpha_i(p) = 1 - \min\left\{1, \left(\frac{\hat{\sigma}^2_{y_i}(p)}{\sigma^2_{w_i,quantile}(p-1)}\right)^{-Q}\right\} \quad (10)$$

Where $\sigma^2_{w,quantile}$ is the noise power in the *i*th subband of the *p*th frame as calculated by using (Eq.5). As will be discussed in Section.4, our experimental results have shown that in most cases setting β=1 and α=0.5 result in the best performance of this new noise estimation technique.

IV. Performance evaluation

The performances of the given algorithms were evaluated on different speech signals sampled at 8 kHz, as acquired from the TIMIT database. The distorted speech frames are overlapped by 50% and different types of noise have been used to test the four noise estimation techniques using the SGWT. The noisy speech signals were decomposed into 6 bands (details) using the dB(7-9) wavelet filter [8], and each of the four techniques was then used to estimate the added noise. The real (solid line) and estimated noise for band 2 of the decomposed signal (0.5-1 kHz) resulting from each technique is shown in Fig.2, for the cases of Pink noise. The second part of the evaluation process deals with the perceptual wavelet decomposition. Fig.3 shows the real and estimated noise for band 8 of the decomposition, for the case of White noise. In Fig's 2 and 3, (a) corresponds to the adaptive smoothing-based method, (b) quantile-based method, and (c) the minimum variance tracking-based method. Also, in (a), (b) and (c), a dashed line is used to mark the estimated noise, while in (d) a dashed line is used to mark the noise estimate obtained by a quantile-based method and a dotted line for that obtained by the proposed method.

To provide an objective performance measure, we also calculated the average relative error factor in the estimated noise defined as:

$$ARE = \frac{1}{N_{frame}}\sum_{p}\frac{\left|\hat{\sigma}^2_{wi}(p) - \sigma^2_{w_i}(p)\right|}{\sigma^2_{w_i}} \quad (12)$$

Where N_{frame} represents the number of frames in the test signal. Using this factor, tables 1 and 2 illustrate the performance of the four presented noise estimation techniques for one SGWT subband (band 3 for the SGWT case and band 7 for the PWPD) over different SNRs. Here, T1, T2, T3 and T4 refer to the first, second, third and the proposed noise estimation techniques in the sequence presented in Section 3. It is obvious from this evaluation that all the four techniques considered here demonstrate capability in tracking various types of noise, but their performance accuracy varies depending on the rate of change of the noise under test. The minimum variance tracking-based method seems to offer the best performance in tracking the average noise variation. On the other hand, the adaptive smoothing-based method noise can track rapid changes of stationary and non-stationary noise depending on the value of smoothing parameter. Presented results also demonstrate that the performance of the quantile-based noise estimation method was improved when combined with the adaptive noise estimation method, as proposed in our new noise estimation technique. In particular, significant improvement was achieved by the proposed method for cases of speech signals of relatively low SNRs. The presented noise estimation methods are of clinical interest and can be employed to track the level of noise from pathological speech signals.

For pathological voices, accurate estimated noise spectrum gives a good indication to the degree of the perceived hoarseness.

V. CONCLUSION

The problem of wavelet-based noise tracking has been investigated in this paper using two different decompositions and three noise estimation approaches. The performance of these approaches have been evaluated and compared under different noisy conditions. Our results demonstrate that all three algorithms are capable of tracking both stationary and non-stationary noise, but with varying degree of accuracy depending on the level and rate of change of the noise under consideration. Reported results also show that by modifying the standard quantile-based algorithm, a new adaptive and robust noise estimation method with relatively superior performance to the above three techniques for cases of high additive noise, can be achieved.

REFERENCES

[1] Rainer, M., "An efficient algorithm to estimate the instantaneous SNR of speech signal ", *Proc. Eurospeech'93*, pp. 1093-1096, 1993.

[2] Rainer, M, "Noise power spectral density estimation based on optimal smoothing and minimum statistics ", *IEEE Trans. Speech and Audio Proc.*, 9(5): pp. 504-512, 2001.

[3] Stahl, V., A. Fischer, A, and Bippus, R., " Quantile based noise estimation for spectral subtraction and weiner filtering*", ICASSP'2000*, Istanbul, 40: pp. 1875-1878, 2000.

[4] Evans, N., and Mason, J., "Time-Frequency quantile–based noise estimation", *Proc.EUSIPCO 2002.*

[5] Lin, L., Holmes, W., and Ambikairajah, E.,"An adaptive noise estimation algorithm from speech enhancement", *Proc. Australian Conf. on Speech Science and Techno.* Melbourne, 2002.

[6] Qiang, F. and Eric, W., " A novel speech enhancement system based on wavelet denoising", *Technical Report, OGI, School of Science and Engineering*, Oregon, 2003.

[7] Lin, L., Ambikairajah, E., and Holmes, W.," Speech enhancement for nonstationary noise environment", Proc. *Asia-Pacific Conference APCCAS '02*, 1: pp.177-180, 2002.

[8] Daubechies, I. and Sweldons, W.," Factoring wavelet transforms into lifting steps", *J. Fourier Analy. Appl.;* 4(3): pp. 247-269, 1998.

[10] Black, M., and Zeytinoglu, M.,"Computationally efficient wavelet packet coding of wide-band stereo audio signals", *Proc. ICASSP'95*, 3075-3078, 1995.

Table 1: Average relative error *ARE* in band-3 SGWT for the four noise estimation methods.

SNR dB	ARE-PINK NOISE			
	T1	T2	T3	T4
10	4.88	2.43	0.24	3.41
5	1.60	0.65	0.24	1.18
0	0.56	0.29	0.22	0.45
-5	0.23	0.58	0.20	0.18

Table 2: Average relative error *ARE* in band-7 PWPD for the four noise estimation methods.

SNR	ARE-WHITE NOISE			
	T1	T2	T3	T4
10	2.56	2.20	0.47	2.30
5	0.9	0.66	0.46	0.87
0	0.43	0.47	0.50	0.31
-5	0.25	0.61	0.42	0.20

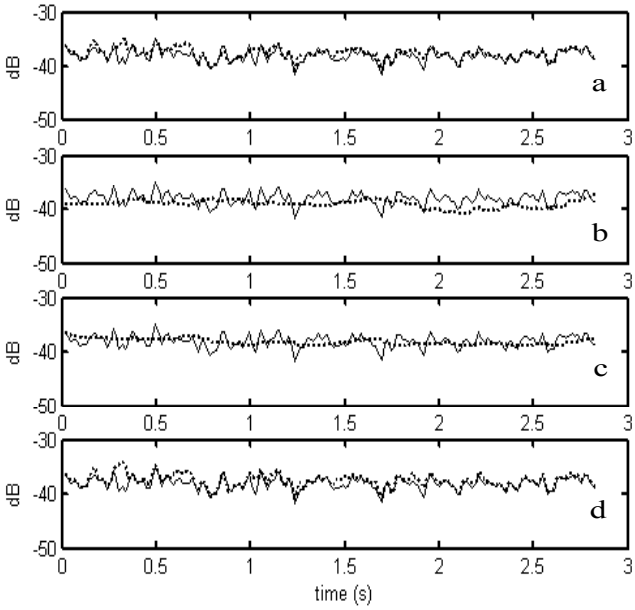


Figure.2: Real and estimated noise using SGWT-based noise estimation with Pink at -5dB

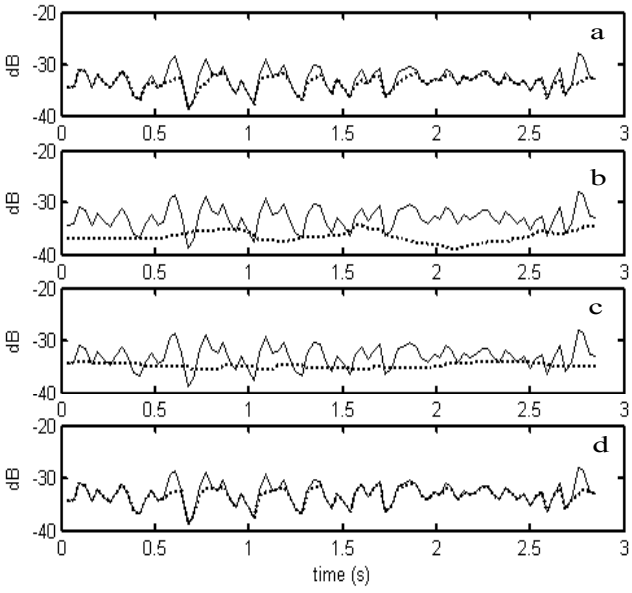


Figure.3: Real and estimated noise using PWPD-based noise estimation with AWGN noise at -5dB

3rd International Workshop MAVEBA 2003, 65-68

HMM MODELLING OF ADDITIVE NOISE IN THE WESTERN LANGUAGES CONTEXT

C. S. Lima[1], J. F. Oliveira [2]
[1]Department of Industrial Electronics, Universidade do Minho, Guimarães, Portugal
[2]Department of Electrical Engineering, Instituto Politécnico de Leiria, Leiria, Portugal

***Abstract*: This paper is concerned to the noisy speech HMM modelling when the noise is additive, speech independent and the spectral analysis is based on sub-bands. The internal distributions of the noisy speech HMM's were derived when Gaussian mixture density distributions for clean speech HMM modelling are used, and the noise is normally distributed and additive in the time domain. In these circumstances it is showed that the HMM noisy speech distributions are not Gaussians, however, fitting these distributions as a Gaussian mixture, only a little bit of loss in performance was obtained at very low signal to noise ratios, when compared with the case where the real distributions were computed using Monte Carlo methods.**

I. INTRODUCTION

In the western languages the intonation does not make part of the linguistic message, so a very fine detail in frequency is not necessary concerning to speech recognition applications, becoming the signal envelope of the most importance. Therefore some spectral components are frequently grouped, for example by sum and each group is known as sub-band.

Recently the importance given to the field of environmental/speaker adaptation has been increased in part to the difficulties in the obtaining of a feature extraction method sufficiently robust against these types of speech variability. The contemporary adaptation algorithms are mostly based on the MLLR algorithm [1], which can't be able to separate speaker mismatch from environmental (additive and convolutional) mismatch. Alternative approaches can deal separately with an additive noise model and a convolutional noise model in both stationary [2] and non-stationary [3] noise conditions in order to separate these two types of distortions. However these algorithms are essentially based on cepstrum based features, which contributes to increase significantly the computational load once that a mapping between the cepstral and linear domains is required. In [4] [5] it is suggested that a proper spectral normalisation can be more useful than the cepstrum derived features in the noisy speech modelling, while [6] proposes an incremental adaptation algorithm based on spectral derived features. The next step is to investigate the drawbacks of using a gaussian mixture to model the internal distributions of the noisy speech HMM's when using power spectral density based features jointly with additive noise in the linear (not cepstral) domain. This is the purpose of this paper.

II. NOISE AND NOISY SPEECH STATISTICS

The use of continuous observation density in HMMs is not restricted to the use of Gaussian mixtures. Although some restrictions must be placed on the form of the model probability density function (pdf) to ensure that the parameters of the pdf can be re-estimated in a constant way, any log-concave or elliptically symmetric density [7] can be used.

Typically the clean speech features are modelled as a Gaussian mixture and generally the existing speech recognisers perform well in clean speech conditions. In noisy conditions the performance degrades in part due to inaccuracies in noise modelling, given that in some situations the noise is artificially generated thus, known. Using power spectral density features and Gaussian distributed additive noise strong evidences exist that the noisy speech distribution can't be Gaussian. In fact if the noise is Gaussian distributed in the time domain it is well known from the statistics theory that it becomes exponentially (chi-square with two degrees of freedom) distributed in the power spectral density domain, which is the feature domain where the distribution of the noisy speech must be computed. Additionally, as usual, some power spectrum density components have to be grouped anyway (in our case by sum) in order to reduce the feature vector dimensionality, which will also be taken into consideration in the obtaining of the noisy speech statistics.

An exponential distribution of parameter λ is defined by the following probability density function

$$f_x(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} U(x) \qquad (1)$$

where U(x) is the unit step function. The exponential distribution is characterised by the fact that its mean is equal to its standard deviation, which is equal to λ. So, the periodogram distribution of a white noise Gaussian stochastic process with zero mean is a white noise exponential stochastic process with zero mean and $\lambda=N\sigma^2$, where σ^2 is the signal variance and N the signal length.

Supposing HMMs with Gaussian sources then the clean speech y has a Gaussian mixture distribution where the distribution of each component of the mixture is given by

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \quad (2)$$

Let y=(y[0], …,y[N-1])^T, x=(x[0], …,x[N-1])^T and z=(z[0], …,z[N-1])^T be, respectively, vectors of clean, noise and noisy signals. If the noise is additive, y[n] is given by

$$z[n] = y[n] + x[n], \text{ n=0, …, N-1.}$$

The autocorrelation function of the noisy speech can be obtained from the autocorrelation functions of the clean speech, the noise the respective cross correlation as follows

$$\begin{aligned}\varphi_{zz}(m) &= E\{z[n]z^*[n+m]\} \\ &= E\{(x[n]+y[n])(x[n+m]+y[n+m])^*\} \\ &= E\{x[n]x^*[n+m]+x[n]y^*[n+m]+ \\ &\quad y[n]x^*[n+m]+y[n]y^*[n+m]\} \\ &= E\{x[n]x^*[n+m]\}+E\{x[n]y^*[n+m]\}+ \\ &\quad E\{y[n]x^*[n+m]\}+E\{y[n]y^*[n+m]\} \\ &= \varphi_{xx}(m)+\varphi_{xy}(m)+\varphi_{yx}(m)+\varphi_{yy}(m)\end{aligned}$$

As the noise is speech independent, the two processes are non-correlated, so the cross-correlations in the above equation are null. Consequently the autocorrelation function of the noisy speech is simply the sum of the autocorrelation functions of the clean speech and noise.

Let Y=(Y(0), …,Y(K-1))^T, X=(X(0), …,X(K-1))^T and Z=(Z(0), …,Z(K-1))^T denote, respectively, vectors of spectral components of clean, noise and noisy signals. As the Fourier transform is a linear operation and the power spectral density is the Fourier transform of the autorrelation sequence, then for additive noise, and considering the analysis window too large the next expression holds

$$|Z(k)|^2 = |Y(k)|^2 + |X(k)|^2$$

Accounting to the nature of the speech signal $|Y(k)|^2$ in the above equation does not represent the true autocorrelation sequence of the speech once that the autocorrelation sequence of an autoregressive process is theoretically infinite. The segment analysis truncates the autocorrelation sequence. However, as this occurs in both the test and training and the autocorrelation of the noise is finite the above equation stays approximately valid.

Therefore each component of the clean speech distribution generates jointly with the Gaussian noise a noisy speech distribution component (z) given by

$$f_z(z) = \int_{-\infty}^{+\infty} f_x(z-y)U(z-y)f_y(y)dy \quad (3)$$

In reference [8] it is proved that the solution for the above integral is

$$f_z(z) = \frac{e^{-\frac{4\lambda z - 4\lambda\mu_y - 2\sigma_y^2}{4\lambda^2}}}{2\lambda}\left(1 + erf\left(\frac{\lambda z - \lambda\mu_y - \sigma_y^2}{\sqrt{2\sigma_y^2}\lambda}\right)\right) \quad (4)$$

where erf stands for error function which is defined by the integral

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \quad (5)$$

For high SNRs equation (4) roughly fits the Gaussian distribution given that the noise distribution approaches the impulse function.

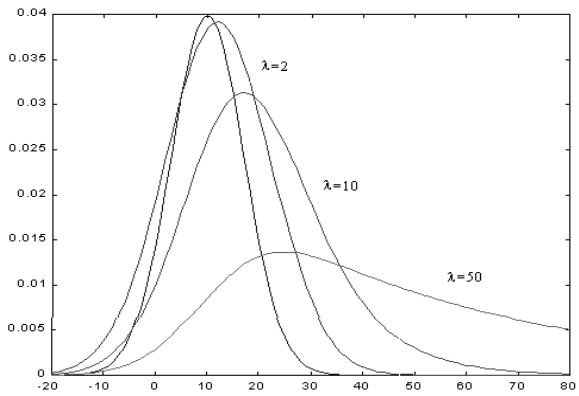


Figure 1. Distribution of the clean and noisy speech for λ=2, 10 (SNR=0dB), 50.

Figure 1 shows the difference between equation (4) and the Gaussian function for λ=2, 10, 50; mean of the Gaussian equals to 10 and variance equals to 100, therefore simulating a SNR=0 dB when λ=10.

For high noises the noisy speech distribution is clearly non-Gaussian and so, the noisy speech distributions have to be changed from Gaussians as usually used, to the function defined by equation (4).

By analysing in the sub-bands context the HMMs for clean speech model the sum of n power spectral contiguous components, instead of only one power spectral component. Therefore, the solution for the noisy

speech sub-band distribution can be obtained from equation (4) taking into account that the means and variances in each model must be divided by *n*, once that they model the sum of *n* random variables with Gaussian distribution and all with the same parameters. Therefore equation (4) holds for the noisy speech distribution, and it would still be necessary develop the distribution of the sum of *n* random variables each one with the distribution given by equation (4).

An easier and equivalent solution is to develop the probability density function of the sum of *n* exponential distributed random variables as shown in equation (1) and perform the convolution of this function with a Gaussian function which models the sum of *n* power spectral components of the clean speech.

Reference [8] shows that the distribution of the sum of *n* random independents and identicaly distributed (according equation (1) variables is

$$f_x(x) = \frac{x^{n-1}\exp\left\{-\frac{x}{\lambda}\right\}}{(n-1)!\lambda^n}U(x) \qquad (6)$$

Equations (2) and (6) allow to derive the probability density function as usual by convolving the two probability density functions

$$f_z(z) = \int_{-\infty}^{+\infty} f_x(x)U(x)f_y(z-x)dx$$

$$= \int_{0}^{+\infty} f_x(x)f_y(z-x)dx$$

$$= \frac{1}{(n-1)!\lambda^n\sqrt{2\pi\sigma_y^2}}\int_0^{+\infty} x^{n-1}e^{\frac{-x}{\lambda}}e^{\frac{-(z-x-\mu_y)^2}{2\sigma_y^2}}dx \qquad (7)$$

The above integral is difficult to calculate due to the term x^{n-1} where *n* is of the order of more than ten, once that the recognition systems nowadays use observation vectors dimensionality from tipically ten to forty (with dinamical characteristics) thus, much smaller than the normally used as FFT length.

III. Aproximated distribution of the noise and noisy speech

By using the Central Limit theorem equation (6) can be approximated by

$$f_x(x) = \frac{1}{4\sqrt{2\pi\lambda}}e^{-\frac{(x-16\lambda)^2}{16\lambda^2}} \qquad (8)$$

The nature of the Central Limit theorem approximation and the required number of variables for a specified error bound, depend on the form of the densities of the summed random variables. For most applications a number of 30 random variables is adequate, however, for smooth distributions a number as low as 5 can be used. In our case we have 16 random variables and no smooth distributions thus, a considerable difference between the real and approximated function can be expected. This difference is shown in figure 2 for λ=10. However, in real situations λ is greater, (order of 10^7 at 10dB), the variance is in order of the square of λ and the function defined by equation (8) fits best to the function defined by equation (6), what is expected by the inspection of figure 2.

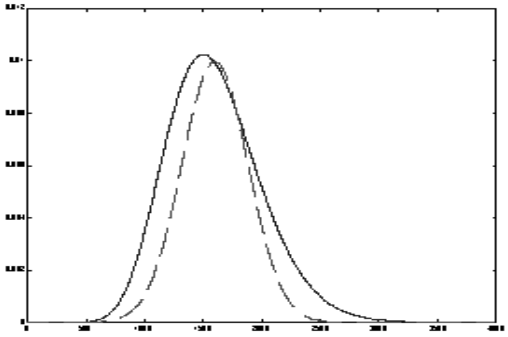


Figure 2. Approximation of the sum of 16 random i. i. d. variables with λ=10, by a Gaussian.

Under this approximation the noisy speech distribution (equation (6)) becomes

$$f_z(z) = \frac{1}{4\sqrt{2\pi}\sqrt{\sigma_y^2+16\lambda^2}}e^{\frac{(z-\mu_y-16\lambda)^2}{2(\sigma_y^2+16\lambda^2)}} \qquad (9)$$

given that the convolution between two Gaussian functions is still a Gaussian function which mean and variance are equal to the sum of the initial means and variances, respectively.

IV. Experimental results

The loss in performance due to the using of equation (9) instead of equation (7), which was computed by numerical integration (exact method), was tested in an Isolated Word Recognition system using Continuous Density Hidden Markov models. The database of isolated words used for training and testing is from AT&T Bell. The used speech was acquired under controlled environmental conditions band-pass filtered from 100 to 3200 Hz, sampled at a 6.67 kHz and analysed in segments of 45 ms duration at a frame rate of 66.67 windows/sec. Only the decimal digits were used. The noise has white noise characteristics, is speech independent and computationally generated at various SNR as shown in table 1. The goal is to compare

the performance of the proposed approximation, exact solution and contemporary speech robust features. Some of these robust features are the OSALPC (One-Sided Autocorrelation Linear Predictive Coding), the conventional cepstrum with liftering (CEPS + liftering) and the well known MFCC (Mel-Frequency Cepstral Coefficients). In table 1, MMC stands for conventional Markov model composition in the power spectrum density domain by using the suggested approximation while NI stands for the numerical integration. Table 1 shows that the suggested approximation is as effective against additive white noise as the exact solution except for very low signal to noise ratios (-5db), where the loss in performance is even so very low. In both cases the noise parameters were learned from the periodogram method in a data segment of 100ms without speech. On the first six entries of the table 1, all the features are 8 static, energy and dynamic features excepting * (12 static + energy + dynamics) and ** (13 static + energy + dynamics).

Table 1 – Performance of the proposed approximation

SNR (dB)	15	10	5	0	-5
LP	56.5	39.5	30	16.25	
OSALPC	98.25	92	65.75	32.25	
CEPS *	97.5	95	72	34.5	
+liftering	98.25	95	75.25	39	
MFCC **	97.75	94.75	72.25	37.5	
OSALPC*	98.5	96.25	74.25	32.5	
MMC	98	96.75	92.5	91	78.5
NI	98	96.75	92.5	91	80.25

V. Discussion

The main advantage of using spectral based features instead of cepstral based features is the decreasing of computational load given that the mapping between the linear and cepstral domains becomes not necessary. In fact, as the noise is considered additive in the linear domain and the features adaptation is performed in the cepstral domain, a mapping from cepstral to linear domain and then an inverse mapping from linear to cepstral domain are needed (Parallel Model Combination). This decreasing in computational load is particularly important on environmental/speaker incremental adaptation where recently some effort has been made in order, for example, to separate speaker mismatch from environmental mismatch or adapting to non-stationary additive noise situations where the channel distortion is stationary. This situation requires training, of the combined HMM's of the clean speech and noise, on the recognising speech (incremental adaptation) which becomes more easy if the internal distributions remain Gaussians. Additionally a proper spectral normalisation [4][5] can be more effective concerned to speech modelling than the cepstral based features, at least for some types of noise. However, the main drawback associated with cepstral based features is related with the difficulty in the modelling of speech dynamics. In fact the adaptation of the dynamic coefficients is not possible, although some approximate solutions have been suggested.

References

[1] C. J. Leggetter, and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech & Language, vol 9, pages 171-185, 1995.

[2] M. J. F. Gales and S. J. Young, "PMC for speech recognition in additive and convolutional noise," Technical Report 154, Cambridge University, 1993.

[3] Z. Wang, P. Kenny, and O'Shaughnessy, "Robust Speech Recognition in Nonstationary Adverse Environments," Proceedings of ICASSP, Seattle, vol. 1, pp 265-268, 1998

[4] C. Lima, L. B. Almeida and J. L. Monteiro, "Improving the Role of Unvoiced Speech Segments by Spectral Normalisation in Robust Speech Recognition," 7th International Conference on Spoken Language Processing (ICSLP'2002), pp 1573-1576, 2002.

[5] C. Lima, L. B. Almeida, A. Tavares and C. Silva, "Spectral Multi-Normalisation for Robust Speech Recognition," *IEEE & ISCA Workshop on Spontaneous Speech Processing and Recognition,* pp 39 - 42, 2003.

[6] C. Lima, L. B. Almeida and J. L. Monteiro, "Continuous Environmental Adaptation of a Speech Recogniser in Telephone Line Conditions," 7th International Conference on Spoken Language Processing (ICSLP'2002), pp 1401-1404, 2002.

[7] S. E. Levinson., L. R. Rabiner and M. M. Sondhi, "An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition," Bell System Tech. J., 62(4): 1035-1074, 1983.

[8] C. Lima, "Speech Recognition in Non-stationary Environments," Ph. D. Thesis, Department of Industrial Electronics, University of Minho, Portugal, 2002.

3rd International Workshop MAVEBA 2003, 69-72

HOARSE VOICE DENOISING FOR REAL-TIME DSP IMPLEMENTATION: CONTINUOUS SPEECH ASSESSMENT

E. Iadanza[1], F. Dori[1], C. Manfredi[1], S. Dubini[1]
[1]Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy

***Abstract*: Voice hoarseness is mainly related to airflow turbulence in the vocal tract. It can be due to vocal fold paralysis, polyps, cordectomisation or other dysfunction, which alter regular speech production, and is commonly treated as a noise component in the speech signal. A denoising approach is proposed, based on low-order singular value decomposition (SVD) of matrices whose entries come from sampled speech data frames, properly organised. A prototype DSP board implementing the procedure was developed. Objective quality indexes are proposed, showing the results achieved with the proposed method both on vowel and consonantal sentences**.
***Keywords:* SVD, hoarse voice, DSP, continuous speech, real-time**

I. INTRODUCTION

This paper deals with the problem of enhancing voice quality for people suffering from dysphonia. This can be due to vocal fold paralysis, cordectomisation or other dysfunction, which alter regular speech production and commonly cause more efforts to be used in speaking than for healthy people. Objective speech quality measures are reliable, easy to implement and have been shown to be good predictors of subjective quality [7], [16]. The main goal of the system presented here is to realise a mobile hardware/software system for real-time voice denoising, to obtain a more intelligible speech with small effort. The method is based on the singular value decomposition (SVD) of matrices whose entries come from sampled speech data frames, properly organised [1]. SVD is widely used for speech enhancement, mainly to improve the performance of speech communication systems in a noisy environment [2], [3], [4]. For the present application, a fixed two-dimensional signal subspace dimension was found sufficient for data filtering, thus allowing real-time implementation. Objective quality measures (PSD ratios, SNR) are defined and evaluated, in order to assess enhancement of voice and compare results. The proposed approach was implemented on a DSP board, by means of properly optimised C and Assembler code. Thus, a simple portable device could be realised, as an aid for dysphonic speakers for diminishing effort in speaking, which is closely related to social problems due to awkwardness of voice.

II. DENOISING WITH SVD

The SVD is a numerically reliable and robust means for estimating the space of clean data (signal subspace) from the white noise corrupted data, and is thus particularly suited for speech denoising [1], [5], [6], [7], [8]. Despite its simplicity, the SVD approach was found effective in increasing voice quality. Extensive simulations were performed and detailed results are reported in [9], [10]. This paper aims at testing the method on continuous speech, to evaluate its performance on consonantal sounds mixed to vocalic ones. Moreover, in order to measure performance, some simple objective quality indexes will be introduced and evaluated.

III. QUALITY MEASURES

Extensive research has been carried out in developing both subjective and objective tests to ascertain quality, but few results are available as far as correlation among them is concerned [16]. In the following, some indexes are proposed, closely related to the signal characteristics. In this work it is assumed that "harmonic" range means frequencies below $f_{th} = 4kHz$, while "noise" range indicates frequencies over this threshold. This threshold is an empiric choice based on analysis of various speech signals; we are currently tuning it using a wider dataset. The subscript "non-filt" refers to the original signal, while "filt" refers to the SVD-filtered signal. The simplest measure is:

$$\mathrm{PSD} = 10\log_{10}\frac{\mathrm{PSD}_{\mathrm{non-filt}}}{\mathrm{PSD}_{\mathrm{filt}}} \qquad (1)$$

representing the ratio of the PSDs, evaluated on the whole frequency range;

$$\mathrm{PSD}_{\mathrm{low}} = 10\log_{10}\frac{\mathrm{PSD}_{\mathrm{non-filt}}(\mathrm{f} \le 4\mathrm{kHz})}{\mathrm{PSD}_{\mathrm{filt}}(\mathrm{f} \le 4\mathrm{kHz})} \qquad (2)$$

measures the ratio of the PSDs evaluated on the "harmonic" range, while

$$\mathrm{PSD}_{\mathrm{high}} = 10\log_{10}\frac{\mathrm{PSD}_{\mathrm{non-filt}}(\mathrm{f} \ge 4\mathrm{kHz})}{\mathrm{PSD}_{\mathrm{filt}}(\mathrm{f} \ge 4\mathrm{kHz})} \qquad (3)$$

is the ratio of the PSDs, evaluated on the "noise" range.

A good denoising procedure should give PSD and $\mathrm{PSD}_{\mathrm{low}}$ values around zero (no loss of power), but high

PSD_{high} values (loss of power due to noise). Finally,

$$SNR = 10\log_{10} \frac{\sum_{n=1}^{M} y^2(n)}{\sum_{n=1}^{M} (y(n) - y_{filt}(n))^2} \quad (4)$$

where: y(n) = noisy signal sample at time n, $y_{filt}(n)$ = filtered signal sample at time n.

Notice that PSD_{low} and SNR have good correlates with NHR [16] and the GIRBAS scale, while being simple and reliable at a very low computational cost. This point will be further exploited in future work.

IV. Experimental results

The denoising procedure was applied here to real data. These concern hoarse pathological voices, coming from adult male subjects that underwent partial cordectomisation, due to T1A glottis cancer. Patients were asked to pronounce the Italian word /aiuole/ (flowerbeds), which is composed of the five principal vowels. This choice is due to the clinical interest in evaluating the effort in speaking made by patients, for surgical and rehabilitation purposes.Besides, the method has been also tested on a pathologic subject pronouncing a 12 sec. sentence taken from Kay Elemetrics disordered voice database, developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab.

The results from SVD filtering procedure applied on the word /aiuole/ were compared to those coming from the complete phrase, by means of the quality indexes described in sect.3 in order to evaluate the method's performance also on non-vocal sounds and silence.

Fig. 1 shows the results relative to one subject (lancet operated) pronouncing the word /aiuole/. The approach lowers the PSD on the whole frequency range (PSD=0.02 dB), and especially on the low frequency range (PSD_{low}=-0.004 dB). This corresponds to a good voice level at the output of the filtering chain. Good value is also found on the high frequency region (PSD_{high}=14.6 dB), and correspondingly a SNR value near to 16 dB (SNR=16.4 dB). Fig. 1 shows the spectrogram of the unprocessed signal (upper plot), as compared to that obtained from the SVD filtering chain (lower plot). For clearness, the frequency range is limited to a maximum of 6 kHz. The lower plot confirms the good denoising properties of the proposed procedures, as the noise level is largely reduced above 4 kHz. As already said, denoising with the proposed SVD approach preserves the temporal and spectral characteristics of the original signal, thus providing a filtered voice of better quality, without distorting effects. Fig. 2-3 plot the results obtained for a 12 sec sentence (hence, not just vowel sounds).

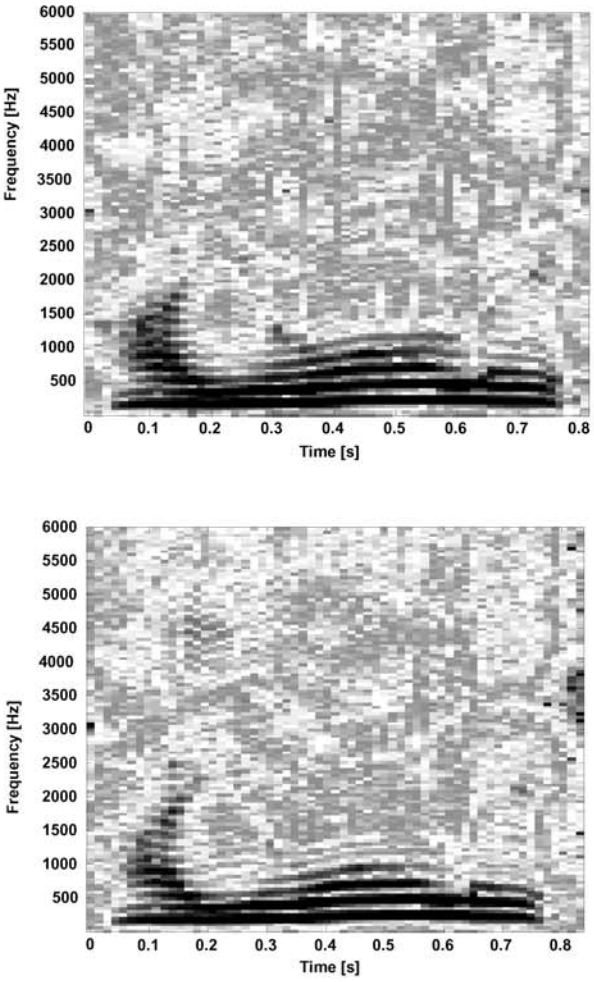


Figure 1 – Spectrogram of the signal before denoising (lower), after denoising (upper).

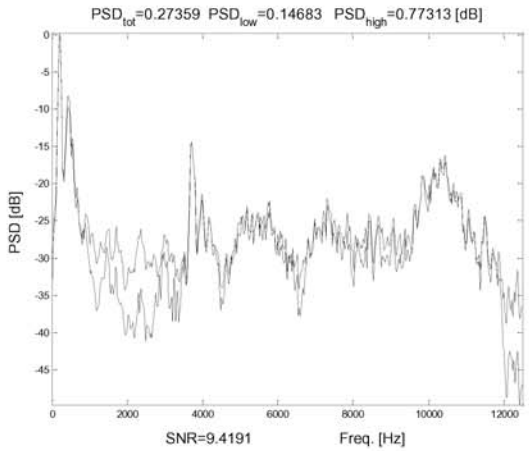


Figure 2 – Comparison of PSD plots for non-filtered (solid line) and for the filtered sentence (dotted line)

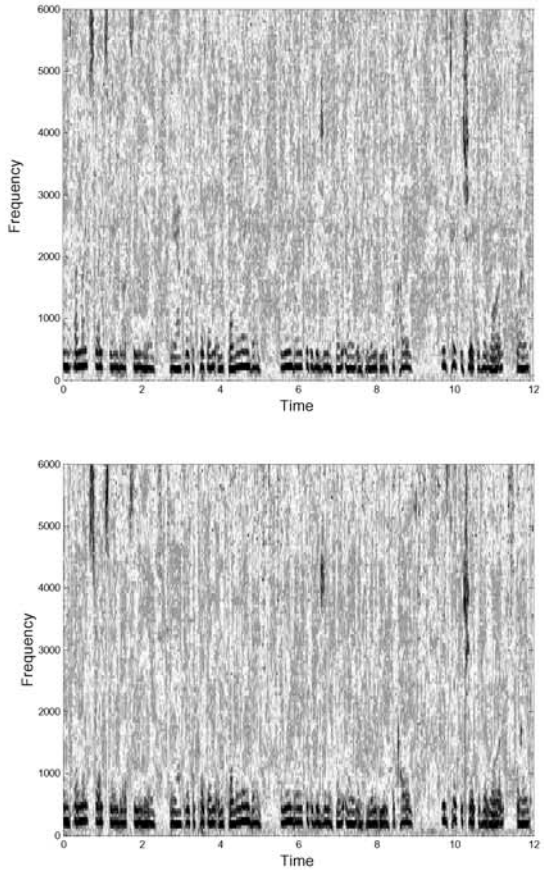


Figure 3 – Spectrogram of the naturally speaking signal before denoising (upper), after denoising (lower).

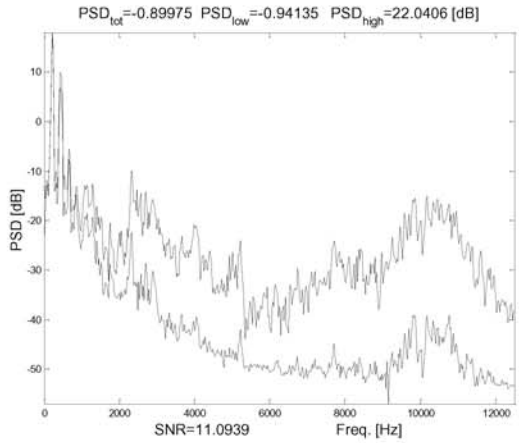


Figure 4 –PSD plots for non-filtered (solid line) and for the filtered naturally speaking signal (dotted line) (/rainbow/).

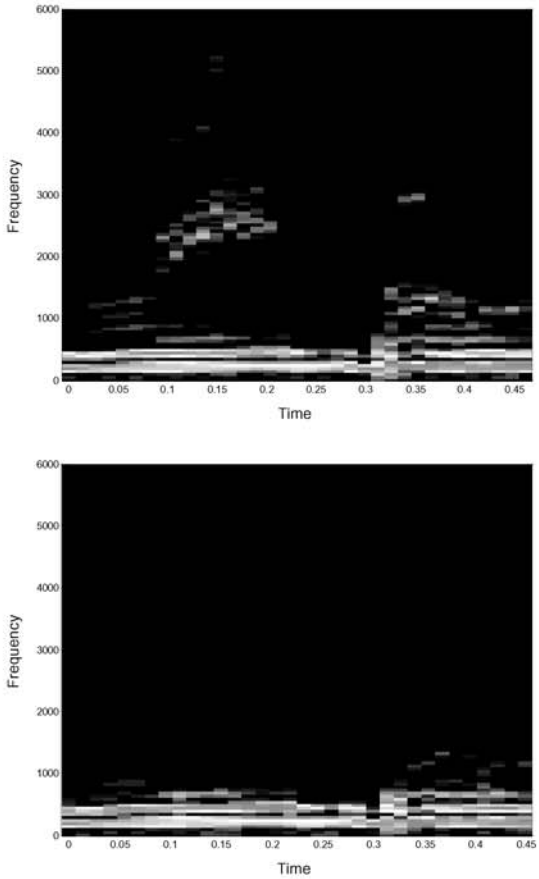


Figure 5 – Spectrogram of the naturally speaking signal before (upper), after denoising (lower) (/rainbow/). Colormap rescaled to fit signal dynamics.

Fig. 2 shows the PSD evaluated for the non-filtered signal (solid line) and for the signal filtered with the proposed method (dashed line). Low PSD values are found both for the PSD on the whole frequency range (PSD=0.274 dB), and for the low and high frequency ranges (PSD_{low}=0.147 dB; PSD_{high}=0.773 dB), and correspondingly a SNR value near to 10 dB (SNR= 9.4191 dB). The results basically correspond to close power values in output and input signals. This means that the system correctly doesn't cut informative signals in unvoiced sounds, even in frequencies above 4 kHz, while it shows strong denoising capabilities in noisy signals. Actually Fig. 3 highlights that noise level is widely reduced above 4 kHz while the so called harmonic range is left nearly unchanged. Specifically, the SVD approach allows lowering the noise component especially with voiced sounds, where the informative content is most in the harmonic range, while has negligible effect for unvoiced sounds, where the informative content is shared out both in the harmonic and in the noise range. Figs. 4-5 point out this aspect. The word /rainbow/ (prevalence of

vocalic sounds) is taken out from the whole sentence, giving good results. Fig. 4 shows very good values both for the PSD on the whole frequency range (PSD=0.9 dB), and especially for the low and high frequency ranges (PSD_{low}=0.941 dB; PSD_{high}=22.041 dB), and correspondingly a SNR value near to 11 dB (SNR= 11.094 dB). Fig. 5 confirms these results, being comparable to those in Fig.1.

V. HARDWARE/SOFTWARE IMPLEMENTATION

The software development tool integrates a C compiler/linker and the DSP/BIOS firmware for implementing a basic kernel with run-time services [11]. The SVD algorithm is implemented by means of a two-step procedure: first, the data matrix A is bi-diagonalised applying a sequence of Householder reflections; second, A is made diagonal using a modified QR algorithm [12-16]. The criteria adopted to implement the hardware platform are:

- High processing performance.
- Low power consumption/Low cost.

The board is supplied with analog front-end, capable to accept the audio signal as input and to furnish the output processed signal at the output stereo jack. The DSP-based board allows to process signals in the 0-48kHz bandwidth. For further details see [17]. The developed hardware was tested with real data in order to reach the real-time processing requirements.

VI. FINAL REMARKS

A simple approach for enhancing voice quality in dysphonic subjects is proposed. The method applies SVD for data filtering, separating the clean signal from its noisy component. The denoised signal is reconstructed along the directions spanned by the principal eigenvectors of the signal subspace. For filtering purposes, the best choice was found that of picking only the two dominant eigenvalues, thus resulting in a low-cost procedure, suitable for on-line implementation on a DSP board. The tests with whole sentences, as well as voiced sounds only, show that this method is suitable both for sustained vowels analysis and for portable application devices.

VII. REFERENCES

[1] Rao B D, Arun K S., "Model based processing of signals: a state space approach", *Proc. IEEE*, vol.80, 1992, pp. 283-309.

[2] Asano F, Hayamizu S, Yamada T, Nakamura S., "Speech enhancement based on the subspace method", IEEE *Trans. Speech Audio Proc.*, vol.8, 2000, pp.497-507.

[3] Ephraim Y, "Statistical model-based speech enhancement systems", *Proc. IEEE*, vol.80, 1992, pp.1526-1558.

[4] Ephraim Y, Van Trees H L., "A signal subspace approach for speech enhancement", *IEEE Trans. Speech Audio Proc.*, vol.3, 1995, pp.251-266.

[5] Klemma V C, Laub AJ. "The singular value decomposition: its computation and some applications", *IEEE Trans. Automat. Control*, vol. 25, 1980, pp.164-176.

[6] Marple S L., "Digital spectral analysis with applications", Prentice Hall, Englewood Cliffs, NJ, 1987.

[7] Deller J R, Proakis J G, Hansen J H L., "Discrete-time Processing of Speech Signals", Maxwell McMillan, New York, 1993.

[8] Manfredi C., "Adaptive noise energy estimation in pathological speech signals", *IEEE Trans. Biomed. Eng.*, vol.47, 2000, pp.1538-1542.

[9] Manfredi C., D'Aniello M., Bruscaglioni P., "A simple subspace approach for speech denoising", *Logopedics Phoniatrics Vocology*, vol.26, p.179-192, 2001.

[10] Manfredi C., Landini L., Faita F., Gemignani V. SVD-based portable device for real-time hoarse voice denoising. Proc. Int. Conf. Digital Signal Processing, Santorini, GR, 2002, pp. 857-860.

[11] Hirano M., "Psycho-acoustic evaluation of voice", In: Hirano M. Clinical examination of voice, Springer-Verlag, New York, 1981.

[12] Golub G.H., Van Loan C.F., "Matrix Computations", 2nd Ed., Johns Hopkins University Press, 1989.

[13] Forsythe G.E., Malcolm M.A., Moler C.B., "Computer methods for mathematical computations", Prentice-Hall, 1977.

[14] Stoer J., Bulirsch R., "Introduction to numerical analysis", Springer-Verlag, 1980.

[15] Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T., "Numerical recipes in C – The art of scientific", Cambridge University Press, 1988.

[16] Dejonckere P H, Remacle M, Fresnel-Elbaz F, Woisard V, Crevier-Buchman L, Millet B, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements", *Rev. Laryngol. Otol. Rhinol*., vol.117, n.3, 1996, pp.219-224.

[17] Manfredi C., Dori F., Iadanza E., "Improvement in hoarse voice denoising for real-time DSP implementation", Proc. Int. Conf. Voice quality: functions, analysis and synthesis, Geneve, 2003

3rd International Workshop MAVEBA 2003, 73-75

AN EFFICIENT METHOD OF SPEECH SIGNAL RECONSTRUCTION BASED ON NEURAL NETWORK AND FAST DECONVOLUTION ALGORITHM

A.M.Krot, H.B. Minervina, V.V. Sarapas

United Institute of Informatics Problems of the National Academy of Sciences of Belarus,
Surganov Str. 6, 220012, Minsk, Belarus

Abstract: **In this paper we propose a new method of speech signal restoration based on a well-known fast deconvolution algorithm and a modern neural network approach. Such a combination inherits the adaptive capability from a neural network as well as the effective inverse filter calculation. According to our expectations, the experimental results reveal the fact that the new method is superior to the traditional ones.**

Keywords: **Neural network, signal restoration, fast inverse deconvolution**

I. INTRODUCTION

There are a number of reasons, which lead to the speech corruption, for example, a vocal tract pathology or pronunciation deficiencies. An efficient reconstruction of such a signal helps to understand and increase the quality of further signal processing. The reconstruction of digital signals can be reduced to the search of a filter, which is inverse to the one that causes the distortion [1], [2]. If the value of the impulse response is known, the distorted signal can be reconstructed with absolute accuracy [2], [3]. There are several different iterative and non-iterative methods for solving the inverse filtering problem (with or without noise) [2]. A neural network filter can solve this problem, but the training phase requires a lot of computation time in order to archive the minimum. There is an efficient fast filtering algorithm for inverting linear convolution by means of sectioning method combined with effective real-valued split radix fast Fourier transform algorithm [4], [5]. In this paper, our purpose is to combine the adaptive capability of neural network for the specification of impulse response of the distorting effect and calculation power of fast deconvolution algorithms for a signal reconstruction.

II. NEURAL NETWORK FOR THE SPECIFICATION OF THE DISTORTION EFFECT

The model of the restoration filter, showing Fig.1, is composed of two components.

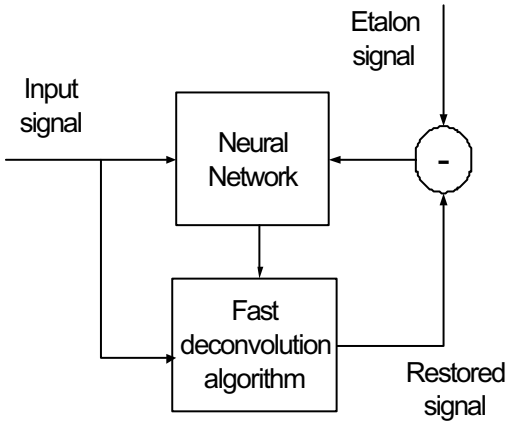


Fig. 1. The restoration structure.

One component is neural network, and another is fast deconvolution algorithm. Neural network estimates the impulse response of the distortion effect. The impulse response could be used for vocal tract pathology diagnostic or pronunciation deficiency classification. Then we use fast deconvolution algorithm for the signal restoration. Difference between standard signal and restored one is used as the training data for the neural network.

A feed-forward three-layer neural network structure can be used to identify the distortion function showing Fig.2.

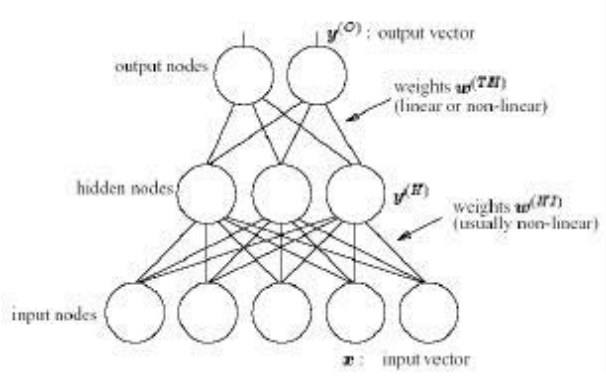


Fig. 2. The structure of a multi-layered neural network

In a most simple case, a feed forward neural network is specified by the following expressions [6]:

$$y_i^{(H)} = f^{(H)}(\sum_j w_{ij}^{(HI)} x_j),$$
$$y_i = f^{(T)}(\sum_j w_{ij}^{(TH)} y_j^{(H)}), \quad (1)$$

where x is the vector of input data, w is the weights coefficients of the neural network, f is neurons activation functions. Sigmoid is typically used as activation function:

$$f(x) = \frac{1}{1+e^{-x}}, \quad (2)$$

The w parameters are determined by means of teaching the neural network, i.e. minimizing the functional suitable for the problem being solved. While building the neural network approximation of a certain function $U = U(x)$ according to the final number of its samples (x, U) with $p = 1,\ldots,P$, the optimization of the difference between the output values of the neural network and the sampled values of the function being modeled takes place as a certain norm, for example:

$$F = \sum_{p=1,\ldots,P} (y_i^{(p)} - U^{(p)})^2, \quad (3)$$

The given problem is solved with the help of standard optimizing methods.

III. Fast restoration algorithm of digital signals

A known overlap-add method is taken as a basis of this algorithm. This method is used for the direct problem - filtering. We will consider the digital filtering of signals by inversion of an LC of the form:

$$y_m = \sum_{n=0}^{m} h_{m-n} x_n, m = 0,1,\ldots,N+M-2, \quad (4)$$

where x_n is the incoming one-dimensional signal, y_m is the distorted signal and h_n is the impulse response which describes a linear FIR-filter. Since the direct method of inverting cyclic convolution (CC) matrix (circulant) seldom gives positive result (there are zeros in Fourier spectrum of impulse characteristics), the relationship of circulant and triangular Toeplitz matrices, which always have inverse matrices, is investigated. According to the proposed approach [3], [4], [5], triangular Toeplitz matrices of LxL size are complemented to the LC matrices of (2L-1)xL size. Then the LC matrices are transformed to the square matrices of CC of 2Lx2L size by complementing them with zeros.

The algorithms based on calculating triangular Toeplitz matrices by means of twice the dimensions can be written as follows [3], [4], [5]:

(1) for i= 1 we compute a 2L-point CC of the form

$$x_l^{(1)} = \sum_{m=0}^{2L-1} \tilde{h}_{((l-m))}^{(-1)} \tilde{y}_m^{(1)}, l = 0,1,\ldots,2L-1, \quad (5)$$

where $\{\tilde{h}_m^{(-1)}\} = \{h_0^{(-1)},\ldots,h_{L-1}^{(-1)},0,\ldots,0\}$, $\{\tilde{y}_m^{(1)}\} = \{y_0^{(1)},\ldots,y_{L-1}^{(1)},0,\ldots,0\}$, are 2L-point sequences, $((l-m)) = (l-m) \bmod 2L;$

(2) for i ≥ 2 we form the (N-1)-point sequence $\{\bar{x}_l^{(i)}\} = \{x_{M+l}^{(i-1)}\}, l = 0,1,\ldots,N-2$ and compute the CC of the form

$$\bar{y}_m^{(i)} = \sum_{l=0}^{2N-1} \tilde{h}_{((m-l))} \tilde{\bar{x}}_l^{(i)}, \; l = 0,1,\ldots,2N-1, \quad (3)$$

where $\{\tilde{h}_n\} = \{h_0,\ldots,h_{N-1},\ldots,0,\ldots,0\}$, $\{\tilde{\bar{x}}_l^{(i)}\} = \{\bar{x}_0^{(i)},\ldots,\bar{x}_{N-2}^{(i)},0,\ldots,0\}$ are 2N-point sequences, $((m-l)) = (m-l) \bmod 2N;$

(3) for i ≥ 2 we form the L-point sequence $\{y_m^{(i)}\} = \{\bar{y}_0^{(i)},\ldots,\bar{y}_{N-2}^{(i)}, y_{N-1}^{(i)},\ldots,y_{L-1}^{(i)}\}$ and then compute the 2L-point CC:

$$x_l^{(i)} = \sum_{m=0}^{2L-1} \tilde{h}_{((l-m))}^{(-1)} \tilde{y}_m^{(i)}, l = 0,1,\ldots,2L-1, \quad (6)$$

where $\{\tilde{h}_m^{(-1)}\} = \{h_0^{(-1)},\ldots,h_{L-1}^{(-1)},0,\ldots,0\}$ and $\{\tilde{y}_m^{(i)}\} = \{\bar{y}(_0^{(i)},\ldots,\bar{y}_{N-2}^{(i)},y_{N-1}^{(i)},\ldots,y_{L-1}^{(i)},0,\ldots,0\}$ are 2L-point sequences, $((l-m)) = (l-m) \bmod 2L;$

The computational complexity of algorithm (3)-(5) is given by expressions:

$$M(R) = O((6\log_2 L + 12/L - 5)R), \quad (6a)$$

$$A(R) = O((18\log_2 L + 20/L - 9)R), \quad (6b)$$

which show a gain over the initial algorithm [7], the computational complexity of which is characterized by relations :

$$M(R) = O(1.25(L + 2/5)R), \quad (7a)$$

$$A(R) = O(1.25(L - 2 + 8/(5L))R). \quad (7b)$$

Thus, despite the view held by the author [7], the cost of solving inverse filtering problems using inversion of an LC by sectioning [2],[3],[4] is reduced by using FFT algorithms (in this case, the real-valued split-radix FFT (RFFT-SR) algorithm [1],[2] one of the best). As a result we comes to the matrix of CC two times larger in size then the initial Toeplitz one, but it can be calculated on the basis of effective fast algorithms.

The proposed fast inverse convolution algorithm for reconstructing distorted signals by sectioning the inverse convolution and using the RFFT-SR algorithm was programmed. A computer experiment was performed in which distorted sequences of P=1024-4096 readings were reconstructed, using impulse responses of various lengths (N=65, 129, 257, 513 elements). When impulse response is quite long (N ≥257), the signal is reconstructed 1.6 times faster on average by the proposed algorithm than with the approach based on the algorithm of [7]. The advantage is especially pronounced when the sequence is fairly long, of length (P≥2048).

IV. Conclusion

In this paper, a new approach of neural network filter combined with the fast algorithm based on the sectioning method and the efficient real-valued FFT algorithm has been proposed. The filter reduces the training time in contrast to traditional neural network filters. Moreover, this filter possesses adaptive capability than the traditional inverse filters.

References

[1] A.M.Krot, *Discrete Models of Dynamic Systems Based on Polynomial Algebra.* Minsk: Navuka i tekhnika, 1990 (in Russian).

[2] A.M.Krot and H.B.Minervina, *Fast Algorithms and Programs for Digital Spectral Processing of Signals and Images*. Minsk: Navuka i tekhnika, 1995 (in Russian).

[3] A.T.Kas'ko, A.M.Krot and H.B. Minervina, "A fast algorithm for calculating the inverse convolution for signal and image restoration", *Comput. Maths Math. Phys.,* vol. 36, no.2, pp.269-277, 1996.

[4] A.M.Krot and H.B.Minervina, "Minimal multiplicative complexity and fast restoration algorithm of digital signals and images", *Proc. of Annual SPIE Symposium "AeroSense",* vol. 3374, Orlando, Florida, USA, pp. 426-435, April 13-15, 1998.

[5] H.B. Minervina, *"An express-reconstruction of distorted speech by inverse filtering method", Proc. of 2nd Intern. Workshop 'Models and Analysis of Vocal Emissions for Biomedical Applications",* Firenze, Italy, pp. 152-156, September 13-15, 2001.

[6] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine Learning, Neural and Statistical Classification,* University of Strathclyde, MRC Biostatistics Unit, Cambridge and University of Leeds, 1994, p.298.

[7] B.V.Titkov, "Algorithm for inversion of discrete convolution based on sectionalization method", *U.S.S.R. Comput. Math. and Math. Phys*., vol.32 (a translation of *Zh. Vychisl. Mat. i Mat. Fiz*., vol.32, no.2, pp.119-207, 1992).

Poster session

3rd International Workshop MAVEBA 2003, 79-82

ON PACKET LOSS CONCEALMENT USING TIME-VARYING SPEECH ANALYSIS

Keiichi Funaki

Computing and Networking Center, University of the RYUKYUS
Senbaru 1, Nishihara, Nakagami, OKINAWA, 903-0213, JAPAN
funaki@cc.u-ryukyu.ac.jp

Abstract: Demand of IP telephone is increasing more and more as broadband IP network is being commonly used by many people. In VoIP, packet loss concealment (PLC) is one of the key subjects to keep speech quality since packet loss occurs in IP network. PLC methods based on Linear Predictive coding (LP) method have been proposed in which LP coefficients and LP residual are repeated to recover speech corresponding to the packet loss. However the repetition would not perform well in any speech frames. This paper presents a novel PLC method based on time-varying speech analysis and synthesis, in which AR parameters can be predicted owing to its time-domain function of AR parameter. Three kinds of AR parameter prediction methods by means of time-varying analysis are evaluated subjectively and objectively and novel PLC scheme switching the AR prediction methods with respect to F0 prediction gain is proposed.
Keywords: VoIP, PLC, Time-varying speech analysis

I. INTRODUCTION

Recently broadband IP network becomes to be commonly available, consequently VoIP (Voice Over IP) is being paid attention more and more since IP telephone makes it possible to cut a cost for telephony. In best effort type of IP network packet loss occurs due to transmission error, packet collisions, or so on while PSTN network keeps its quality. In VoIP, accordingly packet loss concealment (PLC) scheme is required to keep quality of transmitted speech. Several PLC algorithms have already been proposed[1][2][3][4][5][6]. These can be categorized into the following four types. (1) Waveform for the last correctly received frame is repeated on the packet loss frame[1][2]. (2) Linear Predictive coding (LP) analysis is carried out with decoded PCM speech and the residual is calculated on the last correctly received frame. On the packet loss frame speech is recovered by filtering with the repeated LP coefficients and residual[3][6]. (3) Speech coding parameters on the last correctly received frame are used to compensate the LP coefficients and excitation on the packet loss frame[4]. (4) Special transmitted codes are defined to be robust against the packet loss[5].

(1) is low-complexity method and can be realized easily. However discontinuity of speech occurs on frame boundaries. Although (3) is commonly used, it depends on speech coding algorithm. On the other words it can not be applied to other speech coding algorithms. (4) determines all of packet form. Hence, it can not be applied to other speech coding methods as well. (2) is coder independent and maintains backward compatibility. However, the LP coefficients repetition would not always perform well for any speech frames.

We are interested in type (2), namely low delay and receiver-based PLC method which operates decoded PCM speech and can be applied to any speech coding and adds no algorithmic delay.

In this paper, new receiver-based PLC scheme using time-varying speech analysis is proposed. Time-varying speech analysis methods estimating the parameter sample by sample have already been proposed[7][8]. On the other hand complex-valued speech analysis for analytic signal has already been proposed[9][10]. Analytic signal is complex-valued signal. These methods can achieve more accurate spectral estimation due to the attractive feature of analytic signal. However, these can not extract time-varying feature from speech signal. We have already proposed time-varying complex AR (TV-CAR) speech analysis methods for analytic signal[11][12][13][14][15][16].

In these TV-CAR speech analysis methods, AR parameters are modeled by complex basis expansion as a function of time. These methods can estimate the parameters for each sample on the analysis interval as well as can predict the parameters for each future sample. This feature may perform well for PLC. To compensate the parameters on packet loss frame, the estimated parameters at last sample on last correctly received frame and the parameters on current packet loss frame predicted by the analysis parameters on the last frame can be utilized. Hence, it is expected that these can realize better quality of recovered speech than LP coefficients repetition.

II. TV-CAR SPEECH ANALYSIS

A. Speech production model

In time-varying complex AR analysis, target signal is an analytic signal defined as in Eq.(1). Analytic signal is complex-valued signal whose real part is observed speech signal and imaginary one is Hilbert transformed signal of real part.

$$y^c(t) = \frac{y(2t) + jy_H(2t)}{\sqrt{2}} \tag{1}$$

where $y^c(t)$, $y(t)$, and $y_H(t)$ denote an analytic signal at time t, an observed signal at time t, and a Hilbert transformed signal for the observed signal $y(t)$, respectively. Since analytic signals provide the spectra only over the range $(0, \pi)$, analytic signals can be decimated by a factor two. In Eq.(1), analytic signal is divided by the term of $\sqrt{2}$ in order to adjust the power of an analytic signal with that of the observed one.

Speech production model, viz. time-varying complex AR (TV-CAR) model is defined as follows.

$$a_i^c(t) = \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) \tag{2}$$

$$\begin{aligned} y^c(t) &= -\sum_{i=1}^{I} a_i^c(t) y^c(t-i) + u^c(t) \\ &= -\sum_{i=1}^{I} \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) + u^c(t) \end{aligned} \tag{3}$$

where $u^c(t)$, $a_i^c(t)$, I, L are complex-valued input, complex AR parameter, AR order, order of complex basis expansion, respectively. $f_l^c(t)$ is any kinds of complex basis function, such as complex Fourier basis $f_l^c(t) = e^{-j2\pi lt/N}$ or first order polynomial $f_0^c(t) = 1, f_1^c(t) = t$. Complex parameter $g_{i,l}^c$ is estimated for each frame and AR parameters for each sample are calculated by Eq.(2). Moreover, by using Eq.(2) one can not only estimate AR parameters for each sample, especially center sample or last sample of the frame but also can predict AR parameters for future sample due to the function of time.

B. MMSE algorithm

MMSE-based TV-CAR speech analysis[11] is adopted to predict AR parameters for packet loss frame. It is expected that complex analysis can realize better speech quality for wideband speech coding such as[17] due to twice band width of analytic signal. However it is difficult to realize Hilbert transform on the following correct frame after the packet loss frame. In this paper, PLC method is constructed by using the real-valued MMSE algorithm for observed real-valued speech signal.

III. PLC METHOD

Proposed PLC scheme treats AR parameters and excitation separately. Fig.1 shows block diagram of the proposed PLC method.

On correctly received frame decoded PCM speech is analyzed by AR analysis and residual is calculated and then the parameters and residual are stored in buffer. AR synthesis is carried out to generate the speech with AR parameters and residual in this frame.

On packet loss frame AR parameters and excitation are predicted by the stored parameters and residual calculated on the last correctly received packet. AR synthesis is carried out to recover the speech on packet loss frame using the predicted AR parameters and excitation.

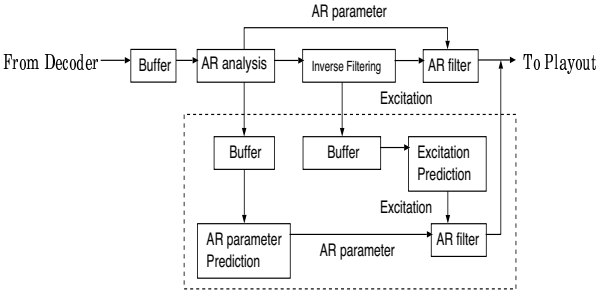


Fig.1: Block diagram of PLC algorithm

A. Prediction method of AR parameter

In LPC method, LP coefficients on last correctly received packet are repeated to predict the coefficients on packet loss frame[3][4][6].

The following three kinds of AR parameter prediction methods using MMSE-based time-varying speech analysis are proposed.

1. Prediction method(1): AR parameter estimation at last sample on the last frame

Time-varying analysis can estimate the AR parameters for each sample. AR parameters at last sample on the correctly received frame can be estimated by Eq.(2). It is thought that the parameters at last sample are more similar to those on next frame. In the prediction method (1), the parameters estimated at the last sample by Eq.(2) are used for the parameters on next packet loss frame.

2. Prediction method(2): AR parameter prediction at the center sample on packet loss frame

Time-varying analysis can predict the AR parameters in future due to the time-domain function of Eq.(2). In the prediction method (2), the parameters predicted by Eq.(2) at the center sample on packet loss frame are used.

3. Prediction method(3): AR parameter prediction at each sample on packet loss frame

In the prediction method (3), the parameters predicted by Eq.(2) at each sample on packet loss frame are used.

B. Prediction of excitation

Excitation is predicted by repeating the residual of the last correctly received frame with the last fundamental period for voiced frame and frame length for unvoiced frame, respectively. In order to avoid unnatural sound amplitude of excitation is reduced to 80 %.

C. Synthesis

On correctly received frame as well as on packet loss frame AR synthesis is done to generate speech signal in order to keep the filter state updating. On packet loss frame bandwidth expansion is operated by using Eq.(4) to avoid unnatural sound as well. AR parameters are linear-interpolated over LSP between frames.

$$a_i'(t) = a_i(t)(0.98)^i \qquad (i = 1, 2, ..., I) \tag{4}$$

IV. EXPERIMENTS

The proposed PLC methods are compared with two conventional methods, waveform repetition and LP coefficients repetition.

A. Analysis Condition

Five kinds of PLC algorithms shown in Table 1 are evaluated. Method(0) is the PLC method repeating speech with the last fundamental period for voiced frame and with frame length for unvoiced frame. Analysis conditions for each analysis are shown in Table 2. Analysis order I is 14 for both methods. In time-varying analysis order of expansion L is 2 and first order polynomial is adopted as a basis function.

Speech data are $8[KHz]$ sampled sentence data converted from ATR database data, whose speakers are male /MYI/ and female /FKN/. Packet length is set to be $20[msec]$. Speech analysis length N and shift length S are set to be the same as packet length. Packet loss is generated randomly at 10%.

Table 1: PLC method

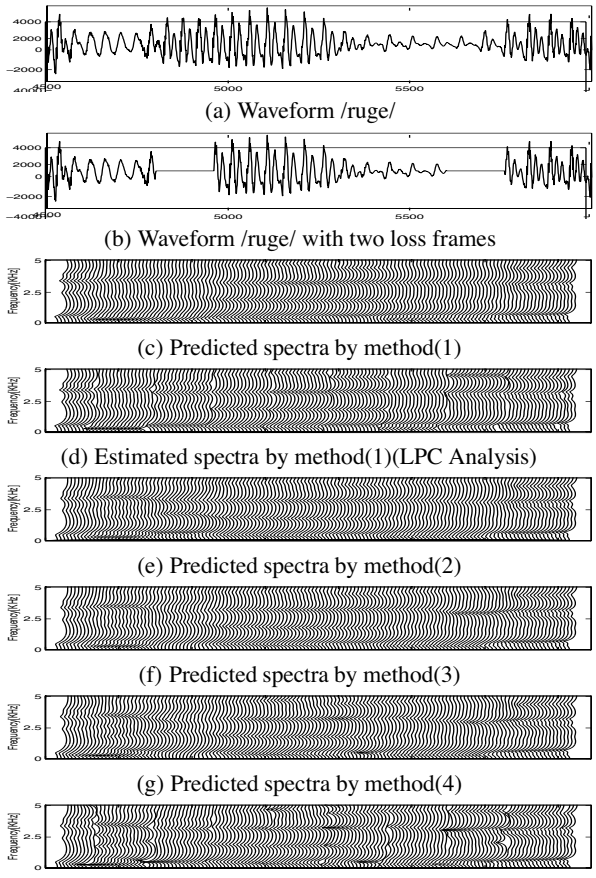
	Speech Analysis	AR Prediction
Method (0)	Repetition	
Method (1)	LPC	Repetition
Method (2)	Time-varying analysis	Prediction method(1)
Method (3)	Time-varying analysis	Prediction method(2)
Method (4)	Time-varying analysis	Prediction method(3)

Table 2: Analysis Conditions

Speech Analysis	I	L	$N[msec]$	$S[msec]$
LPC	14	-	20	20
Time-varying analysis[11]	14	2	20	20

B. Predicted spectra

Fig.2 shows the predicted spectra by proposed AR prediction methods and LP coefficients repetition. (a) shows the speech waveform /ruge/ without packet loss. (b) shows the speech waveform /ruge/ with 2 packet losses, (4800,4960) and (5600,5760). Amplitude corresponding to the loss frames is zero in (b). (d) and (h) indicate the estimated spectra by LPC and time-varying speech analysis, respectively. (c),(e),(f),(g) indicate the predicted spectra by means of method (1),(2),(3),(4), respectively. In (c),(e),(f),(g), the predicted spectra are plotted on the loss frames and the estimated spectra by speech analysis are plotted on the other frames. Note that linear-interpolation over LSP is done on frame boundaries. Spectra are being plotted every $1.25[msec]$. Fig.2 suggests that proposed AR prediction methods are suitable for stationary voiced frame and not suitable for unvoiced frame while LPC method manages to predict appropriate parameters for unvoiced frame.



(a) Waveform /ruge/

(b) Waveform /ruge/ with two loss frames

(c) Predicted spectra by method(1)

(d) Estimated spectra by method(1)(LPC Analysis)

(e) Predicted spectra by method(2)

(f) Predicted spectra by method(3)

(g) Predicted spectra by method(4)

(h) Estimated spectra by method(4)(time-varying analysis)

Fig.2: Estimated and predicted spectra

C. Listening test

In order to evaluate five kinds of PLC methods preference test is organized. Listeners are 5 adult males and 5 adult females. 5 sentences uttered by male speaker and 5 sentences uttered by female speaker are used. Listeners select their preference for each pair sentence. The number of pair sentence is 250 which includes same sentence pair generated by the same PLC method. Fig.3 presents the selected number for each method. Fig.3 demonstrates that LPC method (LP coefficients repetition) can achieve best recovery than other PLC methods. The reason why the proposed methods are not superior to LPC method is that the proposed methods can predict inappropriate AR parameters for unvoiced frame although the proposed methods can predict appropriate AR parameters for voiced frame, as shown in Fig.2.

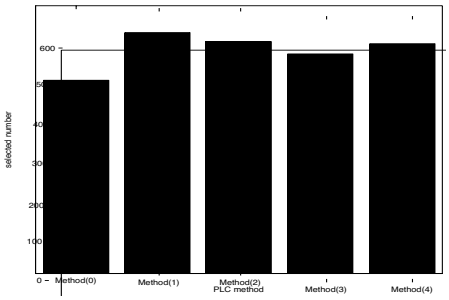


Fig.3: Selected Number

D. Spectral distance

Spectral distance between AR parameters estimated by speech analysis and AR parameters predicted by the speech analysis parameters on the last frame with four kinds of PLC method(1)-(4) is calculated. Speech analysis conditions are the same as in Table 2.

Fig.4 shows 14-th order LPC cepstral distance for method (1) to (4). In Fig.4, solid line, dotted line, dashed and dotted line and dashed line denote the distance for method(1),(2),(3),(4), respectively. X-axis denotes frame number in future and Y-axis denotes cepstral distance. 10 sentences uttered by male speaker and 10 sentences uttered by female speaker are used. Fig.4 demonstrates that LPC method can predict more suitable AR parameters than other methods. It can be expected that the proposed methods can achieve better prediction for voiced frame, especially for stationary voiced frame. Therefore, speech frame is classified into 4 modes by using the following F0 prediction gain PG.

$$PG = 10log_{10}\frac{\sum_{t=N/2}^{N-1} x(t)^2}{\sum_{t=N/2}^{N-1} x(t)^2 - \frac{\left(\sum_{t=N/2}^{N-1} x(t)x(t-T_0)\right)^2}{\sum_{t=N/2}^{N-1} x(t-T_0)^2}} \quad (5)$$

where $x(t)$, N and T_0 are speech signal at time t on the last frame, frame length and estimated fundamental period on the last frame, respectively. Modes are classified into 4 by criteria shown in Table 3.

Table 3: Mode Selection

Unvoiced or $PG < 2$	mode 0	Unvoiced
$2 \leq PG < 5$	mode 1	Voiced
$5 \leq PG < 9$	mode 2	Voiced
$PG \geq 9$	mode 3	Voiced

Mode 0 means unvoiced mode. Mode 1 to 3 mean voiced mode and mode 3 is the stationary voiced mode.

Fig.5 shows 14-th order LPC cepstral distance for each mode in next frame. In Fig.5, each bar means the distance for mode 0, mode 1, mode 2, mode 3, and average from left to right, respectively. Fig.5 demonstrates that method (2) achieves better prediction than LPC except for mode 0. Method (2")

is similar to method (2) in which AR parameters estimated at center sample on the last frame are used as the recovered parameters. Method (2") offers best prediction for mode 0 while method (2) offers best prediction for mode 1 to 3. We can expect that switching of method(2) and method(2") with respect to voiced or not may achieve better prediction than LPC method. The prediction methods for future sample, method (3) and (4) can not achieve good prediction. The reason is that predicted length is too long. In method (3) parameters at $10[msec]$ of future sample are predicted and in method (4) parameters for $20[msec]$ of future samples are predicted. In order to examine the limitation of the prediction in future, spectral distance from 10 to 80 sample in future are calculated. The results are shown in Fig.6. In Fig.6, solid line, dotted line, dashed and dotted line and dashed line denote the distance for mode 0,1,2,3 and average, respectively. Fig.6 demonstrates that in mode 3 better prediction is accomplished and the limitation of prediction is 40 sample($5[msec]$) in future while appropriate AR parameters can not be predicted in mode 0, 1, 2. By taking these facts into account, we can conclude that switching of method (2) and (2") with respect to voiced or not may achieve better prediction than LPC. The method (3) and (4) may achieve better prediction of parameter in only mode 3 up to $5[msec]$ in future.

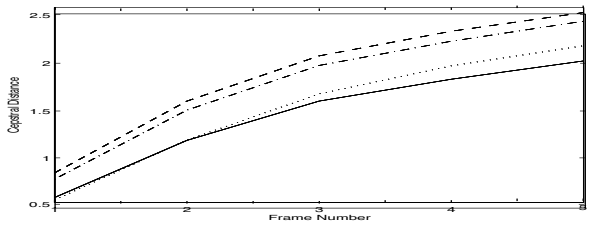


Fig.4: Cepstral Distance for future frame

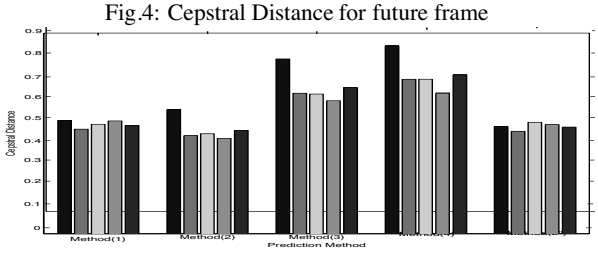


Fig.5: Cepstral Distance for each mode

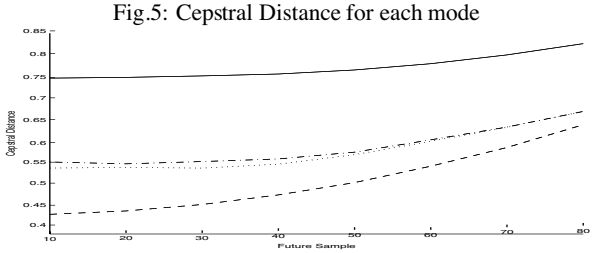


Fig.6: Cepstral Distance in future sample

According to informal listening test the switching PLC method achieves better recovery for packet losses than LP coefficients repetition.

V. Conclusions

This paper has proposed receiver-based PLC algorithms based on time-varying AR analysis. Three kinds of AR prediction using time-varying analysis have been proposed and evaluated by using listening test and spectral distance. Novel PLC methods in which AR prediction is switched with respect to mode for the last frame have been proposed. According to informal listening test the novel PLC method achieves better speech recovery. Formal listening test is continuous way.

VI. Acknowledgements

This work is partially supported by Grant-in-Aid of Japan Society for the Promotion of Science.

VII. References

[1] ITU-T Recommendation G.711 Appendix I, "A high quality low-complexity algorithm for packet loss concealment with G.711," 1999.

[2] O.J.Wasem,et al., "The effect of waveform substitution on the quality of PCM packet communications," IEEE Trans. ASSP. Vol.36, No.3, March 1988.

[3] E.Gunduzhan and K.Momtahan, "A linear prediction based packet loss concealment algorithm for PCM coded speech," IEEE Trans. Speech and Audio Processing, Vol.9, No.8, Nov. 2001.

[4] ITU-T Recommendation G.729, "Coding of speech at 8kb/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)," 1996.

[5] T.Morinaga,et al., "Robust speech coding under packet-loss conditions using recovery sub-codec for broad-band IP network,"Proc. ICASSP-2002, May 2002.

[6] J.Lindblom and P.Hedelin, "Packet loss concealment based on sinusoidal extrapolation,"Proc. ICASSP-2002, May 2002.

[7] Y.Grenier, "Time-dependent ARMA modeling of non-stationary signals," IEEE Trans. ASSP-31, pp.899-911, 1983.

[8] T.Kiryu and T.Iijima, "Estimation of time-varying parameters without assuming local stationary," IEICE Trans. Vol.J68-A No.9, 1985. (in Japanese)

[9] S.M.Kay, "Maximum entropy spectral estimation using the analytic signal," IEEE Trans. ASSP-26, pp.467-469, 1980.

[10] T.Shimamura and S.Takahashi, "Complex linear prediction method based on positive frequency domain," IEICE Trans., Vol.J72-A, pp.1755-1763, 1989. (in Japanese)

[11] K.Funaki,et al., "On a time-varying complex speech analysis," Proc. EUSIPCO-98, Sep. 1998.

[12] K.Funaki,et al., "On robust speech analysis based on time-varying complex AR model," Proc. ICSLP-98, Dec. 1998.

[13] K.Funaki,"A time-varying complex speech analysis based on IV method," Proc. ICSLP-2000, Oct. 2000.

[14] K.Funaki, "A time-varying complex AR speech analysis based on GLS and ELS method," Proc. EUROSPEECH-2001, Sep. 2001.

[15] K.Funaki, "GLS-based TV-CAR speech analysis using forward and backward linear prediction," Proc. IEEE MMSP-2002, Dec. 2002.

[16] K.Funaki, "Improvement of the ELS-based time-varying complex speech analysis," Proc. ICSLP-2002, Sep. 2002.

[17] ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-rate Wideband (AMR-WB)," 2002.

3rd International Workshop MAVEBA 2003, 83-86

VLSI IMPLEMENTATION OF A LMS BASED ADAPTIVE NOISE CANCELLER

F. Gittel [1,2], T. D. Smith [1], A. Th. Schwarzbacher [1] and E. Hilt [2]

[1] Dublin Institute of Technology, School of Electronic and Communications Engineering, Dublin 6, Ireland
[2] Deutsche Telekom University of Applied Sciences, Leipzig, Germany

Abstract: **To better integrate disabled persons is a continuous aim in a modern society. For handicapped people, robots are used to support the personal freedom and provide more convenience. These robots need to be controlled by voice which requires a reliable working speech recognition system. Therefore, algorithms that can improve the quality of speech and thus support the detection of the speech information are highly desirable.**

This paper introduces a hardware implemented and optimised Adaptive Noise Canceller (ANC), which can be utilised in speech detection devices to reduce the noise intensity of the speech to be recognised. In addition, it can also be used to improve the speech quality in information transfer systems. The evaluation results show how the circuit is able to reduce the unwanted components within a speech signal and therefore, the system is able to increase the speech quality. Furthermore, any prior knowledge of the surrounding environmental properties is not needed.
Keywords: **Noise Cancelling, LMS, VLSI**

I. INTRODUCTION

This paper describes the VLSI hardware implementation of an Adaptive Noise Canceller (ANC) which is able to filter an input speech signal to provide noise reduced output speech. To achieve this goal, the digital filter, which is the main section of the ANC, needs to adjust its frequency response continually to the changing conditions of the surrounding environment. Therefore, an update functionality must be introduced. This functionality is based on the Least Mean Square (LMS) algorithm [1]. The method of least mean square adaptive filtering takes advantage of the quasi – periodic nature of the speech signal to form an estimate of the clean speech signal at time *t*. This estimation is derived from the value of the signal at time *t-T* which represents the actual time shifted by one estimated pitch period. The principle of this method is shown in Figure 1. To describe this approach, some considerations have to be taken into account. In practice, an a priori knowledge to adjust the filter response is not available. The output of the FIR filter used for this implementation is given by

$$y(n) = \sum_{i=0}^{L} b_i x(n-i-T) \qquad (1)$$

where x is the noisy speech signal, L is the filter order and T is the analysed pitch period for the speech signal. The b_i represent the filter coefficients updated sequentially according to the LMS algorithm. The filter provides an estimate of the clean input signal $y(n)$. One possibility to extract the necessary reference from the input signal is to estimate the additive noise during the silent speech segments when only the noise occurs. The problem is, that noise is rarely stationary and the detection of silence speech parts is not error free. In addition, this method can not be applied for quantisation noise. The difficulty of forming a reference noise signal is solved by extracting a reference signal from the original speech $x_s(n)$. Due to the quasi periodic nature of speech, a section of speech delayed by its pitch period $x(n-T)$ is highly correlated to the original speech $x_s(n)$ but uncorrelated to the additive noise $x_n(n)$. The derivations of [1] describe that by minimising the energy of the estimation error $e(n)$, the output of the filter, and consequently the system output will be a signal $y(n)$ that is the best least square fit of the input speech signal $x_s(n)$. As can be seen in (2), the error signal is defined as the difference between input signal and estimated filter output.

$$e(n) = x(n) - y(n) \qquad (2)$$

This error signal is used to update the filter coefficients and thus to adjust the filter response.

$$b_{n+1} = b_n + 2\mu \cdot e(n) \cdot X_{n-T} \qquad (3)$$

Each coefficient b_{n+1} is updated using the corresponding present coefficient and a correction term which is formed by the filter tap values shifted by the pitch period (X_{n-T}), the estimation error and the step size μ. This factor controls stability and rate of convergence. The ANC starts with an arbitrary coefficient vector, the algorithm converges in the mean and will remain stable as long as the parameter μ is greater than zero but less than the reciprocal largest eigenvalue λ_{max} of the matrix R [1].

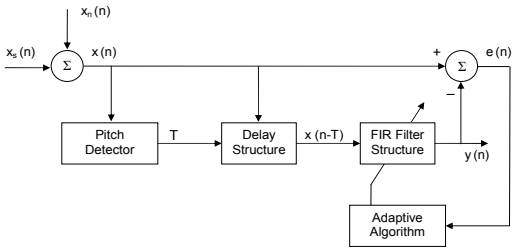


Figure 1: Adaptive Filtering Approach

The correct pitch period of the input speech is extracted using the Average Magnitude Difference Function (AMDF) [6]. This function was chosen because no multiplications are performed and thus, lower area and power consumption can be achieved. In addition, a voiced / unvoiced classification was implemented [2] which is based on a short term energy determination, zero crossings count and a min / max ratio calculation of the values within an AMDF frame. This classification functionality is used to bypass the filter until the first estimated pitch period and to keep the filter coefficients constant during unvoiced sections of the speech.

II. Implementation

For the hardware implementation the ANC was split into different modules with separate functionalities. They were implemented using the ES2 ECPD 0.7µm CMOS technology. All designs were written in abstract VHDL and synthesised using the Synopsys Design Compiler without any design constraints. Figure 2 shows a block diagram which describes the structure. The ANC consists of two main sections, the Pitch Detector [2] and the Adaptive Filter [3]. It uses the incoming speech samples to provide the noise reduced output signal. Two different clock frequencies are used, an 8kHz clock to read in the input samples and a clock frequency of 22MHz to perform all the necessary calculations within one slow clock period. The filter order was chosen to be 10 and its coefficients are represented by a 23 bit vector to ensure sufficient accuracy of the filtering process. A converter was implemented to change the default two's complement number system to a signed magnitude system. This procedure lowers the amount of switching in the Adaptive Filter section and thus it reduces the power consumption.

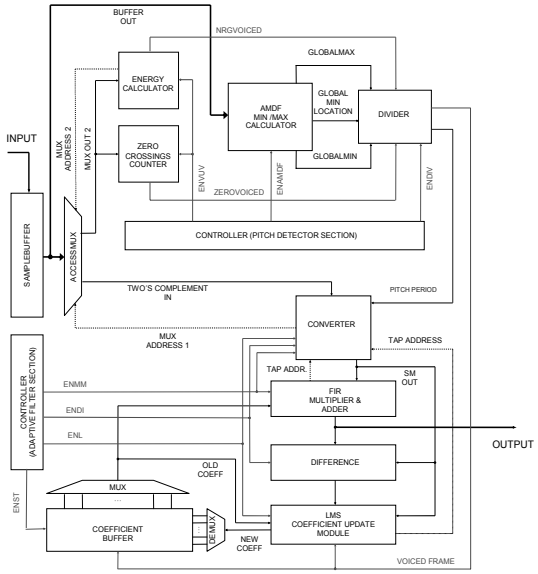


Figure 2: Block Diagram of the ANC

III. System Performance & Results

The performance of the Adaptive Noise Canceller is demonstrated and benchmarked using noisy synthetic speech signals which were filtered by the ANC. The synthetically produced input signals that were used are a vowel 'A' with a fixed pitch frequency of 125Hz and a real speech phrase "Her wardrobe consists of only skirts and blouses" with variable pitch frequencies as well as voiced / unvoiced sections. The formant frequencies which form the vowel 'A' are f_1=730Hz, f_2=1090Hz, and f_3=2440Hz. All signals are distorted by White Gaussian Noise (WGN) and have a signal to noise ratio (SNR) of 5dB to 10dB. The results of those tests are presented in Figures 3 to 12. The magnitude spectra of a noisy and filtered vowel (after detecting the pitch and convergence of the adaptive filter) with a SNR of 10dB are shown in Figures 3 and 4. For reasons of clarity, they have been normalised and only the range from 0 to 0.5 is presented. It can be seen that the filtered version retains the spectral shape of the input signal with the formant frequencies remaining prominent. Hence, the perceptual characteristics of the signal are, in the main, unchanged. Furthermore, the noise component is reduced in the filtered signal, being particularly noticeable in the higher frequencies from 2000Hz to 4000Hz where it is nearly completely reduced. However, in the region 0Hz to 1500Hz, although the adaptive filter manages to reduce the noise, remnants of the noise component and some attenuation of the lower harmonics can be observed.

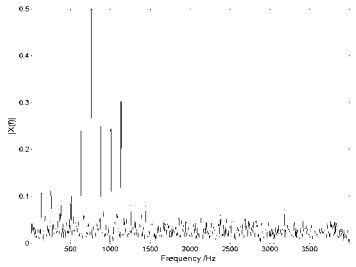


Figure 3: Enlarged Spectrum of a noisy Vowel ‘A’ with Pitch Freq. 125Hz and SNR=5dB

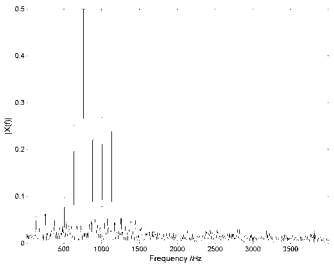


Figure 4: Enlarged Spectrum of a filtered Vowel ‘A’ with Pitch Freq. 125Hz and SNR=5dB

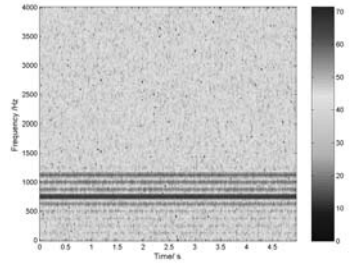


Figure 5: Spectrogram of a noisy Vowel ‘A’ with Pitch Freq. 125Hz and SNR=5dB

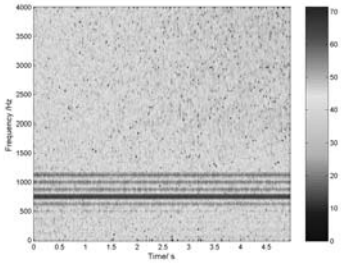


Figure 6: Spectrogram of a filtered Vowel ‘A’ with Pitch Freq. 125Hz and SNR=5dB

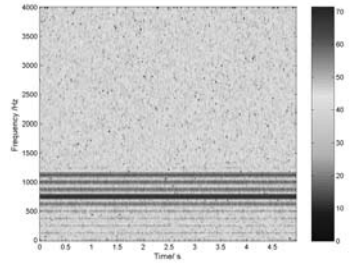


Figure 7: Spectrogram of a noisy Vowel ‘A’ with Pitch Freq. 125Hz and SNR=10dB

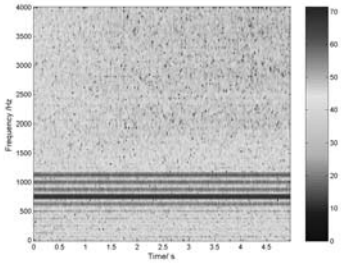


Figure 8: Spectrogram of a filtered Vowel ‘A’ with Pitch Freq. 125Hz and SNR=10dB

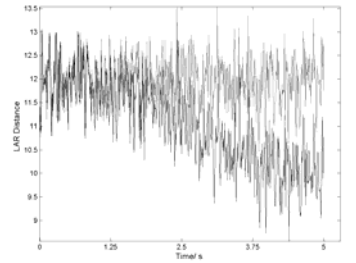


Figure 9: LAR Distance Measure Result, Pitch=125Hz, Input Signal SNR=5dB

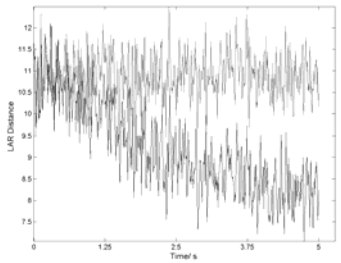


Figure 10: LAR Distance Measure Result, Pitch=125Hz, Input Signal SNR=10dB

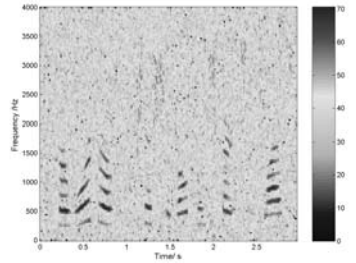


Figure 11: Spectrogram of noisy Synthetic Real Speech, SNR=10dB

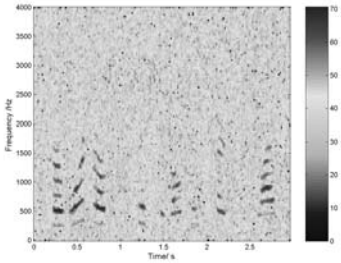


Figure 12: Spectrogram of filtered Synthetic Real Speech, SNR=10dB

Figures 5 to 8 show the performance from the spectrogram perspective of the distorted signal and of the filtered signal respectively. Again, the figures show clearly that the noise component of the higher frequencies 2000Hz to 4000Hz is well reduced but that in the region of 0Hz to 1500Hz some noise remains and the harmonic attenuation persists over time. Thus, it must be concluded that the filter reduces the noise component while not significantly changing the perceptual characteristic of the speech signal. Figure 9 and 10 present the corresponding results of applying the Log Area Ratio (LAR) distance speech quality measure [4] to the distorted and filtered signals. This measure is based on finding a set of Linear Predictive Coefficients (LPC) for each frame of the distorted/filtered speech signals and the original clean speech, transforming them into Log Area Ratio (LAR) coefficients [4] and then calculating the difference between them. This measure was shown to have a correlation coefficient of 0.62 with subjective speech quality assessment data [5]. Figures 9 and 10 demonstrate the speech quality improvement using vowels with SNRs of 5dB and 10dB. It is shown that the LAR distance is shortened by 17% (5dB) and in the case of 10dB SNR the distance is even reduced by 28%. Finally, the performance of the ANC is visualised using a distorted real speech phrase. The spectra of the original and distorted signal are shown in Figures 11 and 12. The spectral shape of the voice information remains after the filtering process and the broad band noise energy is noticeably reduced. Furthermore, it can be seen that the energy of the speech information containing spectral sections of the speech signal are kept after filtering.

IV. Conclusions

The objective of this paper was to describe an Adaptive Noise Canceller which was successfully developed and implemented in hardware using VLSI design techniques in conjunction with a VHDL development environment. Two components, a Pitch Detector and an Adaptive Filter were incorporated into the ANC with additional hardware optimisation of the structure. It has been shown that the developed structure is able to reduce noise in a distorted speech signal using objective speech measures. Furthermore, the frequency components of the signal, which are the bearers of information, are almost unaffected. Additionally, subjective listening tests have shown that the audibility of a noisy speech signal is significantly improved after processing. An insertion into hearing aids, speech recognition systems that aid the handicapped or mobile telephony devices is unproblematic as the silicon area of the whole system is only 14.5mm^2 (based on the 0.7µm library) and is therefore suitable for such purposes.

In summary, this paper has presented that an effective filtering performance under real conditions is given by the ANC device. It is able to adapt its behaviour to suit different input signals and environments without the need to provide an additional reference source.

V. References

[1] Marvin R. Sambur, “Adaptive noise cancelling for speech signals,” IEEE Transactions on acoustics, speech, and signal processing, vol. ASSP-26, No. 5, pp. 419-423 October 1978.

[2] T.D. Smith, F. Gittel, A.Th. Schwarzbacher, E. Hilt and J.T. Timoney, “VLSI implementation of an AMDF pitch detector,” Irish Systems and Signals Conference, Limerick, Ireland, pp 500 – 505, July 2003.

[3] A.Th. Schwarzbacher, M. Herz, F. David and J.T. Timoney, “A hardware optimised CMOS adaptive noise canceller Implementation”, Electronic Devices and Systems Conference, Bruno, Czech Republic, pp. 173-176, September 2002.

[4] John H. L. Hansen and Bryan L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms”, *http://www.ee.duke.edu/Research/Speech*, Dec.1998

[5] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective Measures of Speech Quality*, Prentice Hall , Englewood Cliffs, NJ, 1988.

[6] Myron J. Ross, Harry L. Shaffer, Andrew Cohen, Richard Freudberg, Harold J. Manley, “Average Magnitude Difference Function Pitch Extractor”, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-22, No. 5, October 1974.

3rd International Workshop MAVEBA 2003, 87-90

SPEAKER STRESS DETECTION BY ANALYSIS OF GLOTTAL EXCITATION

M. Bostik, M. Sigmund
Department of Radio Electronics, Brno University of Technology, Brno, Czech Republic

Abstract: **In this contribution the recognition of stress and emotional state is analysed by speech signal analysis using Liljencrant-Fant's model. It is based on the knowledge that some parameters of glottal pulses, obtained by this model, are changed owing to stress, hence they are suitable for the detection of speaker's stressed ("abnormal") state. Two procedures for the analysis of these parameters are described in detail. The first of them is an analysis of parameters of randomly chosen speech parts (of phonetically constant length) that makes fewer demands on segment selection, the second is an analysis of speech parts going one by one in time. The methods were applied to sound recordings made at "stressed" oral examinations at a university. The results obtained show the applicability of these parameters and methods especially for speech analysis when we have at our disposal a signal recorded in the "normal" (steady) state of speaker.**
Keywords: **stress, glottal excitation**

I. INTRODUCTION

The usually used methods for identifying stress and other emotional states [1] usually start from the time distribution of single phonetic parts of words or sentences. Speech influenced by psychical stress can be identified e.g. by different time lengths of the concrete phonemes or by different time lengths of speech pauses between words [2], [3]. Statistical evaluation is also often used to examine e.g. the distribution function of the first two formants or the distribution of time samples. Also used are classifiers based on the pitch period detection and its variation in time. All procedures described above have one common factor, namely that a long time record has to be processed (for statistical methods it is necessary). In the present contribution the method of recognizing stress and some other emotional states, based on the analysis of one or a few period of speech signal is discussed. The low time requirements (from the viewpoint of the length of speech signal not computation) of the method are paid for by the need to own a sound record of the speaker at "normal" state and if it is possible of the some phonetic content. The description of the analysis of the speech signal using Liljencrant-Fant's (LF) model can be found in [4]. This model estimates parameters of glottal pulses (E_e, ω_g, α and ε in Fig. 5) and can also be used for speech signal synthesis and the parameters of this model it is possible to imitate the voice of a specific person. Some parameters of glottal pulses, obtained by the LF model, are especially suitable for "abnormal" speaker state identification [5]. In the following sections two procedures for processing the obtained LF parameters are described and their results are compared in the conclusion. The first procedure is an analysis of the parameters computed from randomly chosen parts (of phonetically constant length) that makes fewer demands on the selection of segments, and the second procedure is an analysis of the parameters obtained from speech parts going one by one in time. The methods were applied to sound recordings made at a diploma work defence, under the influence of speakers' examination stress.

II. SPEECH DATA

It is really difficult to obtain realistic voice samples of speakers in various stressed states, recorded in real situations. There are not many corpora designed to allow the study of speech under stress. A typical corpus of stressed speech from a real case is extracted from the cockpit voice recorder of a crashed aircraft. The only publicly available corpus is the SUSAS database of stressed American English. Two of our own databases [10] were created for use in our experiments; a database of stressed speech and a database of alcoholic speech.

However, for our studies conducted within the research of speech processing in noise and stress we used our own database, namely the SZZ database, consisting of data collected during oral final examinations at our Institute of Radio Electronics. The recorded utterances were manually examined (including both examination of the waveform and parameter contour, and listening) and then endpoints of words were determined. In this way, a number of pauses and irrelevant extraneous voices were eliminated. This material contains stressful phases (improvisations relating to unknown technical problems) and other phases with lower stress (during discussions relating to known technical problems). The hardware and software were hosted by a PC hooked up to the local net for automatic backup of the recorded speech files. The recording platform is set up to store the speech signals „live" in 16-bit coded samples at a sampling rate of 22 kHz. Thus, the acoustic quality of the records is determined by the speaking style of the students and the background noise in the room. For the experiment, only voiced speech segments were used because of our previous experience.

III. METHODS USED

Fig. 1 shows the block diagram of a system for obtaining the LF parameters from continuous speech.

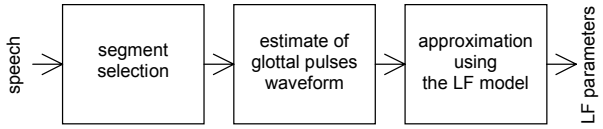


Fig. 1 *Block diagram for the LF parameters estimation.*

The function of the single blocks in Fig. 1 is described in detail in the following.

- *Segment selection*

Before we can start to analyse speech segments it is necessary to choose, by some suitable method, speech signal parts that are suitable for analysis. It is, for example, unsuitable to analyse unvoiced parts if our aim is to obtain and to describe glottal pulses of the vocal apparatus. The next criterion for the selection can be e.g. the difficulty of selection and selection effectivity (effectivity is to be understood in this case as the ratio of the sum of time lengths of the chosen segments from a concrete set of speech data and the whole time length of the set). If we choose a concrete phoneme from the speech data, the selection effectivity is small and the time length of speech data increases (if we want to preserve the level of statistical reliability).

The main aim of this work is to find a suitable procedure for segment selection for "abnormal" speaker state identification. Two methods were used and tested. The first method assumes that the LF parameters of the glottal pulses are rather constant and do not change much during the speech due to coarticulation. Then it is possible to choose voiced segments randomly, independently of the position in the utterance, see Fig. 2.

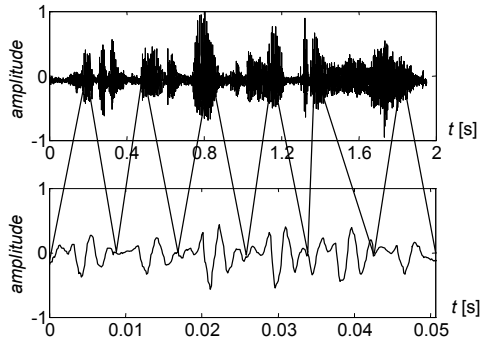


Fig. 2 *Random segment selection, up – continuous speech, down – chosen segments situated one by one in time.*

The second method assumes that the LF parameters of the glottal pulses are changed during the speech. Then it is necessary to choose voiced segments one by one in time, in dependence on the position in the speech, see Fig. 3.

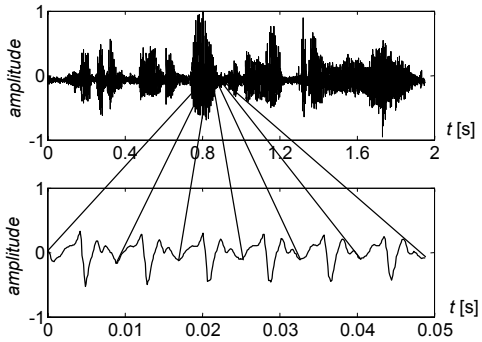


Fig. 3 *Segment selection one by one in time, up – continuous speech, down – chosen segments.*

A further limitation is that we have to have phonetically identical utterances of "normal" and "abnormal" speech.

- *Estimation of glottal pulse waveform*

For the glottal pulse estimation, several methods exist [5]. The well known and effective method is the transfer function estimation of vocal tract with subsequent inverse filtering. This algorithm is one of the basic methods of speech signal processing, further information can be found e.g. in [5], [6], another similar algorithm is presented in [7]. For illustration, Fig. 4 shows a primary speech signal and its excitation signal obtained by inverse filtering.

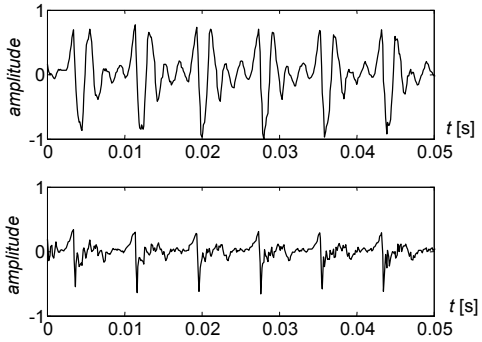


Fig. 4 *Time waveform, up – speech signal (phoneme "a", down – speech signal after inverse filtering.*

- *Approximation using the LF model*

Now, it remains to mention the computation and properties of LF parameters. Glottal pulse approximation using the LF model uses, as the approximation curve, the exponential function combined with harmonic function. That can be seen in Eq. (1) and Eq. (2). Vectors $v_{g1}(n)$ and $v_{g2}(n)$ are two parts of the approximation curve and together they form approximation function v_g, see Fig. 5.

$$v_{g1}(t) = -E_e \frac{\sin\left[\omega_g\left(t - T_{op}\right)\right]}{\sin\left[\omega_g\left(T_e - T_{op}\right)\right]} e^{\alpha(t-T_e)} \quad (1)$$

for $T_{op} \le t \le T_e$

$$v_{g2}(t) = \frac{-E_e}{\varepsilon T_a}\left[e^{\varepsilon(T_e - t)} - e^{\varepsilon(T_e - T_c)}\right] \quad (2)$$

for $T_e < t < T_c$

Variables T_{op}, T_e, T_c and time interval T_a are important parameters and their meaning can be clear from Fig. 5. Approximation is limited to the time interval $T_{op} \le t \le T_c$. The remaining variables E_e, ω_g, α and ε are the LF parameters sought. It is possible to obtain them by some of the iterative methods. The parameters are determined by criteria of the minimal average quadratic deviation of the approximating and the approximated function. All procedures described here were implemented using mathematical software Matlab on the modified PC with a professional sound card.

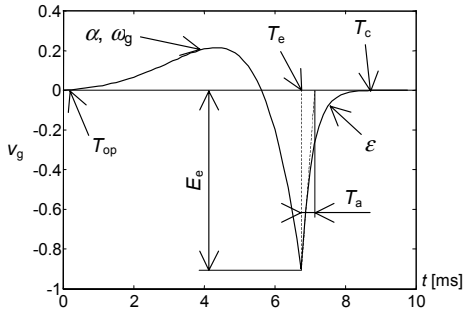


Fig. 5 *Time waveform of the approximation function and the meaning of individual parameters.*

IV. Data Evaluation

The methods described above were applied to speech data recorded at "normal" and "abnormal" state of the speaker. Records of both states were phonetically identical. The results presented in [5] show that only some of the LF parameters E_e, ω_g, α and ε are suitable for speaker state recognition. As mentioned above the main aim was to show the dependence of analysis results on the methods for segment selection. The procedures described in the previous section (see Fig. 2 and Fig. 3) were used for the selection of segments and the results were evaluated by the following procedure:

- for both methods of segment selection ten sets were created, each set contains six segments, see Fig. 2 and Fig. 3. Recordings of one male speaker were used. Six parameters were deduced from the fact that a phoneme 40 ms long contains just six fundamental periods (thus in this case segments too) with frequency 150 Hz. In the case of more segments in the set, the selection effectiveness will decrease below admissible limits, because longer-time phonemes occur in speech less frequently.
- for each segment of the speech the glottal pulses were estimated by using an estimation of linear prediction error, by cepstral coefficients [8] or by ARMA modelling [9].
- for each set of segments the LF parameters were computed. In Fig. 6 the parameter α is shown in dependence on the segment from which it was computed.
- for each set of segments the average value of the corresponding parameter (μ_{Normal}, μ_{Stress}) and dispersion (R_{Normal}, R_{Stress}) were computed. So, we obtained ten average values and ten dispersion values for ten sets of segments.

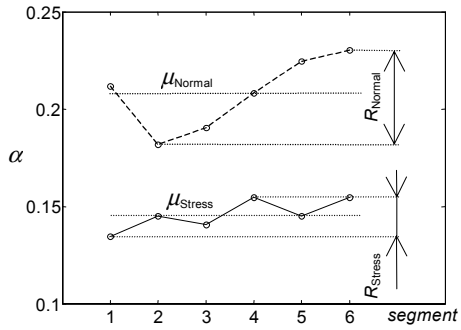


Fig. 6 *Values of the parameter α for one segment – one set is shown, phoneme "a", segment selection one by one in time.*

In Fig. 7, the values μ and R are shown in dependence on the set from which they were computed.

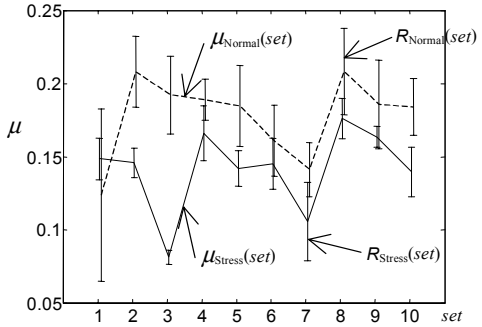


Fig. 7 *Average values of the parameter α and its scatter for individual sets - ten sets, phoneme "a", segment selection one by one in time.*

V. Results

The results of the described algorithms with the final output shown in Fig. 7 are plotted in the diagrams in Fig. 8 and Fig. 9.

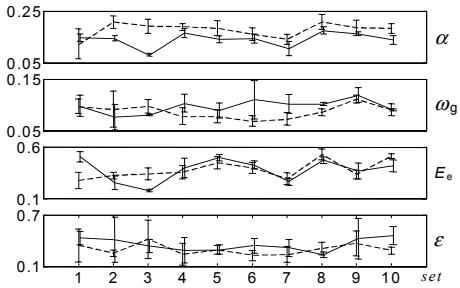


Fig. 8a *Average values of the LF parameters and their scatter for individual sets, phoneme "a", selection one by one in time (upper diagram is identical with Fig. 7). Dashed line is "normal" state.*

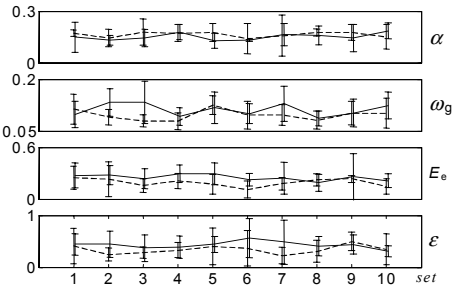


Fig. 8b *Average values of the LF parameters and their scatter for individual sets, phoneme "a", randomly chosen segments. Dashed line is "normal" state.*

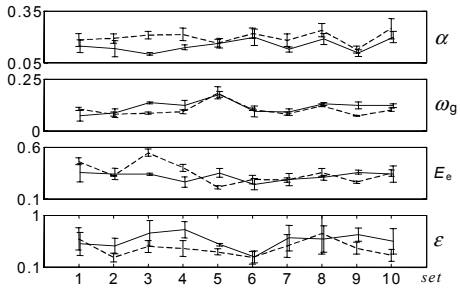


Fig. 9a *Average values of the LF parameters and their scatter for individual sets, phoneme "e", selection one by one in time. Dashed line is "normal" state.*

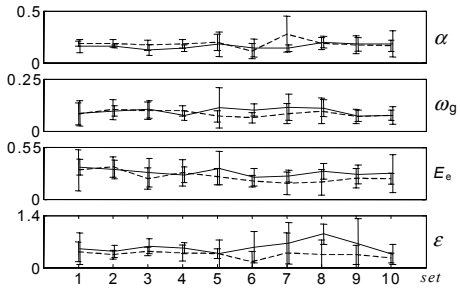


Fig. 9b *Average values of the LF parameters and their scatter for individual sets, phoneme "e", randomly chosen segments. Dashed line is "normal" state.*

VI. Conclusion

By comparing of the diagrams in Fig. 8a and Fig. 8b, upper diagram for parameter α, it was found that the results are better if the segments were chosen one by one in time (greater differences between the parameters for "normal" and "abnormal" speaker state). Similar conclusion can be drawn for phoneme "e" too, parameter α. On the other hand, the parameter ε is almost independent of the method of segment selection, for both analysed phonemes. Generally, it can be said that the analysis of the segments going one by one in time provides better results. The parameters are not only more different for single states than in the case of the random segment selection, but the results are also less scattered.

References

[1] C. Vloeberghs at al.: *The Impact of Speech Under „Stress" on Military Speech Technology.* Technical Report 10, RTO, Canada Communication Group Inc. 2000.

[2] R. J. Baken: *Clinical Measurement of Speech and Voice.* Singular Publishing Group, Inc., New York 1996.

[3] B. Boyanov and G. Baudoin: Acoustical Analysis of Pathological Voice. *Proc. of the 3rd Slovenian-German Workshop Speech and Image Understanding.* Ljubljana 1996, pp. 157-166.

[4] M. R. Iseli and A. Alwan: Inter- and Intra-speaker Variability of Glottal Flow Derivative Using the LF Model.
svr-www.eng.cam.ac.uk/~glm20/ICSLP/pdf/01551.pdf

[5] M. Boštík: *Analýza hlasu pro diagnostické účely.* Diploma thesis, BUT, Brno 2002.

[6] J. Psutka: *Komunikace s počítačem mluvenou řečí.* Academia, Praha 1995.

[7] A. El-Jardouli and J. Makhoul: Discrete All-Pole Modelling. *IEEE Transaction on Signal Processing*, vol. 39, No. 2, February 1991, pp. 411-418.

[8] M. Boštík: Influence of Speaker Stress on the Glottal Pulses Form. *Proc. of the 13th International Czech – Slovak Conference RADIOELEKTRONIKA 2003*, Brno, pp. 451-455.

[9] M. Boštík and M. Sigmund: Methods for Estimation of Glottal Pulses Waveforms Exciting Voiced Speech. *Proc. of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003*, Geneva, pp. 2389-2392.

[10] M. Sigmund: *Voice Recognition by Computer.* Tectum Verlag, Marburg 2003.

3rd International Workshop MAVEBA 2003, 91-93

AN INVESTIGATION OF THE SPEECH PRODUCTION

Jana Kleckova
Department of Computer Science
Faculty of Applied Sciences
University of West Bohemia
Univerzitni 22, CZ – 306 14 PLZEN, Czech Republic
Email: kleckova@kiv.zcu.cz

Abstract— **Processing spontaneous speech deals with problems that are influenced by the several facts. This paper reports an investigation of the production of real and non-words in two normal speaker groups. Group 1 consists of 10 young people - 5 women and 5 men (mean age 23 years) and group 2 consists of 5 older women and 5 older men(mean age 52 years). The speech material used in study consisted of two repetitions of 10 real, 10 pseudo-real and 10 non-words. The speech data were subsequently digitized (16 KHz) and the following were measured: response latency, utterance duration and duration. The results are presented and discussed within a dual-route model of speech production.** [1]

Keywords: **Spontaneous speech, phonetic and phonological representation, direct and indirect route.**

I. Introduction

Processing spontaneous speech deals with problems that are influenced by the following facts: (i) speakers make mistakes and correct themselves, produces false starts and use ungrammatical constructions; (ii) the acoustic signal produced by a human speaker is mapped onto a written form by a speech recognizer - this mapping is rarely completely correct. This introduces two levels of uncertainty into the processing of speech, which make the task of linguistically analyzing a spoken utterance in a speech processing system doubly hard. In addition, the dialog context imposes strict time constraints. Some psycholinguistic research suggests that may be two routes which employed in phonetic encoding. One route involves storage of frequently used syllables in a mental syllabary (”direct” route) and second is used for novel or low frequency syllables (”indirect” route). The former encoding route is more dependent on on-line computational resources. Dual route models have been proposed for other cognitive functions such as reading aloud [2]. Some of measures that have been used to gauge the employment of direct and indirect routes have included response latencies and the duration of utterances, where greater values for both measures would be interpreted as a sign of the greater planning and encoding demanded by the ”indirect” route. The current study aims to investigate whether dual routes may be encoding of real and non-monosyllabic words elicited via a repetition task, in two groups of speakers. This is done by investigating the response latencies, utterance and word durations of monosyllabic real words, pseudo- words and non-words, elicited via a repetition task. Experiments and first results are given.

[1]The work presented in this paper was partly supported by the Grant Agency of Czech Republic under contract number 201/02/1553.

II. Methods

Two groups of subjects participated in the study. Group 1 were all students in tertiary education. Group 2 consisted of 10 adult women and men speakers ranging in the age from 45 to 61 years, who worked in tertiary education. All speakers had no speech, language or hearing difficulties.

A. Speech Material

The speech material used in the experiments consisted of two repetitions of 10 monosyllabic real, 10 monosyllabic pseudo-real (containing articulatory sequences that are likely to have been encountered before in real Czech words and conforming to Czech phonotactic constraints) and 10 monosyllabic non-words (containing articulatory sequences that are unlikely to have been encountered before in real Czech words. For example (English) monosyllabic real word ”soap” ['seup], pseudoreal word ”sote” ['seut] and monosyllabic non/word ”soekf” ['seukf]. This gave a total of twenty tokens for each of the word groups, which were randomized into a single list. Subjects were instructed to repeat each word on the list after the experimenter.

B. Durational measures

Speech pressure waveforms, wideband FFT spectrograms and LPC analyses were used to obtain the durational acoustic measures. The measures that were taken were:

- response (or repetition) latencies - these were measured from the end of the experimenter's prompting utterance to the utterance start of the participant's utterance,
- utterance durations - these were measured form start to the end of the entire utterance,
- word durations - these were measured from the start to the of the stimulus word.

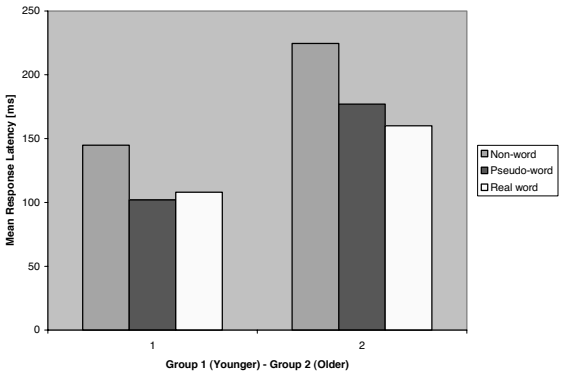


Fig. 1. Mean response latency [ms].

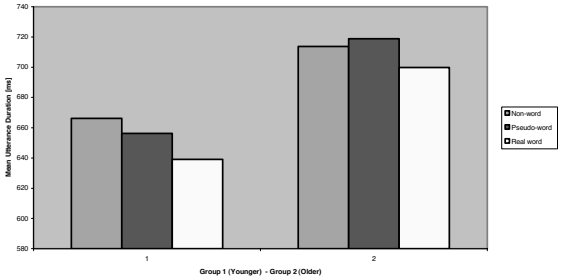


Fig. 2. Mean utterance duration [ms].

III. Results

The response latency, utterance duration and word duration values for the monosyllabic real words, pseudo- words and non-words are given in Table 1 for the both Group1 and Group 2.

TABLE I
MEAN AND STANDARD DEVIATION VALUES FOR RESPONSE LATENCY, UTTERANCE DURATION, WORD DURATION (ALL GIVEN IN MILLISECONDS) BY GROUP AND WORD FREQUENCY FOR REAL WORDS, PSEUDO-WORDS AND NON-WORDS.

Group 1 (number 10)			
Measure	**Non**	**Pseudo**	**Real**
Response latency	144,8 (90,7)	102,0 (80,6)	108,1 (75,9)
Utterance duration	666,2 (89,9)	656,2 (63,0)	639,1 (71,4)
Word duration	519,0 (95,0)	513,1 (58,1)	500,0 (71,1)

TABLE II
MEAN AND STANDARD DEVIATION VALUES FOR RESPONSE LATENCY, UTTERANCE DURATION, WORD DURATION (ALL GIVEN IN MILLISECONDS) BY GROUP AND WORD FREQUENCY FOR REAL WORDS, PSEUDO-WORDS AND NON-WORDS.

Group 2 (number 10)			
Measure	**Non**	**Pseudo**	**Real**
Response latency	224,5 (116,2)	177,1 (93,9)	160,1 (86,9)
Utterance duration	713,8 (147,6)	718,9 (97,9)	699,9 (95,0)
Word duration	572,0 (148,0)	578,2 (91,0)	562,9 (85,9)

A series of repeated measures was carried out on the data for combined data of Group 1 and Group 2, for measures: response latency, utterance duration and word duration a repeated measures indicated that there were significant differences. A series of post-hoc parried T-tests indicated significant differences in the response latencies of non-words and pseudo-words. Both significant comparisons showed longer response latencies for the non-words. No significant differences were found between the response latencies of the pseudo-words and real words. In addition, significant group differences were found between the response latencies of Group 1 and Group2. In the case of utterance duration, there were also significant differences. A series of post-hoc parried T-tests indicated significant differences between the utterance durations of pseudo-words and real words,with the pseudo-words being longer than real words. No significant differences were found between the utterance durations of the non-words and pseudo-words, or the non-words and real words. Again significant group differences were found between the utterance durations of Group 1 and Group 2.

IV. Discussion

There is some evidence in the data reported here to suggest-that there may be differences in the phonetic encoding of the real words, pseudo-words and non-words. These differences are illustrated by the significantly slower response latencies of the non-words when compared to those of the pseudo-words and real words. Response latency, is a difficult parameter to interpret. It is difficult to ascertain to what degree responce latency is determined by either auditory recognition or motor encoding, or endeed both. Therefore it not clear wheter the results reported here are evidence for differences in motor encoding and/or auditory recognition. Although not significant, hovewer, the word and utterance duration for the non-words and pseudowords for both groups displayed trends og being longer than those of the words. these findings could be interpreted as some evidence for a greater areliane on ”indirect” route mechanism in the motor encoding of the non-words and pseudo-words that were elictid in this study, with greater time required for their production. It has been suggested that dual-routes may be operating in speech encoding [3], [5], with novel and low frequency word/syllables being largely reliant in ”indirect” mechanisms. There were some differences between the data of non-words versus pseudo-words and real words. Although there were some trends in the data, there was a general lack of significant differences between the utterance and word durations of real words and pseudo-words, and those of real words and non-words. This finding could be

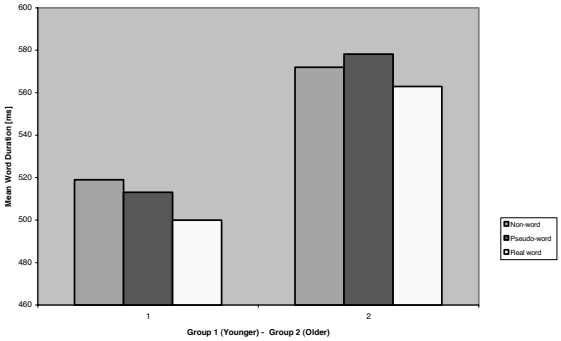


Fig. 3. Mean word duration [ms].

explained by the fact that the real words used in the study were relatively low frequency words, and were therefore, probably phonetically encoded using some degree of ”indirect” route resources. The non-words consisted of articulatory sequences that were less likely to have been encourtered in a word context, and were therefore also more likely to have been reliant on ”indirect” route mechanisms than either the real or pseudo-words. The word stimuli used in this study were all, therefore, more likely to have been encodedusing some level of ”indirect” route mechanisms. However, the issue of how much auditory recognition affected the pattern of results should not be overlooked in this study. Some of patterns in the response latency data may reflect fast lexical access in the case of real words, but failed lexical access in the case of the non-words. This issue merits some futher investigation within the model of dual/route phonetic encoding.

V. Conclusion

The results of repeated measures indicated significant age-effects in the response latencies, utterance durations and word durations of Group 1 and Group 2. The older subjects displayd longer response latencies compared to the younger subjects. This could be interpreted as evidence either a greater level of planning time, or less auditory recognition to motor encoding processing, or indeed some degree of both, was required by the older subjects in the production of the non-words, pseudo-words and real words. The utterance and word duration were also significantly longer for the older subjects compared to the younger subjects. This suggests a slower articulation rate in the older subjects and could be interpreted either as evidence for some degree of atrophy in the efficiency of motor speech production, with increasing age. The real and non-words phonetic encoding is one of importance to a number fields, including psycholinguistics, linguistics, and artificial speech recognition. The expiriences of this study will be used in the project of the develpoment and design of an user-friendly communication interface enabling an easy interaction of handicapped persons with information systems.

References

[1] Kleckova J., Matousek V.: ”Developing the Database of the Spontaneous Speech Prosody Characteristics”. In: Proceedings of the Int. Conference EUROSPEECH ’99, Volume 2, pp. 731 – 734,Budapest, Hungary, September 1999.

[2] Kolinsky,R.: ”Spoken Word recognition: A Stage -processing Aproach to Language Differences.” In: European Journal of Cognitiove Psychology, vol.10, pp.1-40.

[3] Levelt, W.J.M., Wheldon, L.: ”Do speakers have to access to a mental syllabary?” In: Cognition, 50, pp. 239 - 269, 1994.

[4] Selkirk E.: ”Sentence Prosody: Intonation, Stress, and Phrasing”. In: Handbook of Phonological Theory. Ed. by J.A. Goldsmith, Oxford: Basil Blackwell, 1995, pp.550-569.

[5] Whiteside, S.P., Varley, R.A.: ”A new conceptualisation of apraxia of speech: a synthesis of evidence.” In: Cortex, 34, pp. 221 - 231, 1998.

3rd International Workshop MAVEBA 2003, 95-98

TOWARDS CHAOTIC MODELING OF SPEECH SIGNALS

A. Petry[1], D. A. C. Barone[2]
[1]Unidade de Gestão do Conhecimento Computação, Universidade Luterana do Brasil, Canoas, Brazil
[2]Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Abstract: **This paper shows how the chaotic systems theory can be applied to the modeling of speech signals, whose dynamics is highly complex. We verify that, when using a theory that is able to model nonlinear features, speech signals present a highly nonlinear behavior, which could not be inferred from a linear theory.**
Keywords: **Chaos theory, speech modeling, nonlinear dynamical systems, Lyapunov exponents, time series**

I. INTRODUCTION

Many physical phenomena present a complex behavior with fluctuations over time. Biological signals, such as electroencephalograms (EEGs), electrocardiograms (ECGs), vocal sounds, and measures of arterial blood pressure, represent a great challenge for analysis and modeling.

A detailed model of the vocal tract should consider the time variation of vocal tract shape, the vocal tract resonances, losses due to heat conduction and viscous friction at the vocal tract walls, nasal cavity coupling, softness of the vocal tract walls, the effect of subglottal (lungs and trachea) coupling with vocal tract resonant structure and radiation of sound at the lips [1]. A time-varying linear filter can model the effects of some of these factors, but the remaining ones are very difficult to model. Some techniques have been proposed in the literature to analyze the non-linearities of dynamical systems. The Chaos theory offers a set of techniques that can perform this analysis in complex signals that present deterministic chaos. Many nonlinear dynamical features can be extracted from chaotic signals, such as fractal dimension, entropy and Lyapunov exponents. These features may be used with speech processing systems and potentially improve their accuracy [2-13].

However, the application of Chaos theory techniques assumes that the signal under analysis is stationary and comes from a system with chaotic components. This assumption must be criteriously verified. In this paper, we will explore the possibility of using Chaos theory techniques in speech signals. Furthermore, we will verify when and under which conditions they can be applied. Our initial goal is to use the techniques shown in this paper to analyze non-pathologic speech signals. Once we have seized the effectiveness of the techniques application, we want to toward our focus to analyze problematic speech production.

II. CHAOS IN TIME SERIES

Any time series is considered chaotic when it is obtained from a stationary state of a dynamic system that presents nonlinearities and sensitivity to initial conditions. Sensitivity to initial conditions means that a small variation in the conditions that the system is embedded will produce a significant modification in the system behavior. The sensitivity to initial conditions is directly related to chaotic systems.

There are many ways of verifying the existence of chaos associated to a time series. Initially, the trajectory of the possible attractor associated to the time series must be reconstructed in a proper state space. Attractor is a contraction in certain areas in state space, such that all trajectories nearby converge to it. Chaotic time series have chaotic attractors. It is possible to know if an attractor is chaotic or not by evaluating pairs of trajectories whose initial conditions are very close. If they diverge, on average, at a positive exponential rate given by the largest Lyapunov exponent, the attractor is chaotic. Thus, the existence of a positive Lyapunov exponent is a certain evidence of existence of chaos in the time series analyzed [8] [2] [14-15].

The analysis that must be used to evaluate speech signals, in a search for chaotic characteristics, assumes that data were obtained from a system's stationary state. It is known that speech is not stationary during a time window of seconds, since the vocal articulatory apparatus is continuously changing its configuration to produce different sounds that compose the speaker utterances and sentences. On the other hand, small time windows of speech (few dozens of milliseconds) can be considered as stationary, because the variation of vocal tract configuration is slow [1][16-17]. Thus, the search for chaotic components must be accomplished using successive small windows of speech.

III. ANALYSIS OF CHAOS IN SPEECH

In this paper, we explore the possibility of using an important nonlinear dynamic feature in order to improve traditional modeling of phonological speech production. This characteristic, known as Lyapunov exponents, quantifies the sensitivity of a dynamical system to initial conditions. When an attractor associated to a time series is chaotic, the average exponential divergence of nearby

trajectories is quantified by estimating the largest Lyapunov exponent. For time series produced by a dynamical system, the presence of a positive value for the Lyapunov exponents indicates the presence of chaos. Furthermore, in many applications it is sufficient to estimate only the largest value of the Lyapunov spectrum.

Rosenstein *et. al.* [15] proposed a method to estimate the largest Lyapunov exponent (λ_1) from time series composed by a very limited number of available samples. Good results were obtained for estimating the largest Lyapunov exponent of known systems using just 100 to 1000 samples. This characteristic is quite important when dealing with speech, once a speech signal can be considered stationary only during a small window of approximate 30ms.

The first step is the reconstruction of the attractor's trajectory in an appropriate state space. After, the nearest neighbor of every vector of the reconstructed trajectory is found. A constraint that two nearest neighbors have a temporal separation greater than the mean period of the time series must be satisfied. Doing this, it is possible to consider the pair of neighbors as belonging to different trajectories. When considering two trajectories whose initial conditions are very similar, the trajectories diverge, on average, at an exponential rate characterized by the largest Lyapunov exponent (λ_1), as follows

$$d_j(i) = C_j e^{\lambda 1 (i\Delta t)} \quad (1)$$

where $d_j(i)$ is the distance between the *j*th pair of nearest neighbors after *i* steps (equals to $i\Delta t$ seconds where Δt is the time series sampling period) and C_j is the initial separation between the neighbors.

Applying the natural logarithm to both sides, the previous equation becomes

$$\ln d_j(i) = \ln C_j + \lambda_1 (i\Delta t) \quad (2)$$

If the logarithm of the distance evolution between every pair of neighbors is monitored, they will appear as a set of approximately parallel lines, each with a slope proportional to λ_1. The largest Lyapunov exponent is then estimated by applying least-squares method to best model the mean line. Fig. 1 shows the logarithm of the mean distance evolution between every pair of neighbors from the reconstructed state space vectors of a 30ms window of speech. It is easy to verify its positive slope, which indicates a positive value for the correspondent largest Lyapunov exponent.

The process of estimating largest Lyapunov exponent from an approximate stationary speech signal, with duration of tens of milliseconds, can be repeated to every window of a long term speech signal, no matter its length. Thus, a complex, long term and not stationary speech signal can still be analyzed by Chaos theory, and its Lyapunov exponents (one for every window) can be estimated. These time-dependent largest Lyapunov exponents may show regions where the speech signal can or can not be considered chaotic.

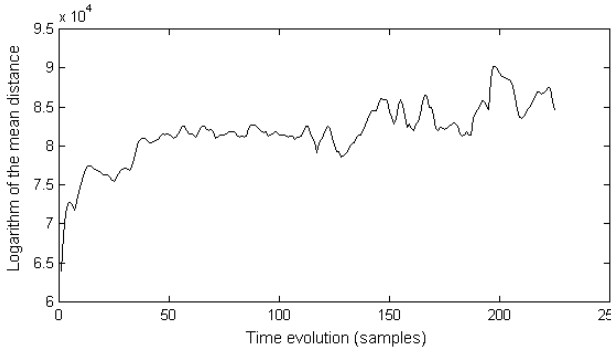


Fig. 1: Logarithm of the mean distance evolution between every pair of neighbors from the reconstructed state space vectors of a 30ms window of speech

IV. Experimental Results

In order to verify the existence of chaotic components in speech signals, it was accomplished an experiment that uses speech samples from different speakers. The experiment used 30 ms windows and from every window the trajectory of a possible associated attractor was reconstructed. After that, the largest Lyapunov exponent was estimated from every reconstructed trajectory, using the Rosenstein method [15]. Fig. 2 illustrates the process of largest Lyapunov estimation in a speech window. The repetition of this process in every window of speech signal provides the time variation of the largest Lyapunov exponent.

The estimation of time variation of the largest Lyapunov exponent values from speech signal in fig. 2 is shown in fig. 3. Fig. 3 shows the largest Lyapunov values as black dots in the figure and, in order to maintain the relation with time, the waveform of the speech is also plotted using gray color in background. It is possible to note that not all windows of speech have positive largest Lyapunov exponents. This occurs mainly in the transition of words where coarticulation or silence between words can produce negative Largest Lyapunov exponents. We have also noted that, if we don't dispose of an adequate number of samples to estimate the Lyapunov exponent, this can increase negative exponents estimation, due to the implementations of computational algorithms. This occurs mainly with signals sampled at reduced rates.

In order to deal with the analysis of chaotic nature of speech signals, we have to process more than just one speech data file. More truthful results can be obtained when the Largest Lyapunov estimation is applied to a large set of different speakers. So, this process was repeated, producing the estimation of 1000 Largest Lyapunov exponent values, using data from 50 different

speakers of varied ages. In this case, these speakers have produced the utterance composed by a random sequence of different numbers. The aim of using a random sequence of numbers is to avoid a "mechanical" repetition of speech, since the next number in the sequence is not memorized. Furthermore, the numbers in an unknown sequence are usually spoken slowly and correctly, providing a better combination of phonemes.

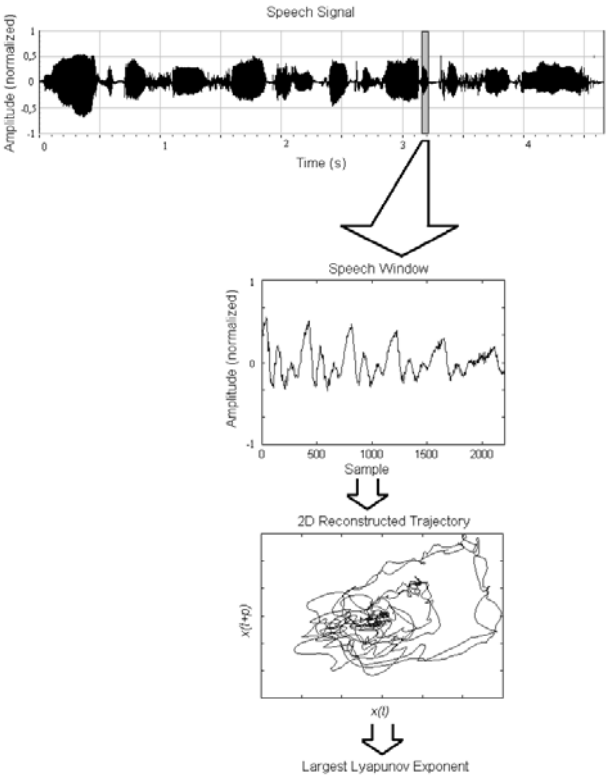


Fig. 2: Largest Lyapunov estimation process from a window of speech

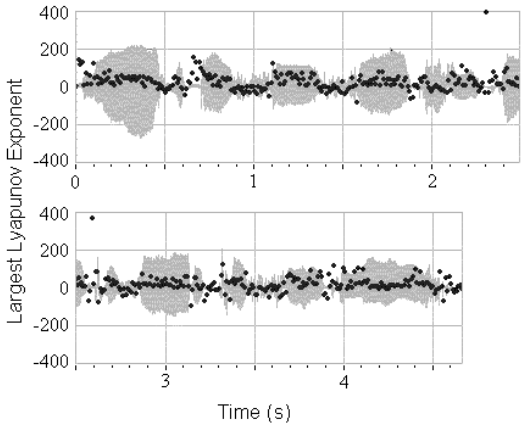


Fig. 3: Largest Lyapunov exponent estimation from 30 ms windows (applied every 10 ms) of speech signal in fig. 2

The generated Largest Lyapunov exponents' histogram is shown in fig. 4. The speech data were recorded at 44100 Hz sampling rate, and we have used 30 ms windows, extracted at every 10 ms. Analyzing fig. 4, we can notice the existence of chaos in the majority of reconstructed attractors. The amount of time required to estimate the largest Lyapunov values was expressive. Fig. 5 shows the variation of CPU processing time for speech windows of varied lengths. The processor used in the measures was an Athlon processor, running at 1.1GHz.

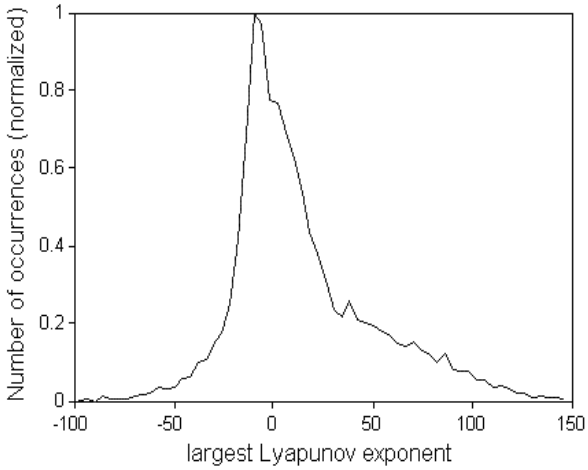


Fig. 4: Largest Lyapunov exponents histogram, using 50 different speakers

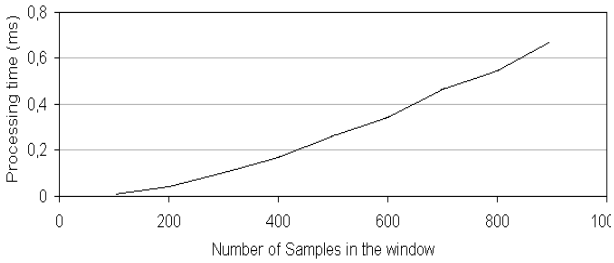


Fig. 5: CPU Processing time required to extract largest Lyapunov exponents from speech windows of varied lengths

V. CONCLUSION

We have presented in this paper an approach to characterize speech signals by introducing the Largest Lyapunov exponents estimation in the analysis of data from paired people. Chaotic components were detected in speech signals, which validate the use of many nonlinear dynamical features, such as fractal and multifractal dimension, entropy, etc, in several speech classification systems. This method also presents high potential to be used in the characterization of unpaired people speech production, which will be done in the sequence of the presented work.

Acknowledgments - The authors would like to acknowledge Tibério S. Caetano for important comments and suggestions, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support

REFERENCES

[1] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Prentice Hall, 1987.

[2] A. Kumar, *Nonlinear Dynamical Analysis and Predictive Coding of Speech,* Ph.D. Thesis in Electrical Engineering, Indian Institute of Technology, Kanpur, 1994.

[3] A. Kumar and S. K. Mullick "Nonlinear Dynamical Analysis of Speech," *J. Acoust. Soc. Am.*, [S.l.], v. 100, n. 1, July 1996.

[4] R. Port, F. Cummins and M. Gasser "A dynamic approach to rhythm in language: toward a temporal phonology," *Technical Report* n. 150, Bloomington, Indiana: University Cognitive Science Program, 1995.

[5] S. Sabanal and M. Nakagawa "The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model," *Chaos, Solitons & Fractals*, v.7, n.11, p. 1825-1843, 1996.

[6] L. P. L. Oliveira, W. L. Roque and R. F. Custódio "Lung Sound Analysis with Time-Dependent Fractal Dimensions," *Chaos, Solitons & Fractals*, v. 10, n. 2, p.1419-1423, 1999.

[7] D. Sciamarella and G. B. Mindlin "Topological Structure of Chaotic Flows from Human Speech Data," *Phys. Rev. Letters*, v. 82, n. 7, Feb. 1999.

[8] M. Banbrook, S. Mclaughlin and I. Mann "Speech Characterization and Synthesis by Nonlinear Methods," *IEEE Transactions on Speech and Audio Processing*, New York, v. 7, n. 1, Jan. 1999.

[9] A. M. Chan and H. Leung "Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach," *IEEE Transactions on Speech and Audio Processing*, v. 7, n. 3, May 1999.

[10] E. L. J. Bohez and T. R. Senevirathne "Speech recognition using fractals," *Pattern Recognition*, v. 34, p. 2227-2243, 2001.

[11] C. Guo et al. "A study on fractal properties of Mandarin speech," *International Journal of Non-linear Mechanics*, New York, v. 37, p. 409-417, 2002.

[12] A. Petry and D. A. C. Barone "Speaker Identification Using Nonlinear Dynamical Features," *Chaos, Solitons & Fractals*, v. 13, n. 2, p. 221-231, Feb. 2002.

[13] A. Petry and D. A. C. Barone "Preliminary Experiments in Speaker Verification Using Time-dependent Largest Lyapunov Exponents," *Computer Speech and Language*, v. 17, p. 403-413, 2003.

[14] H. D. I. Abarbanel et al. "The Analysis of Observed Chaotic Data in Physical Systems," *Reviews of Modern Physics*, v. 65, n. 4, p. 1331-1392, Oct. 1993.

[15] M. T. Rosenstein, J. J. Collins and C. J. De Luca "A practical method for calculating largest Lyapunov exponents from small data sets," *Physica D*, v. 65, p. 117-134, 1993.

[16] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Prentice Hall, 1978.

[17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

3rd International Workshop MAVEBA 2003, 99-102

BLIND SOURCE SEPARATION BY INDEPENDENT COMPONENT ANALYSIS APPLIED TO ELECTROENCEPHALOGRAPHIC SIGNALS

C. S. Lima[1], C. A. Silva[1], A. C. Tavares[1] and J. F. Oliveira[2]
[1]Department of Industrial Electronics, Universidade do Minho, Guimarães, Portugal
[2]Department of Electrical Engineering, Instituto Politécnico de Leiria, Leiria, Portugal

Abstract: **Independent Component Analysis (ICA) is a statistical based method, which goal is to find a linear transformation to apply to an observed multidimensional random vector such that its components become as statistically independent from each other as possible.**
Usually the Electroencephalographic (EEG) signal is hard to interpret and analyse since it is corrupted by some artifacts which originates the rejection of contaminated segments and perhaps in an unacceptable loss of data. The ICA filters trained on data collected during EEG sessions can identify statistically independent source channels which could then be further processed by using event-related potential (ERP), event-related spectral perturbation (ERSP) or other signal processing techniques. This paper describes, as a preliminary work, the application of ICA to EEG recordings of the human brain activity, showing its applicability.

I. INTRODUCTION

An important application of multichannel EEG is to try to find the location of a epileptic focus (a small spot in the brain where the abnormal activity originates and then spreads to other parts of the brain) or of a tumor, even when they are not visible in a x-ray or CT scan of the head.

Blind Source Separation (BSS) concerned to signal processing applications is an application area which main goal is the recovering of independent source signals, after they are linearly mixed by an unknown medium. This source separation is achieved by using recordings of several sensors. A classical example of blind source separation is the cocktail party problem, where several people are speaking simultaneously in the same room. The problem is to separate the voices of the different speakers, by using recordings of several microphones in the room.

Some acceptable solutions for the blind source separation problem have been found in the neural network and statistical signal processing fields. The classical application of the ICA model is blind source separation. In contrast with decorrelation techniques such as Principal Component Analysis (PCA), which ensures that output pairs are uncorrelated, the ICA maximizes the degree of statistical independence among outputs using contrast functions approximated by the Edgeworth expansion of the Kullback-Leibler divergence [1]. Therefore when compared with the PCA, ICA imposes the much stronger criterion that the multivariate probability density function of output variables factorizes. Finding such a factorization requires that the mutual information between all variable pairs go to zero. While decorrelation only takes account of second-order statistics, the mutual information depends on all higher-order statistics of the output variables. Although ICA can be seen as an extension of the PCA and factor analysis it is really a more powerful technique, capable of finding the underlying sources when these classical methods fail completely.

As the problem of determining brain electrical source from patterns recorded on the scalp surface is mathematically undetermined the joint problem of EEG source identification, segregation, localization and removing artifacts becomes very difficult. Recent efforts to identify EEG sources have focused mostly on performing spatial segregation and localization of source activity. The problem of both source localization and source identification have been investigated by using the ICA algorithm. Independent sources can be derived from highly correlated EEG signals and without regarding to the physical location or configuration of the source generators, by using the ICA algorithm, however, canceling these noise sources is a central, and as yet unsolved problem in EEG signal processing.

One of the most successful method is mainly based on ICA of an artificial neural network by using an adaptive algorithm. In the adaptive case, the algorithms are obtained by stochastic gradient methods. When all the independent components are estimated simultaneously, the most popular algorithm in this category is natural gradient ascent of likelihood, or related contrast functions like "Infomax". The experiments described in this paper were obtained by using a kind of extended "Infomax" algorithm for the EEG analysis.

II. RELEVANT ICA THEORY

The ICA algorithm allows to separate N independent sources from N sensors under the constraints that the propagation delays of the unknown "mixing medium" are negligible, and the sources are non-log and have probability density functions (pdf's) not too unlike the gradient of a logistic sigmoid. Therefore the EEG signal must be recorded by N scalp electrodes and the correlated signals are used to separate N unknown "independent brain sources" that generated these mixtures.

Before proceeding we have to make a clear distinction between ICA, which is a theoretical method with different applications, and blind source separation, which is an application that can be solved using various theoretical approaches, including but not limited to ICA. One of these approaches is the PCA, which is a decorrelation technique, so ensuring that output pairs are uncorrelated <y_i, y_j=0>, for all *i* and *j*. Decorrelation only takes account of second-order statistics. In contrast the ICA is based on the much stronger criterion of statistical independence which requires all higher-order correlations of y_i to be zero. The relation between Principal Component Analysis and ICA is evident. Both methods formulate a general objective function that define the 'interestingness' of a linear representation, and then maximize that function. A second relation between PCA and ICA is that both are related to factor analysis, though under the contradictory assumptions of Gaussianity and non-Gaussianity, respectively. The affinity between PCA and ICA may be, however, less important than the affinity between ICA and other methods. This is because PCA and ICA define their objective functions in quite different ways. PCA uses only second-order statistics, while ICA is impossible using only second-order statistics. PCA emphasizes dimension reduction, while ICA may reduce the dimension, increase it or leave it unchanged. However, the relation between ICA and nonlinear versions of the PCA criteria is quite strong.

Suppose y_1, y_2, …, y_N random variables with joint pdf given by f(y_1, y_2, …, y_N). If the random variables y_i are statistically (mutually) independents then the joint pdf can be factorized since

$$f(y_1,...y_N)=\prod_{i=1}^{N} f_{y_i}(y_i) \qquad (1)$$

where $f_{y_i}(y_i)$ denotes the marginal density of y_i. If the random variables y_i are statistically independents, then for any functions g_1 and g_2 one has

$$E\{g_1(y_i)g_2(y_j)\}-E\{g_1(y_i)\}E\{g_2(y_j)\}=0, i\neq j \qquad (2)$$

which is clearly a stricter condition than the condition of uncorrelatedness given by

$$E\{y_i y_j\}-E\{y_i\}E\{y_j\}=0, i\neq j \qquad (3)$$

However for the special case of joint Gaussian distribution, independence and uncorrelatedness are equivalent [2] and ICA becomes in these cases not interesting or impossible.

A simple neural network algorithm based on information maximization (Informax) was derived by Bell and Sejnowski [3] and is able to separate super-Gaussian (sparse) independent components. A source s_i can be distinguished from mixtures x_i by considering the activity of each source statistically independent of the other sources. This means that their joint probability density function, measured across the input time ensemble factorizes. Therefore the mutual information between any two sources, s_i and s_j is zero:

$$I(y_1,y_2,...,y_N)=E\left\{\ln\frac{f_y(y)}{\prod_{i=1}^{N} f_{y_i}(y_i)}\right\}=0 \qquad (4)$$

where $E\{.\}$ denotes mathematical expectation. The sources s_i are assumed to be temporarily independent, while the observed mixtures of sources, x_i are statistically dependent on each other, therefore the mutual information between pairs of mixtures, $I(x_i,x_j)$ is in general positive. The problem of blind source separation consists in finding a matrix **W** such that the linear transformation

$$I=Wx=Was \qquad (5)$$

re-establishes the condition $I(y_i,y_j)$=0, for all i≠j.

Consider the joint entropy of two non-linearly transformed components of **u**:

$$H(u_1,u_2)=H(u_1)+H(u_2)-I(u_1,u_2) \qquad (6)$$

where $u_i=g(y_i)$ and $g(.)$ is an invertible, bounded nonlinearity. The nonlinear function provides, through its taylor series expansion, higher order statistics which are necessary to establish independence.

The maximization of the joint entropy is obtained by maximizing the individual entropies, $H(u_1)$ and $H(u_2)$ and minimizing the mutual information $I(u_1,u_2)$. In general the maximization of $H(u)$ minimizes $I(u)$ and when the mutual information reaches the value zero the two variables become statistically independents. The algorithm attempts to maximize the entropy by iteratively adjusting the elements of the square matrix **W**, by using small batches of data vectors drawn randomly from {**x**}. Without substitution, one has

$$\Delta W\propto\frac{\partial H(u)}{\partial W}W^TW=\left[I+\phi y^T\right]W \qquad (7)$$

where

$$\phi_i=\frac{\partial}{\partial y_i}\ln\frac{\partial u_i}{\partial y_i}$$

The term (W^TW) is the natural gradient and avoids matrix inversions speeding up the convergence. The form of the nonlinearity g(u) is crucial in the performance of the algorithm and its ideal form is the cumulative density function (cdf) of the distributions of the independent sources.

Assuming that the complexity of the EEG dynamics can be modelled as a relatively small number of independent brain processes, the EEG source analysis problem satisfies ICA assumption. The foremost problem in interpreting the output of ICA is determining the number of input channels, and the physiological and/or psychophysiological significance of the derived source channels.

III. Experimental results

The extended ICA algorithm was tested in both simulated data, as shown in figure 1, and in real data as shown in figure 2. Figure 1a) shows four independents generated signals that are then linearly mixed resulting the signals shown in figure 1b). Figure 1c) shows the result of the extended ICA decomposition algorithm applied to the signals shown in figure 1b), which obviously does not take into consideration the linear transform from which the signals obtained in figure 1b) were obtained from the ones shown in figure 1a).

By comparing figures 1a) and 1c) we can conclude that the result of the decomposition is satisfactory since the order, polarity and amplitude of the output only have a simple changing.

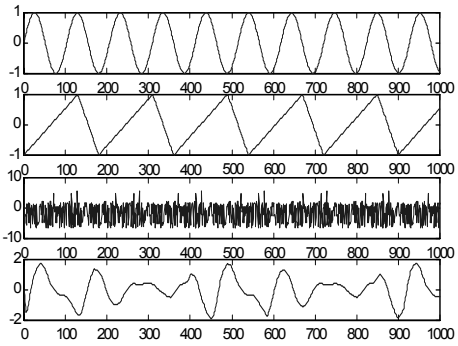


Figure 1a). Four signals generated independently

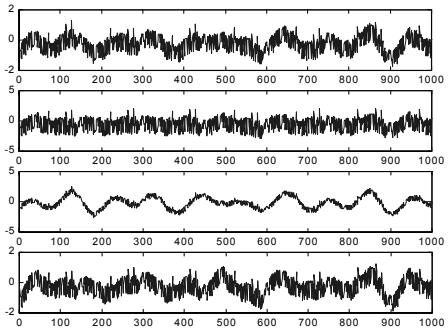


Figure 1b). The Signals shown in figure 1a) after passed through a random mixed matrix.

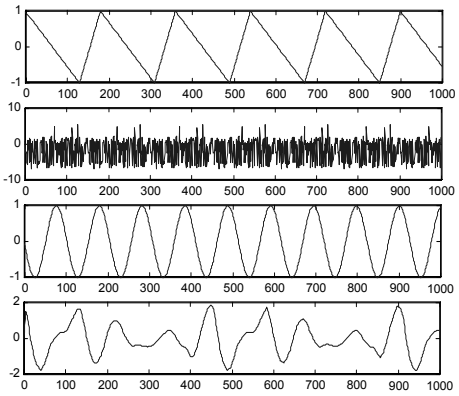


Figure 1c). Signals after ICA decompose

The extended ICA algorithm was also applied to the analysis of 10 EEG recordings of the human brain activity. To ensure signal stationarity the time index was permuted, and the 10-dimensional time vectors were presented to a 10->10 ICA network one at a time. First and second order statistics were removed in order to speed up the convergence, so the data were first pre-whitened. The learning rate was annealed from 0.03 to 0.0001 during convergence. After each pass through the whole training set, the value of correlation between the ICA output channels and the value of change in the weight matrix were checked, and the training was stopped when the mean correlation among all channel pairs was bellow 0.06 and the ICA weights had stopped changing appreciably.

EEG recordings of the human brain generally include either super-Gaussians signals (ERPs for example), or sub-Gaussian signals (for example working frequency disturb and EOG). So ICA appears suited for this kind of applications as shown in figure 2 where the experimentation was done in real EEG data.

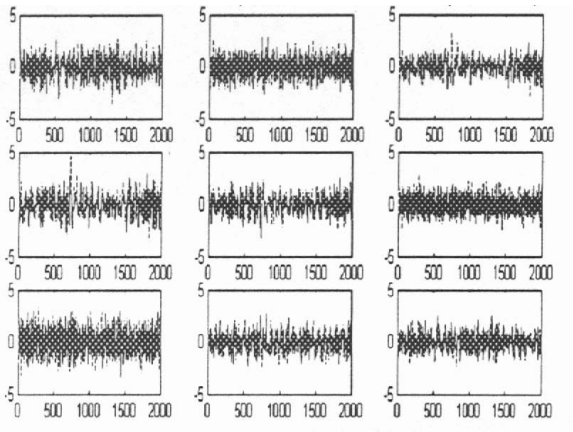


Figure 2. EEG real data separated by ICA

The first row on the left of figure 2 shows a normal EEG data, the second row is a close and open eye's EEG data and finally the third row is a working frequency disturbing. These original signals were mixed as in the last case of synthetic data and the ICA algorithm realized the blind source separation. The results are very promising taking into consideration that the target signals include both super-Gaussian and sub-Gaussian sources.

IV. Discussion

This paper has focused on the application of ICA to the analysis of EEG, which proved a reasonable efficiency.

Apart from the brain signals, signals from other organs, as for example from the heart system have similar problems with artifacts and could also benefit from ICA techniques. In general biomedical signals are a rich source of information about physiological processes, but they are often contaminated with artefacts or noise and are typically mixtures of unknown sources summing differently at each sensor. Besides other interesting questions such as to understand the nature of the sources, ICA seems to hold a great promise, for blindly separating artifacts and decomposing the mixed signals into subcomponents that may reflect the functionality of distinct generators of physiological processes, which must also be interpreted in the near future.

References

[1] p. Comon, "Independent Component Analysis- a new concept?".Signal Processing, 36:287-314, 1994.

[2] A. Papoulis, "Probability random variables and stochastic processes," McGraw-Hill International Editions,1994.

[3] A.J. Bell, and T.J. Sejnowski, "The independent components' of natural scenes are edge filters", Vision Research, 37:3327-3338, 1997.

[4] D. Yellin, and E. Weinstein, "Multichannel signal separation: Methods and analysis," IEEE Trans. on Signal Processing, 44:106-118, 1996.

[5] J.-F. Cardoso, "Super-symmetric decomposition of the fourth-order cumulant tensor. blind identification of more sources than sensors," In Proc. ICASSP'91, pages 3109-3112, 1991.

[6] A. Hyvärinen, R. Cristescu, and E. Oja, "A fast algorithm for estimating overcomplete ICA bases for image windows," In Proc. Int. Joint Conf. on Neural Networks, Washington, D.C., 1999.

3rd International Workshop MAVEBA 2003, 103-106

SPECTRAL BI-NORMALISATION FOR SPEECH RECOGNITION IN ADDITIVE NOISE

C. S. Lima[1], J. F. Oliveira[2]
[1]Department of Industrial Electronics, Universidade do Minho, Guimarães, Portugal
[2]Department of Electrical Engineering, Instituto Politécnico de Leiria, Leiria, Portugal

***Abstract*: The changing on peaks structure of the speech spectrum is perhaps the most important cause of degradation of speech recognition systems under adverse conditions. Another drawback concerned to the additive noise effect occurs on the flat spectral zones which are usually raised proportionally to the noise level. These combined effects on both the peaked and the flat spectral zones can be alleviated by trying to restore its original structure, which assumes noise knowledge. However, the random nature and the variability of the noise, the difficulty in discriminating speech pauses, among others, discourage the use of noise estimates as the basis of robust speech recognition algorithms. Alternative approaches based on normalisation procedures become very promising since the noise effect can be alleviated without any knowledge regarding to its existence. This paper suggests a spectral normalisation that though being different can be viewed as a noise estimation procedure in a frame by frame basis, so assuming the clean database as lightly corrupted. This speech normalisation is used to restore the normalised speech spectrum. This normalised spectrum is then re-normalised by a baseline spectrum normalisation method, which concentrates essentially in the speech regions of small energy, since in these regions the noise is more dominant, so they require a better degree of robustness.**

I. INTRODUCTION

In [1] it is argued that a proper spectral normalisation, which concentrates essentially on the speech regions of less energy, could improve significantly the robustness of speech recognition systems when operating under additive noise conditions. From a theoretical point of view, the spectral regions with small energy would need more noise robustness, given that for the same noise level they are more corrupted. The spectral regions of small energies usually correspond to unvoiced sounds regions, which are spectrally not very well defined. Roughly speaking nearly half of the consonants can be classified as unvoiced, while the other half and the vowels are generally classified as voiced. Generally the importance of the vowels in classification and representation of written text is very low; however, most practical automatic speech recognition systems rely heavily on vowel recognition to achieve high performance. Consequently, the spectral regions which contains higher speech energy seems to be usually more important in speech recognition under difficult conditions once they are generally less corrupted. On the other hand, the spectral regions with small energy are more corrupted, thus they need a larger degree of robustness.

Others authors [2] have also given an increasing importance to the spectral regions of small energy of the speech signal, although by using alternative approaches.

The algorithm proposed in [1] does not take into consideration the properties of the voiced speech regions, which are usually characterised by "peaked" spectral zones. These portions of spectrum are flattening, as the noise becomes more and more dominant which degrades the system performance.

The algorithm proposed in [3] tries to cope with this limitation by restoring partially both the original spectral "peaks" and the flat spectral regions where the signal power is increased by the wide band noise effect. This approach assumes the clean database lightly contaminated and the noise power is estimated in a frame-by-frame basis by the lowest power of all the sub-bands in each segment. The algorithm does not assume noise existence, in the sense that the features are extracted exactly in the same way in both noisy and noise free conditions. One drawback associated with this algorithm is concerned to the noise estimate which includes a significant amount of speech characteristics that is proportional to the number of spectral components that constitute a sub-band. This can mean that to many speech characteristics can be disregarded in the restoration of the clean speech normalised features. Another drawback of the algorithm proposed in [3] is that the spectral peaks classification is based on heuristics, which is obviously undesirable. In order to overcome these drawbacks the algorithm proposed in this paper differs from the algorithm proposed in [3] essentially in the following aspect:

The frame by frame spectral normalisation is done before the baseline normalisation instead of after it, assuring that the spectrum that will be processed by the baseline spectral normalisation is always the normalised spectrum (by the small spectral component), which is not very dependent on the noise level.

The results show a significant improvement in performance when compared with the baseline method when used alone [1] and an interesting improvement in performance when compared with the algorithm proposed in [3].

II. BASELINE SPECTRAL NORMALISATION

The baseline spectral normalisation defined in [1] is motivated by the fact that the additive noise is not a narrow band noise, thus its spectrum is reasonably dispersed in frequency. Additionally a mechanism adequate to dealing with non-stationary additive noise situations, which frequently occurs in practical situations, is needed. One solution can be trying to extract the distribution of the speech energy along the spectrum, normalised by the total energy of the speech within the segment. Therefore noise variations can be attenuated once that which is really measured is the relative and not the absolute distribution of the spectral energy of the speech signal.

The baseline normalisation process consists in a division of the frequency band in sub-bands given that usually a very fine detail in frequency is not required for western languages speech recognition applications. The method is based on the power spectral density components and consists in dividing the speech power inside each sub-band by the total short-time speech power. The power in each sub-band is obtained by summing the components of the power spectral density inside the sub-band. All the sub-bands have the same number of spectral components and any spectral component is shared by different sub-bands, thus avoiding increases of statistical dependence between sub-bands (feature components). The background noise contributes simultaneously to increase the sub-band and total power, which contributes for stabilising the amplitudes of the feature vectors.

To best understand this reasoning, consider S_i denoting the speech power in sub-band i and S denoting the short time speech signal power of the considered segment. Similarly, let N_i and N denote the power of the noise in sub-band i and the short time noise power, respectively. So, the i^{th} component of the observation vector for clean and noisy speech are given respectively by

$$c_i = \frac{S_i}{S} \quad , \quad c_i = \frac{S_i + N_i}{S + N} \tag{1}$$

Figure 1 shows the clean speech and noisy speech spectral power normalisation features for 240 ms of the word "zero" where each sub-band has 16 power spectral components. The SNR is 0 dB.

If the noise is stationary then its short time power equals its long time power. Note that this is not true for the speech due to its non-stationary property, but as an approximation we will consider that the short time speech signal power equals the long time speech signal power. Under this constraint, S and N can be related by the signal to noise ratio (SNR). Therefore the next expression holds

$$S + N = S\left(1 + \frac{1}{10^{\frac{SNR}{10}}}\right) \tag{2}$$

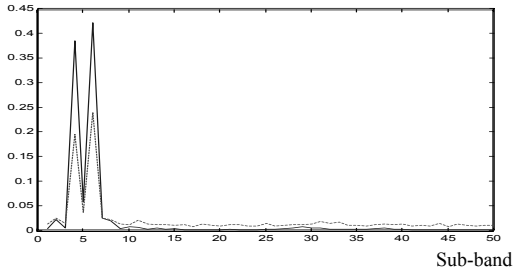


Figure 1. White noise effect in the power spectrum density normalization domain in the beginning of digit "zero". Dashed line represents noisy speech features.

If the noise has white noise characteristics the environment will shift the clean speech vector by a noise dependent vector $C_i(N)$, which can be calculated by subtracting equations (1).

Let l, the number of components in each sub-band and L the FFT length. Then N and N_i, considering flat noise spectrum, are related by the quotient l/L. By using these considerations, the calculation of the shift vector imposed by the environment is accomplished by subtracting equations (1) and becomes [1]

$$C_i(N) = \left(\frac{S_i}{S} - \frac{l}{L}\right)\frac{1-k}{k}, \quad k = 1 + \frac{1}{10^{\frac{SNR}{10}}} \tag{3}$$

Equation (3) shows that if the speech has a flat power spectrum density, the means of $C_i(N)$ become null as S_i/S equals l/L. Thus, this normalisation process becomes optimal in the sense that the environment does not affect the means of the speech features. This means that this normalisation procedure provides some noise robustness to unvoiced speech segments, where neither the speech nor the noise are spectrally well defined. More details can be found in [1]

III. ADDITIVE WHITE NOISE EFFECT AND PRE-PROCESSING APPROACH

Figure 1 shows that the noise effect, in the proposed power spectral baseline normalisation domain, is raising the "flat" spectral zones while the "peaked" spectral ones are "flatten". In fact equation (1) in noisy conditions (equation shown on the right) shows that, for sub-bands with high speech power, as the amount of noise in the sub-band is much smaller than the total amount of noise, the speech features in that regions are decreased proportionally to the amount of contaminating noise. For sub-bands with small speech power the opposite happens, given that the sum of all the coefficients extracted in each segment is unitary. As the spectral flattening is proportional to the amount of contaminating noise, for low signal to noise ratios the "peaked" spectral regions almost disappear, which is the main origin of degradation in performance under noisy conditions.

The main goal of a robust features extraction method is providing robustness against noise or other sources of variability by ignoring its presence. Although the noise can be compensated, the effectiveness of this approach becomes very dependent on the accuracy of the noise estimate, which is a very hard task in practical situations. Hence our main goal was searching for a compensation process independent of the noise level or characteristics, although the proposed baseline normalisation assumes a wide band additive noise for maximal performance. More details can be found in [1].

In this context we propose the following two steps approach:

1) For task uniformity in clean and in noisy conditions the clean database must be considered lightly contaminated. Trying to clean completely the database, which can be viewed as another kind of normalisation, represents a procedure compatible with the noise compensation paradigm, however if the procedure is not particularised for any kind of noise, it can be used without concerning to the noise existence. Hence, under noisy conditions the features extraction method can compensate for the noise existence taking into account the noise level, which can be estimated in a frame-by-frame basis, becoming the procedure compatible with real time applications.

2) The estimated noise level, which really constitutes a spectral normalisation by the smallest spectral component. This speech component, which has small significance and is proportional to the amount of noise must be used to alleviate the noise effect. Then the baseline spectral normalisation algorithm [1] can be more efficient since the noise effect was *a priori* reduced.

IV. Proposed noise compensation

To cope with the additive noise effect we propose estimating the noise power in each segment, which can be viewed as a secondary normalisation procedure (the first normalisation procedure is behind the normalisation proposed in the baseline system [1]) by taking the value of the lowest component of the power spectrum density in each speech frame.

We propose alleviating the noise effect by subtracting the estimated noise level from all the others components of the feature vector. Therefore the power spectral components of the speech must be changed so that

$$c_i = \begin{cases} P_i - \min\{P_i\}, P_i \neq \min\{P_i\} \\ P_i, otherwise \end{cases} \quad (4)$$

where P_i denotes the amplitude of the ith component of the power spectral component of the speech, and c_i denotes the ith component of the normalised spectrum (observation vector) that will be processed by the baseline spectral normalisation algorithm proposed in [1]. The spectral normalisation procedure described by equation (4) reduces clearly the noise effect since a factor (lower spectral component in each segment) that is proportional to the noise level is subtracted from all the others spectral components. Additionally the speech characteristics described by the smallest spectral component are maintained since this component is included in the observation vector. However these mathematical operations involving all the spectral components can increase the statistical dependence among them, which is undesirable regarding to the HMM modelling. In this context the baseline spectral normalisation procedure helps to decorrelate the data since the data are grouped and processed inside the group independently of the data inside the other groups.

Therefore considering wide band noise its effect is reduced in terms of means. It is obvious from equation (1) that the variance effect is also reduced by the baseline normalisation procedure once that each observation is divided by the power of the speech segment.

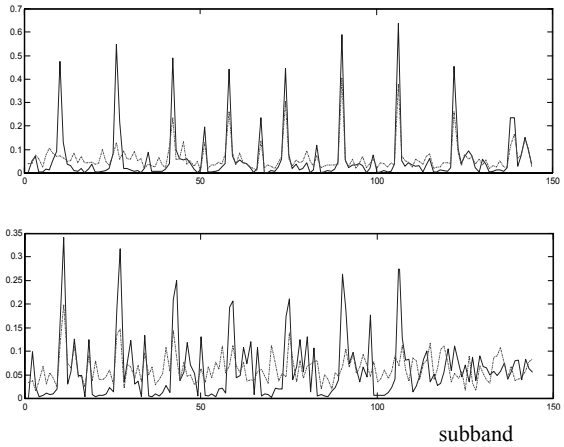


Figure 2. Spectral speech structure recovered by the algorithm proposed in [3] for the first half of the word "eight" at an SNR of 0 dB. Normal line stands for clean speech.

This *a priori* noise effect attenuation obtained by spectral normalisation in each frame shows better effectiveness than the *a posteriori* noise effect attenuation described in [3] as can be observed by comparing figure 2 and figure 3. It is clear that in figure 3 the recovered peak structure is more closed to the peak structure of the clean speech than the recovered peak structure in figure 2.

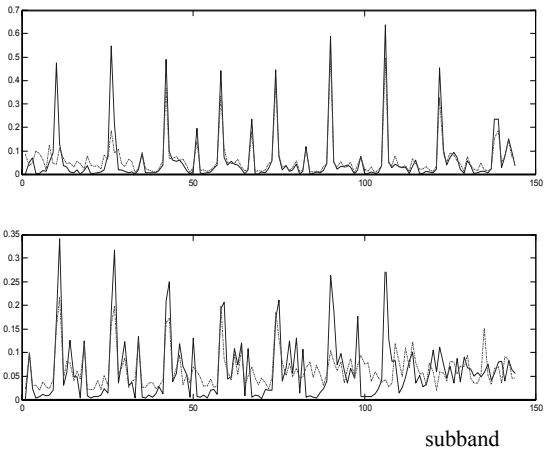


Figure 3. Spectral speech structure recovered by the algorithm proposed in this paper for the first half of the word "eight" at an SNR of 0 dB. Normal line stands for clean speech.

V. Experimental results

The proposed algorithm was tested in an Isolated Word Recognition system using Continuous Density Hidden Markov models. The database of isolated words used for training and testing is from AT&T Bell. The used speech was acquired under controlled environmental conditions band-pass filtered from 100 to 3200 Hz, sampled at a 6.67 kHz and analysed in segments of 45 ms duration at a frame rate of 66.67 windows/sec. Only the decimal digits were used. The noise has white noise characteristics, is speech independent and computationally generated at various SNR as shown in table 1. The goal is to compare the performance of the proposed and contemporary speech robust features. Some of these robust features are the OSALPC (One-Sided Autocorrelation Linear Predictive Coding), the conventional cepstrum with liftering (CEPS + liftering) and the well known MFCC (Mel-Frequency Cepstral Coefficients). In table 1, MMC stands for conventional Markov model composition in the power spectrum density domain, Norm. stands for the baseline normalisation procedure described in [1], N. + MMC stands for Markov model composition in the baseline power normalisation domain [1], PR stands for the post-processing spectral restoration procedure proposed in [3] and BN stands for the bi-normalisation proposed in this paper. Table 1 shows that the suggested spectral multi-normalisation features are more effective against additive white noise than both the baseline normalisation, which is more effective than some robust features used nowadays, and the PR algorithm proposed in [3]. For SNR greater than or equal to 5 dB the baseline spectral normalisation outperforms the conventional Markov model composition (MMC) when the noise parameters are learned from the periodogram method in a data segment of 100ms without speech. As in the Parallel Model Combination, the distortion can be integrated (compensated) in the composite model increasing thus the recogniser performance [1]. On the first six entries of the table 1, all the features are 8 static, energy and dynamic features excepting * (12 static + energy + dynamics) and ** (13 static + energy + dynamics).

Table 1 – Performance of the spectral normalisation

SNR (dB)	15	10	5	0	-5
LP	56.5	39.5	30	16.25	
OSALPC	98.25	92	65.75	32.25	
CEPS *	97.5	95	72	34.5	
+liftering	98.25	95	75.25	39	
MFCC **	97.75	94.75	72.25	37.5	
OSALPC*	98.5	96.25	74.25	32.5	
MMC	98	96.75	92.5	91	78.5
Norm.	98.5	97.75	93.75	88	42.5
PR	99.25	98.25	95	89.75	61.5
BN	99.25	98.5	95.75	90.75	64.25
N.+ MMC	99.5	98.75	97.25	92.25	84.75

VI. Discussion

The main advantage of this bi-normalisation process is the recognition performance obtained when no knowledge of the noise statistics exists. As a robust extraction features, the suggested method seems to be superior to the most used nowadays. Additionally, for white noise and at SNR greater than or equal to 5 dB it presents better performance than a standard noise compensation technique, which assumes integral noise knowledge. In fact for high Signal to Noise Ratios the spectral normalisation where the distortion is ignored outperforms the Markov model composition where the distortion is learned from a small amount of isolated noise samples and incorporated into the system. If isolated noise samples exist, the noise can be estimated and this knowledge can be incorporated into the system, and consequently increasing the recogniser performance.

References

[1] C. Lima, Luís B. Almeida, and João L. Monteiro, "Improving the Role of Unvoiced Speech Segments by Spectral Normalisation in Robust Speech Recognition," *proceedings of the 7th International Conference on Spoken Language Processing,* vol. pp 1573 - 1576 , 2002.

[2] Biksha Raj, "Reconstruction of Incomplete Spectrograms for Robust Speech Recognition", *Ph. D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University*, 2000.

[3] C. Lima, Luís B. Almeida, A. Tavares, and C. Silva, "Spectral Multi-Normalisation for Robust Speech Recognition", *IEEE & ISCA Workshop on Spontaneous Speech Processing and Recognition,* pp 39 - 42, 2003.

3rd International Workshop MAVEBA 2003, 107-110

ALGORITHM OF PHONEME IDENTIFICATION USING FAST MEASUREMENT OF WIENER KERNELS OF SPEECH SIGNALS

A. M. Krot[1], P. P.Tkachova[2]
[1]United Institute of Informatics Problems of the National Academy of Sciences of Belarus 6, Surganov Str., 220012, Minsk, Belarus. Tel.: (375 17) 2842086; Fax.: (375 17) 2318403. E-mail: alxkrot@newman.bas-net. by
[2]Belarusian State University, 4, Skorina Av., 220050, Minsk, Belarus. Tel.: (375 17) 220 30 77

Abstract: **The nonlinear speech signal decomposition based on Volterra-Wiener functional series is described. The solution of speech recognition problem by means of measuring Wiener kernels is proposed. The recognition system of speech signal is considered for speech phoneme identification.**
Keywords: **Nonlinear signal decomposition, Wiener kernels, phoneme recognition**

I. INTRODUCTION

For speech signal recognition problem solving there are a variety of paradigms and approaches. Among them, we can mention a statistical approach [1], [2], [3], a nonlinear dynamic method using neural networks [4], a dynamic programming method and so on. These methods include modeling of an acoustic processor's performance. In different systems, the acoustic processor varies in complexity. Nevertheless, it is desirable for the acoustic processor to take into account analyzed acoustic signal peculiarities. This reason require, in general case, implementation of the acoustic processor in the form of a nonlinear model.

Among the different approaches to synthesis of a nonlinear model is that it should be based on the Volterra-Wiener functional series [5], [6]. This approach allows identification and modeling of systems without additional preliminary information about their structure. This method was developed for solving problems of identification of nonlinear dynamical systems (NDS) in control theory and of analysis of physiological systems in biology. These facts are premise for the usage of this approach to solve the speech signal recognition problem.

This paper presents the nonlinear model of the acoustic processor based on Volterra-Wiener functional series. We show the usage of this nonlinear decomposition for speech phoneme identification.

II. LINEAR AND NONLINEAR DECOMPOSITION OF SIGNAL INTO SERIES OF FUNCTIONS AND FUNCTIONALS

It is well known that linear signal $y(t)$ (for example, music signal) in the time t can be represented as an output of a linear dynamic system (LDS) of the kind:

$$y(t) = \sum_{k=-\infty}^{\infty} H(\omega_k) X(\omega_k) e^{i\omega_k t}. \quad (1)$$

where the input signal $x(t)$, or its Fourier image $X(\omega_k)$, acting on LDS generates the output signal $y(t)$, and the transfer function $H(\omega_k)$ is an impulse response LDS function $h(t)$ in frequency domain:

$$H(\omega_k) = \frac{1}{T}\int_0^T h(t) e^{-i\omega_k t} dt. \quad (2)$$

As concerns a majority acoustic signals (including speech signal), these signals are the product of strongly nonlinear dynamical systems, i.e. they are nonlinear processes.

According to [5], [6], [7], [8] the output NDS signal can be represented by means of the Volterra-Wiener series as follows:

$$y(t) = h_0 + \sum_{k_1=-\infty}^{\infty} H_1(\omega_{k_1}) X(\omega_{k_1},\theta) e^{i\omega_{k_1} t} +$$

$$+ \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} H_2(\omega_{k_1},\omega_{k_2}) X(\omega_{k_1},\theta) X \times$$

$$\times X(\omega_{k_2},\theta) e^{i(\omega_{k_1}+\omega_{k_2})t} -$$

$$- D_x \sum_{k_1=-\infty}^{\infty} H_2(\omega_{k_1},-\omega_{k_1}) +$$

$$+\sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty}\sum_{k_3=-\infty}^{\infty} H_3(\omega_{k_1},\omega_{k_2},\omega_{k_3})X(\omega_{k_1},\theta)X(\omega_{k_2},\theta)$$

$$\times X(\omega_{k_3},\theta)e^{i(\omega_{k_1}+\omega_{k_2}+\omega_{k_3})t} -$$

$$-3D_x\sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty}H_3(\omega_{k_1},-\omega_{k_1},\omega_{k_2})X(\omega_{k_2},\theta)+\ldots, \quad (3)$$

where θ is a state parameter, $\theta\in[0,1]$, $H_m(\omega_{k_1},\omega_{k_2},\ldots,\omega_{k_m})$ are Wiener kernels of m-order (the first order Wiener kernel is the transfer function (2) of linearized NDS)

Comparing (1) with (3), we can see that above-mentioned linear decomposition is a particular case of this nonlinear decomposition, similarly LDS is the specific case of NDS.

III. Identification algorithm based on measuring the Wiener kernels on finite intervals

When NDS model based on Volterra-Wiener series is realized on a computer, discrete input x_n and output y_n signals and Wiener kernels $h_m[n_1,\ldots,n_m]$ have a finite duration in time; that is why a refinement of the relations (3) is required.

To represent one-dimensional sequences y_n and x_n with the finite length N in frequency domain, we use coefficients of discrete Fourier transform (DFT) Y_k and X_k respectively. Analogously, for a frequency representation of Wiener kernels $h_m[n_1,\ldots,n_m]$, their multidimensional analogs are needed, i.e. the coefficients of m-dimensional DFT's $H_m[k_1,\ldots,k_m]$.

The identification scheme of discrete NDS is similar to Wiener's circuit for determining m -order kernel [5]. This scheme can be presented as follows: white Gaussian noise with zero mean and variance D_x, is given by the inputs of the unknown NDS and a known system as a bank of m complex exponential filters with the multiplying outputs. Then the output signals from the unknown and known systems and bank are multiplied and the result signal is averaged [6].

The DFT-image of kernel $h_1[n]$ can be calculated as follows [7], [8]:

$$H_1[k]=\frac{Y_k X_k^*}{ND_x}. \quad (4)$$

According to relation (4), $H_1[k]$ is the sample of transfer function $H_1(\omega)$ for the stationary LDS identified on the basis of stationary white noise $\{x_n\}$.

The DFT-image of kernel $h_2[n_1,n_2]$ can be estimated in an analogous manner [7], [8]:

$$H_2[k_1,k_2]=\frac{Y_{k_1+k_2}X_{k_1}^*X_{k_2}^*}{2ND_x^2}-\frac{h_0}{2D_x}\delta_{k_1,N-k_2}. \quad (5)$$

The work [8] shows that DFT-image of kernel $h_3[n_1,n_2,n_3]$ is:

$$H_3[k_1,k_2,k_3]=\frac{Y_{k_1+k_2+k_3}X_{k_1}^*X_{k2}^*X_{k3}^*}{6ND_x^3}-$$

$$-\frac{N(H[k_1]\delta_{k_2,N-k_3}+H[k_2]\delta_{k_1,N-k_3}+H[k_3]\delta_{k_1,N-k_2})}{6D_x}. \quad (6)$$

IV. The recognizer of phonemes of Belarusian lanquage based on its Wiener kernels measuring

The nonlinear decomposition (3) can be used for identification of a group of phonemes (in particular, sonorous phonemes of Belarusian language) by means of the *m*-order multidimensional nonlinear filters.

As concerns a phonetic structure of Belarusian language, the following classification is used [7], [8]. All the phonemes of Belarusian language are divided into two groups: the first group has vocal (vowel) ones and the second group has consonant ones.

The vocal phonemes are again divided into labial once (Î, Ó) and nonlabial once (À, Ý, I{Û}); the sound Û is not considered as individual phoneme since in Belarusian language it appears only after hard consonants and is modification of the phoneme I.

The consonants are classified by the two blocks involving the ten groups [7],[8]:

The labial (the block A):

1. Labial-labial, hard: Á, Ï, Ì, Â, ¡.
2. Labial-labial, politicization (soft): Á', Ï', Ì', Â'.
3. Labial-dental, hard: Ô.
4. Labial-dental, soft: Ô'.

The lingua (the block B):

5. Front, dental, hard: Ä, Ò, Ç, Ñ, (Z), Ö, Ë, Í.
6. Front, dental, soft: Ç', Ñ', Z', Ö', Ë', Í'.
7. Front, alveolar, hard: Æ, Ø, Ž, ×, Ð.
8. Middle, soft: J(É).
9. Back, hard: (Ã), (Ê), ã, Õ.
10. Back, soft: (Ã'), (Ê'), (â'), (Õ').

According to this classification we have a recognizer in the form of the nonlinear filter banks consisting of 10 Volterra-Wiener filters which can include from 1 to 7 nonlinear multidimensional (m-order) filters (or Wiener kernels). Each of them can be stimulated by the white noise or another type of testing signal. Then each phoneme (there are from 1 to 7 ones for each corresponding Volterra-Wiener filter) can be recognized as one from Wiener kernels based on the afore mentioned identification scheme (see Sect. III). As the final result, each phoneme corresponds to itself functionally, i.e. to a certain Wiener kernel [7], [8].

In the case of the other type of signal (colored noise, tone plus noise, etc.), the identification scheme can be built by analogy [9].

V. Computer realization of nonlinear decomposition for speech phoneme signal recognition

It is well-known the main point in designing automatic speech recognition systems is the modeling of speech signal variability. There exist two kinds of variability: temporal and acoustic. In the best manner, the temporal variability can be modeled by hidden Markov models [3], [10]. The acoustic variability is more complicated for modeling because of its nonlinear nature [10].

In this connection we will use the nonlinear decomposition based on Volterra-Wiener functional series for modeling and recognition of speech phoneme signals.

The recognition system (acoustic processor) of speech phoneme signals based on functional Volterra-Wiener decomposition operates in two stages. The first stage permits finding of speech phoneme standards, while the second stage realizes an acoustic recognition procedure.

During the first stage (Fig. 1), speech phoneme signal samples enter the input of recognition system (a phoneme belonging from the phoneme alphabet is supposed to be measured) while a test signal (a white noise) enters on the second input. The input speech signal decomposition into functional is carried out in the block for finding m Wiener kernels. Then an average of obtained kernels for several samples of the chosen phoneme is fulfilled in the block of sampling average. As a result, a final estimate of Wiener kernel is obtained from the output of the block of input signals characterizing a phoneme [8]. The obtained set of kernels for enough large number of samples is a *phoneme signal standard*. Such a set of kernels is found for all phonemes belonging to a phoneme alphabet of the recognition system [8].

During recognition stage (Fig.2), a real speech signal, as well as white noise, enter as input to the system (let us note that the signal belonging to a phoneme is not known) [8]. The Wiener kernel estimations are calculated based on input signals into the speech signal decomposition block (they characterize a chosen speech signal relative to input white noise). The obtained kernel estimations are given to a classifier, together with phoneme standards found earlier. The classifier makes a decision on what kind of phoneme the input signal belongs to. A sequential or parallel classifier can be used as the classifier; among these classifiers are based on neural networks or applied in statistical pattern recognition (for example, Bayes or Wald classifiers) [8].

From the description of stages functioning, it follows that operation of input signal decomposition into Volterra-Wiener functional series is carried out both the speech recognition signal stage and the phoneme standard finding one. As a result, the Wiener kernels are measured. Unlike training stage, which requires a large enough training sample (with a view of proximate Wiener kernels finding), the speech recognition stage permits us to estimate the Wiener kernels only approximately , i.e., with errors, because a real (workable) speech signal is limited by the short length of sample. It is obvious that the computational error is greater for the second case. In this connection, it is important to develop efficiency both in time and in accuracy methods for Wiener kernels measuring in the case of small lengths of the speech signal.

References

[1] L.R. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs NJ: Prentice Hall Inc., 1993.

[2] H.A Bourland, and N. *Morgan Connectionist Speech Recognition: A Hybrid Approach*. Boston MA: Kluwer Academic Publishers, 1994.

[3] E.I. Bovbel, P.P.Tkachova, and I.E. Kheidorov, "Autoregressive Hidden Markov Models for Isolated Words Recognition", *Recent Advances in Information Science and Technology.* N. Mastorakis, Ed. Singapore etc.: World Scientific, 1998, pp. 211-214.

[4] P. Mehra, and B.W. Wah, (eds), *Artificial Neural Networks. Concept and Theory,* Los Alamitas Ca: IEEE Computer Society Press, 1992.

[5] N. Wiener, *Nonlinear Problems in Random Theory.* New York: John Wiley and Sons, 1958.

[6] A.S. French, and E.G. Butz, "Measuring the Wiener Kernels of Nonlinear System Using the Fast Fourier Algorithm", *Int. J.Control,* no. 17, pp. 529-539, 1973.

[7] A.M. Krot, and P.P. Tkachova, "The Nonlinear Signal Decomposition in Voice Recognition System Constructing", *Proc. of International*

Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Firenze, Italy, pp.54-58, September 1-3, 1999.

[8] A.M. Krot, P.P. Tkachova, and B.A. Goncharov, "New Approach to Speech Signal Recognition Using Nonlinear Signal Decomposition by Measuring Wiener Kernels", *Smart Engineering System Design,* vol.4, pp. 265-276, 2002.

[9] A.M.Krot, and M.A.Shcherbakov, "Identification of Discrete Input Nonlinear Systems for Digital Chaotic Signal Processing", *Proc. of 2nd IMACS International Conference "Circuits, Systems and Computers (CSC'98)",* vol.2, Athens, Greece, pp. 795-797, October 26-28, 1998.

[10] A.M. Krot, P.P. Tkachova, and H.B. Minervina, "On Algorithm for Phoneme Speech Recognition Using Nonlinear Signal Decomposition", *Proc. of 8th IEEE International Conference on Electronics, Circuits and Systems (ICECS 2001),* vol. 3, Malta, pp.1251-1254, September 2-5, 2001.

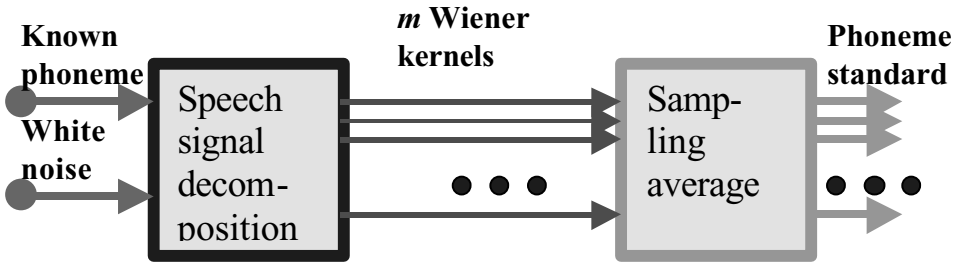


Fig.1 Recognition system functioning for the stage of phoneme standard finding

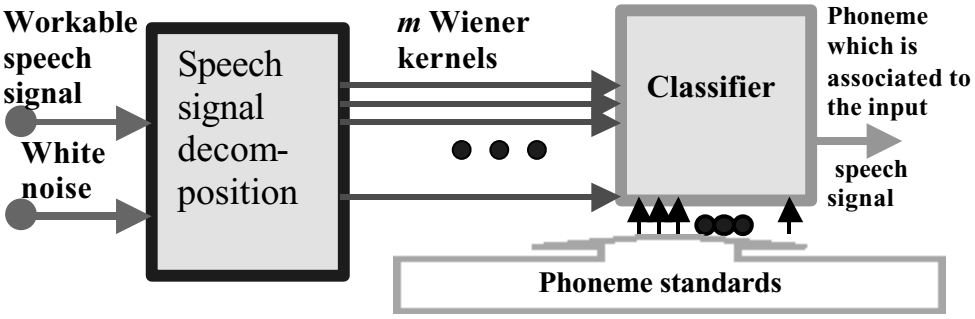


Fig.2 Recognition system functioning for the stage of speech recognition

3rd International Workshop MAVEBA 2003, 111-114

DISTINGUISHING AND RECOGNITION OF PATHOLOGICAL SPEECH BASED ON ESTIMATION OF CONTROL PARAMETER OF CHAOTIC ATTRACTOR

A.M.Krot[1], H.B.Minervina[1], P.P.Tkachova[2]

[1] United Institute of Informatics Problems National Academy of Sciences of Belarus
Surganov Str.,6, Minsk, 220012, Belarus
[2]Belarusian State University
Skorina Av.,4, Minsk, 220050, Belarus

Abstract: **This paper investigates the approach for revealing pathological speech signal based on estimating specific geometric structure of Lorenz attractor in a chaotic regime. Analysis of the Lorenz attractor on the basis of proposed nonlinear decomposition into matrix series is developed. This analysis permits to estimate the values of characteristic parameters (including control one) of Lorenz attractors and predict their evolution in time. This paper shows that estimation of control parameter of Lorenz attractor in the chaotic regime permits to distinguish even very similar speech signals.**
Keywords: **pathological speech signal, attractor, matrix decomposition**

I. INTRODUCTION

The nonlinear dynamical systems (NDSs) with self-organization named as complex systems are investigated with great activity in last decades. A complex NDSs functioning is closely connected with the presence of chaos in their behavior. The NDSs behavior can be described on the basis of construction of chaotic attractor in *m*-dimensional Euclidean state-space. Chaotic behavior occurs for many various processes in different natural and engineering objects. In particular, dynamic model of Lorenz [1] describes well-known Rayleigh-Benard convection phenomenon. Investigation of system of Lorenz model equations permitted to reveal so-called *control parameter* whose a specific value leads to chaotic solution of state of this model. Phase trajectories of Lorenz equation system in chaotic regime are characterized strange alternative properties: on the one hand, they diverge (because of positive Lyapunov exponents), on the second hand, they attract to the limited domain of phase space called an *attractor*. Strange attractor of Lorenz demonstrates chaotic behavior of fully deterministic system of nonlinear equations [2]. At the same time, Lorenz attractor has a specific geometrical structure and can be characterized by means of fractional fractal dimension. Thus, analysis of Lorenz attractors for different values of its control parameter gives a possibility to develop high sensible measurement method for recognition of pathological speech signals. In this connection one of the aims of this report is development of analysis of Lorenz attractor based on proposed nonlinear decomposition into matrix series [3], [4]. This analysis permits to estimate the values of characteristic parameters (including control one) of Lorenz attractors and predict their evolution in time. Using results of this quantitative analysis it is proposed an approach to distinguishing and recognition of pathological speech signals from normal ones.

II. THE MATRIX DECOMPOSITION FOR OPERATORS OF NONLINEAR DYNAMICAL SYSTEM INTO STATE-SPACE

With point of view of behavior analysis the continuous NDS is described in state-space by the relations [5], [6]:

$$\dot{u}^{(m)}(t) = \vec{f}_1\left(u^{(m)}(t), x(t)\right), \qquad (1a)$$

$$y(t) = f_2\left(u^{(m)}(t), x(t)\right), \qquad (1b)$$

(where $\vec{f}_1(\cdot)$ is a nonlinear vector function, $f_2(\cdot)$ is nonlinear scalar function, $u^{(m)}(t)$ is a state-space vector belonging the state-space U, t denotes by a time, m is a dimension of U, $x(t)$ and $y(t)$ are input and output signals respectively). In general, we suppose that $x(t) \neq 0$, i.e. we consider the NDS with a nonzero input signal. We study behavior of the solution for the relation (1) near to a specific standard state $u^*(t)$ being considered as an undisturbed one permanently disturbed by external actions or internal fluctuations on a value $\vec{v} = \vec{v}(t)$ [5]. For this NDS we will linearize the function $\vec{f}_1(\cdot)$ near the state $\vec{u}^* = \vec{u}^{*(m)}(t)$. In this case we have to use the matrix *nonlinear decomposition*

proposed in [3] for expansion in matrix series of the vector function $\vec{f}_1(\cdot)$ into state-space. According to [3] a change of vector-function into state-space can be decomposed into matrix series of the kind:

$$\Delta\vec{f}(\vec{v},\vec{u}^{*(m)},x(t)) = \vec{f}_1(\vec{u}^{*(m)}+\vec{v},x(t)) - \vec{f}_1(\vec{u}^{*(m)},x(t)) =$$

$$= L_m^{(1)}\vec{v} + \frac{1}{2!}L^{(2)}_{m\times m^2}(\vec{v}\otimes\vec{v}) + \frac{1}{3!}L^{(3)}_{m\times m^3}(\vec{v}\otimes\vec{v}\otimes\vec{v}) + \cdots, \quad (2)$$

where

$$L_m^{(1)} = L^{(1)}_{m\times m} = \left(\frac{\partial}{\partial\vec{v}^T}\otimes\vec{f}_1\right)_{\vec{0}^T} =$$

$$\left[\left(\frac{\partial}{\partial v_1}\vec{f}_1\right)_0 \cdots \left(\frac{\partial}{\partial v_m}\vec{f}_1\right)_0\right],$$

$$\vec{f}_1 = \begin{bmatrix} f_{11} \\ \vdots \\ f_{m1} \end{bmatrix},$$

$$L^{(2)}_{m\times m^2} = \left(\frac{\partial}{\partial\vec{v}^T}\otimes\left(\frac{\partial}{\partial\vec{v}^T}\otimes\vec{f}_1\right)\right)_{\vec{0}^T}, \ldots,$$

$$L^{(3)}_{m\times m^3} = \left(\frac{\partial}{\partial\vec{v}^T}\otimes\left(\frac{\partial}{\partial\vec{v}^T}\left(\frac{\partial}{\partial\vec{v}^T}\otimes\vec{f}_1\right)\right)\right)_{\vec{0}^T},$$

$\otimes$ denotes by the symbol of the Kronecker matrix product [3]. As a result, instead of $\vec{u}^*$ there is a new solution $\vec{u}^{(m)} = \vec{u}^{*(m)} + \vec{v}$, $\|\vec{v}\|/\|\vec{u}^{*(m)}\| << 1$. In view of this one rewrites the relation (2) as follows :

$$\vec{f}_1(\vec{u}^{(m)},x(t)) = \vec{f}_1(\vec{u}^{*(m)},x(t)) + L_m^{(1)}(\vec{u}^{(m)}-\vec{u}^{*(m)}) +$$

$$+\frac{1}{2!}L^{(2)}_{m\times m^2}(\vec{u}^{(m)}-\vec{u}^{*(m)})\otimes(\vec{u}^{(m)}-\vec{u}^{*(m)}) +$$

$$+\frac{1}{3!}L^{(3)}_{m\times m^3}(\vec{u}^{(m)}-\vec{u}^{*(m)})\otimes(\vec{u}^{(m)}-\vec{u}^{*(m)})\otimes(\vec{u}^{(m)}-\vec{u}^{*(m)}) + \ldots \quad (3)$$

We also suppose a zero state $\vec{u}^{*(m)} = \vec{0}$ for the NDS under investigation. Taking into account this condition the decomposition (3) becomes :

$$\vec{f}_1(\vec{u}^{(m)},x(t)) = \vec{f}_1(\vec{0},x(t)) + L_m^{(1)}\vec{u}^{(m)} +$$

$$+\frac{1}{2!}L^{(2)}_{m\times m^2}\vec{u}^{(m)}\otimes\vec{u}^{(m)} +$$

$$+\frac{1}{3!}L^{(3)}_{m\times m^3}\vec{u}^{(m)}\otimes\vec{u}^{(m)}\otimes\vec{u}^{(m)} + \ldots \quad (4)$$

III. Results of numerical analysis of Lorenz attractor based on matrix decomposition

In [1] Lorenz investigated convective movement of liquid by means of numerical solutions of respective differential equations. Bénard experiment considered by Lorenz is such that a horizontal liquid layer of infinite length in the gravity field (with a positive coefficient of volume extension) is warming from below [2] , [7].

Lorenz attractor of continuous complex NDS can be written by a system of three ordinary differential equations [1], [2] :

$$\begin{cases} \dot{u}_1 = au_2 - au_1; \\ \dot{u}_2 = -u_1\cdot u_3 + cu_1 - u_2; \\ \dot{u}_3 = u_1\cdot u_2 - bu_3, \end{cases} \quad (5)$$

where a and b are dimensionless constants to characterize the system (for example, a = 10 and b =8/3), c is an external control parameter proportional to ΔT [2] : c = Rà/Rà$_c$ ~ ΔT, besides, Rà is a Rayleigh's number [7] and Ra$_c$ is its critical value [2]. The variable u_1 is proportional to the velocity of the circulating liquid, u_2 characterizes a difference of temperatures between the ascending and descending flows of liquid, u_3 is proportional the deviation of temperature profile from equilibrium value.

Because in the general case the Lorenz's model is nonintegrable, its solutions can be found by means of numerical methods if three parameters a , b and c are fixed. The parameter c (connected directly with the Rayleigh number Ra in the Rayleigh-Bénard experiment, i.e. with the temperature difference) is a *bifurcation or control* parameter [1], [2], [5], [6], [7]. It has been programmed on the basis of Java the numerical integration of Lorenz system of three ordinary nonlinear differential equations (5) with the following parameters:

a = 10;
b = 2.66;
c= 24.27.

This program also reproduces geometric locus of Lorenz attractor in the projection on two-dimensional state-subspace (see Fig.1). The domains (marked by different colors) correspond various regimes of movement of the point on the Lorenz's attractor: those regions, where acceleration of point movement is positive, are drawn by the rose-coloured, while regions corresponding to the negative acceleration of the point are marked by the yellow or green colors (besides, green color points to the large deceleration).

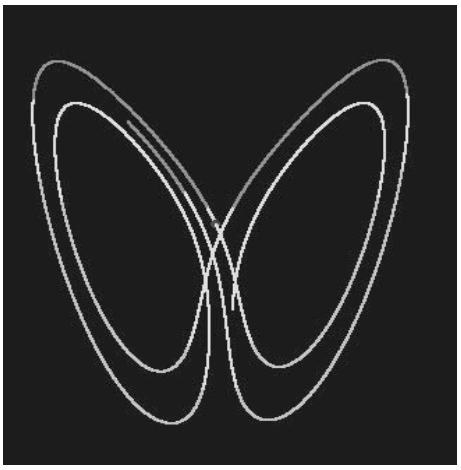


Fig.1 Geometric locus of Lorenz attractor on the plane and matrix analysis its current state

Using the matrix notation the system of equations (5) can be represented by means of the following vector functions:

$$\dot{\vec{u}} = \begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \end{bmatrix}; \ \vec{f}(\vec{u}, x(t), \vec{u}_0) = \begin{bmatrix} au_2 - au_1 \\ -u_1 \cdot u_3 + cu_1 - u_2 \\ u_1 \cdot u_2 - bu_3 \end{bmatrix}. \quad (6)$$

According to Sect. II we shall study how the vector function $\vec{f}$ depends upon the considerable variable

$$\vec{u} = \vec{u}^* + \vec{v}. \quad (7)$$

Taking into account (7) we can find the change of the vector function (6) as follows :

$$\Delta\vec{f}(\vec{v}, \vec{u}^*) = \vec{f}(\vec{u}^* + \vec{v}) - \vec{f}(\vec{u}^*) =$$

$$= \begin{bmatrix} av_2 - av_1 \\ -u_1^* v_3 - v_1 u_3^* - v_1 v_3 + cv_1 - v_2 \\ u_1^* v_2 + v_1 u_2^* + v_1 v_2 - bv_3 \end{bmatrix}. \quad (8)$$

Applying the matrix decomposition (2) to (8) we can evaluate the following terms of matrix series :

$$L_{3x3}^{(1)} \vec{v} = \left[\begin{array}{c|c|c} -a & a & 0 \\ \hline -u_3^* + c & -1 & -u_1^* \\ \hline u_2^* & u_1^* & -b \end{array}\right] \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} =$$

$$= \begin{bmatrix} av_2 - av_1 \\ -v_1 u_3^* + cv_1 - v_2 - u_1^* v_3 \\ v_1 u_2^* + u_1^* v_2 - bv_3 \end{bmatrix}. \quad (9a)$$

$$L_{3x9}^{(2)}(\vec{v} \otimes \vec{v}) = \left[\begin{array}{c|c|c|c|c|c|c|c|c} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ \hline 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \text{x}$$

$$\text{x} \begin{bmatrix} v_1^2 \\ v_1 v_2 \\ v_1 v_3 \\ v_2 v_1 \\ v_2^2 \\ v_2 v_3 \\ v_3 v_1 \\ v_3 v_2 \\ v_3^2 \end{bmatrix} = \left[\begin{array}{c} 0 \\ \hline -2v_1 v_3 \\ \hline 2v_1 v_2 \end{array}\right]. \quad (9b)$$

By substituing (9a) and (9b) in (2) it is not difficult to see that vector function (8) can be approximated by only linear (9a) and quadratic (9b) terms :

$$\Delta\vec{f}(\vec{v}, \vec{u}^*) = L_{3x3}^{(1)} \vec{v} + \frac{1}{2!} = \begin{bmatrix} av_2 - av_1 \\ -v_1 u_3^* + cv_1 - v_2 - u_1^* v_3 \\ v_1 u_2^* + u_1^* v_2 - bv_3 \end{bmatrix} +$$

$$+ \frac{1}{2} \left[\begin{array}{c} 0 \\ \hline -2v_1 v_3 \\ \hline 2v_1 v_2 \end{array}\right]. \quad (10)$$

It has been mentioned above, Fig. 1 illustrates the numerical analysis of a Lorenz attractor on the basis of matrix decomposition in accord with (9a,b), (10). Because the values of the first and second order derivatives can be calculated by means of numerical methods (for example, based on Runge-Kytta method) we can estimate $\Delta f^{est}(\vec{v}, \vec{u}^*)$ from a computational experiment. In result, as it follows from (10), we can estimate the values of parameters of Lorenz's attractor:

$$a = \frac{\Delta f_1^{est}}{v_2 - v_1}; \quad (11a)$$

$$c = \frac{\Delta f_2^{est} + v_1 u_3^* + v_2 + (u_1^* + v_1)\cdot v_3}{v_1}; \quad (11b)$$

$$b = -\frac{\Delta f_3^{est} - v_1 u_2^* - (u_1^* + v_1)\cdot v_2}{v_3}. \quad (11c)$$

Formulas (11a)-(11c) solve the task of identifying the current dynamical state of a Lorenz attractor; in particular, the relation (11b) determines the value of control parameter permitting to reveal chaotic regimes in Lorenz NDS functioning.

IV. Very accurate measurement based on comparing model and reconstructed chaotic attractor

Let $\{s_n \mid n=0,1,2,\ldots,N-1\}$ is a known speech phoneme signal obtained from a healthy person under investigation (i.e. $\{s_n\}$ is a personal phoneme standard), n is a discrete time, N is a duration of signal. Let $\{x_n \mid n=0,1,2,\ldots,N-1\}$ is a measured signal obtained from the same person during a period of medical observation of this person. When the "true" samples of signal are known, the fitted values can be compared with them by defining an error measure e as follows :

$$e = \frac{1}{N}\sum_{n=0}^{N-1}\frac{[x_n - s_n]^2}{s_n},$$

where s_n is a member of the "true" samples of known signal. We also examine the sensitivity of the fit to noise added to the data.

The main idea of this method is an estimation of the measure of proximity e through changing structure of chaotic attractor under observation (in particular, Lorenz attractor). First of all, we choose a chaotic attractor as Lorenz attractor with the control parameter $c = 24.27$. Such value of control parameter, as it has been mentioned above, corresponds to a chaotic regime of Lorenz attractor (see Fig.1). Therefore, we can store the phase portrait of Lorenz attractor with $c = 24.27$ in the memory as a standard and then compare it with a reconstructed Lorenz attractor. The reconstructed Lorenz attractor is built on the basis of numerical integration of Lorenz system (5) by means of respective program tool with the value of control parameter c equal to :

$$c = 24.27 + e .$$

Taking into account the highest sensibility of chaotic attractor from its control parameter we can expect a variation of the standard structure even if $|e| \ll c$.
Really, computer experiments with usage of the mentioned Java program based on Runge-Kutta scheme give us the following results :

The total number of points for solution: 100000;
The initial point $(u_1, u_2, u_3) = (1,1,1)$;
The sampling value in time : $h = 0.0001$;
The parameters of Lorenz system :
$a = 10.00$;
$b = 2.66$;
$c = 24.27$.
The local error : $\leq 8 * 10^{-6}$.

Thus, even a measured signal x_n is very similar to the standard one s_n for the same person then the proposed approach permit us to reveal this error through the variation of the phase portrait of standard chaotic attractor.

References

[1] E.N. Lorenz, "Deterministic Nonperiodic Flow", *Journal of Atmospheric Sciences*, vol. 20, pp.130-141, 1963.
[2] P. Berge, Y.Pomeau and C. Vidal, *L'ordre dans le Chaos: Vers une Approche Deterministe de la Turbulence.* Paris : Hermann, 1988.
[3] A.M.Krot, "Matrix decompositions of vector functions and shift operators on the trajectories of a nonlinear dynamical system", *Nonlinear Phenomena in Complex Systems,* vol. 4, no 2, pp. 106-115, 2001.
[4] A.M.Krot, "Analysis of attractors of complex dynamical system and output signals based on matrix decomposition", *Proc. of 2nd Intern. Workshop "Models and Analysis of Vocal Emissions for Biomedical Applications",* Firenze, Italy, pp. 77-79, September 13-15, 2001.
[5] G.Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuation*, New York etc.: John Willey & Sons, 1977.
[6] H. Haken, *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices,* Berlin etc.: Springer-Verlag ,1983.
[7] L.Landau and E.M. Lifshitz, *Hydrodynamics,* Moscow: Nauka, 1986 (in Russian).

3rd International Workshop MAVEBA 2003, 115-117

BLIND SIGNAL SEPARATION OF VOCAL SIGNALS TAKEN IN NOISY ENVIRONMENT

Ajay Somkuwar and R. P. Singh
Department of Electronics and Computer Science, Maulana Azad National Institute of Technology, (Formerly MACT), BHOPAL -462007

ABSTRACT: The separation of independent sources from mixed observed data is a fundamental and challenging signal processing problem. A method for directly extracting clean speech features from noisy speech is implemented. This process is based on independent component analysis (ICA) and a new feature analysis technique to reduce the computational complexity of the frequency-domain ICA. For noisy speech signals recorded in real environments, this method yielded consider-able performance improvement. Thus the process for extracting clean speech features can be performed without recovering the actual source signal.

I. INTRODUCTION

Noise robustness is a very important issue in the field of automatic speech recognition. Microphones have been used to achieve noise robust-ness, and blind source separation has been implemented to enhance the noisy speech signal. For the speech recognition process, however, only clean

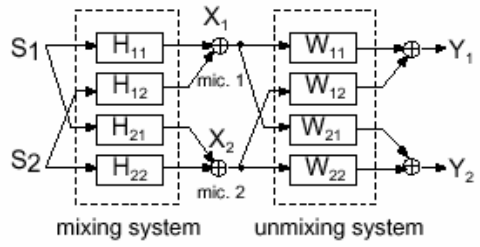


Figure 1: BSS system configuration.

speech features are required. Therefore, instead of de-noising the noisy speech signal in the preprocessing step, it is computationally more efficient to directly extract the clean speech features from noisy speech. The blind signal separation (BSS, fig.1) is an approach to estimate original source signals using only the information of the mixed signals observed in each input channel. This technique is applicable to the realization of noise robust speech recognition and high quality hands-free telecommunication systems.

It may also become a cue for auditory scene analysis. In many practical situations, one or more desired signals need to be recovered from the mixtures only. A typical example is vocal signal or speech signal observations made in an acoustic environment in the presence of background noise. Other examples include Biomedical signals, sonar applications and cross talk in data transmission. The vocal signal separation problem is sometimes referred to as the cocktail party problem. When several people in the same room are conversing at the same time, it is remarkable that a person is able to choose to concentrate on one of the speakers and listen to his or her speech flow unrestrained. A signal separation pre-process would be desirable in such circumstances. The terminology 'blind source separation problem' has been coined by [1], they have done their work on adaptive blind signal processing .and blind source separation technique based on second order statistics .The possibility of noise corrupted sources raises the issue of robustness. A statistical procedure is called robust if it still works well reasonably well when model the model assumptions from which it is designed is more or less violated .In this respect it is of interest to consider the independent component analysis (ICA) introduced by [2]. The separation process is based on ICA. In which Three signals are linearly separated from three mixed speech microphone recordings. The Technique given by [3] have been applied to calculate ICA directly to feature level. A "small-band" approach is implemented to average out fast Fourier transform (FFT) points in a frequency range and apply ICA directly to feature levels. To remove the mixed noise, this requires only one un-mixing network for each small band. This technique shows that the method yielded considerable performance improvement for weak signals also. Mixture of signals which is in the analog form is converted into digital form by using software Cool-Edit (Syntrillium™ software USA) for further processing This technique is implemented in Matlab™ environment.

II. METHODOLOGY

The signals recorded by M microphones are given by

$$x_i(n) = \sum_{i=1}^{N} \sum_{p=1}^{P} h_{ij}(p) s_i(n - p + 1)$$

j =1……M

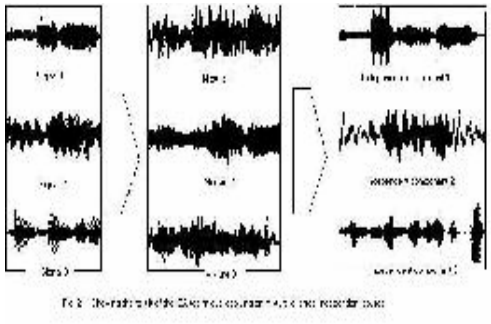
Where s_i is the source signal from a source i, x_j is the received signal by a microphone j, and h ji is a P point impulse response from source i to microphone j . In this paper, we consider a three-input, three-output convolutive BSS problem, i.e., N = M = 3.the convoluted mixture can be obtained according to figure1, where two signal can be mixtured by microphone arrey. The frequency domain approach t convolutive mixtures is to transform the problem into an instantaneous BSS problem in the frequency domain. The most basic and necessary preprocessing of signal is centering, i.e. subtract its mean vector so as to make a zero-mean variable. Another useful preprocessing strategy in ICA is to whiten the observed variables. This means that before the application of the ICA algorithm (and after centering), we transform the observed vector linearly so that we obtain a new vector which is white, i.e. its components are uncorrelated and their variances equal unity. In other words, the covariance matrix of equals the identity matrix. Spectral analysis, the k^{th} band energy y (k) can be expressed as

$$y(k) = \sum_{n=Fk}^{lk} \left[\mathrm{Re}(\bar{x}(n))2 + \mathrm{Im}(\bar{x}(n))2 \right]$$

Where k = 1 .. . . .K and X (n) is the value of the n^{th} FFT point. F_k and l_k denote the index of the first and last point of the kth band and K denotes the number of bands respectively. Use of this method can improve the recognition performance in noisy environments by smoothing the spectrum components. Additionally, fever-unmixing network are required in frequency domain approach. This method can improve recognition performance in noisy environments by smoothing spectrum components. Additionally results in much less number of unmixing networks.

III. Result

The result is that the individual signals could be recovered from the mixture of signals and then the problem of receiving the individual signals from the mixture of signals that is the cock tail party problem is solved.



IV. Discussion

Independent component analysis aims at extracting unknown components from multivariate data using only the assumption that the unknown factors are mutually independent. Since the introduction of ICA concepts in the early 80s in the context of neural networks and array signal processing, many new successful algorithms have been proposed that are now well-established methods. Since then, diverse ICA applications in telecommunications, biomedical data analysis, feature extraction, speech separation, time-series analysis and data mining have been reported. Biomedicine is one important research area where the above techniques has proven their success. The use of ICA in electroencephalography, magnetoencephalography or in the extraction of the fetal electrocardiogram (ECG) from maternal recordings are some examples of it. In the ECG, ICA also has been applied to the separation of breathing artifacts, and other disturbances. The above technique can be used in Voice Extraction by on-line Signal Separation.

V. Conclusion

By applying the fast ICA technique in frequency domain for speech signal more robust performance is obtainable. Thus the process for extracting clean speech features can be performed without recovering the actual source signal. Also, the frequency-domain approach is implemented with less number of un-mixing networks for noisy speech signal recognition.

References

[1] Amari, S., Cichocki, A., and Yang, H., ‘A New Learning Algorithm for Blind Signal Separation’, *Advances in Neural Information Processing Systems,* 1996, 8, pp.757-763.

[2] P. Comon, Independent component analysis - a new concept? , *Signal Processing*, 36:287-314, 1994.

[3] Hyung-Min Park, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee, Subband-Based Blind Signal Separation for Noisy Speech Recognition, Electronics letter, (1991)

[4] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3): 626-634, 1999.

Plenary lecture

3rd International Workshop MAVEBA 2003, 121-135

A GENERALIZED CONCEPT OF PROSODY

Osamu Fujimura
The Ohio State University
Fujimura.1@osu.edu

Abstract: **According to the C/D model, the base function for each utterance comprises a skeletal structure represented by a syllable-boundary pulse train and melodic control functions that are linked to individual syllables. The melody includes vocalic, tonal, and some other control variables. The concept of prosody may be generalized to include all base function aspects according to this concept. Melodic time functions are represented by phonetic status contours, dimension by dimension, *i. e.* syllable-based pseudo-step functions with occasional interpolations for phonological underspecification. The quasi-stationary target values for each syllabic segment are enhanced or reduced according to the syllable magnitude. Consonantal perturbation functions represented by elemental gestures are superimposed onto these control variables of the base function and their ballistic movement patterns as impulse responses to each syllabic excitation pulse have amplitudes according to the syllable magnitude. Jaw opening cotains a prosodic component that directly reflects the syllable magnitude, which determines an abstract syllable duration. Some examples of mandibular, vocalic and tonal variables associated with durational variation are discussed with empirical data, referring to two recent PhD dissertations by Caroline Menezes and Patrizia Bonaventura.**
Keywords : **C/D Model, phonetics, prosody, syllable, boundary.**

I. INTRODUCTION

Prosody is traditionally considered as suprasegmental characteristics of speech signals that are outside the phonemic segmental description [Lehiste 1970]. Phenomena such as voice fundamental frequency (F0) variation (often called intonation), segmental durations, and acoustically silent periods (pauses) are typical variables that are treated as prosodic characteristics of speech signals. Recent research in speech production points out the inaccuracy of this traditional concept of prosody. Experimental studies on articulatory movement patterns as well as acoustic analyses have revealed, for example, significant variability of vocal tract filter functions or corresponding formant frequency patterns, due solely to prosodic variables such as syllable prominence. Such prosodic variables are controlled as a function of various communicative meaning or expressiveness, including focus or contrastive emphasis, in realistic conversational speech [Gu *et al*. 2003, Fujimura 2000a, Erickson 1998, Maekawa 1996, Fujimura 1990, Laver 1980]. In terms of the voice source signal, in addition to the fundamental frequency and intensity, other quasi-stationary spectral properties have been discussed as variables representing voice quality in relation to prosodic control [Fant *et al*.1985, Pierrehumbert 1989, Fant *et al*. 2000]. Besides quasi-stationary speech parameters, temporal fluctuation of the source signal which varies within an utterance, as well as variation among utterances depending on the type of phonation has also been studied (see Estill *et al*. [1996] and Kawahara *et al*. [2001]).

The C/D model [Fujimura 1992, 2000a, 2002] takes a new view, defining a generalized concept of prosody to be represented by the base function as a whole, separated from the temporally local functions representing consonantal perturbation [Öhman 1967]. Fig. 1 shows a block diagram of this model. In this depiction of speech organization patterns, phonemic segments no longer play any role; the «segments» as the basic concatenative units of speech organization are syllables. Each syllable comprises phonological features and corresponding gestural manifestations in its phonetic implementation. Syllable boundaries eventually become obscure as the phonetic signals are implemented [Leben 1999], but in acoustic signals, there are apparent discontinuities frequntly, due to the nonlinearity of the mapping from articulatory movement variables to acoustic signal parameters. Such discontinuities observed in the acoustic signals correspond to the traditional acoustic segmental boundaries. The base function (see below) has its skeletal structure and melodic variables. The latter, variable by variable for different physiological control dimensions separately, are temporally linked to each syllable of the skeletal structure..

II. Base function: Skeleton *vs.* Melody

According to the C/D model, a speech utterance is phonetically represented by its base function and superimposed consonantal perturbation. The base function has a skeleton and melody. The skeleton of the base function embodies the rhythmic organization of the utterance, and it is represented by a syllable-boundary pulse train. Each pulse, representing either a syllable or a boundary, has its own magnitude. A unique characteristic of the C/D model is to associate the syllable or boundary magnitude directly to the temporal property of the concatenative unit, *i. e.,* an abstract duration of the syllable or boundary. Some temporal modulation of this basic organization of articulatory movement patterns is added, as seen in phrase-final elongation phenomena, exhibiting interaction between syllables and boundaries.

The syllable, as an abstract concatenative unit, has a target value to represent the phonetic status in each dimension of the base function, usually, but not always, as a stationary (time-free) scalar value. Typically, the concatenated string of syllables, with intervening boundaries, forms, in each of its dimensions, a pseudo-step function of time called a phonetic status contour, with some (abstractly) inserted function (such as interpolation) for the intervening boundary. Such effects of boundaries may be observed for a syllable string as a whole simultaneously (*e. g*. phrase-final elongation or pause), or only in some phonetic variables of the syllable string (*e. g*. tonal features for a yes-no question). Such manipulation of control variables is often observed in tonal control in relation to phonological feature underspecification (see, *e. g*. Shih & Kochanski [2000] and Xu [1999, 2001]). Part of such boundary effects was traditionally discussed as Sandhi rules, in terms of discrete alteration of phonological features as contextual effects.[i]

A boundary generally has its dynamic gestural manifestation in the temporal vicinity of the occurrence of the boundary pulse. A boundary may also manipulate the time scale, common for all control dimensions, by a continuous temporal modulation function, for example, manifesting a period of silence or a phrase-final elongation of articulatory and/or phonatory gestures [Fujimura 1990].

III. Voice quality including F0 control

Speech production control, according to the source-filter theory [Fant 1960], has two aspects: source and filter. If the source signal is produced by vocal fold vibration, it controls properties of voice quality. Loudness and pitch are the most widely recognized psychoacoustic characteristics of voice. Physically, we often consider the voice fundamental frequency and acoustic pressure signal intensity as primary variables that determine pitch and loudness, respectively, with some interaction both in production and perception. However, there are many other physical correlates of independently controllable voice quality; some are used commonly in conversational speech, for conveying specific communicative meanings. Stress, as a measure of phonetic control of each syllable, is an abstract concept, primarily related to the respiratory effort in the concrete process of speech production [Ladefoged 2001].

Stress has many physical correlates. A higher subglottal pressure typically raises voice pitch and loudness together and, unless specific voice quality control is executed, a higher intensity of the voice signal is accompanied by a higher F0 and *vice versa*. It should be noted, however, that this default correlation could be reversed on purpose, even in routinely observed conversational phrases. For example, as illustrated by a CD diagram in Fig. 2, a sarcastic expression of a sentence 'That's wonderful!' may exhibit a deliberately lowered F0 with stress attached to the word 'wonderful', affecting its main-stressed first syllable. In such an utterance, the default rise of F0 due to an enhanced respiratory effort is overridden by a special suppression of F0 accompanied by an alteration of the voice source spectrum. Such an alternation of voice quality would be observed as a boosting up of the high frequency components of the voice source signal [Pierrehumbert 1989, Fujimura *et al*. 1995]. In this situation, however, the extended syllable duration of the main-stressed syllable of the emphasized word remains to be perceived as a strong indication (prominence) of the stress. Increased syllable duration is another primary manifestation of phonetic stress. In terms of the C/D model, the skeleton of the sentence utterance with an emphasis on 'wonderful' remains with an enlarged syllable triangle (see Fig. 2), enhancing all syllable gestures including, in particular, jaw opening. Voice quality control also affects temporal perturbation of voice periodicity. It is understood generally that voice quality changes considerably according to the expressive style in conversational speech. The sarcastic utterance mentioned above is just one example. Almost any emotion ranging from sorrow to anger or retreat to aggression, as well as delight to frustration, is expressed by the choice of marked voice quality along with a choice of particular linguistic forms.

In Japanese (*e. g*., Tokyo dialect), which does not use stress but pitch accent for lexical distinction, stress pattern is controlled only in phrasal phonetics, in terms of manipulation of syllable magnitudes, for example, for marking focus placed on a word [Fujimura *in press*]. Manifestation of emotion in Japanese also may be reflected in voice quality of the entire utterance. Sometimes, it may be observed most clearly in a particular part of an utterance, such as toward the end of an utterance. It may be associated with a choice of a particular sentential particle, but often the emotion is expressed just by changing the «tone» using a regular sentential particle, such as 'ka' for questioning. Thus, for example, 'Soo desu ka' (Is that so?) can be a question simply verifying the dialogue partner's statement, or an expression of incredulity, or perhaps an expression of indifference, all using syntactically the same question form. The difference in the communicative meaning, which typically is quite obvious to the listener even in a relatively calm conversation with limited F0 variation, can be observed reliably by measuring the voice source spectrum. Maekawa studied the amplitude ratio between the fundamental and second harmonics near the end of the utterance (vowel [a] in the example above) [Maekawa, personal communication]. Erickson [2002] reported articulatory characteristics of sad speech in a recorded telephone conversation in comparison with simulated sad and normal utterances of the same sentences in English.

IV. Prominence effects on articulatory and phonatory gestures

Emphasizing a word in an utterance affects various physical correlates including (spectral) voice quality as discussed above. In addition to the change of the vocal fold vibration pattern, which determines the voice source characteristics, the vocal tract configuration changes, related to the effect of stress on jaw opening and the vowel-proper gesture for the syllable; the F-pattern (formant frequencies) is also systematically affected accordingly. The change in position of the larynx, related to the tonal phonological features, also may affect the F-pattern. Usually, the first formant is the most noticeably affected by prominence control of different kinds. Erickson [2002] reports on the effect of contrastive emphasis due to correction of a word on high and low vowels, in American English sentence utterances. Her data show that jaw opening is increased for high vowels as well as low and mid vowels, while the tongue surface is not elevated for high vowels by emphasis [Erickson 2002], Fig. 3 illustrates such effects of emphasis on tongue body position for the American English tense high front vowel /i/, comparing emphasized *vs*. unemphasized words. In this case, the tongue body position is affected by emphasis mainly in its front-back position.

Presumably, the lingual and labial articulatory gestures that characteristically deviate for the given vowel, relative to a neutral vowel articulation, are enhanced when the syllable magnitude is increased (see also Fujimura [2002]). In the case of low vowels, as in the Pine Street data, the enhancement of the inherent vowel gesture of the tongue cooperate with increased jaw opening in lowering the tongue surface for more prominent syllables. Therefore, jaw opening can be taken directly as an indication (proportional measure) of the syllable magnitude, even though the proportionality coefficients as contributing factors cannot be assessed. In the case of high vowels, the two effects of syllable magnitude enhancement more or less cancel each other between the jaw gesture and the tongue proper gesture relative to mandible position [Fujimura *in press*-a], as we see in Fig. 3. In such a case, the primary effects of prominence we can observe are the enhancement of advancing/retracting gesture (and lip protrusion gesture for a rounded vowel) along with the durational effects (see Fujimura [2000b]).

An interpretation by the C/D model of such an articulatory effect of syllable magnitude is depicted in Fig. 4. This speculative figure illustrates an example sentence of 'That's the most important', uttered in a neutral intonation pattern without particularly emphasizing a word. The tongue body is retracted for the back vowel /o/ in the nominal vocalic gesture (dot-dashed curve), according to the phonological feature specification {back}. This inherent gesture for phonological backness is implemented as a tongue body retracting, as a deviation from the neutral advanced/retracted posture of the vowel articulation. According to the syllable magnitude, which, for this syllable, with nuclear stress, exceeds a reference level (indicated by a horizontal solid line in the next-to-top panel going across syllable triangles), the tongue retraction gesture is shown to deviate more strongly (dashed curve for «retracting») than the nominal gesture (dash-dotted curve).

The mandible lowering (dashed curve, lowest panel) in this figure reflects the syllable magnitude control, which is represented by the height of each syllable triangle in the next-to-top panel of the figure. Note that the syllable triangles are similar to each other with the same (symmetric) angles regardless of their size. This implies, according to the assumption of the C/D model, that the (abstract) syllable duration is proportional

to the (abstract) syllable magnitude. In this figure, phrase-final lengthening is caused by a boundary (the small half triangle in the middle of the next to top panel) between two consecutive syllables, which may be implemented in part as a silent period, if the boundary is large enough in magnitude. For vocalic gestures, the increment of the phonetic implementation in each relevant dimension is assumed to be proportional to the excess of syllable magnitude (see above). For consonantal gestures, shown in the top panel, impulse response functions (IRFs) for elemental gestures are amplified according to the syllable magnitude, since the syllable pulse excites each IRF, as a linear system response.

Menezes [2003] analyzed formant frequency variation in the low vowel of /aJ/ in 'five' and 'nine' in different phrasal positions of street addresses such as '599 Pine Street', when one of the digit is corrected repeatedly (Blue Pine data, see Erickson [1998] and Erickson *et al*. [1998]). This dissertation demonstrated by detailed data analyses including some perceptual evaluation, that jaw opening maneuvering, along with syllable duration, as predicted by the C/D model, is a robust measure of syllable magnitude that is controlled by the correction, confirming previous reports. Fig. 5 illustrates, for one of four speakers, this mandibular effect of contrastive emphasis, which is consistently observed in all speakers. In other words, given the same low vowel, maximum jaw opening during the syllable is a reliable measure of intended emphasis (by correction). This was shown to be consistent with perceived emphasis as well. In addition, Menezes & Honda [2002] showed that raised F0 was not a reliable measure of emphasis. Fig. 6 demonstrates some variability of F0 patterns with respect to the effect of contrastive emphasis due to focus in correcting utterances. Menezes *et al*. [2002] also found that listeners did not judge emphasis consistently based on pitch patterns.

V. Temporal modulation and phonetic phrasing

As mentioned above, according to the C/D model, the syllable magnitude directly relates the magnitude of articulatory gestures to the temporal span of each syllable, even though the observed jaw opening and acoustic syllable duration are both affected also by other concomitant factors. Maximum jaw opening is determined, according to this theoretical prediction, by the abstract syllable magnitude in the prosodically controlled component of this articulatory gesture, while the inherent vowel gesture, according, in particular, to the phonological vocalic feature high *vs*. low, also contributes to the mandibular position. On the other hand, the syllable magnitude dictates the underlying syllable duration, while the observed acoustic syllable duration, however it may be defined, is affected also by boundary effects, which cause, notably, what is called phrase-final lengthening [Lehiste 1980].

Even the movement of the crucial articulator for the given demisyllable (combination of the consonants and the vowel for the initial or final half of the acoustic syllable pattern, see Fujimura [1976, 1979]) shows somewhat affected temporal characteristics depending on the adjacent boundary. Patrizia Bonaventura, analyzing the iceberg patterns of the crucial articulators in the Pine Street data, addresses this issue in her forthcoming Ph.D. dissertation. Menezes [2003] also discussed a tentative interpretation of phrasing strategies used by different speakers in the Blue Pine data, as manifestation of (repeated) corrections. Mitchell in his MA thesis [2000] (also see Mitchell *et al*. [2000]) provided an early discussion of such issues.

By analyzing a Red Pine database more recently acquired by Erickson, Bonaventura examined the iceberg movement patterns of the consonantally crucial articulator in /faJv/ and /naJn/ in detail. In the time function display of the vertical position of a selected flesh point of the crucial articulator, the slope (speed of movement) when the pellet position passes a fixed vertical position (iceberg threshold, see Fujimura [1986]) was examined in large numbers of utterances by three speakers. In Fig. 7, the time origin for each curve was shifted to make the group of curves to pass a fixed iceberg threshold coincide in time (see Fujimura [1986]). The syllable in this example is /naJn/. The slope of the demisyllabic movement of the tongue blade (about 1 cm behind the tip) varies slightly depending on the extent of vertical movement for each demisyllable. Fig. 7 compares the initial and final demisyllables, [na] and [aJn], respectively.

Fig. 8 shows, correspondingly, scatter plots of the speed of tongue tip movement at the iceberg crossing point against the total vertical distance of excursion for each demisyllable. It is seen in this scatter plot that for the initial demisyllables in this set of data, there is a clear linear relation between the iceberg threshold crossing speed and the distance of the demisyllabic movement. That means the larger the distance of descent from consonant to vowel, the faster the speed of (downward) movement. On the other hand, for the final demisyllable, the scatter plot shows a much more dispersed pattern. The correlation varies somewhat from speaker to speaker and depending on other factors, but this tendency is consistently observed in our data.

Depending on whether the (digit) is emphasized or not, by correction, the excursion is larger or smaller, but there are other factors that affect the excursion distance: the separation of the two conditions, emphasized *vs*. not emphasized, is not very clear-cut. In the case of final demisyllabes, the boundary of various magnitudes immediately follows. A boundary of sufficient magnitude may cause deceleration of the movement in the immediately preceding time domain. Since the extent of this deceleration varies depending on the boundary magnitude, and the boundary magnitude varies depending on various factors from utterance to utterance as well as speaker strategies, the movement speed is not as consistent as in initial demisyllables. It seems, however, that emphasis conditions does not determine the speed of demisyllabic movement directly, while it does affect jaw opening consistently as a tendency. In any case, the iceberg analysis provides us, as demonstrated earlier [Fujimura 1986, 1990, 2000a, Mitchell 2000, Menezes *et al*. 2002, Menezes 2003], with a useful tool for evaluating the timing of each demisyllable, and thereby a reliable measure of duration of each syllable. Boundary magnitudes can be evaluated based on the syllable timing and duration, and the pattern of phonetic phrasing for the given intention of prosodic control can be examined.

Phonetic phrasing, in the C/D model, is a numerical phenomenon with continuously variable boundary magnitude, as observed in the syllable-boundary pulse train. In contrast, the phonological phrase structure, as discussed, for example, by Selkirk [1984], is a discrete organization of a linguistic form as its inherent property. The phonological structure, when it is implemented, is continuously modified by the phonetic environment of the utterance. Phonetic phrasing patterns can not be discussed without quantitatively defined boundaries. Considering only acoustically observed silent period as the phrase boundary is a naive concept, but there has been no theory to the author's knowledge that provides a descriptive framework for defining abstract general boundaries with their quantitatively defined strengths

The C/D model proposes such a descriptive framework of phonetic structure. It defines an abstract syllable duration, which can be inferred from observable articulatory signals, at least under favorable phonetic situations, *e. g*., Pine Street data using a very limited phonologic materials uttered in widely varied prosodic conditions. By using such simple materials with respect to vowels and consonants, we can infer, at least approximately, syllable magnitudes of syllables as the function of stress. By using a fixed low vowel, as in 'five', 'nine', and 'Pine', we can assume that jaw opening directly reflects syllable magnitude to first approximation. By calculating syllable duration using the assumption that it is proportional to the syllable magnitude, we can evaluate the gaps between consecutive syllables.

Once we understand the basic nature of the temporal organization of speech using simple lexical materials, we then can proceed to study properties of specific and varied vowels and consonants, according to the theoretical framework of the C/D model. Then using a large variety of syllabic (segmental) as well as prosodic conditions in natural or semi-natural conversational speech, we can test the validity of the theory in terms of consistency. Original assumptions can be revised according to empirical findings and refinement of system parameters, as a process of successive approximation can be obtained. This seems to be the only methodology to discover the abstract organization principles of infinitely complex conditions that determine observable properties of speech signals. We are only at the beginning stage of this ambitious research program. Without such a basic exploration of the newly described phonetic principles, speech will never be understood, and speech technology will never see a true breakthrough.

---

[1] Note that, according to the C/D model, phonetics describes properties of phrases in an utterance, while the lexicon is a phonological system, which must be implemented as phrases in utterances.

REFERENCES

Erickson, D. 1998. Effects of contrastive emphasis on jaw opening. Phonetica 55, 147-69.

Erickson, D. 2002. Articulation of extreme formant patterns for emphasized vowels. Phonetica 59, 134-49.

Erickson, D., Fujimura, O. & Pardo, B. 1998. Articulatory correlates of prosodic control: Emotion and emphasis. *Language & Speech* **41**, 395-413.

Estill, J., Fujimura, O., Sawada, M. & Beechler, K. 1996. Temporal perturbation and voice qualities. In P. J. Davis & N. H. Fletcher (*eds*.), *Vocal Fold Physiology: Controlling Complexity and Chaos*. San Diego: Singular Publishing Group.

Fant, G 1960, 1970[II]. *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G., Liljencrants, J. & Ling, Q. 1985. A four-parameter model of glottal flow. *KTH/STL QPSR* **4**, 1-13.

Fant, G, Kruckenberg, A. & Liljencrants, J. 2000. The source-filter frame of prominence, *Phonetica* **57**, 113-27.

Fujimura, O. 1976. Syllable as concatenated demisyllables and affixes. *J. Acoust. Soc. Am.* **59**, Supplement 1, S55 (abstract).

Fujimura, O.1979. An analysis of English syllables as cores and affixes. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* **32**, 471-476.

Fujimura, O. 1986. Relative invariance of articulatory movements. In J. S. Perkell & D. H. Klatt (*eds*.), *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum.

Fujimura, O. 1990. Methods and goals of speech production research. *Language & Speech* **33**,195-258.

Fujimura, O. 1992. Phonology and phonetics -- A syllable-based model of articulatory organization. *J. Acoust. Soc. Japan (E)* **13**, 39-48.

Fujimura, O. 2000a. The C/D model and prosodic control of articulatory behavior. *Phonetica* **57**, 128-38.

Fujimura, O. 2000b. C/D model prediction of CVC segmental duration for varied syllable prominence. *The Phonetician* **82**, 9-21.

Fujimura, O. 2002. Temporal organization of speech utterance: a C/D model perspective. *Cad.Est.Ling., Campinas* **43**,9-36.

Fujimura, O. *in press*. Stress and tone revisited: Skeletal *vs*. melodic and lexical *vs*. phrasal. In S. Kaji (*ed*.), *Proc. International Symposium on Cross-Linguistic Studies of Tonal Phenomena, Historical Development, Phonetics of Tone, and Descriptive Studies.* Tokyo University of Foreign Studies, ILCAA.

Fujimura, O., Cimino, A. & Sawada, M. 1995. Voice quality control within a sentence: Expressive effects of source spectral envelope change. In O. Fujimura and M. Hirano (eds.), *Vocal Fold Physiology: Voice Quality Control*, pp. 201-15.

Gu Z., Mori, H. & Kasuya, H. 2003. Prosodic variations in disyllabic meaningful words focused with different stress patterns in Mandarin Chinese. *Acoust. Sci. & Tech.* **24**,111-9.

Kawahara, H., Estill, J. & Fujimura, O. 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT, *Second International Workshpo on Models and Analysis of Vocal Emissions for Biomedical Applications, Firenze*.

Ladefoged, P. 2001. *A Course in Phonetics* (4th Edition). Orlando, FL: Harcourt.

Laver, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Leben, W. R. 1999. Weak vowels and vowel sequences in Kwa: Sounds that phonology can't handle. In O. Fujimura, B. Joseph & B. Palek (*eds*.), *Proceedings of LP'98*. Prague: Charles University Press, pp. 717-732.

Lehiste, I. 1970. *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. 1980. Phonetic manifestation of syntactic structure in English. *Annual* Bulletin *of the Research Institute of Logopedics and Phoniatrics (U. Tokyo*)**14**, 1-28.

Maekawa, K. 1996. Effects of focus on duration and vowel formant frequency in Japanese. In Y. Sagisaka, N. Campbell & N. Higuchi (*eds*.), *Computing Prosody*. New York: Springer Verlag, pp. 129-53.

Menezes, C. 2003. *Rhythmic Pattern of American English: An Articulatory and Acoustic Study*. Doctoral dissertation, Dept. Speech and Hearing Science, The Ohio State University.

Menezes, C., Pardo, B., Erickson, D. & Fujimura, O. 2002. Changes in syllable magnitude and timing due to repeated correction. *Speech Communication* **40**, 71-85.

Menezes, C. & Honda, M. 2002. Rhythmic pattern of semi-spontaneous dialogues in General American English. LP2002, Tokyo.

Mitchell, C. J. 2000. *Analysis of Articulatory Movement Patterns according to the Converter/Distributor Model*. Master's thesis, Dept. Speech & Hearing Science, The Ohio State University.

Mitchell, C., Menezes, C., Williams, J. C., Pardo, B., Erickson, D. & Fujimura, O. 2000. Changes in syllable and boundary strengths due to irritation. In R. Cowie, E. Douglas-Cowie & M. Schröder (*eds*.), *Proceedings of ISCA Workshop on Speech and Emotion*. Belfast: Textflow, pp. 98-103.

Öhman, S. E. G. 1967. Numerical model of coarticulation. *J. Acoust. Soc. Am*. **41,** 310-20.

Pierrehumbert, J. B. 1989. A preliminary study of the consequences of intonation for the voice source. *Speech Transmission Laboratory Quarterly Progress and Status Report 4***,** 23-36.

Selkirk, E. O. 1984. Phonology and Syntax: The Relation between Sound and Structure. Cambridge, MA: The MIT Press.

Shih, C. & Kochanski, G. 2000. Chinese tone modeling with STEM-ML. *Proc. 6th ICSLP,* Beijing, Vol. II, 67-70.

Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F0 contours. J. *Phonetics* **27**, 55-105.

Xu, Y. 2001. Sources of tonal variations in connected speech. *J. Chinese Linguistics, Monograph Series* 17, pp. 1-31.

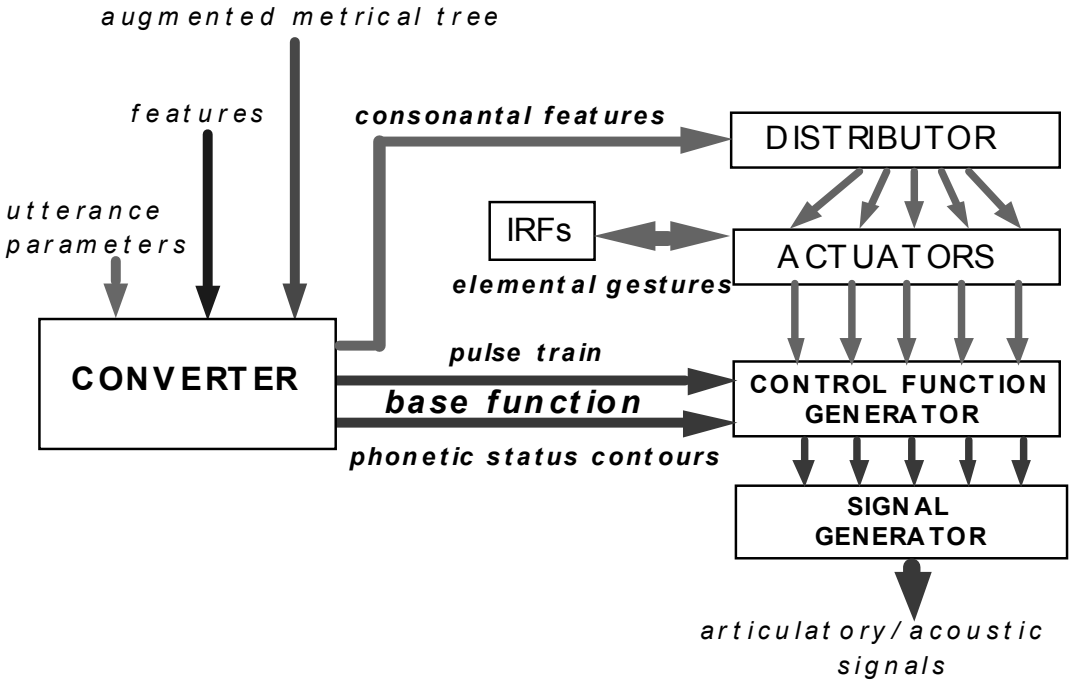
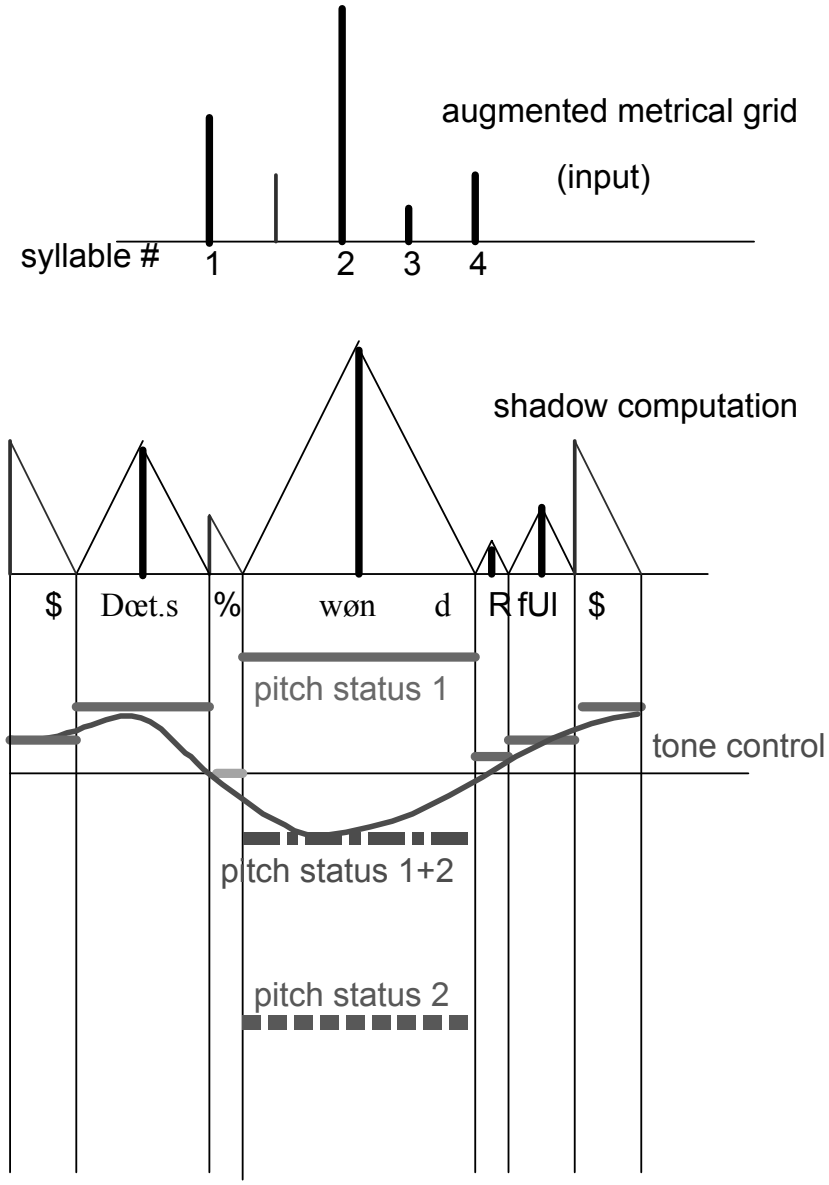


Fig. 1: C/D Model (block diagram)



Pitch Status 1 shows the default pitch levels according to the syllable magnitudes. Pitch status 2 shows the marked pitch control for pitch lowering. The smooth curve is the result of implementing the pitch status step function (a kind of coarticulation).

Fig. 2: Low pitch for stressed syllable in a sarcastic utterance of ‘That’s wonderful’.

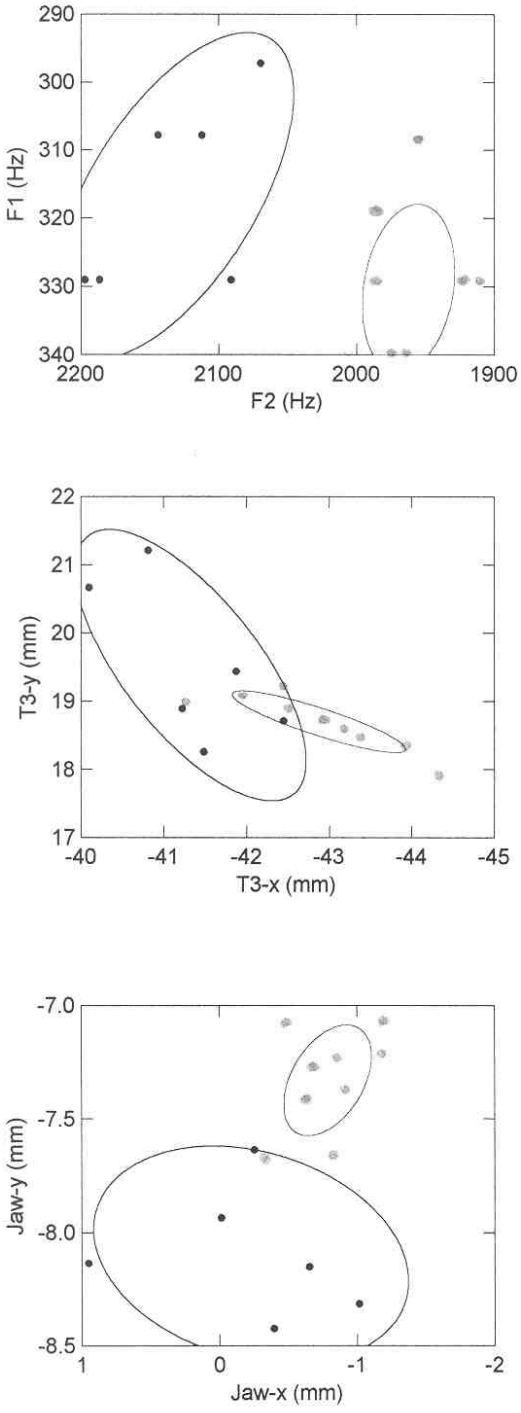


Fig. 3: Emphasized (black) vs. neutral (red) (see Erickson [2002])

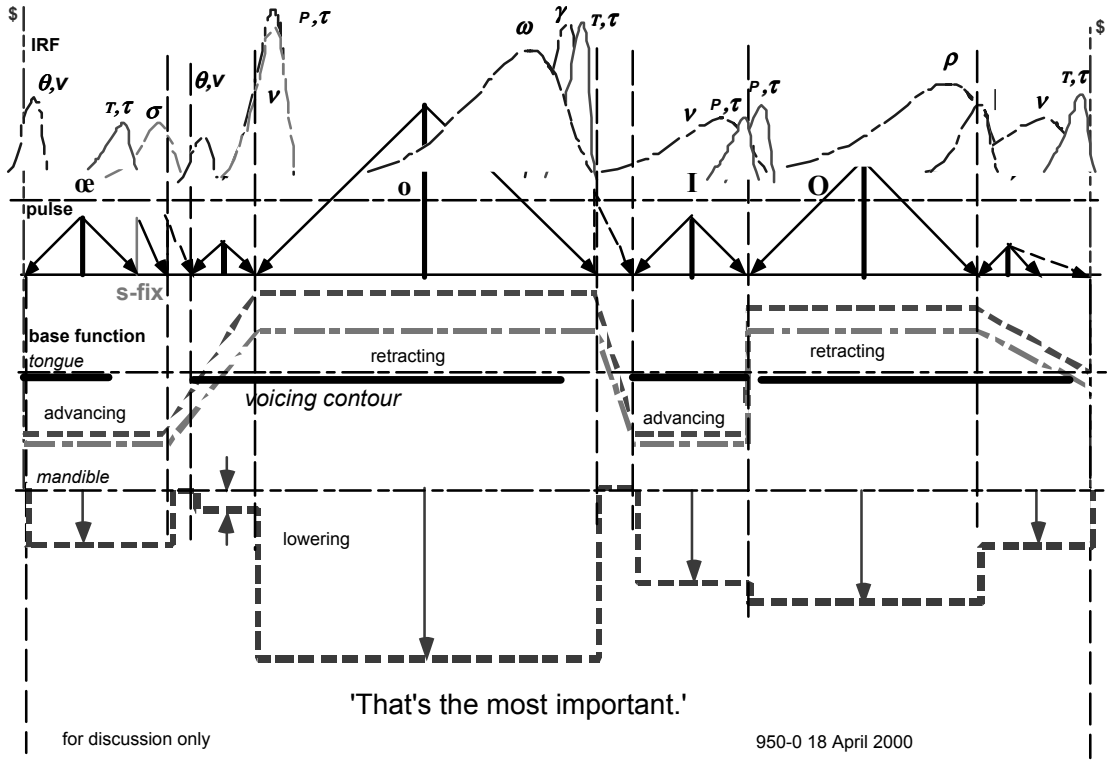


Fig. 4: ‘That’s the most important’ (CD diagram)

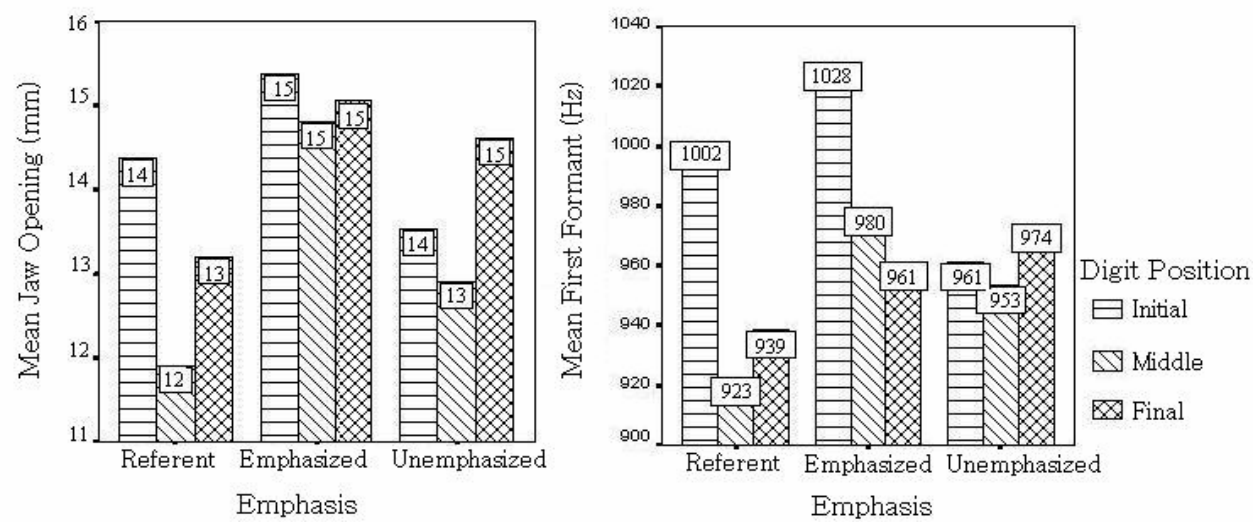


Fig. 5: Mean maximum jaw opening and F1 for different emphasis conditions and digit position in a 3-digit sequence (female speaker [Menezes, personal communication]). All digits in (repeated) correcting utterances have larger jaw opening and F1 than those in reference. For middle digit, the direct effect of emphasis on the corrected digit is particularly strong in this speaker.

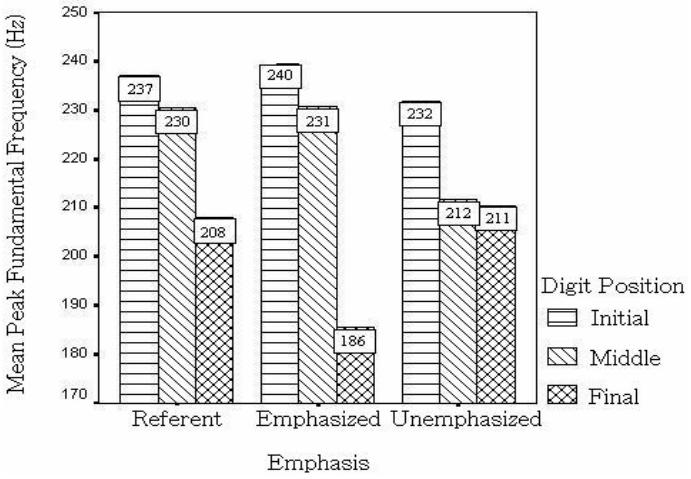
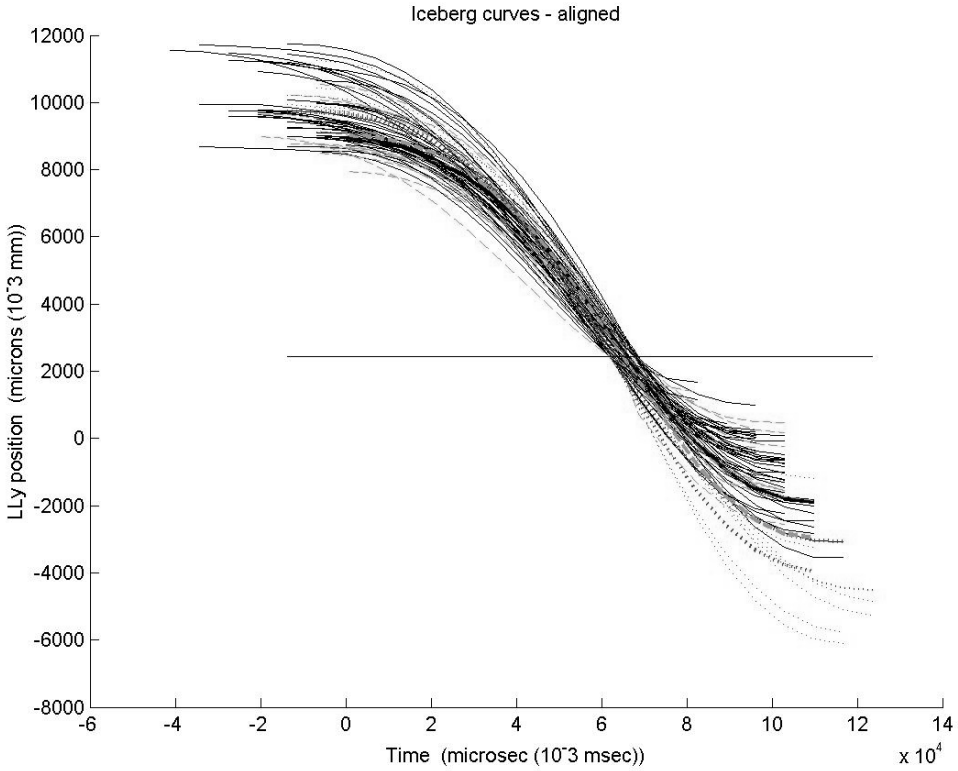


Fig. 6: Mean peak F0 for different emphasis conditions and positions in 3-digit sequence (same speaker/session as in Fig. 5 [Menezes, personal communication]). Note that for the final digit, in this speaker, F0 is considerably lowered when the digit is emphasized by (repeated) correction (the right-most bar of the middle 3-bar group) compared with the neutral utterances (in the left group) or in the unemphasized digits in the correcting utterances.

70 Iceberg curves – initial demisyllable
(Tot. 72, 2 missed, because all values above threshold)
Speaker 3, word 'nine'

LEGEND: Reference curves = black and solid;
Non-emphasized curves = green and dashed;
Emphasized ones = red and dotted



The set of curves includes emphasized and unemphasized digit in correcting utterances of three-digit street addresses in different intra-phrase positions of the digit strings, spoken by the same speaker.

Fig. 7 (a): C-V movements (time functions) for the initial demisyllable of 'five', temporally aligned by the iceberg threshold crossing time.

72 Iceberg curves – final demisyllable
Speaker 3, word 'nine'

LEGEND: Reference curves = black and solid;
Non-emphasized curves = green and dashed;
Emphasized ones = red and dotted

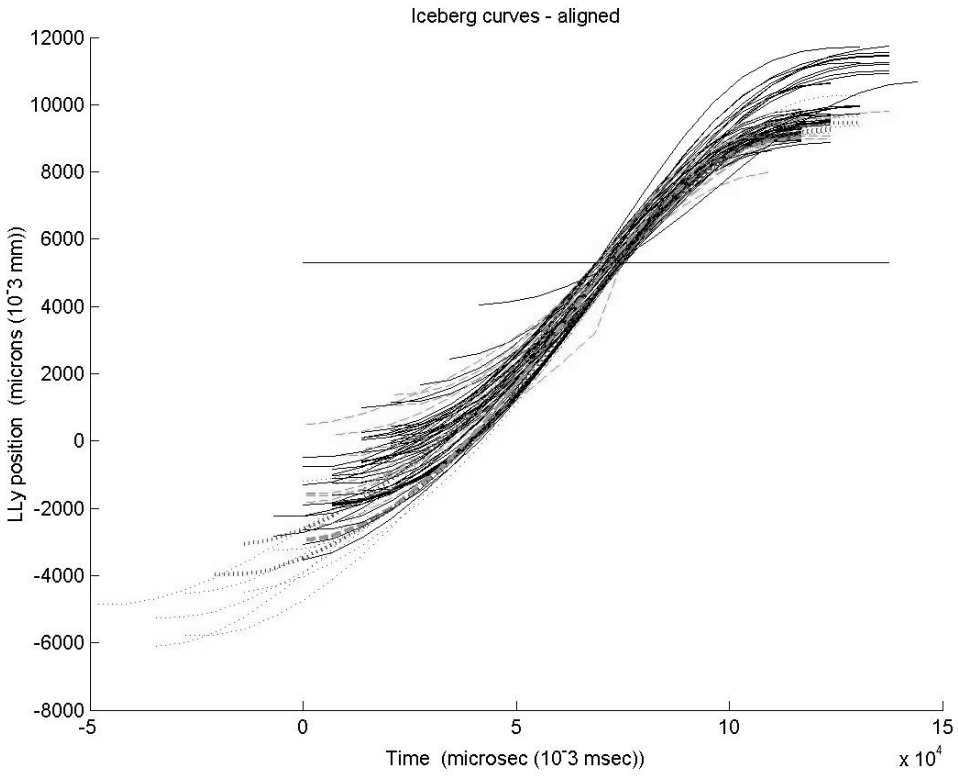
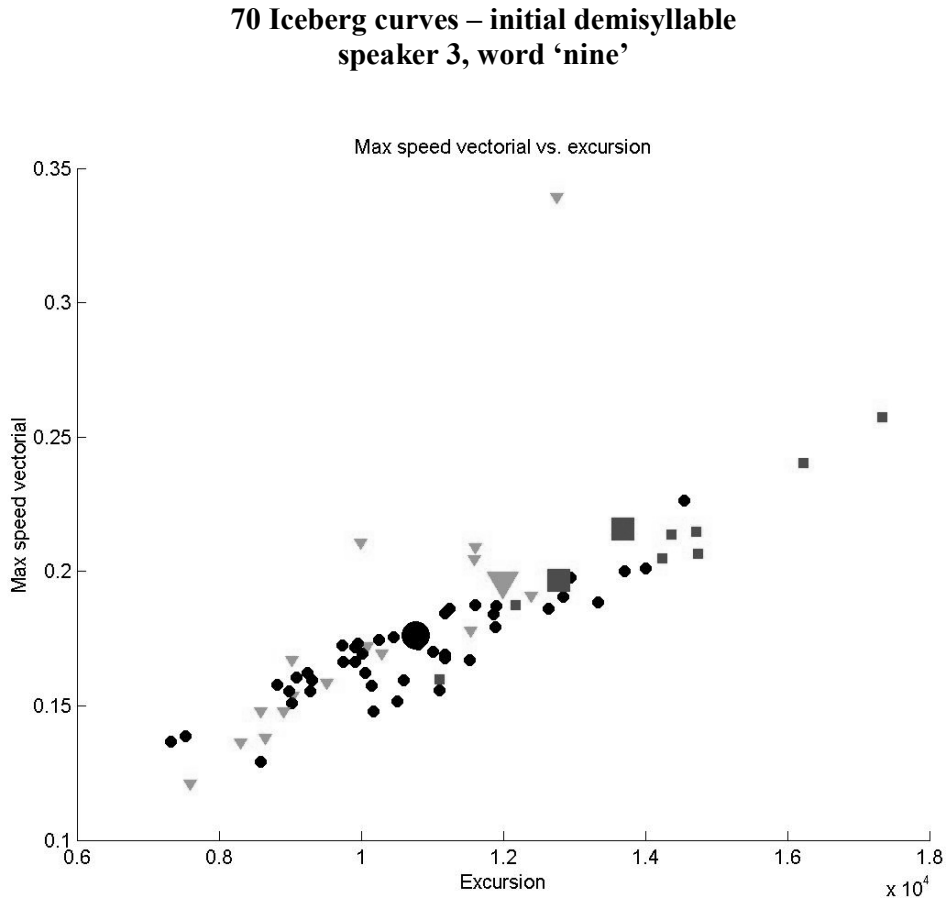


Fig. 7 (b): V-C movements (time functions) for the final demisyllable of 'five', temporally aligned by the iceberg threshold crossing time.



The large marks indicate those cases where an obvious pause follows. Different shapes of the data points indicate different position in the three-digit string.

Fig. 8 (a): Scatter plot showing the movement speed at iceberg threshold crossing *vs*. excursion distance. There is a slight linear dependence of speed on excursion.

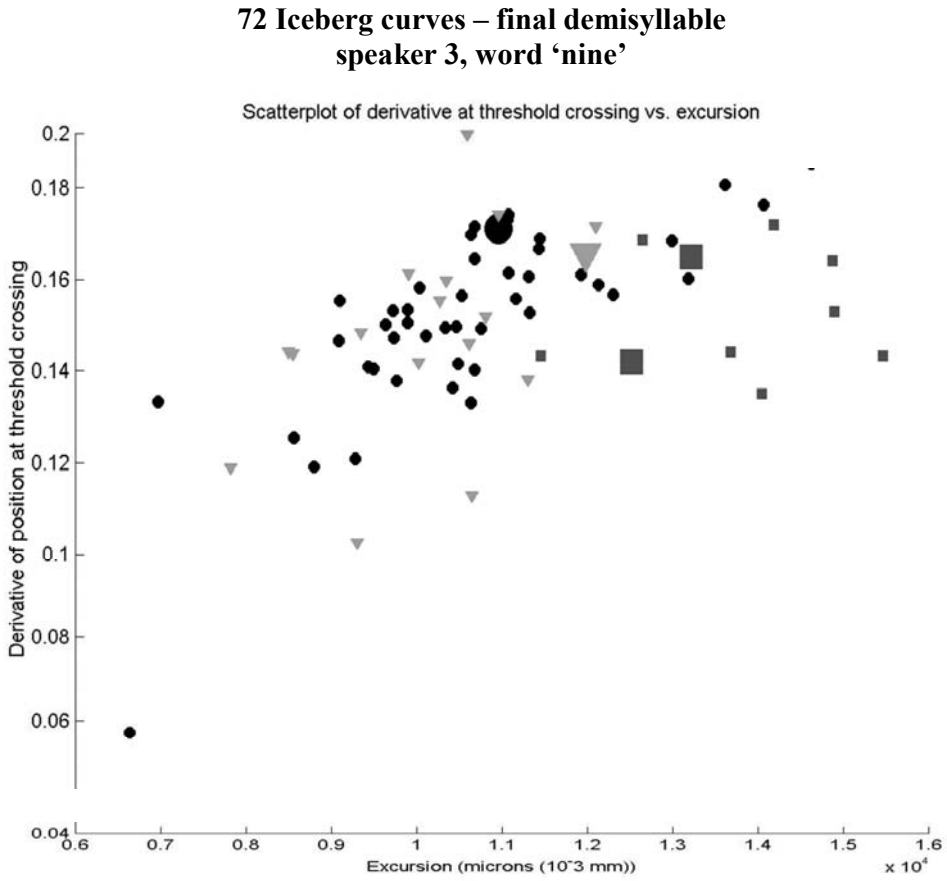


Fig. 8 (b): Scatterplot showing derivative at crossing *vs*. excursion for the final demisyllable, which are often followed immediately by boundaries of various magnitudes.

Mechanical models

3rd International Workshop MAVEBA 2003, 139-142

ASYMMETRIC AND SYMMETRIC VOCAL FOLD OSCILLATION IN THE EXCISED SQUIRREL MONKEY LARYNX

C.H. Brown[1], F. Alipour[2]
[1]Department of Psychology, University of South Alabama, Mobile, Alabama, U.S.A. 36688
[2]Department of Speech Pathology and Audiology, The University of Iowa, Iowa City, Iowa, U.S.A. 52242

Abstract: The larynges of nine squirrel monkeys were harvested, dissected, mounted on a tapered pseudotracheal tube, and phonated using heated and humidified air. The patterns of oscillation of the vocal folds were videotaped with stroboscopic illumination, and simultaneous measurements of airflow, subglottal pressure, and audio signal were obtained. The pressure wave and audio signal were subjected to spectral and phase portrait analysis methods. It was found that the left vocal fold tended to oscillate at lower subglottal pressure compared to the right vocal fold. This resulted in unilateral oscillation. Bilateral oscillation was seen at higher subglottal pressures. Patterns of symmetric and asymmetric bilateral oscillations were observed.

I. INTRODUCTION

The squirrel monkey larynx exhibits at least four different regimes of oscillation including biphonation, staccato phonation, and aperiodic phonation, as well as periodic phonation with overtones [1]. These various regimes of oscillation are exhibited "naturally" in the excised squirrel monkey larynx without any attempt to manipulate differentially the stiffness, mass, or elongation of the left and right vocal folds. In the present study we examined selected cases from our data set in which bifurcations between symmetric and asymmetric patterns of vocal fold oscillation were observed as a function of changes in subglottal pressure, while vocal fold elongation and adduction were held constant. Of particular concern here is the observation that as subglottal pressure is incremented and the threshold for phonation is achieved, the pattern of vocal fold oscillation is frequently unilateral, the right vocal fold is relatively immobile and oscillation is nearly confined to the left vocal fold. It is as if the mechanical properties of the left and right vocal folds differ with the right vocal fold exhibiting greater stiffness. At other levels of subglottal pressure bilateral motion of the two folds is observed, and the bilateral oscillations may be either synchronized or asynchronized. The goal of the present study was to examine the acoustic significance of transitions between asymmetric and symmetric patterns of oscillation.

II. METHODOLOGY

Subjects: Excised squirrel monkey larynges were obtained from the Squirrel Monkey Breeding and Research Resource, University of South Alabama. The Squirrel Monkey Breeding and Research Resource, housing approximately 500 animals, is the largest squirrel monkey colony in the United States, with a low annual mortality of about 5%. The larynges of nine monkeys were harvested from animals which suffered a natural spontaneous death. No monkeys were killed for the purpose of conducting this research. Larynx ID 1630, 4510, 90780 and 2618 were extracted from adult female Bolivian squirrel monkeys (*Saimiri boliviens boliviensis*). Larynx ID 1232 was removed from an adult female Guyanese squirrel monkey (*Saimiri sciureus sciureus*). The remaining three larynges were harvested from the Peruvian subspecies (*Saimiri boliviensis peruviensis*). Of these, Larynx ID 410 was harvested from an adult male, while larynx ID 742 and 90004 were obtained from adult female specimens.

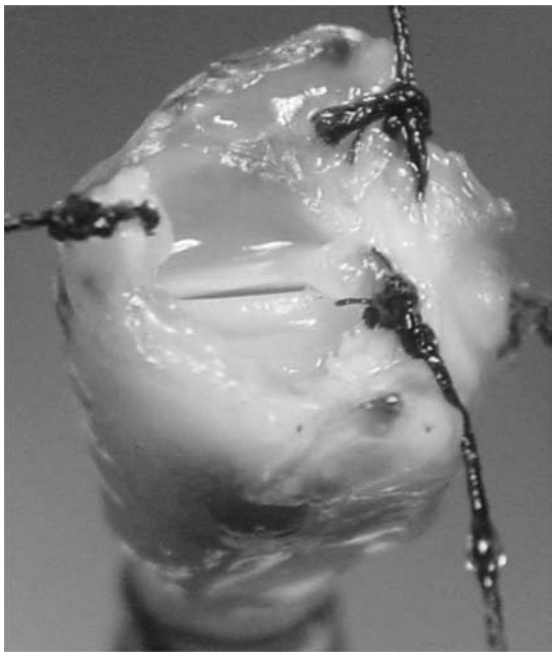


Figure 1: Mounted larynx with control sutures.

Apparatus and Procedure: Experiments were conducted in an IAC single-walled both in which the

interior surfaces were covered with Sonex foam to reduce acoustic reflections. Each larynx was dissected and trimmed, and the false vocal folds were removed. The tracheal tissue and larynx was mounted on a pseudotracheal tapered rigid tube (3mm average diameter), positioned on a laboratory bench, and the tracheal axis was oriented in the vertical position, exposing the glottis to a camera and recording apparatus. Adduction was controlled by two sutures attached to micrometers which pulled together the arytenoid cartilages. In one condition, the length of the vocal folds was not manipulated, and the larynx was permitted to phonate freely without any attachments to the thyroid tissue. In the second condition, vocal fold length was manipulated. A surgical suture pulling the thyroid cartilage against the cricoid cartilage controlled the length changes. No attempt was made to apply asymmetrical adjustments to differentially lengthen the left and right vocal folds**.** Figure 1 shows a squirrel monkey larynx mounted on the pseudotracheal tube.

The pseudotracheal tube received air from the building's oil and water free compressed air supply. The air was heated to 37^{o} C via a Concha Therm III Servo Control Heater (RCI laboratories, Arlington Heights, IL), and was humidified to approximately 100% relative humidity. The mean air pressure below the glottis was monitored with a wall-mounted water manometer (Dwyer No. 1230-8), and the mean flow rate was monitored with an in-line flowmeter (Gilmont rotameter model J197). The top view of the larynx and vocal folds was videotaped (Sony model DC-102) for later image analysis, and for stroboscopic images, a Pioneer DS-303ST stroboscope was employed. The audio recordings of the signal were obtained with an Shure (model 48) microphone also positioned 10 cm above the glottis, the analog signals were recorded on a Sony model PC-108M Digital Audio Tape (DAT) recorder, and simultaneously filtered, sampled, digitized (12-bit A/D, 44.2 kHz sample rate) and stored on a Gateway personal computer. The digitized time series data were analyzed with MATLAB or TFR signal processing software.

III. RESULTS

Each larynx was readily phonated, and each larynx exhibited samples of both stable phonation, and samples of irregular phonation characterized by non-linear phenomena. In the present study we kept subglottal pressure to 40 cm-H_20 or less. In this preparation, variations in subglottal pressure of 40 cm H20 or less produce fluctuations in the amplitude of voicing that matches the range in amplitude of vocalizations recorded from nonhuman primates [1]. Summed across all nine larynges we recorded 546 samples of phonation.

At the onset of phonation over half of the samples exhibited unilateral phonation where oscillation was virtually confined to the left vocal fold and the right vocal fold was nearly immobile. As subglottal pressure was incremented, airflow increased and motion in the right vocal fold was initiated. We did not encounter examples were unilateral oscillation was observed in the right vocal fold, and the left vocal fold was nearly stationary. Figure 2 shows the audio waveform FFT for larynx ID 2618 at subglottal pressures of 29 and 39 cm-H_20 respectively. At a subglottal pressure of 29 cm-H_20 oscillation was nearly unilateral with good oscillation in the left vocal fold and very little motion in the right vocal fold. At a subglottal pressure of 39 cm-H_20, synchronized bilateral oscillation was observed. The amplitude of the second harmonic increased markedly when bilateral oscillation was established.

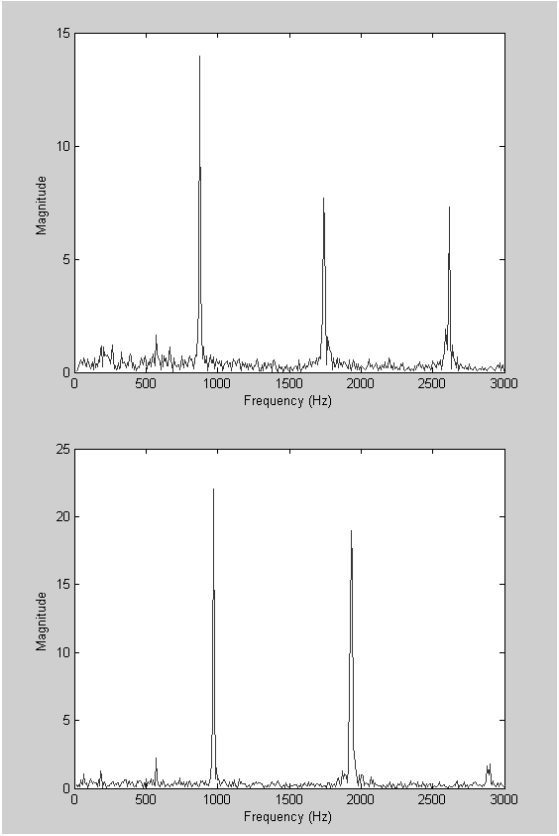


Figure 2: A (top panel) FFT of a unilateral oscillation, B (bottom panel) FFT of a synchronized bilateral oscillation.

Figure 3 shows a similar example of this phenomenon for larynx ID 2683. At a subglottal pressure of 23 cm-H_20 oscillation was unilateral, and synchronized bilateral oscillation of the vocal folds was observed at a subglottal pressure of 32 cm-H_20. At intermediate subglottal pressures this larynx exhibited bilateral oscillation, but the left and right folds oscillated out of phase with one fold "closing" as the other fold was "opening". In Figure 3 bilateral phase shifted oscillation is shown for a subglottal

pressure of 24 cm-H_20. As was observed in Figure 2, synchronized bilateral oscillation was associated with a prominent second harmonic of the fundamental frequency, and a third and fourth harmonics were also evident.

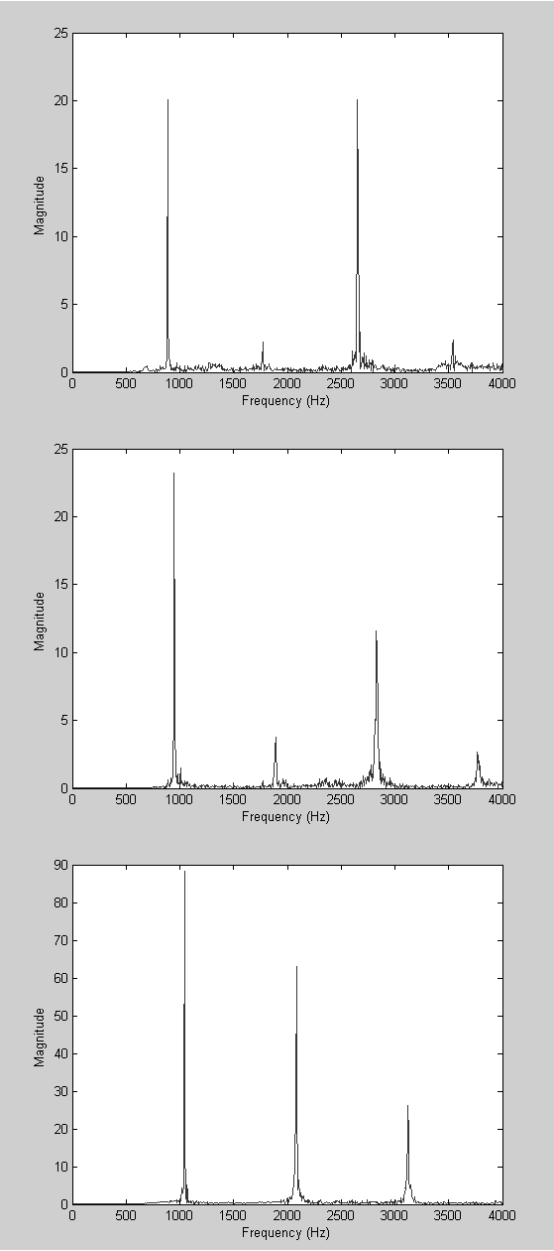


Figure 3: A (top panel) FFT of unilateral oscillation, B (middle panel) FFT of a bilateral phase shifted oscillation, C (bottom panel) FFT of bilateral synchronized oscillation.

These findings suggest that in the squirrel monkey the vibrations in the tissue in the left half of the larynx are not strongly coupled to those in the right side of the larynx, and this increases the possibility that different frequencies or modes of oscillation may be established simultaneously within the laryngeal complex. Figure 4 shows the waveform recorded from a pressure transducer, the FFT of this waveform and the phase portrait of this waveform for complex bilateral oscillation for larynx ID 1232. In this case the vibration patterns in the left and right vocal folds were not synchronized or coupled with each other. This is shown by the fact that the frequency peaks in the FFT were not harmonically related, and the phase portrait was elliptical.

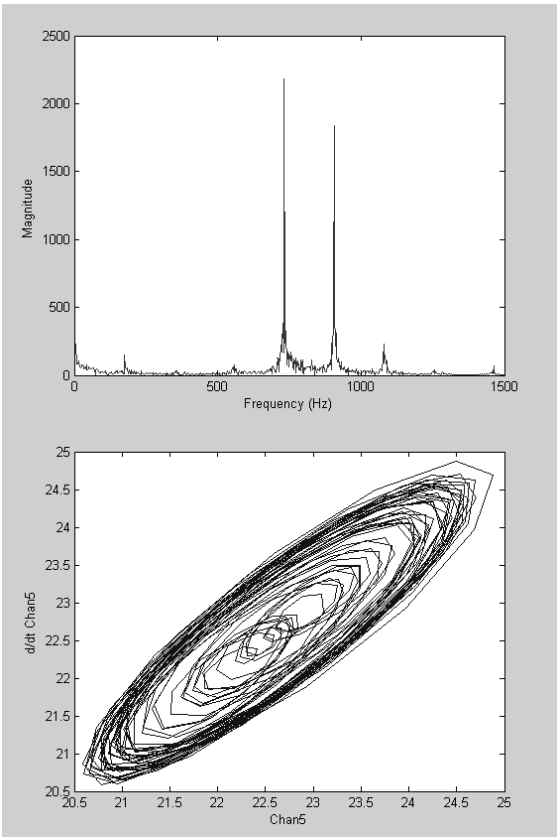


Figure 4: A (top panel) FFT of a low-pass filtered pressure signal measured 2 inches below the vocal folds during asymmetrical bilateral oscillation, B (bottom panel) phase portrait for the above case. The *x-axis* displays the position of the signal, the *y-axis* the derivative.

IV. Discussion

In three previous studies that focused on the nonlinear behavior of vocal fold vibration, nonlinear behavior was experimentally induced by asymmetrically manipulating the stiffness of the two vocal folds [2,3], or the tension and length of the two folds [4]. Berry and his colleagues noted that only occasionally were asymmetric patterns of oscillation observed for symmetric folds [4]. In contrast, in the present study, several patterns of asymmetric oscillation were observed, and these phenomena occurred readily without any experimental manipulation to induce asymmetries in the stiffness,

length or tension of the vocal folds. These observations are consistent with the idea that the magnitude of the coupling between the left and right vocal folds may differ prominently between species, and anatomical studies may help elucidate the differences between the laryngeal tissues of species that exhibit lax or tight coupling.

The present findings were also consistent with the idea that the membrane characteristics of the left and right vocal fold differed such that the right vocal fold exhibited functionally greater mass or stiffness compared to the left fold. Unilateral oscillation was almost entirely confined to the left vocal fold, and symmetrical bilateral oscillation tended to require greater airflow and subglottal pressures. Careful anatomical studies may reveal left-right asymmetries in the tissue complex that may account for this phenomenon.

As shown by Giovanni, Ouaknine, Guelfucci, Yu, Zanaret, and Triglia, if the two vocal folds differed in their relative stiffness, then the frequency and amplitudes of oscillation of the vocal folds would differ, and could result in a signal characterized by the nonlinear phenomenon of an asymmetric attractor [3]. In this case the amplitude of the signal should wax and wane according to the nonlinear combination of the oscillations of the two folds. These mechanisms may account for the characteristics of the sample shown in Figure 4.

V. CONCLUSION

The data suggest that the mechanical properties of the left and right vocal fold differ with the right vocal fold exhibiting greater stiffness. The data also suggest that the coupling between the left and right vocal folds is comparatively weak in the squirrel monkey. The left vocal fold tends to oscillate with greater amplitudes and at lower subglottal pressures compared to that observed for the right vocal fold. These phenomena result in unilateral and bilateral patterns of oscillation of the vocal fold. In some cases the absence of coupling results in asymmetric patterns of bilateral oscillations.

REFERENCES

[1] C.H. Brown, F. Alipour, D.A. Berry, and D. Montequin, "Laryngeal biomechanics and vocal communication in the squirrel monkey (*Saimiri boliviensis*)," J. Acoust.Soc. Am., 113, pp. 2114-2126, 2003.

[2] E. Yanagi,, and T.V. McCaffery, "Study of vibratory pattern of the vocal folds in the excised canine larynx," Arch Otolaryngol Head Neck Surg, 118, 30-36, 1992.

[3] A. Giovanni, M. Ouaknine, B. Guelfucci, P. Yu, M. Zanaret, and J-M. Triglia, "Nonlinear behavior of vocal fold vibration: the role of coupling between the folds," J. Voice, 13, 465-476, 1999.

[4] D.A. Berry, H. Herzel, I.R. Titze, and B.H. Story, " Bifurcations in excised larynx experiments," *J Voice* 10, 129-138, 1996.

3rd International Workshop MAVEBA 2003, 143-146

NUMERICAL MODELLING OF LEAKAGE-FLOW-INDUCED VIBRATIONS OF HUMAN VOCAL FOLDS WITH HERTZ IMPACT FORCES

Horáček J.[1], Šidlof P.[1], Švec J.G.[2]
[1]Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[2]National Center for Voice and Speech, The Denver Center for the Performing Arts, Denver, USA and Medical Healthcom, Ltd., Centre for Communication Disorders, Prague, Czech Republic

Abstract: **Mathematical model of the vocal folds self-oscillations excited by aeroelastic mechanism is presented. A two-degrees-of-freedom element on an elastic foundation with a generally defined shape vibrating in the glottal airflow approximates the vocal fold. The Hertz impact model is considered for the contact forces during the vocal folds collisions. The model's vibratory patterns and resulting flow values are similar to those of the real vocal folds. The model is expected to be helpful in design of artificial voice prosthesis.**

Keywords: **Biomechanics of voice, aeroelastic instabilities, flutter, divergence, post-critical behaviour of the aeroelastic system, nonlinear vibrations, impact oscillator.**

I. INTRODUCTION

A linear two-degrees-of-freedom aeroelastic model was originally developed by the authors in order to study the influence of different geometrical and elastic properties of the vocal folds on phonation thresholds [1,2]. The inviscid incompressible 1-D fluid flow theory is used in the model for expressing the unsteady aerodynamic forces. The numerical solution yields the natural frequencies, damping, mode shapes of vibration and the instability thresholds of the system directly by solving an eigenvalue problem. The thresholds are given by aeroelastic instabilities of divergence or flutter type. The developed aeroelastic model is able to provide qualitative information on conditions for a soft voice onset or for breathy voicing [2]. In order to study also the conditions of small glottal openings and large vibration amplitudes, the model was generalised by taking into account the non-linear aerodynamic terms [3]. Results of numerical simulations in the time domain were in good agreement with the previous solution in the frequency domain, when a linear approximation of the aerodynamic forces for calculation of the stability boundaries was used.

The present paper introduces the Hertz model for modelling the impact forces between the vocal folds. Nonlinear dynamic and aerodynamic forces are implemented into an aeroelastic model of the vocal folds and the *postcritical* behaviour of the system after loosing the stability is simulated. This allows complete numerical simulations of self-oscillations of the vocal folds during phonation. The parameters of the model, i.e., the mass, stiffness and damping matrices are approximately related to the geometry, size and material density of real vocal folds as well as to the known or prescribed fundamental natural frequencies and damping.

II. MATHEMATICAL MODEL

A vibrating element of the length L with mass m and moment of inertia I with two-degrees-of-freedom (rotation and translation ${}^{\mathrm{T}}\mathbf{V}=(V_1(t),\ V_2(t))$ $=((w_2-w_1)/2l,\ (w_1+w_2)/2)$ supported by an elastic foundation and vibrating in the wall of a channel conveying air is used to approximate the vocal fold oscillations (Fig. 1).

Vibrations of one vocal fold are modelled by the equations of motion of an equivalent three mass system on two springs [1,2]:

$$\overline{\mathbf{M}}\ddot{\mathbf{V}} + \overline{\mathbf{B}}\dot{\mathbf{V}} + \overline{\mathbf{K}}\mathbf{V} + \mathbf{F} = \mathbf{0}\,, \tag{1}$$

where $\overline{\mathbf{M}}, \overline{\mathbf{B}}, \overline{\mathbf{K}}$ are the mass, damping and stiffness (2x2) matrices, respectively, and the aerodynamic excitation forces $\mathbf{F}$ are given by the perturbation pressure $\tilde{p}(x,t)$ of the fluid flow in the glottis:

$$\begin{aligned} F_1(t) &= \frac{h}{2}\int_0^L \left(1-\frac{x}{l}+\frac{L_1}{l}\right)\tilde{p}(x,t)\,dx\,,\\ F_2(t) &= \frac{h}{2}\int_0^L \left(1+\frac{x}{l}-\frac{L_1}{l}\right)\tilde{p}(x,t)\,dx\,; \end{aligned} \tag{2}$$

h is the channel depth and the distances l, L_1 define the two springs positions; and $\overline{\mathbf{B}} = \bar{\varepsilon}_1\,\overline{\mathbf{M}} + \bar{\varepsilon}_2\,\overline{\mathbf{K}}$ is assumed.

Aerodynamic forces are calculated from the unsteady Euler and continuity equations:

$$\frac{\partial A}{\partial t} + \frac{\partial (AU)}{\partial x} = 0\,, \tag{3}$$

$$\frac{\partial (AU)}{\partial t} + \frac{\partial (AU^2)}{\partial x} + \frac{A}{\rho_f}\frac{\partial P}{\partial x} = 0\,, \tag{4}$$

where $A(x,t) = h\,H(x,t)$ is the channel cross-sectional area; ρ_f and $U(x,t)$ are the fluid density and velocity; $P(x,t)$ is the pressure.

After separation of steady and unsteady components:

$$\begin{aligned} U(x,t) &= \bar{U}_0(x) + \tilde{u}(x,t),\ P(x,t) = P_0(x) + \tilde{p}(x,t),\\ H(x,t) &= H_0 - w(x,t) - a(x) \end{aligned} \tag{5}$$

the equation (4) yields the following equation for the perturbation velocity and pressure:

$$\frac{\partial \tilde{u}}{\partial t}+\frac{\partial\left(\overline{U}_0(x)\,\tilde{u}\right)}{\partial x}+\tilde{u}\frac{\partial \tilde{u}}{\partial x}=-\frac{1}{\rho_t}\frac{\partial \tilde{p}}{\partial x}. \quad (6)$$

Introducing the velocity potential $\tilde{u}=\frac{\partial \Phi}{\partial x}$ the perturbation pressure can be expressed as

$$\tilde{p}(x,t)=-\rho_t\left[\frac{\partial \Phi}{\partial t}+\overline{U}_0(x)\frac{\partial \Phi}{\partial x}+\frac{1}{2}\left(\frac{\partial \Phi}{\partial x}\right)^2\right]. \quad (7)$$

Considering small vibration amplitudes: $|w| << H_0$, the boundary conditions at the channel inlet and outlet:

$$\tilde{u}=\frac{\partial \Phi}{\partial x}=0\;\big|_{x=0}\;,\quad \tilde{p}=0\;\big|_{x=L} \quad (8)$$

and following the derivations in [1-3], the nonlinear perturbation pressure can be expressed as

$$\begin{aligned}\tilde{p}(x,t)=-\rho\{&K_1(x)\left[V_1(t)\right]^2+K_2(x)\,V_2(t)+\\ &K_3(x)\left[V_2(t)\right]^2+K_4(x)\left[V_1'(t)\right]^2+K_5(x)\,V_2'(t)\\ &+K_6(x)\,V_2(t)\,V_2'(t)+K_7(x)\left[V_2'(t)\right]^2+K_8(x)\,V_1'(t)+\\ &K_9(x)\,V_1'(t)V_1(t)+K_{10}(x)\,V_1'(t)\,V_2(t)+\\ &K_{11}(x)\,V_1'(t)\,V_2'(t)+K_{12}(x)\,V_1(t)+\\ &K_{13}(x)\,V_1(t)\,V_2(t)+K_{14}(x)\,V_1(t)\,V_2'(t)+\\ &K_{15}(x)\,V_1''(t)+K_{16}(x)\,V_2''(t)\},\end{aligned} \quad (9)$$

where the coefficients $K_i(x)$ (i=1,2,…,16) given by complicated algebraic expressions can be found in [3].

Hertz model [4] of impact is implemented in the aeroelastic model for the vocal folds collisions. The impact force F_H is generally considered as

$$F_H=k_H y^{3/2}\left(1+b_H\dot{y}\right),\; k_H\cong\frac{\sqrt{2}}{3}\frac{E}{1-\mu^2}\sqrt{r}, \quad (10)$$

where E is Young modulus, μ is Poisson ratio and r is the radius of the impacting body surfaces.

The *input parameters* for numerical analysis were considered to be approximately of the same order as the data known for the vocal folds from literature. The geometry of the vocal fold was approximated by the following concave function

$$a_i(x)=a_1x+\frac{a_2}{2}x^2=1.858x-159.861x^2\ \text{[m]}. \quad (11)$$

From here the co-ordinates of the contact point can be determined as

$$\begin{aligned}x_{\max}&=\min\left(L,\max\left(0,-\frac{V_1+a_1}{a_2}\right)\right),\\ y_{\max}&=y(x_{\max})=a(x_{\max})+(x_{\max}-L_1)V_1+V_2.\end{aligned} \quad (12)$$

The impact Hertz force can be expressed as

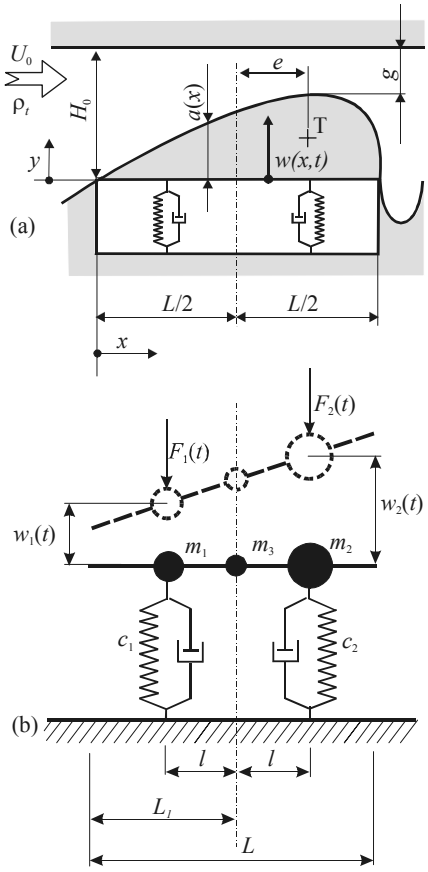


Fig. 1 - Two-degrees of freedom model.

$F_H=k_H\left(y_{\max}-H_0\right)^{3/2}$, where $k_H\cong$ 730 Nm$^{-3/2}$ for E=8 kPa and μ=0.4 [5] and where the damping coefficient was neglected (b_H=0); $H_0=\max_{x\in<0,L>}a(x)+g$ is the height of the channel and g is the glottal half-width (see Fig. 1a). A correction on the static subglottal pressure:

$$p_{sub}=\frac{\rho}{2}U_0^2\left(\frac{a(L)}{H_0-a(L)}\right)^2, \quad (13)$$

which is constant during the vocal folds collision ($U_0=\overline{U}_0(0)$), gives after integration of the pressure p_{sub} in the interval $x\in\langle 0,x_{\max}\rangle$ the resulting forces in Eq. (1) during vocal folds contact:

$$\begin{aligned}F_1&=F_H\frac{L_1+l-x_{\max}}{2l}+p_{sub}h\,x_{\max}\frac{L_1+l-\frac{x_{\max}}{2}}{2\,l},\\ F_2&=F_H\frac{x_{\max}-L_1+l}{2l}+p_{sub}h\,x_{\max}\frac{\frac{x_{\max}}{2}-L_1+l}{2\,l}.\end{aligned} \quad (14)$$

For numerical simulations the equation of motion (1) was transformed into the system of four 1st order ordinary differential equations:

$$\begin{aligned} Z_1' &= f_1(Z_1, Z_2, V_1, V_2) \\ Z_2' &= f_2(Z_1, Z_2, V_1, V_2) \\ V_1' &= Z_1, \quad V_2' = Z_2, \end{aligned} \tag{15}$$

and 4th order Runge-Kutta method was used for the calculations. The functions f_1, f_2 are determined differently for contact (Hertz forces and static pressure - see Eq. (14)) and non-contact (aerodynamic forces – see Eq.(2)) regimes.

The density, thickness and length of the vocal folds were taken as follows: ρ_h =1020 kg/m^3, L = 6.8 mm, h =10 mm [1-3]. From these data there were calculated: the eccentricity e, the total mass m and the moment of inertia I; the air density was considered as ρ =1.2 kg/m^3, L_1=L/2 and l=0.344 L. A tuning procedure was used to adjust the stiffness (c_1, c_2) of the elastic foundation and the damping coefficients $\bar{\varepsilon}_1, \bar{\varepsilon}_2$ in order to approximate the natural frequencies f_1, f_2 and 3dB half-power bandwidths Δf_1 =23 Hz and Δf_2 =29 Hz of both resonances by values measured on true vocal folds [1].

III. Results

Typical simulation output is demonstrated in Fig. 2, where the motions $w_1(t)$ and $w_2(t)$ of the masses m_1 and m_2 are shown in the phase planes (left part) and in time domain (right upper part with marked impact duration) as well as the glottis opening $S(t)$ and perturbation subglottal pressure $\tilde{p}(t)$ at x=0 (right lower part of Fig. 2). The motion of the vocal fold is regular with one impact during one period and with calculated open quotient OQ=0.71 for the oscillation cycle. The motion of the vocal fold is animated in Fig. 3. The spectrum of the airflow velocity at the outlet $\tilde{u}(L,t)$ is shown in Fig. 4. The resulting vibrational frequency F0, which is determined by the flutter frequency, is between the natural frequencies f_1=100 Hz and f_2=105 Hz. The forces loading the mass m_2 during the self-oscillations are shown in Fig. 5, where F_{el}, F_a, F_{in}, F_p, F_H denote the elastic, aerodynamic, inertial, subglottal pressure and Hertz forces, respectively. When the second natural frequency f_2 was increased from 105 to 150 Hz and the flow velocity U_0 increased from 2.4 to 3.5 m/s, everything else unchanged, the glottal area $S(t)$ showed subharmonic oscillations (Q=0.40 l/s, OQ=0.87 - see Fig. 6). Oscillations without impacts (OQ=1) were observed for the same parameters when U_0 was further increased to 4 m/s (Q=0.46 l/s - see Fig. 7). In general, the results were influenced by variation of the coefficient k_H in the interval 100-10 000 Nm$^{-3/2}$ only very slightly.

IV. Conclusions

The presented model based on aeroelastic theory enables to model the vocal fold self-oscillations in time domain after crossing the phonation thresholds given by the critical flow velocities (volume flux) needed for loosing the system stability. The input geometrical and material parameters of the model

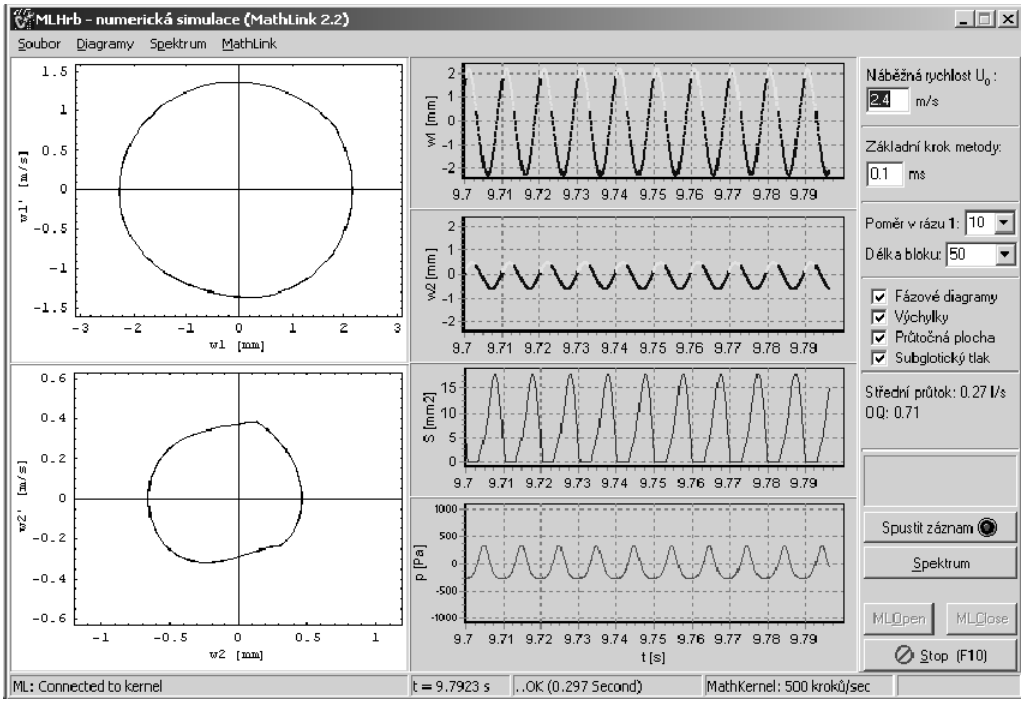


Fig. 2- Numerical simulation for fluid velocity U_0=2.4 m/s (Q=0.27 l/s), g=0.3 mm, f_1=100 Hz, f_2=105 Hz.

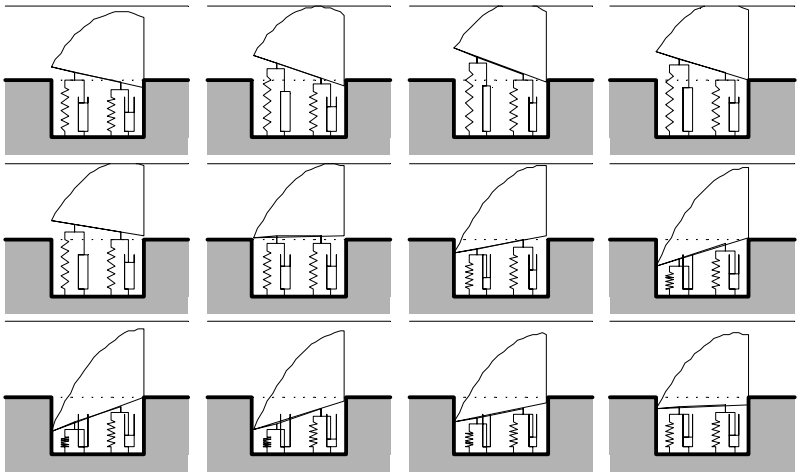


Fig. 3 - Numerical simulation of the vocal fold motion during one oscillation cycle for data as in Fig.2.

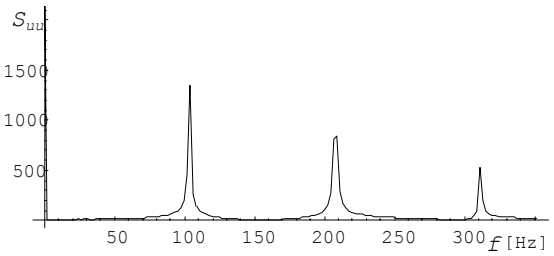


Fig. 4 – Spectrum of the outlet airflow velocity $\tilde{u}(L,t)$.

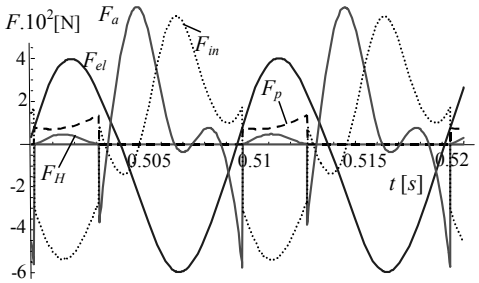


Fig. 5 – Forces loading the mass m_2 during self-oscillations for data as in Fig.2 and U_0=3.5 m/s.

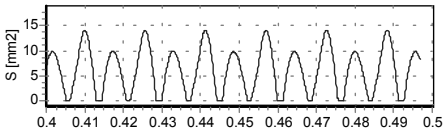


Fig. 6 – Oscillations with subharmonic frequency.

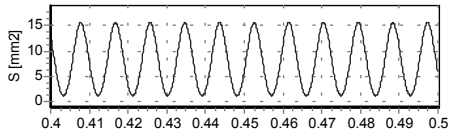


Fig. 7 – Impactless oscillations of the vocal fold.

are closely related to the properties of real vocal folds [6].

The preliminary results show that the model reflects the basic vibration regimes and general types of dynamic behaviour of real vocal folds known from clinical observations such as modal and non-modal phonation, regular, stationary motions with collisions as well as, in special cases, the subharmonic and chaotic oscillations.

ACKNOWLEDGEMENT

The authors acknowledge the financial support of the Grant Agency of the Czech Republic by the project No 106/98/K019 *Mathematical and Physical Modelling of Vibroacoustic Systems in Biomechanics of Voice and Hearing Focused on Development of Artificial Replacements and Prostheses.*

REFERENCES

[1] J. Horáček and J. G. Švec, 2002 "Aeroelastic model of vocal-fold-shaped vibrating element for studying the phonation". *Journal of Fluids and Structures* **16** (7), 927-951.

[2] J. Horáček and J. G. Švec, 2002 "Instability boundaries of a vocal fold modelled as a flexibly supported rigid body vibrating in a channel conveying fluid". In: *Proc. of the 5th Int. Symp. on FSI, AE & FIV+N, 2002 ASME Int. Mech. Eng. Congr.*, Nov. 17-22, 2002, New Orleans, USA, 12p.

[3] J. Horáček and P. Šidlof, 2002 "Numerical simulation of human vocal folds oscillations at phonation onset caused by aeroelastic instability". In: *Proc. Interaction and Feedbacks'2002* (Ed. Zolotarev I.), Institute of Thermomechanics ASCR, Prague, Nov. 26-27, 2002, pp. 27-36 (in Czech).

[4] L. Půst, and F. Peterka, 2003 "Impact oscillator with Hertz's model of contact", *Meccanica* **38**, 99-114.

[5] M.P. de Vries, H.K. Schutte and G.J. Verkerke "Determination of parameters for lumped parameter models of the vocal folds using a finite-element method approach", *J. Acoust. Soc. Am.* **106** (6), 3620-3628.

[6] X. Pelorson, A. Hirschberg, A. P. J. Wijnands, and H. Bailliet, 1995 "Description of the flow through in vitro models of the glottis during phonation", *Acta Acustica* **3**, 191-202.

3rd International Workshop MAVEBA 2003, 147-150

PHYSICAL AND NUMERICAL FLOW-EXCITED VOCAL FOLD MODELS

S. L. Thomson, L. Mongeau, S. H. Frankel
School of Mechanical Engineering, Purdue University, IN, USA

Abstract: **Self-oscillating physical and numerical models of the vocal folds were investigated. The physical model was cast into an idealized shape of the vocal folds, on a 1:1 length scale with the human vocal folds, using a flexible polyurethane rubber. The model in a hemilaryngeal configuration experienced flow-induced oscillations at a frequency of 90 Hz and onset pressure of 1.2 kPa. The numerical model was a two-dimensional finite element model of the vocal folds and vocal tract. The flow was calculated throughout the flow domain using the incompressible, two-dimensional Navier-Stokes equations. The aerodynamics and vocal fold dynamics were fully coupled. Regular, self-sustained oscillations were predicted at a frequency of approximately 275 Hz. The influence of supraglottal duct length on vocal fold motion is discussed. The capabilities and limitations of the models are discussed, and areas for further development are identified.**
Keywords: **Physical model, finite element analysis, vocal fold models**

I. INTRODUCTION

The primary source of sound in the vocal tract is the modulation of the glottal airflow by the vocal folds opening and closing periodically. The motion of the vocal folds depends on the pressure loading on their surfaces due to the airflow. In turn, the airflow through the vocal tract is altered by the presence and motion of static and dynamic laryngeal structures. Consequently, there is a continual exchange of energy between the vocal folds and the airflow. The flow-structure interaction is of utmost importance in the region near the glottis.

To study the interaction between airflow and structural dynamics in a vocal tract model (either physical or numerical), the motion of the vocal folds must be coupled to the airflow within the vocal tract, and not externally driven (using a vibration generator, for example). This is achieved by developing models of the vocal folds which are flow-excited; that is, which move solely due to the glottal airflow.

Previous experimental studies of flow-induced oscillations of the vocal folds have been performed using human larynges [1], excised canine larynges [2], a vocal fold cover model [3], and membranous-type models [4,5]. Studies using canine or human subjects have the advantage of physiological realism, but have disadvantages of limited subject acquisition, and high maintenance requirements. Differences between subjects also limit the scope of potential parametric studies.

Self-oscillating numerical models include the original two-mass model [6] and subsequent modifications [7,8]. More complex vocal-fold models have been developed using two- and three-dimensional finite element methods [9]. This approach is promising because of the greater accuracy in predicting complex flow and structural behavior, such as flow separation, turbulence, and vocal fold collision. The insight gained from two- and three-dimensional models can be used to improve predictions of the computationally inexpensive multi-mass models.

This paper summarizes recent efforts to develop self-oscillating physical and numerical vocal fold models, which exhibit similarity with the human vocal folds, for use in studying laryngeal flow-structure interactions and in predicting dynamic vocal fold behavior.

II. PHYSICAL VOCAL FOLD MODEL

A. Methodology

The physical model was constructed using a three-part polyurethane rubber compound (Evergreen™ with Everflex™, available from Smooth-On, Inc.). The stiffness of the cured rubber could be varied by adjusting the mixing ratios of the three compounds. A tensile test on the rubber used in the experiments yielded a Young's modulus of approximately 4 kPa. This modulus is in the range of that found in vocal fold tissue [10].

The rubber was cast into the shape of the rigid model used by Scherer et al. [11]. This shape was chosen for eventual comparison of dynamic data with published static results. Note that the length scale of the model of ref. [11] was increased by a factor of 7.5; the flexible model reported here was on the same size scale as the human vocal folds.

A coronal cross section of the model, which was uniform along its 1.5 cm anterior-posterior length, is shown in Fig. 1. The vocal fold was mounted over a rigid circular tube, which was connected to an upstream

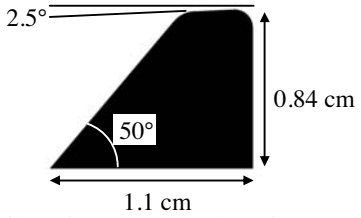


Figure 1. Outline of the cross section of the physical and vocal fold models.

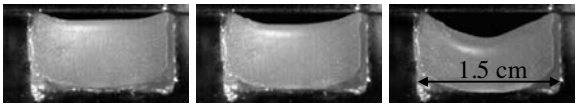


Figure 2. Images of the self-oscillating physical model of the vocal fold at three times during the cycle.

compressed air source. A rigid plate was placed opposite to the vocal fold (laterally) in a hemilarynx configuration.

B. Results

Fig. 2 shows three images obtained during one cycle of oscillation with a 1.5 kPa subglottal pressure. The images were obtained using a high-speed digital camera (Memrecam fx K3, NAC Image Technology) at a rate of 3000 frames per second. The maximum lateral displacement of the rubber was approximately 2.3 mm. Fig. 3 shows the glottal area over four cycles, calculated using the high-speed images; the maximum area is indicated by the dashed line. The open quotient (time open/period) was 0.79 and the skewing quotient (time of increasing area/time of decreasing area) was 1.17.

The vocal fold model oscillated at an onset pressure of 1.2 kPa with a fundamental frequency of 89 Hz. It was found that lower frequencies and oscillating pressures were obtained by either casting the physical model to a length longer than 1.5 cm, and/or using a different compound mixing ratio to produce a material with lower stiffness. Reducing the stiffness, however, tended to lower the tear strength of the material. The fundamental frequency was measured for subglottal pressures ranging from 0.9 kPa (which was just above the offset pressure) to 1.5 kPa. The frequency increased with increasing subglottal pressure at a rate of approximately 10 Hz/kPa.

C. Discussion

The fundamental frequency of 89 Hz was in the range measured in excised human and canine larynx studies; the onset pressure was slightly greater than values reported in the literature [12].

The increase in frequency with subglottal pressure was slightly lower than reported values in the range of 30–70 Hz/kPa obtained in different studies using excised canine and human larynges [12]. This difference may have been due to either the homogeneity and/or isotropy of the rubber material. The human vocal folds are non-homogeneous, and much of the tissue layers are transversely isotropic. The influence of including different layers with differing mechanical properties and including transverse isotropy in the physical model is a subject of current efforts.

The physical model demonstrated several capabilities. First of all, it operated at a frequency and subglottal pressure encountered in human phonation. The differences in onset pressure and frequency-subglottal pressure relationship were relatively small. Thus these

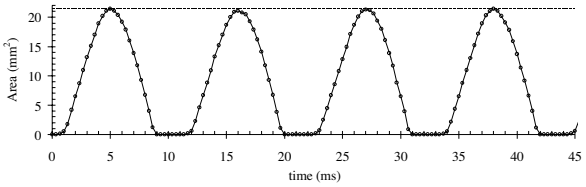


Figure 3. Area vs. time data from the physical model.

preliminary results indicate that further results obtained using this model may be reasonably applied to human phonation. Current studies are underway to obtain data for further comparison with data obtained using excised larynges. Another benefit of the similarity between the model and the vocal folds is that the Reynolds, Strouhal, and Mach numbers are on the same order as those encountered in human phonation, eliminating the need for trade-offs commonly necessary when using models of different length scales.

A further advantage is the relative accessibility of the materials and ease of construction. The process is repeatable, and small changes in the geometry can be incorporated for purposes of parametric studies. This is difficult, if not sometimes impossible, in excised larynges. Studies into the long-term life of the model are underway.

III. Numerical Vocal Fold Model

A. Computational Domain and Method

A two-dimensional numerical model of the vocal folds and surrounding airflow was developed. The flow was modeled using the Navier-Stokes equations for two-dimensional (planar), laminar, incompressible, isothermal flow. The vocal folds were represented by a two-dimensional continuum. The commercial software ADINA (ADINA R&D, Inc.) was used to generate the model. ADINA has been used in other studies involving biological fluid-structure interactions [13].

The domain, illustrated in Fig. 4, consisted of fully-coupled fluid and structural sub-domains, and included the subglottal, glottal, and supraglottal regions. For computational efficiency, it was assumed that the flow was symmetric about the centerline (line BC), and the flow over only one vocal fold was modeled. L_d denotes the length of the duct downstream of the vocal folds, which was varied between 4.9 and 18.9 cm.

The vocal fold shape and dimensions were the same as shown in Fig. 1. The pre-phonatory glottal half-width d_g was 0.5 mm. The vocal fold was allowed to move in the

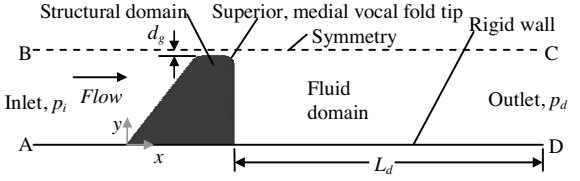


Figure 4. Schematic of the computational domain.

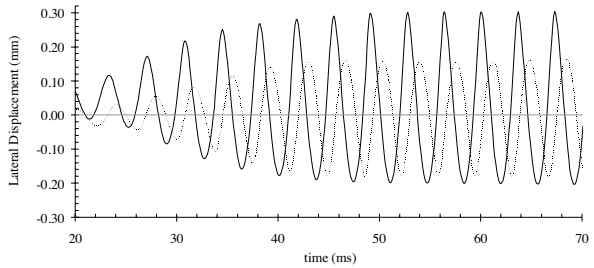


Figure 5. Lateral displacement vs. time of the inferior, medial vocal fold tip (— —) and the superior, medial vocal fold tip (· · ·). L_d = 18.9 cm.

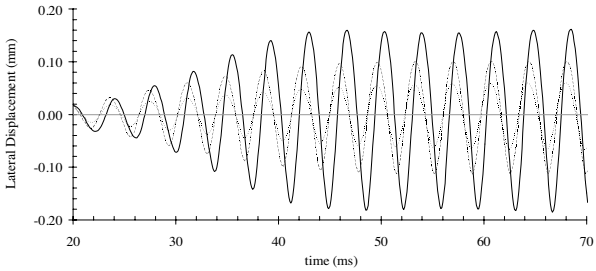


Figure 6. Lateral displacement of the superior, medial vocal fold tip vs. time. — — : L_d = 18.9 cm; – – – : L_d = 8.9 cm; · · · : L_d = 4.9 cm.

x- and *y*-directions, except where it was rigidly attached to the vocal tract wall (line AD). A constant, uniform pressure p_i = 0.8 kPa was applied at the inlet (AB). The outlet pressure p_d was set to zero along CD. The flow was air with density 1.2 kg/m^3 and viscosity 1.8×10^{-5} kg/m·s. Different material properties were assigned to different regions representing the cover, ligament, and body layers of the human vocal folds. The layers were assigned the following properties: Poisson's ratio = 0.45; density = 1070 kg/m^3; Young's modulus = 4 kPa (cover), 2.74 kPa (body), 2.26 kPa (ligament). For reference, the corresponding values defined in the finite element model used in [9] were: Poisson's ratio = 0.9; Shear modulus = 0.53 kPa (cover), 1.05 kPa (body), 0.87 kPa (ligament).

The layers of the two-dimensional models were isotropic. To simulate transverse isotropy, as well as the out-of-plane stiffness encountered by the three-dimensional vocal folds when deformed, spring elements were connected between the element nodes and ground. The springs constants were 11.7 N/m (cover), 98.4 N/m (body), and 18 N/m (ligament).

B. Computational Simulation Results

Modal analysis predicted the first three frequencies to be f_0 = 224 Hz, f_1 = 255 Hz, and f_2 = 269 Hz. Fig. 5 shows the displacement of the inferior and superior medial vocal fold tips vs. time for L_d = 18.9 cm. Regular self-sustained oscillations at a frequency of approximately 275 Hz were achieved by time $t \sim 50$ ms. This is slightly greater than the third modal frequency. The finite element model of Alipour et al. [9] exhibited flow-induced oscillations at the frequency corresponding approximately to roughly the average of f_0 and f_1. The reason for the present model oscillating at a frequency just greater than f_2 is not clear. The amplitudes of the inferior and superior points were approximately 0.25 and 0.17 mm, respectively. The motion of the points was approximately 108° out of phase due to the alternating converging-diverging orifice shape.

The influence of duct length was investigated by performing simulations with L_d = 4.9, 8.9, and 18.9 cm. Results showing the lateral displacement of the superior medial vocal fold tip are shown in Fig. 6. The amplitude significantly decreased when the duct length was shortened. Decreasing the duct length from 18.9 to 4.9 cm resulted in the frequency decreasing by ~ 1%.

Plots of the vorticity in and immediately downstream of the glottis over the glottal cycle are shown in Fig. 7 for L_d = 8.9 cm. The converging-diverging shape is evident, and it is seen that the medial surface altered between concave and convex shapes. The flow separated near the glottal exit radius when the glottis was convergent, straight, or only slightly divergent, but separated further upstream in the glottis when the glottis was more divergent. The vorticity plots in these images are qualitatively similar to those obtained using driven-wall direct numerical simulations [14].

C. Discussion

The numerical model oscillated at a frequency and pressure comparable to that found in human phonation, with material properties and boundary conditions similar to those of the vocal folds. In the current model, the Young's modulus of the cover was greater than that of the ligament. In the human folds the cover is more flexible; current efforts are focused on investigating the influence on vocal fold vibration of the relative stiffness of the different layers. (Allowing for such a study highlights the value of numerical flow-excited vocal fold models with fully-coupled flow and structural domains.)

The two-dimensionality of the model provided greater predictive capabilities of flow phenomena, such as flow separation, while avoiding the prohibitive cost of three-dimensional simulations. Vocal fold collision was not allowed in the simulations, although efforts are underway to include this effect.

The results demonstrated the importance of duct length in the simulations. The decrease in amplitude with decreasing duct length was attributed to: (1) differences in pressure distribution due to different duct lengths, resulting in different pressure loadings on the vocal folds, and (2) increased inertia from the added mass in the longer ducts. Further investigation of this topic is planned, as well as of trends such as frequency-subglottal pressure dependence (as discussed in Section II) for comparison with available measured data.

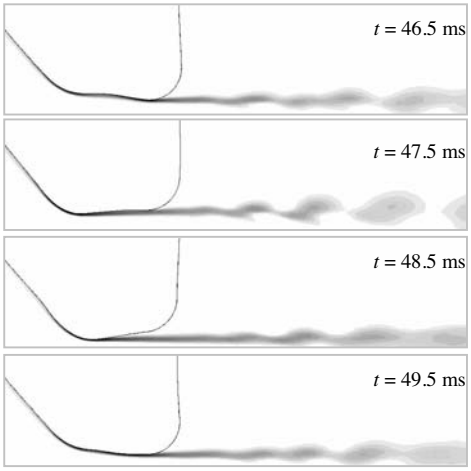


Figure 7. Vorticity in the glottal jet over one cycle. L_d = 8.9 cm.

IV. Conclusions

Two self-oscillating models, one physical and the other numerical, were introduced and results were presented. Because of the their dynamic similarity with the human vocal folds, they demonstrated the potential for further studies involving laryngeal fluid-structure interactions.

The physical model of the vocal folds was presented which demonstrated dynamic behavior similar to data obtained using excised larynges. The model was cast to the size and idealized shape of the vocal folds using an isotropic flexible polyurethane rubber. Flow-induced oscillations were observed at an onset pressure of 1.2 kPa and frequency of 89 Hz. The frequency increased at a rate of approximately 10 Hz/kPa, which was lower (but on the same order of magnitude) as previously measured data using excised larynges. Studies underway include further characterization of the model and comparison with available data, adding layers of different stiffness to the model, and investigating methods of making the model transversely isotropic.

The level of detail in the fluid and structural domains of the numerical model allowed for superior predictive capabilities over one-dimensional multi-mass models, while avoiding the prohibitive computational cost of three-dimensional models. Several areas of current interest were discussed, including studying the influence of the relative stiffness of the different layers, including vocal fold collision, and comparing further results with available measured vocal fold data.

V. Acknowledgements

This study was supported by Research Grant R01 DC03577 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, and by a National Science Foundation Graduate Research Fellowship.

References

[1] B. Cranen and L. Boves, "Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production," *J. Acoust. Soc. Am.*, 77, 1543-1551, 1985.

[2] F. Alipour and R.C. Scherer, "Pulsatile airflow during phonation: An excised larynx model," *J. Acoust. Soc. Am.*, vol. 97, pp. 1241-1248, 1995.

[3] I.R. Titze, S.S. Schmidt, and M.R. Titze, "Phonation threshold pressure in a physical model of the vocal fold mucosa," *J. Acoust. Soc. Am.*, 97, pp. 3080-3084, 1995.

[4] Y. Isogai, S. Horiguchi, K. Honda, Y. Aoki, H. Hirose, and S. Saito, "A dynamic simulation model of vocal fold vibration," in *Vocal Physiology: Voice Production, Mechanisms, and Functions*, O. Fujimura, Ed. New York: Raven Press, Ltd., pp. 191-206, 1988.

[5] Y. Kakita, "Simultaneous observation of the vibratory pattern, sound pressure, and airflow signals using a physical model of the vocal folds," in *Vocal Physiology: Voice Production, Mechanisms, and Functions*, Ed. O. Fujimura. New York: Raven Press, pp. 207-218, 1988.

[6] K. Ishizaka and J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, pp. 1233-1268, 1972.

[7] X. Pelorson, A. Hirschberg, R.R. van Hassel, A.P.J. Wijnands, and Y. Auregan, "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," *J. Acoust. Soc. Am.*, vol. 96, pp. 3416-3431, 1994.

[8] B.H. Story and I.R., "Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 97, pp. 1249-1260, 1995.

[9] F. Alipour, D.A. Berry, and I.R. Titze, "A finite-element model of vocal-fold vibration," *J. Acoust. Soc. Am.*, vol. 108, pp. 3003-3012, 2000.

[10] I.R. Titze, *Principles of Voice Production*, 2nd Printing, National Center for Voice and Speech, 2000.

[11] R.C. Scherer, D. Shinwari, K.J. De Witt, C. Zhang, B.R. Kucinschi, and A.A. Afjeh, "Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees," *J. Acoust. Soc. Am.*, vol. 109, pp. 1616-1630, 2001.

[12] J.J. Jiang and I.R. Titze, "A Methodological Study of Hemilaryngeal Phonation," *Laryngoscope*, vol. 103, pp. 872-882.

[13] E.B. Shim and R.D. Kamm, "Numerical simulation of steady flow in a compliant tube or channel with tapered wall thickness," *J. Fluids Structures,* vol. 16, pp. 1009-1027.

[14] W. Zhao, C. Zhang, S.H. Frankel, and L. Mongeau, "Computational Aeroacoustics of Phonation, Part I: Numerical Methods, Acoustic Analogy Validation, and Effects of Glottal Geometry," *J. Acoust. Soc. Am.*, vol. 112, pp. 2134-2146.

3rd International Workshop MAVEBA 2003, 151-154

NON-MODAL VOICE SYNTHESIS BY LOW-DIMENSIONAL PHYSICAL MODELS

C. Drioli[1], F. Avanzini[2]
[1]Institute of Phonetics and Dialectology, ISTC-CNR, Padova, Italy
[2]Department of Information Engineering, University of Padova, Padova, Italy

***Abstract*: The synthesis of different voice qualities by means of a low-dimensional glottal model is discussed. The glottal model is based on a one-mass model provided with a number of enhancements that make it suitable to the aim of the study. The simulation of modal and non-modal phonatory regimes is discussed. Both symmetric and non-symmetric configurations are explored. The class of models under consideration is shown to be able to reproduce a broad range of phonation styles and to provide interesting control properties.**
Keywords: physical models of vocal emission; non-modal phonation types.

I. INTRODUCTION

The possibility of reproducing different voice qualities by means of a voice synthesis tool has been explored for different applications such as emotive and natural-sounding speech synthesis [1], pathologic voice assessment [2], analysis of voice quality [3], [4], [5]. Many of the acoustic and perceptual features of an individual's voice are believed to be due to specific characteristics of the quasi-periodic excitation signal (glottal flow waveform) provided by the vocal folds. Accordingly, source models have received considerable attention and they come today in a number of versions, the most important ones being the parameterization by analytical functions, such as the LF-model [6], and the physiological modeling of the glottis, such as the multi-mass models [7], [8].

Most source models come with a set of controls to manipulate the pulse shape. The LF-model is provided with parameters for the control of the glottal pulse open phase, return phase, and closed phase durations, with parameters for the control of spectral tilt and the high-frequency content of the spectrum, and with parameters to control the diplophonia observed, for example, in laryngalized or harsh voice [3]. As for physical models, the direct control of the pulse shape is usually less simple, due to the large amount of parameters which are physically motivated but not always connected in a clear way to the characteristics of the glottal pulse. On the other hand, many authors have explored the effect of asymmetries in the mechanical components with respect to non-modal and pathological phonation types (e.g., [9]).

In this paper we explore the use of a class of low-complexity physical models loosely based on the Ishizaka&Flanagan's one- and two-mass models, and on Titze's mucosal wave model, with the specific aim of reproducing non-modal phonation modalities. The use of simplified physical models is justified by the interest raised recently in the field of natural-sounding speech synthesis, in which the possibility of generating a wide range of phonatory styles and voice qualities is highly desirable.

The paper is organized as follows. Section II gives an overview of the voice production model under investigation. In Section III the experimental setting is introduced and results from the simulation of the model are presented for both balanced and imbalanced configurations. In Section IV the conclusions are given.

II. VOICE PRODUCTION MODEL

The voice production model assumed is a source-filter scheme in which the volume velocity at the glottis is produced by a physical model and the vocal tract is represented by a parallel of four formant filters. The glottis model adopted here is a low-dimensional body-cover model in which the lower edge of the folds is represented by a single mass-spring system k; r; m and the propagation of the displacement is represented by a delay line of length T [10], [11]. The coronal cross-section of the model is illustrated in Fig. 1. The equations of the aerodynamics of the model can be found in the referenced papers and will be not repeated here. Briefly, the structure is a one-mass model with a propagation line aimed at simulating the propagation of the motion along the thickness of the fold. A second-order resonant filter represents the oscillating folds, an impact model reproduces the impact distortions on the fold displacement and adds an offset x_0 (the resting position of the folds). The driving pressure Pm acting on the folds is computed from the flow and the fold displacement using Bernoulli's law. A flow model converts the glottis area given by the fold displacement into the airflow at the entrance of the vocal tract. The glottis area is computed as the minimum cross-sectional area between the areas at lower and the upper vocal fold edge, and the flow is assumed proportional to the glottal

area. The propagation line is an approximation of the vocal cord along the thickness (vertical) direction and reproduces the vertical phase difference of the vibration of the cord edges, and it is an essential element for the production of self-sustained oscillations without a vocal tract load.

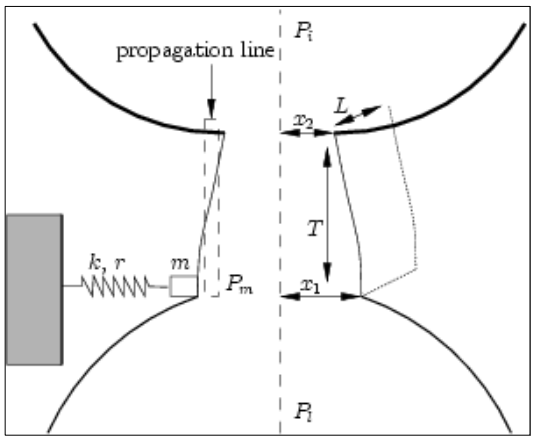


Fig. 1: Low-dimensional body-cover model of the vocal folds. From bottom to top, P_l is the lung pressure, P_m is the driving pressure acting on the vocal folds, m, k, and r represent respectively the mass, stiffness, and damping of the fold, T represents its thickness, x_1 and x_2 are the fold displacements at entrance and exit of the glottis, and P_i is the pressure at entrance of the vocal tract.

III. SYNTHESIS OF VOICED SOUNDS WITH DIFFERENT VOICE QUALITIES

The model adopted here has demonstrated to be successful in reproducing the essential dynamics of voice source and has shown to be able to reproduce real glottal flow waveforms, when extended with an opportune data-driven parametric component [10], [12]. Here we focus on the control of the phonation quality offered by this class of physical models. In particular, we look at the possibility of reproducing convincing 1) breathy, 2) pressed or creaky, and 3) bifurcated phonation types.

The differentiated glottal volume velocity produced by the model is convolved with a vocal tract filter to provide a lip pressure signal for perceptual evaluation of the synthesis.

A. Symmetric structure

A bilaterally symmetric one-delayed mass model is assumed for this section. Model refinements and strategies to produce the target voice quality modifications are described in the following.

Breathy phonation is characterized by the presence of a turbulent aspiration noise combined with the periodic component. The rendering of this phonation type is not always trivial due to the fact that the noise component has a precise phase relation with the periodic voiced component, and a white noise source added to the airflow can sometimes lead to the perception of two distinct sources. n improvement to this aspiration noise model is that of reproducing the amplitude modulation given to the noise by the opening and closing of the glottis. A noise component modulated by the airflow amplitude is thus added to the airflow. Figs. 2 b) and c) shows the result of the simulation for increasing noise component. Fig. 2 d) shows a typical situation of breathy phonation in which the glottis is never completely closed at back, and a DC component is summed to the periodic flow.

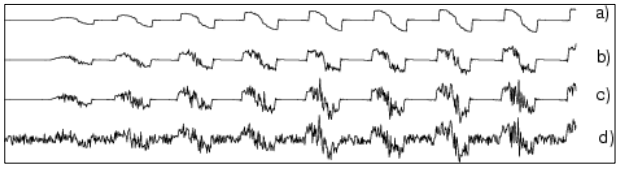


Fig. 2: Differentiated glottal flow waveforms generated by the symmetric model: a) normal; b) and c) increasing breathiness; d) breathy with dc component.

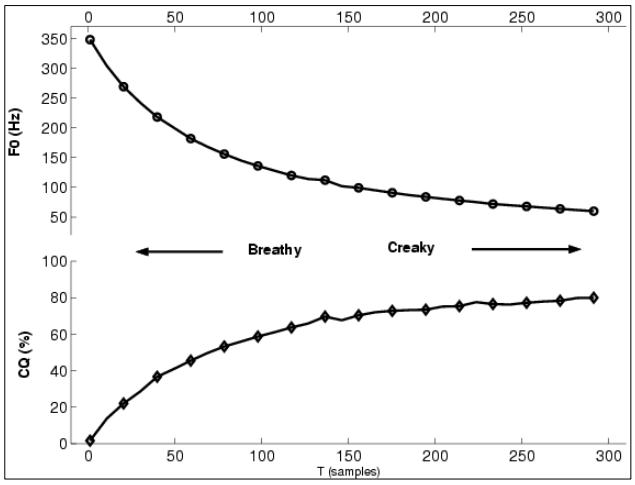


Fig. 3: Relation between the propagation line length (in samples), the resulting pitch, and the resulting closed quotient CQ.

Pressed or creaky phonation is characterized by a narrow airflow pulse (small open quotient) and by a low fundamental frequency. An action on the propagation line length showed to be an effective mean to control the pulse closed phase duration. A physiological interpretation of this parameter can be easier if we look at the propagation line length as the part of the fold actually involved in the oscillation, instead of as the thickness of the whole vocal fold. It is also to note that for all the model configurations tested, the parameter T affects the closed phase duration of the pulse as well as the pitch of the glottal pulse. Fig. 3 shows the relation between the propagation line length (in samples), the resultant pitch, and the resultant closed quotient CQ, defined as the ratio of the closed phase duration to the period length. The fold mass, damping and tension were respectively m = 0.1 g, r = 0.085 Nsm^{-1}, k = 40 Nm^{-1}, resulting in a resonance frequency f_c = 100 Hz and selectivity factor Q = 0.7441 at sampling rate 22050 Hz.

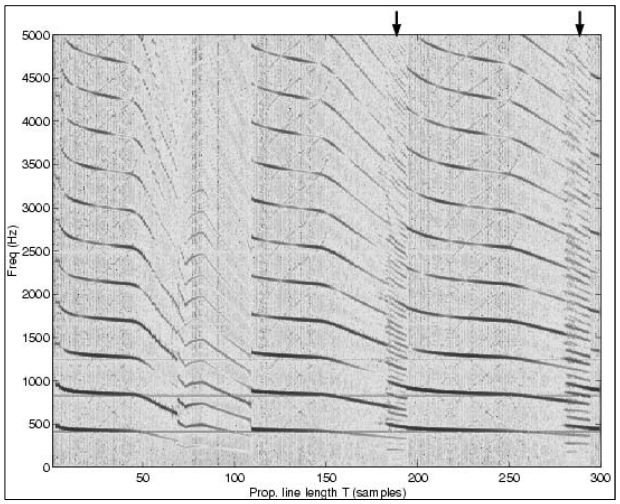


Fig. 4: Bifurcations occurring in the low-dimensional model with symmetric configuration when the length T of the delay line representing the propagation of the fold displacement is slowly varied over time.

Bifurcated phonation is characterized by the presence of period-doubling or sub-harmonics which result in large irregularities in the time domain, usually perceived as a ”rough” voice quality. Bifurcated phonation and irregularities appear occasionally in normal phonation and speech, but is often symptomatic in voice pathology. Often instabilities and sub-harmonic components are the result of tension and mass imbalance of the left and right vocal fold. Asymmetric configurations of the glottal model are explored in the next section. Even with a symmetric configuration, however, we observed the presence of such dynamic behavior when the length of the propagating delay-line was set to values extremely high with respect to the pulse duration. The fold mass, damping and tension were respectively m = 0.05 g, r = 0.002 Nsm^{-1} , k = 80 Nm^{-1} , resulting in a resonance frequency f_c = 200 Hz and selectivity factor Q = 31 at sampling rate 22050 Hz. An empirical rule for the production of bifurcations in the balanced configuration turned out to be the adoption of a considerably high Q factor. Fig. 4 shows the spectrogram of the voiced sound generated by continuously rising the propagation line length. Two clear bifurcation regions can be observed for values of T around 200 and 300.

B. *Asymmetric structure*

In this section, asymmetries are included in the low - dimensional model described in Section II. Earlier studies have already observed the phenomena arising in multi-mass models when the mechanical properties of the folds are made asymmetric [13], [14]. In particular, imbalance in bilateral tension and mass, a configuration usually related to unilateral paralysis, has been extensively explored. Typically, non-stationary regimes are observed when the mass and tension of the two folds are imbalanced. An asymmetry parameter Q_i ? (0, 1] is introduced, which is used to compute the right-hand fold mass and stiffness as $k_r = Q_i k_l$ and $m_r = m_l / Q_i$. Fig. 5 shows the simulated fold displacements for different values of Q_i. The values used for the mass-spring system for this examples were m = 0.17 g, r = 0.02 Nsm^{-1}, k = 64 Nm^{-1}, corresponding to a fundamental period of 100 Hz for the balanced configuration.

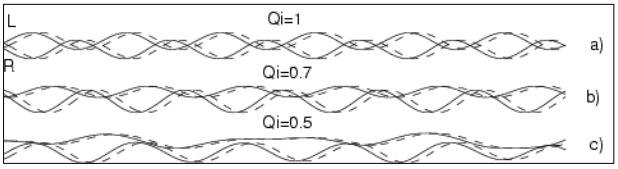


Fig. 5: Fold displacements (left and right; lower edge: solid line, upper edge: dashed line) for various imbalanced configurations. a) symmetrical; b) and c) asymmetrical.

Fig. 6 shows the spectrogram of the synthesis when the asymmetry parameter Q_i is slowly varied over time. Two bifurcation regions are clearly visible as Q_i approaches the values 0.51 and 0.55.

IV. CONCLUSIONS

The dynamic behavior of a low-dimensional one-mass model with delayed mass has been investigated, both for balanced and imbalanced configurations. In the balanced configuration normal, pressed, breathy

phonation styles were obtained, as well as bifurcation phenomena in some regions of the parametric space. In general the synthesis results were judged convincing on the basis of informal perceptual evaluations. In the imbalanced configuration, typical non-stationary and bifurcated regimes were observed. The class of low-complexity models presented is characterized by a wide variety of dynamical behaviors and offers in some cases a simple control interface to switch from modal phonation to non-modal phonatory regimes. The computational efficiency of the model suggests that this could be useful in real-time speech synthesis application.

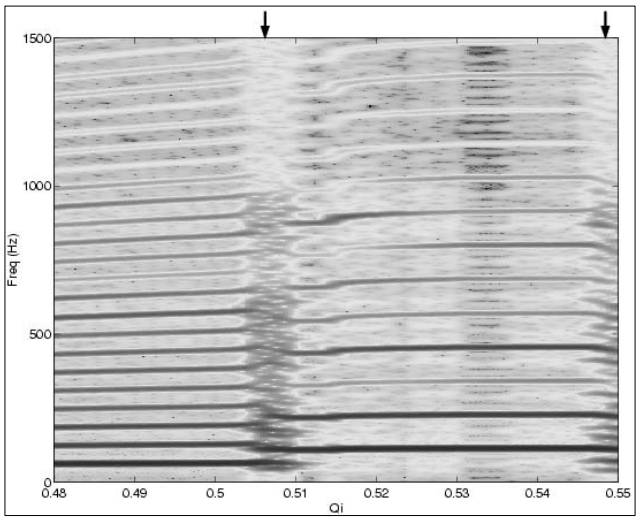


Fig. 6: Bifurcations occurring in the low-dimensional model with asymmetric configuration when the asymmetry parameter Q_i is slowly varied over time.

REFERENCES

[1] C. Gobl and A. N. Chasaide, "The role of the voice quality in communicating emotions, mood and attitude," Speech Communication, vol. 40, pp. 189–212, 2003.

[2] P. Bangayan, C. Long, A. Alwan, J. Kreiman, and B. Gerratt, "Analysis by synthesis of pathological voices using the klatt synthesizer," Speech Communication, vol. 22, pp. 343–368, 1997.

[3] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variation among female and male talkers," Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 820–857, February 1990.

[4] D.G. Childers and C.K. Lee, "Vocal quality factors: analysis, synthesis, and perception," J. Acoust. Soc. Am., vol. 90, no. 5, pp. 2394–2410, November 1991.

[5] D.G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," J. Acoust. Soc. Am., vol. 97, no. 1, pp. 505–519, January 1995.

[6] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPS, pp. 1–13, 1985.

[7] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," The Bell Syst. Tech. J., vol. 51, no. 6, pp. 1233–1268, July-August 1972.

[8] I. R. Titze, "The physics of small-amplitude oscillations of the vocal folds," J. Acoust. Soc. Am., vol. 83, no. 4, pp. 1536–1552, April 1988.

[9] K. Ishizaka and N. Isshiki, "Computer simulation of pathological vocal-cord vibration," The Bell Syst. Tech. J., vol. 60, pp. 1193–1198, 1976.

[10] F. Avanzini, C. Drioli, and P. Alku, "Synthesis of the voice source using a physically-informed model of the glottis," in Proc. of the Int. Symposium on Musical Acoustics (ISMA), pp. 31–34, 2001, available at http://www.dei.unipd.it/~avanzini/papers.html.

[11] F. Avanzini, P. Alku, and M. Karjalainen, "One-delayed-mass modelfor efficient synthesis of glottal flow," Proc. of Eurospeech Conf, pp. 51–54, September 2001, available at http://www.dei.unipd.it/~avanzini/papers.html.

[12] C. Drioli, "A flow waveform adaptive mechanical glottal model," TMH-QPSR, vol. 43, pp. 69–79, 2002, available at http://www.speech.kth.se/qpsr/tmh/.

[13] K. Ishizaka and N. Ishiki, "Computer simulation of pathological vocal-cord vibrations," J. Acoust. Soc. Am., vol. 60, no. 5, pp. 1193–1198, 1976.

[14] H. Herzel, I. Titze, and I. Steinecke, "Nonlinear dynamics of the voice: signal analysis and biomechanical modeling," CHAOS, vol. 5, no. 1, pp. 30–34, 1995.

Pathology classification

3rd International Workshop MAVEBA 2003, 157-160

THE EFFECTS OF INTER AND INTRA SPEAKER VARIABILITY ON PATHOLOGICAL VOICE QUALITY ASSESSMENT

Juan I. Godino-Llorente[1], Tim Ritchings[2], Carl Berry[2]

[1] Dept. Teoría de la Señal y Comunicaciones, Universidad de Alcalá, Campus Universitario, N II Km 33,6, 28871 Madrid, Spain.
[2] School of Computing, Science and Engineering, University of Salford, The Crescent, Salford, M5 4WT, UK

Abstract: **This paper describes some methodological issues to be considered while facing the task of the objective assessment of voice quality from patients with laryngeal cancer. Earlier research works showed that the automatic assessment of voice quality could be addressed by means of short-term and long-term time-domain, and frequency-domain parameters extracted from electroglotographic (EGG) signals, and using Artificial Neural Networks (ANN) such as Multi-layer Perceptron (MLP) [1]. However, despite the good results, further research has showed that the choice of cross-validation techniques used for the pattern recognition can greatly influence the ability of the system to learn and to generalise. In particular, this paper is concerned with the effects of intra and inter speaker variability during cross-validation and hence on the reliability of pathological voice quality assessment. For this study, a database of male subjects steadily phonating the vowel /i/ was used, and the quality of their voices was independently assessed by a speech and language therapist (SALT) according to their 7-point ranking of subjective voice quality. Although it is found that by carefully selecting the datasets used to train and validate the ANN to minimise intra speaker variability reduces the classification accuracy, most of the time the ANN only misclassifies by only one point.**

Keywords: **Voice quality, classification, neural networks, cross-validation.**

I. INTRODUCTION

The effectiveness and importance of the acoustic and EGG analysis of pathological voices have been proven by many experimental researches. The starting point of this work is that carried out in [1]. This work proposed an Artificial Neural Network (ANN) based framework to evaluate the voice quality into a 7 point scale using short term and long term parameters extracted from the EGG signal with an accuracy over 90%. Such work used 50 speakers whose EGG signal were recorded before the treatment. However, due to the limited number of patients, the training and validation datasets used to develop the ANN used multiple frames taken from the signals recorded for some of the patients. As a result the ANN learnt both the intra and inter speaker variations in the data. This could lead to artificially high classifications with small datasets, with the system effectively recognising a speaker in the dataset, rather than assessing voice quality from the parameters derived from signal recorded from different speakers.

This study has reconsidered the training and validation of ANNs used for voice quality assessment in the light of these intra and inter speaker variations

II. CROSS VALIDATION

Pattern recognition techniques by themselves do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data should be removed before training begins. Then when training is done, the data that was removed could be used to test the performance of the model on "new" data. This is the basic idea for cross validation.

The **holdout method** is the simplest kind of cross validation. The data set is separated into two sets: the training and the validation set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the unseen data in the validation set. The advantage of this method is that it takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the validation set, and thus the evaluation may be

significantly different depending on how the division is made.

K-fold cross validation improves the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the validation set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a validation and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

Leave-one-out cross validation is to split the P patterns into a training set of size (P-1) and a validation set of size 1 and average the squared error on the left-out pattern over the P possible ways of obtaining such a partition. This is called leave-one-out (LOO) cross-validation. The advantage is that all the data can be used for training -none have to be held back in a separate validation set. The evaluation given by leave-one-out cross validation error is good, but it is computationally expensive.

For this work the k-fold cross validation method has been used, because the leave-one-out was considered very time consuming.

III. DATA PROCESSING

The procedures used in extracting the parameters in this paper are broadly similar to those used in [1] which contains more detail than we will look at here, the main changes are due to the nature of the two different systems. In the earlier study there was a large amount of manual judgement and adjustment at various stages to obtain the best set of extracted parameters, because the long term aim for this work is to be used in a system used by non-technical users it was necessary to fully automate the extraction procedures, thus losing some accuracy in the process. The signal was first stationarised to remove drift, split into 50 ms. frames (Hanning windows overlapped by 25 ms.) then the autocorrelation was taken to remove some of the noise components. In the next stage silent frames were removed by comparing zero point crossing and short term amplitudes with that of a sample of silence (recorded at the same time as the data samples). Following on from this voiced and unvoiced frames were separated using a cepstral based approach described in [2], here we are looking for a pronounced peak indicating the presence of a fundamental frequency. This was originally done by a user but in the current work we attempt to detect this peak with a fairly simple peak finding algorithm, in any such attempt a trade off has to be made and after much experimentation the system errs on the side of rejecting frames as to be sure that all passed frames do actually contain speech.

After the Power Spectrum Density (PSD) is calculated for each frame the frames are pooled to create the Fundamental Harmonic Normalisation (FHN) from which we can extract some of the long term parameters, again some user interaction was previously required here but this has been replaced with a peak finding algorithm. Once both the PSD and FHN have been calculated they are both fitted with Gaussian Mixture Models (GMM) in order to reduce the number of parameters needed to describe the signals. This proved difficult to automate, especially with the more damaged voices, and the algorithm still has a tendency to try to fit Gaussians to peaks that prove not to be harmonics. Once the GMMs are fitted the parameters are extracted as in the previous work [1]. The parameters extracted are 15 short term parameters of the mean, standard deviation and amplitude of the Gaussians fitted to the fundamental frequency and the first 4 harmonics (if they exist) and 4 other short term parameters, those of the value of the fundamental frequency in each frame (Fo), the noise threshold (No), the FHN Noise Energy (FHNNE) -based on the Normalised Noise Energy (NNE) [4], but derived from the FHN spectrum- and the Residual Harmonic Error (RHE). Along with these are extracted 3 long term parameters, those of the mean fundamental frequency for a sample (MFo), a measure of the jitter of the fundamental frequency between frame (JFo) and the percentage of voiced versus unvoiced frames (V+). The data extracted from the speech data and used for the ANN classification tests comprised of 3 long-term parameters (MFo, JFo, V+) and 17 short-term parameters. Full details of the data processing and extraction of these parameters can be found in [3].

IV. THE CLASSIFIER

A widely used architecture has been used for this purpose: a three layered feedforward perceptron (MLP). The Learning algorithm used is backpropagation with adaptive momentum [4]. The training was carried out along 400 epochs. The activation function used at each node is sigmoidal, and the number of neurons of the input layer is 20, the same number as the parameters extracted. This input data was normalised to between [-1,1] before being supplied to the net. The output layer has 7 neurons that are activated for every single class.

V. DATABASE USED

The data used to develop the system was captured off-line under clinical conditions at the Christie and Withington Hospitals in Manchester, using an Electrolaryngograph PCLX system [6]. This system is used to capture electrical impedance (EGG) signals using pads placed either side of the neck synchronously with acoustic signals captured using a microphone. Both EGG and acoustic data channels were captured synchronously at 20 kHz for up to 3 seconds while the subject phonated the vowel /i/ as steadily as possible. Although speech data was recorded for both male and female patients, the largest pathological group was male, so it is these speech signals that were used in the study, For each patient the SALT made a subjective voice quality assessment using a 7-point ranking.

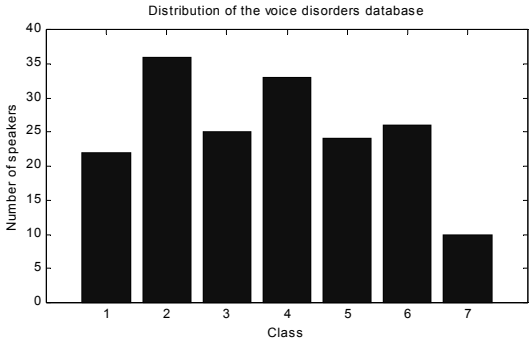


Fig. 1: Distribution of the voice disorders database

The database contains about 50 males voices recorded just before treatment, six months after treatment and one year after treatment (150 files), showing in general an improvement of the voice quality. Fig 1 represents the number of files for every class.

In the earlier study only the pre-treatment registers were used and the training and validation datasets were developed extracting the 30% of the frames for validation and 70% for training. So both datasets contained information about all the speakers stored in the database.

In this research, the pre and post-treatment registers were used. This approach ensure that the same speaker belongs to different categories, depending on the stage of the treatment. It enforces the ANN to learn the speaker independent features, and so minimise the effects of the intra-speaker variations.

The results have been obtained cross validating using the **k-fold** cross method. It is less time-consuming than the leave-one-out, but provides a good idea of the ability of the system to classify according to our criteria. 25 different datasets have been developed for every MLP size. The training datasets contain frames that belong to 7 speakers, whereas the validation dataset contains frames belonging to 3 different speakers. The pre-treatment and post-treatment recordings were mixed together in order to ensure that the system is not able to recognise speaker dependent features. This approach ensures that the ANN is forced to classify according the quality of the voice, keeping aside the features inherent to the speaker, due to the fact that the same speaker belongs to different categories depending on the stage of the treatment

VI. RESULTS

Fig 2 shows the results obtained. In the left column are shown the frame and file (the whole recording) accuracy of the system using the EGG signal parameterised as explained above. The right column shows the results using the same parameterisation approach over the glotogram extracted from the acoustic data by means of pitch asynchronous inverse filtering techniques [7]. The file accuracy has been obtained by aggregating the assessments for every frame in the file. The results have no significant variation on the MLP size, showing a better behaviour with EEG signal than with the glotogram extracted by inverse filtering.

As was expected results are worse than in the earlier study, but it can be seen that when the ANN misclassifies a speaker it generally does so by only one point on the SALT's 7 point scale.

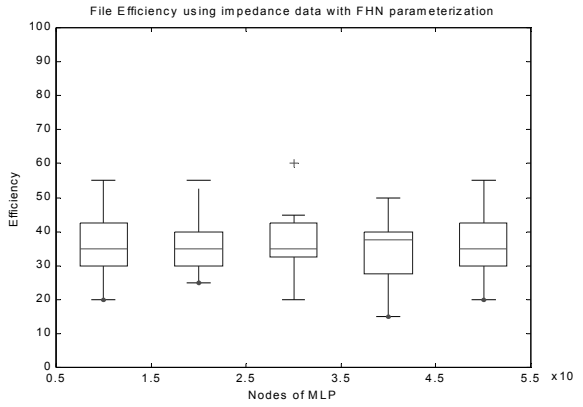
VII. CONCLUSIONS

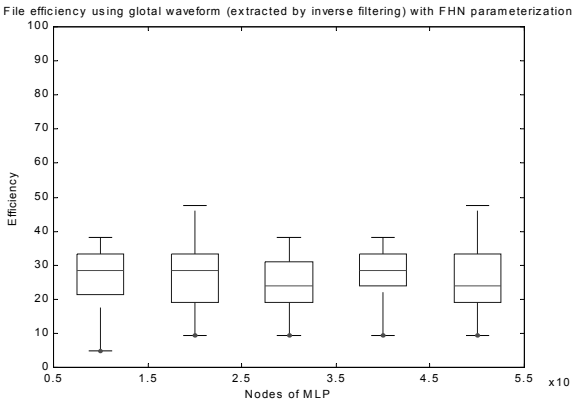
The modest scores (~40%) could either be due to the ability to discriminate the features extracted, or due to the MLP not being able to separate the prototypes correctly. However it is shown that most of the time, the classifier misclassifies by only one class. When interpreting these scores it has to be kept in mind that the SALT classifications were made by perceptual evaluation, and sometimes the experts do not agree on the evaluation of a voice sample. It is well known that there is an intra and inter-evaluator judgement variability, due to the fact that it depends a lot on their own expertise and subjective criteria about how a normal voice should be.

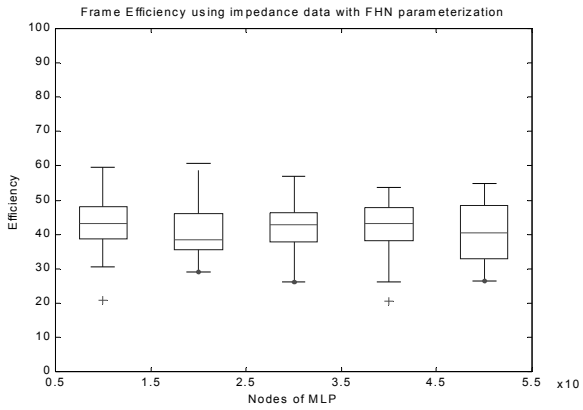
Despite of the modest scores, this system is able to provide an objective approach to the assessment of voice quality. For the future work, it should be tested with a larger database to improve the accuracy of the system, and it has to be tested using on-line with a Clinic.

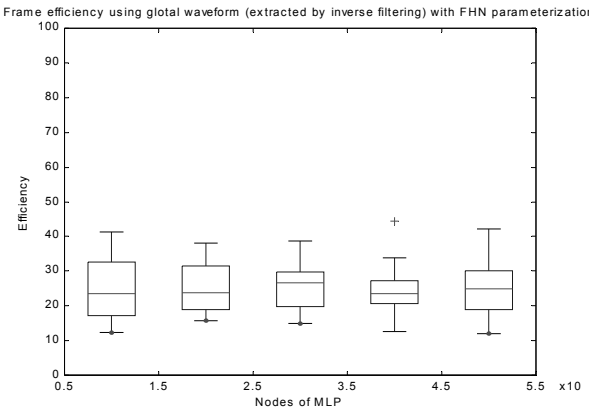
VIII. ACKNOWLEDGEMENTS

This work was supported under the grants: EPSRC GR/L51546 (UK), and PR2002-0239 from the *Secretaría de Estado de Educación y Universidades* (Spain).









Class \ Predicted Class	1	2	3	4	5	6	7
1	30	10	8	5	7	0	0
2	15	29	9	3	4	0	0
3	1	7	7	16	18	5	6
4	3	4	6	17	24	5	1
5	8	3	5	16	22	4	2
6	0	0	54	8	12	26	9
7	0	0	0	6	8	12	14

Class \ Predicted Class	1	2	3	4	5	6	7
1	4	16	18	20	2	0	0
2	2	10	15	25	8	0	0
3	0	9	16	24	9	2	0
4	4	6	14	22	12	2	0
5	3	8	10	21	11	7	0
6	0	4	5	12	14	20	5
7	0	0	1	1	6	27	25

Fig. 2: Results using EGG signal (left column) and glotogram waveform (right column) extracted from acoustic data by means of inverse filtering. It is represented the frame and file accuracy. The performance matrix shows the number of files that have been classified into each class

REFERENCES

[1] Ritchings, T., McGillon, M., and Moore, C., "Pathological voice quality assessment using artificial neural networks," Medical Engineering and Physics, vol. 24, no. 8, pp. 561-564, 2002

[2] Rabiner, L. and Juang, B. H., "Fundamentals of speech recognition" New Yersey: Prentice Hall, 1993.

[3] McGillion, M., Automated Analysis of Voice Quality. Ph.D. Thesis, UMIST, 2000

[4] Kasuya, H.; Ogawa, S.; Mashima, K.; Ebihara, S. "Normalised noise energy as an acoustic measure to evaluate pathologic voice". Journal of the Acoustic Society of America. 80(5):1329-1334, 1986

[5] Lipmann, R. "An Introduction to computing with neural nets"; IEEE ASSP Magazine April 1987

[6] Fourcin, A.J.; Abberton, E.; Miller, D.; Howell, D. "Laryngograph: Speech pattern element tools for therapy, training and assessment". European Journal of Disorders of Communication 30(2)101-115, 1996

[7] Childers, D., "Speech processing and synthesis toolboxes", 1 ed. New York: John Wiley & Sons, 2000

3rd International Workshop MAVEBA 2003, 161-164

DO ORAL CONTRACEPTIVES REALLY HAVE AN ADVERSE EFFECT ON VOICE QUALITY?

O. Amir[1], T. Biron-Shental[2]
[1]Department of Communication Disorders, Sackler Medical School, Tel Aviv University, Tel-Aviv, Israel
[2]Department of Obstetrics and Gynecology, Sapir Medical Center, Kfar Saba, Israel

Abstract: **Traditionally, oral contraceptives are considered to have adverse effect on women's voice quality. The purpose of this study was to evaluate the impact of oral contraceptives on voice quality, using acoustic analysis. Acoustic vocal parameters of seven women who use oral contraceptives and seven women who do not were measured repeatedly during the menstrual cycle. Repeated-measure analyses-of-variance were performed to test for group differences. Results did not reveal an adverse effect in the oral contraceptive users group. Moreover, amplitude and frequency perturbation, as well as noise-to-harmonics ratio values within the test group were found to be significantly lower than those observed among the control group; indicating a more stable voice quality.**
Keywords: **voice, vocal-quality, perturbation, hormones, oral-contraceptives**

I. INTRODUCTION

The interaction between the human larynx and ovarian hormones has been previously demonstrated. Several researchers have discovered receptors for androgen, estrogen and progesterone in the gingival epithelium [1] and in the vocal folds [2,3]. The effect of these sex hormones on the human voice has been previously demonstrated in different cases of endocrine dysfunction. Such vocal changes include increase in vocal instability, hoarseness and pronounced lowered pitch [4].

Vocal changes related to hormonal balance were also reported in relation to the menstrual cycle in healthy women. Specifically, vocal changes were observed either at the premenstrual phase [5,6] or shortly prior to ovulation [7]. It should be noted that these changes in vocal quality were reported primarily among professional voice users and less so among non-professionals voice users. While the mechanism underlying these voice changes is not clear yet, it is assumed that the physiological changes which occur during the menstrual cycle impact voice quality. For example, the lowered pitch during the premenstrual phase is assumed to be the result of the edema and venous dilatation observed in the vocal folds at that time [4]. It was also suggested that changes in ovarian hormones levels could affect laryngeal neuromotor control [7], which in turn, could affect vocal stability and quality. It is interesting to note that the two phases along the menstrual cycle in which vocal changes were reported in the literature (premenstrual phase and prior to ovulation) are also marked by a significant and abrupt change in hormonal balance.

Oral contraceptives present an exciting setting in which hormones could interact with voice production. Most modern birth control pills are designed to maintain constant levels of estrogen and progesterone along the menstrual cycle therefore preventing ovulation. Because birth control pills maintain a steady hormonal balance, it seems logical to expect that women who use birth control pills will show diminished voice changes along the menstrual cycle, in comparison with women who do not use the pill. However, review of the literature on the effect of birth control pills on voice quality revealed only a limited number of studies which addressed this question. These studies included occasional reports of adverse androgenic voice changes (e.g. virilization, and mainly lowered pitch) among women who use the pill [8], and were explained by the relatively high hormonal doses used in birth control pills in the 1960s and 1970s [9]. Based on these studies, otolaryngologists, voice specialists and speech-language pathologists generally suggest that professional voice-users refrain from using oral contraceptives [10]. In addition, using the pill is typically regarded as a risk factor when clinically evaluating voice disorders.

Modern birth control pills, however, contain markedly lower doses of estrogens and progesterones with less androgenic metabolites. Thus, smaller androgenic effect can be expected. In a previous study that evaluated voice quality among women who use low-dose oral contraceptive formulations, using subjective measures [8], no voice changes were reported as a result of using the pill. Recently, two preliminary studied were reported, that compared acoustic parameters of voices produced by women who use the pill with voices of women who do not [11,12]. Results of these studies revealed no adverse effect on the voices of those women who use the pill. Moreover, amplitude perturbation (shimmer) and frequency perturbation (jitter) values among the pill-users were reported to be lower than those observed among the non-users. The purpose of the present study was to expend on the scarce knowledge regarding the effect of modern low-dose oral contraceptives on the voices of healthy women, through the use of acoustic-analysis evaluation.

II. METHODOLOGY

A group of young and healthy women volunteered to serve as participants in this study. Seven women who

used birth control pills (Pill group) and Seven women who did not (Control group) were selected from the initial group of 30 students at Tel-Aviv University according to the criteria described bellow. The Pill group had a mean age of 23.96 years (range: 22-26), a mean weight of 58.29 kg (range: 52-70) and a mean height of 166.8 cm (range: 160-174). Three of the seven women in the Pill group used the oral contraceptive Meliane© (Schering AG, Berlin, Germany) with 0.075 mg gestodene and 0.02 ethinylestradiol; and two women used Harmonet© (Newbridge CO. Kildare, Ireland) which has identical formulation to Meliane©. One women used Gynera© (Schering) with 0.075 mg gestodene, and 0.03 mg ethinylestradiol, and one used Microdiol© (Oragon International Inc., Roseland, NJ) with 0.15 mg desogestrel and 0.03 mg ethinylestradiol. Since the four preparations used by the women in the Pill group are so similar in composition, and because all these compositions are considered low-dose formulations, it was decided to regard them as one group. All women in this group reported no omission in pill taking during the time of the study and the three preceding months. The Control group consisted of seven women who did not use any hormonal contraceptive prior or at the time of the study. Mean age for this group was 22.00 years (range: 20.3-24.5), mean weight was 54.57 kg (range: 45-65) and mean height was 165.6 cm (range: 155-173).

All women ruled out any speech or voice disorder and were also assessed by two experienced speech-language pathologists to confirm the diagnosis. None of the women had a history of formal voice or singing training, as well as smoking or substance abuse. In addition, no history of pregnancies, hormonal imbalances and neurological problems were reported. All women reported regular menses and menstrual cycle of 28-32 days.

All women were recorded repeatedly over a period of 35-40 days. While our preliminary observations did not reveal a significant effect for menstruation-cycle phase, we still decided to consider it as a possible confounding factor, based on previous reports in the literature [5-7]. Thus, each subject's menstruation cycle was divided into six consecutive and equal intervals. The six intervals were defined such that interval 1 includes the days of the menses and interval 6 includes the four days preceding the following menses. The remaining days of the menstrual cycle were divided equally to four intervals: 2 to 5 respectively. Each woman was recorded at least twice during each interval, totaling approximately 20 recording sessions for each participant.

Each recording session consisted of two recordings of the Hebrew vowel /i/ (similar to the vowel in the word "heed") and two recordings of the vowel /a/ (similar to the vowel in the word "father") in isolation. Each vowel was produced for 3-5 seconds, in a random order that was changed between subjects and sessions. Recordings were performed individually in a quiet room. Signal was recorded onto TDK (Tokyo, Japan) data cartridges, using a Sony TCD-D100 (Tokyo, Japan) digital audio tape recorder and a Sony ECM-T150 headset-microphone. Sampling rate for the recording was set at 44.1 kHz. Acoustic analyses were performed after each recorded vowel was fed to a voice analysis computer program (Multi Dimensional Voice Profile-MDVP, model 5105, Ver. 2 [Kay Elemetrics, Lincoln Park, NJ]).

Four acoustic parameters were measured for each vowel production: *Mean fundamental frequency* (F0), which quantifies the number of complete cycles produced by the vocal folds per second; *Jitter,* which quantifies frequency instability (perturbation) along the voice sample; *Shimmer*, which quantifies amplitude instability (perturbation) along the voice sample; and *Noise-to-Harmonics Ratio* (NHR) which compares the ratio between the aperiodic to periodic components in the voice signal. Note that for *Jitter*, *Shimmer* and *NHR*, lower values typically represent a healthier voice, whereas higher values are generally associated with less stable and lower quality voice [13].

Statistical analyses were performed using four separate repeated-measure analyses of variance; one for each acoustic parameter. The two vowels (/i/ and /a/) and the six menstrual-cycle intervals (1 through 6) were treated as repeated factors, while Group (Pill versus Natural) was regarded as the between-subject factor.

III. Results

A. Group Differences

Based on the individual data collected, group means were calculated for each acoustic parameter at all six intervals and two vowels. Table 1 presents these data. As can be seen, jitter, shimmer and NHR values in the Pill group were generally lower than those observed in the Control group, while F0 values were generally higher in the Pill group.

Statistical analysis revealed a significant group difference across all intervals and vowels for *jitter* ($F_{1,12} = 6.29$, $P = 0.027$), *shimmer* ($F_{1,12} = 7.32$, $P = 0.019$) and for *NHR* ($F_{1,12} = 5.47$, $P = 0.037$). Group differences for the F0 parameter were found to be non-significant ($P > 0.05$); yet in most conditions, the Pill group had a slightly higher F0 mean values than the Control group.

B. Menstrual-Cycle Interval Differences

The effect of menstrual-cycle interval was tested across the six intervals for each parameter. No statistically significant differences were found among the six intervals for either of the acoustic parameters measured ($P > 0.05$). In addition, no significant Group X Interval was found for any of the parameters ($P > 0.05$).

Table 1. Mean and standard deviation (in parentheses) for Fundamental Frequency (F0), Jitter, Shimmer and Noise-to-Harmonics Ratio (NHR) of the Pill (P) and Control (C) groups for the Vowels /a/ and /i/ at Each of the Six Menstruation-Cycle Interval

			Interval					
Vowel	Parameter	Group	1	2	3	4	5	6
/a/	F0 (Hz)	P	214.92 (17.29)	216.81 (19.98)	217.32 (23.31)	218.59 (23.55)	212.57 (15.52)	211.95 (21.15)
		C	211.63 (29.37)	214.93 (26.49)	213.17 (29.23)	212.20 (28.68)	215.97 (.97)	214.02 (30.85)
	Jitter (%)	P	1.00 (.52)	.76 (.35)	.81 (.37)	.89 (.29)	.83 (.24)	.89 (.29)
		C	1.34 (.35)	1.51 (.38)	1.26 (.35)	1.25 (.22)	1.43 (.32)	1.39 (.36)
	Shimmer (%)	P	3.33 (1.13)	2.37 (.50)	2.66 (.46)	2.70 (.24)	3.10 (1.18)	2.77 (.44)
		C	4.09 (1.44)	3.99 (1.26)	3.96 (.95)	3.93 (1.13)	4.10 (1.03)	4.01 (1.23)
	NHR	P	.127 (.013)	.118 (.012)	.122 (.013)	.122 (.019)	.133 (.018)	.13 (.019)
		C	.133 (.010)	.129 (.010)	.135 (.013)	.136 (.012)	.140 (.014)	.13 (.008)
/i/	F0 (Hz)	P	223.19 (26.15)	227.07 (25.99)	228.60 (27.66)	228.39 (28.33)	222.40 (20.81)	218.89 (26.96)
		C	220.06 (25.67)	222.54 (20.86)	221.21 (25.09)	221.18 (21.99)	224.70 (27.03)	219.81 (30.61)
	Jitter (%)	P	1.39 (.43)	1.46 (.41)	1.23 (.53)	1.25 (.35)	1.37 (.52)	1.27 (.67)
		C	1.87 (.40)	1.54 (.70)	1.69 (.37)	1.77 (.70)	1.62 (.53)	1.89 (.91)
	Shimmer (%)	P	2.80 (.62)	2.50 (.41)	2.68 (.76)	2.39 (.32)	2.44 (.26)	2.42 (.51)
		C	3.88 (1.56)	3.44 (1.53)	3.78 (1.57)	3.38 (1.22)	3.52 (1.68)	3.59 (1.46)
	NHR	P	.128 (.024)	.117 (.019)	.109 (.015)	.117 (.027)	.125 (.013)	.11 (.017)
		C	.129 (.026)	.123 (.015)	.124 (.009)	.123 (.018)	.125 (.017)	.13 (.028)

C. Vowel Differences

Vowel difference between /a/ and /i/ was not defined as a primary research question in this study. However, since differences in fundamental frequency and other acoustic measures were previously demonstrated to be vowel related, it was decided to include Vowel as a possible confounding factor in the analyses. As expected *F0* was significantly higher for the vowel /i/ compared to /a/ ($F_{1,12}$ = 15.00, P = .002). This vowel difference is in keeping with previously established data [13]. Statistically significant vowel differences were also found for *jitter* ($F_{1,12}$ = 20.56, P = .001) and for *NHR* ($F_{1,12}$ = 20.962, P = .001). Specifically, jitter values for the vowel /i/ were significantly higher than for the vowel /a/, while NHR values were significantly lower for the vowel /i/. These results are also in keeping with the literature [13], thus help to validate the current results. No vowel difference was found for the shimmer parameter ($P > 0.05$).

In order to approximate the follicular and secretory phases within menstrual cycle, data were rearranged, collapsing intervals 1 through 3, and 4 through 6. Statistical analyses of the modified set of data yielded identical results for all comparisons reported above, for group, interval and vowel differences.

IV. Discussion

The results of the present study did not reveal an adverse effect of birth control pills on voice. Moreover, our results indicate that in all four acoustic parameters tested, women who use the pill performed better than the women in control group. Specifically, women in the pill group demonstrated reduced amplitude and

frequency perturbation (*shimmer* and *jitter*) and had lower *NHR* values which represent a clearer voice. Lower values of these three parameters are regarded as indication of a more healthy voice [13]. These results can be interpreted to show that the more stable voice quality presented by the women in the pill group could be attributed to a more stable hormonal balance which is maintained by the oral contraceptives they use. In contrast, women in the control group are affected by the natural changes in serum levels of estrogen and progesterone which occur during the menstrual cycle. The hormonal changes along the menstrual cycle induce histological changes in muscles, mucus and laryngeal glandular cells; hence these women's voice quality is less stable [4].

Oral contraceptives are traditionally viewed by voice professionals as potentially hazardous for the female voice [10]. The main reason for this view is the concern from androgenic effect of progesterone derivates on the female larynx. The most common effect caused by androgens to the female voice is virilization, which is primarily characterized by lowered pitch (F0). Our results indicate that women who use the pill did not exhibit any lowering in fundamental frequency. In fact, *F0* values for the Pill group were generally higher than those observed in the control group, although these group differences did not reach statistical significance. The reason for the contradiction between the current results and the traditional view of oral contraceptives as a potential hazard, stem probably from the difference between the formulations used in pills in the past and the low-dose formulation which are commonly used presently. Based on these results, it is suggested that the traditional approach towards oral contraceptives as a potential risk factor for voice, should be reevaluated. It should be kept in mind, though, that our participants were not professional voice users or performers, hence it is possible that somewhat different results could be observed within that specific population.

The results presented here are in agreement with the Wendler et al study [8] who reported no adverse voice effect associated with low-dose pills. However, while their results were drawn from subjective evaluation made by listeners, the current results are based on acoustic measurements that are more reliable and are potentially sensitive to small physical differences. The current results are also in agreement with the two preliminary studies that were conducted using similar methodologies but utilizing a different voice analysis program [11,12]. The relation between the acoustic results presented here and subjective evaluation of voice quality should be also further explored.

V. Conclusion

The present study utilized acoustic tools to examine the effect of oral contraceptives on voice quality. Results challenge the traditional approach which views oral contraceptives as a potential risk-factor for voice. Based on the results presented here and in two recently published studies that used similar methodologies, it appears that low-dose monophasic oral contraceptives were not found to negatively affect voice quality. Instead, the four parameters that were included in the analysis improved among the women who used the pill. Obviously, further study is needed to better understand the interaction between female hormonal balance and voice quality, as well as the effect of different oral contraceptive formulations (for example, monophasic versus multiphasic) on voice production and quality.

References

[1] J. Vittek, M.R. Hernandez, E.J. Wenk, S.C. Rappaport and A.L. Southren, "Specific estrogen receptors in human gingival," *J. Clin. Endocrin & Metab.* Vol. 54, pp. 806-612, 1982.

[2] J. Abitbol, P. Abitbol and B Abitbol, "Sex hormones and the female voice", *J. Voice*, vol. 13, pp. 424-46, 1999.

[3] S.R. Newman, J. Butler, E.H. Hammond and S.D. Gray, "Preliminary report on hormone receptors in the human vocal fold", *J. Voice*, vol. 14, pp. 72-81, 2000.

[4] J. Abitbol , J. de Brux, G. Millot, M.F. Masson, O.L. Minoun, H. Pau and B. Abitbol, "Does a hormonal vocal cord cycle exist in women? Study of vocal premenstrual syndrome in voice performers by vidoestroboscopy-glottography and cytology on 38 women", *J. Voice*, vol. 3, pp. 157-62, 1989.

[5] J. Abitbol and B. Abitbol, "The voice and menopause: The twilight of the divas", *Contracept. Fertil. Sex*. Vol. 26, pp. 649-55, 1988.

[6] C.B. Davis and M.L. Davis, "The effect of premenstrual syndrome (PMS) on the female singer", J. *Voice*, vol. 7, pp. 337-53, 1993.

[7] M.B. Higgins and J.H. Saxman, "Variation in vocal frequency perturbation across the menstrual cycle", *J. Voice*, vol. 3, pp. 233-243, 1989.

[8] J. Wendler, C. Siegert, P. Schelhorn, G. Klinger, S. Gurr, J. Kaufmann, S. Aydinlik and T. Braunschweig, "The influence of Microgynon® and Diane-35®, two sub-fifty ovulation inhibitors on voice function in women". *Contracep*. Vol. 52, pp. 343-8, 1995.

[9] L. Speroff, R.H. Glass and N.G. Kase NG, *Clinical gynecologic endocrinology and infertility*, 6th ed. Baltimore: Lippincott, Willimans & Wilkins, 1999.

[10] D.C. Rosen and R.T. Sataloff, *Psychology of voice disorders,* San Diego: Singular Publishing, Inc., 1997.

[11] O. Amir, L. Kishon-Rabin and C. Muchnik, "The effect of oral-contraceptives on voice: Preliminary observations", *J. Voice*, vol. 16, pp. 267-273, 2003.

[12] O. Amir, T. Biron-Shantal, C. Muchnik and L. Kishon-Rabin, "Do oral contraceptives improve vocal quality? Limited trial on low-dose formulations", *Obstet. Gynecol*., vol. 101, pp. 773-777, 2003.

[13] R.J. Baken, *Clinical measurement of speech and voice,* 2nd ed. London: Singular Pub., 1997.

3rd International Workshop MAVEBA 2003, 165-168

A SUGGESTED METRIC FOR CEPSTRAL ARMA BASED SPEECH CLASSIFICATION

F. Martínez, A. Guillamón and J.J. Martínez
Department of Applied Mathematics, Universidad Politécnica de Cartagena , Spain

Abstract: **In this paper, we purpose a theoretical development of a metric for speech classification based on cepstral features obtained from ARMA models. Thus working with an *ARMA model* as a complex rational function, is possible to define a metric *d(M,M´)* between two stable ARMA models *M, M´* by means of the *cepstrum coefficients* of the models. This metric may be calculated algorithmically as a finite sum in the pole-zero domain. We suggest that the metric can be used in at least two circumstances: first, we might a large number of signals that come from various types of pathological sources and we wish to classify them; alternatively, we might the underlying models M_i corresponding to several pathological voices and we wish to classify a voice (modeled as *M*, say) from one of those. In that case, we compute *d(M,M$_i$)* for each *i* and we guess the *(M$_i$)* closest to the model *M*.**

Keywords: **ARMA model, cepstrum, distance measure, classification, pathological voice**

I. INTRODUCTION

All speech/speaker classifier include a signal processing that converts a speech waveform into features useful for further processing and a decision rule based on a metric [1].

Recent suggestions show that speech production may be a nonlinear process, see [2, 3, 4]. These authors assume the rather natural hypothesis that nonlinear processes occur in speech production, due to: turbulent air flow produced in the vocal tract; nonlinear neuro-muscular processes that should occur at the level of vocal cords and the larynx; nonlinear coupling, during speech production, between different parts of the vocal tract. On the other hand, the "cepstrum" represents a transformation on the speech signal with two important properties:

- Representatives of component signals are *separated* in the cepstrum.
- Representatives of component signals are *linearly combined* in the cepstrum.

In this way, the cepstral coefficients provide an efficient computation of the log-spectral distance of two frames [5].

To study the problem of to obtain a decision rule we suggest a decision techniques based on the computation of a distance, which quantifies the degree of dissimilarity between the features vector associated with pairs of events.

Taking into account this consideration, in this paper, we purpose a theoretical development of a metric for speech classification based on cepstral features obtained from ARMA models.

It is worth pointing out that the application of methods of classical speech processing to the analysis of medical speech signals during the last years and to date have been dealt by many research groups.

Section II deals with the methods to obtain the previous metric. This metric may be calculated algorithmically as a finite sum in the pole-zero domain. Section III presents the summary and conclusions of the paper.

II. METHODOLOGY

Cepstral analysis is used in a variety of applications such as speech processing, radar and sonar, etc. Another area in which cepstra shown up, is that of distance measures between models and/or signals. For requirements of invariance with respect to the measurement scale, it is desirable that distance to be a function of the ratio between the spectra of processes, i.e., of the difference between the cepstra. Several cepstral distances for ARMA models were defined in [6], [7], [8].

It is said the time series (x_n) follows *ARMA(p,q) model* if

$$x(n) = -\sum_{k=1}^{p} a_k \cdot x(n-k) + G \cdot \sum_{l=0}^{q} b_l \cdot u(n-l), b_0 = 1 \quad (1)$$

where (u_n) are unknown input elements, G is the gain of the model and (a_k) and (b_l) are the ARMA parameters with $b_0=1$. Alternatively, if we work in the z-domain the system function is

$$F(z) = G\frac{\sum_{l=0}^{q} b_l \cdot z^{-l}}{\sum_{k=0}^{p} a_k \cdot z^{-k}} = G\frac{\prod_{l=1}^{q}(1-\beta_l z^{-1})}{\prod_{k=1}^{p}(1-\alpha_k z^{-1})}, a_0 = 1 \quad (2)$$

where $a_0=1$ and , (α_k) and (β_l) are the poles and zeros of the model, respectively.

We can associate an ARMA model with a nonzero complex rational function. The cepstrum is the inverse Fourier Transform of the logarithm of the power spectrum and, for ARMA models, this will reduce to

$$Log\left[F(z)F^*(1/z)\right] = \sum_{n=-\infty}^{+\infty} c_n z^{-n} \quad (3)$$

where $F(z)$ is system function of ARMA model, [6]. The (c_n), that form a hermitian sequence $c_n^* = c_{-n}$, are the *cepstrum coefficients*.

Note 1: An alternative interpretation of the cepstrum of an ARMA model is given by the next result: If $F(z)$ is the transfer function of a stable and minimum phase ARMA model, then the cepstrum coefficients are the coefficients of the Laurent expansion of function $Log\left[F(z)F^*(1/z)\right]$ that is valid on an open annulus $r_1 < |z| < r_2$, with $0<r_1<1$ and $r_2>1$, [11].

Note 2: For stable and minimum phase ARMA models, the cepstrum is causal, i.e., $c_n = 0, \forall n < 0$.

The definition for the distance between two stable and minimum phase AR models given by [6] can be adapted to the ARMA case.

Definition 1: *For stable and minimum phase ARMA models M, M′ with system functions F and F′ and cepstrum coefficients (c_n) and (c'_n), respectively the metric d is defined as*

$$d(F,F') = \left(\sum_{n=1}^{\infty} n^2 \left|c_n - c'_n\right|^2\right)^{\frac{1}{2}} \quad (4)$$

Note 3: *d is a pseudometric,* because if (c_n), (c'_n) are such that $c_n = c'_n \quad \forall n \in N$ and $c_0 \neq c'_0$, then $d(F.F')=0$. There is a standard method of turning a pseudometric space into a metric space, [9].

The associated cepstral norm for an ARMA model is defined as follows.

Definition 2: *For stable and minimum phase ARMA model M with system function F and cepstrum coefficients (c_n), the cepstral norm of this model is defined as*

$$\left\|LogF\right\|_{cep^2} = \left(\sum_{n=1}^{\infty} n^2 \left|c_n\right|^2\right)^{\frac{1}{2}} \quad (5)$$

Note 4: Let (c_n) the cepstrum coefficients of ARMA model M and consider the double infinite Hankel matrix H

$$H = \begin{pmatrix} \sqrt{1}c_1 & \sqrt{2}c_2 & \sqrt{3}c_3 & \dots \\ \sqrt{2}c_2 & \sqrt{3}c_3 & \sqrt{4}c_4 & \dots \\ \sqrt{3}c_3 & \sqrt{4}c_4 & \sqrt{5}c_5 & \dots \\ \dots & ,,, & \dots & \dots \end{pmatrix} \in C^{\infty\times\infty} \quad (6)$$

The Hilbert-Schmidt norm of H is given by

$$\left\|H\right\|_{HS} = \left(tr\left(HH^T\right)\right)^{\frac{1}{2}} = \left(\sum_{n=1}^{\infty} n^2 \left|c_n\right|^2\right)^{\frac{1}{2}} \quad (7)$$

Because the right –hand side is equal to the cepstral norm defined in (5), we obtain $\left\|LogF\right\|_{cep^2} = \left\|H\right\|_{HS}$.

The Hilbert-Schmidt norm of that Hankel matrix, i.e., the Hilbert-Smchmidt-Hankel norm of $F(z)$, denoted by $\left\|F(z)\right\|_{HSH}$, is related to the poles and zeros of system function $F(z)$ in the following result.

Theorem 1: *Let F(z) be the system function of a stable and minimum phase ARMA(p,q) model M. The cepstral norm of F(z) is then equal to*

$$\left\|LogF\right\|_{cep^2} = \left(\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\alpha_i\alpha_j^*}{1-\alpha_i\alpha_j^*} + \sum_{i=1}^{q}\sum_{j=1}^{q}\frac{\beta_i\beta_j^*}{1-\beta_i\beta_j^*} - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\mathrm{Re}\left(\frac{\alpha_i\beta_j^*}{1-\alpha_i\beta_j^*}\right)\right)^{\frac{1}{2}} \quad (8)$$

where (α_k) and (β_l) are the poles and zeros of the model, respectively, and $\mathrm{Re}(\cdot)$ denotes the real part.

Proof: An equation, relating the cepstrum coefficients with the poles of an AR model, is well know in the speech processing literature, [6]. A similar equation, for an ARMA model can also be derived in a straightforward manner as follows. This equation may be stated as follows. For a stable and minimum phase ARMA model with system function given by (2), $Log[F(z)]$ is an analytic function in the open annulus $r < |z|$, $0<r<1$, [10], and can be represented by the Laurent expansion

$$Log[F(z)] = \sum_{n=0}^{\infty} c_n z^{-n} \quad (9)$$

Expressing the system function *F(z)* in terms of its poles and zeros and using the identity:

$$\sum_{n=1}^{\infty} \frac{\alpha^n}{n} = -Log(1-\alpha), |\alpha| < 1$$

we obtain

$$Log[F(z)] = LogG + \sum_{k=1}^{p}\sum_{n=1}^{\infty} \frac{\alpha_k^n}{n} z^{-n} - \sum_{l=1}^{q}\sum_{n=1}^{\infty} \frac{\beta_l^n}{n} z^{-n} \quad (10)$$

By comparing equations (9) and (10) we get

$$c_n = \begin{cases} LogG & n = 0 \\ \sum_{k=1}^{p} \frac{\alpha_k^n}{n} - \sum_{l=1}^{q} \frac{\beta_l^n}{n} & n > 0 \end{cases} \quad (11)$$

Now we can express the cepstral norm of *F(z)* that is defined in (2) in terms of its poles and zeros

$$\|LogF\|_{cep^2}^2 = \sum_{n=1}^{\infty} n^2 |c_n|^2 = \sum_{n=1}^{\infty} n^2 c_n c_n^* =$$

$$= \sum_{n=1}^{\infty} n^2 \left(\sum_{k=1}^{p} \frac{\alpha_k^n}{n} - \sum_{l=1}^{q} \frac{\beta_l^n}{n} \right) \left(\sum_{k=1}^{p} \frac{\alpha_k^{*n}}{n} - \sum_{l=1}^{q} \frac{\beta_l^{*n}}{n} \right) =$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=1}^{p}\sum_{j=1}^{p} (\alpha_i \alpha_j^*)^n + \sum_{i=1}^{q}\sum_{j=1}^{q} (\beta_i \beta_j^*)^n - \right.$$

$$\left. - \sum_{i=1}^{p}\sum_{j=1}^{q} (\alpha_i \beta_j^*)^n - \sum_{i=1}^{p}\sum_{j=1}^{q} (\alpha_i^* \beta_j)^n \right)$$

Finally, using the identity $\sum_{n=1}^{\infty} \alpha^n = \frac{\alpha}{1-\alpha}, |\alpha < 1|$ and the properties of complex conjugation, we obtain equation (8).

It is important that the expression (5) reduces to a finite sum in the pole-zero domain because it shows that the infinite sum (5) converges.

Note 5: Consider two stable and minimum phase ARMA models of order *p, q* and *p´, q´* with system functions *F(z)* and *F´(z)* and cepstra c_n and c_n', respectively. The cepstral distance between the ARMA models *M and M´* is defined in (4) as

$$d(F, F') = \left(\sum_{n=1}^{\infty} n^2 |c_n - c'_n|^2 \right)^{\frac{1}{2}}$$

The sequence $c_n - c_n', \forall n \in Z$ is the cepstrum of stable and minimum phase system function $\frac{F_1(z)}{F_2(z)}$. Consequenly, the distance between the ARMA models *M and M´* is

$$d(F_1, F_2) = \left\| Log \frac{F_1(z)}{F_2(z)} \right\|_{cep^2}$$

and applying (8) the distance value can be obtained by means of a finite sum in the domain pole-zero.

As an example, we take the fifth-order ARMA model with poles $0.9 \pm 0.1i$, $0.2 \pm 0.8i$ and -0.95 and zeros $-0.5 \pm 0.82i$, $0.1 \pm 0.7i$ and 0.92. The cepstral norm has been calculated by computing the finite sum (8), obtaining a value of 5.089. In the other hand, this cepstral norm may be calculated algorithmically as a truncated sum. Table 1 shows the magnitude of resulting error in the sum of this series when is truncated after *N* terms.

Table 1.

N	*50*	*75*	*100*	*125*
Error	*0.0281*	*0.0033*	*0.0004.*	*0.0001*

III. CONCLUSION

In this paper, we have shown that the ceptral distance between two stable and minimum phase ARMA models that was introduced by [6] may be calculated algorithmically as a finite sum in the pole-zero domain. We suggest that the metric can be used in the area of modeling and analysis of pathological voice in at least two circumstances. First, we might a large number of signals that come from various types of pathological sources and we wish to classify them. Having first fitted ARMA models to each signal, we could construct a distance matrix, that is, a matrix *D* whose *(i,j)th* element is the distance between the models of *ith* and *jth* signals. By performing the cluster analysis on *D,* the signals are classified. Alternatively, we might the underlying models M_i corresponding to several pathological voices and wish to classify a voice (modeled as *M*, say) from one of those. In that case, we compute $d(M,M_i)$ for each *i* and we guess the (M_i) closest to the model *M*.

REFERENCES

[1] R.L. Klevans and R.D. Rodman, *Voice Recognition*., Artech House, 1997

[2] M. Banbrook and S. McLughlin, "Is speech chaotic? Invariant geometrical mesures for speech data", *IEEE Colloquium on Exploiting Chaos in Signals Processing,* 8/1-8/10, 1994.

[3] A. Kumar and S.K. Mullick, "Attractor dimension, entropy and modeling of speech time series", *Electronics Letters*, Vol.26, No.21, 1990, pp. 1790-1791.

[4] H.N. Teodorescu,, F. Grigoras and V. Apppei, "Nonlinear and nonstacionary Processes in Speech Production", *Int. J. of Chaos Theor. And Appl,* Vol. 5, No. 3, 1996, pp. 1453-1457.

[5] Papamichalis, P.E. *Practical Approaches to Speech Coding.* Prentice Hall, Englewood Cliffs, NJ, 1987

[6] L.B. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1998.

[7] R. J. Martin, "A Metric for ARMA Processes", *IEEE Transition and signal processing,* vol. 48, pp. 1164-1170, 2000.

[8] K. D. Cock and cols., "On a cepstral norm for an ARMA model and the polar plot of the logarithm of its transfer function", *Signal Processing,* vol. 83, pp. 439-440, 2003.

[9] I.J. Madox, *Elements of functional analysis,* Cambridge University Press, 1970.

[10] A. Kaderli and A. Salim Kayham., "Spectral Estimation of ARMA Processes using ARMA-Cepstrum Recursion"*, IEEE Signal Processing Letters,* vol. 7, pp. 259-261, 2000.

[11] M.K. Ablowitz and A.S. Fokas, *Complex Variables. Introduction and Applications,* Cambridge University Press, 1997.

3rd International Workshop MAVEBA 2003, 169-172

PERCEPTUALLY-BASED OBJECTIVE MEASURE FOR NON-INTRUSIVE SPEECH QUALITY ASSESSMENT

D. Picovici and A.E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

Abstract: **This paper proposes a new perceptually-based method for assessing speech quality and evaluates its performance. The method is based on comparing the received speech to an appropriate reference representing the closest match from a pre-formulated codebook. The codebook holds a number of optimally clustered speech parameter vectors extracted from a large number of various undistorted clean speech records. The objective auditory distances between vectors of the distorted speech signal and their corresponding matching references are then measured and appropriately converted into an equivalent subjective score. The optimal clustering of the reference codebook is achieved by using a dynamic k-means method. Efficient data mining technique known as Self-Organising Map is used to match the distorted speech vectors to the references. Speech parameters derived from Bark spectrum analysis, and Mel-Frequency Cepstral coefficients (MFCC) are used to provide speaker independent parametric representation of the speech signals as required by an output-based quality measure.**
Keywords: **Speech Processing, Perceptually-Based Speech Quality, Perceptual Quality Measure.**

I. INTRODUCTION

Most existing objective assessment methods for speech quality in modern voice communications systems require measuring some form of distortion between the input (transmitted) and output (received) speech signals. Processing steps typically include normalisation of signals powers, time alignment between input and output records, and determining a distance value which is used to estimate the equivalent subjective quality score. In practice the input speech record may not be available in all situations. For these situations an alternative technique is necessary to evaluate the quality of the transmitted speech using only the received signal. Such an approach could have numerous applications. The most practical application is non-intrusive monitoring the performance of communications systems. However this approach is not easy to realize due to the wide-ranging variability of the transmitted speech resulting from different speakers with different vocal tract and pitch characteristics.

In an attempt to consider this problem, this paper proposes a new perceptually based technique for objective prediction of speech quality, which utilizes a new efficient data-mining algorithm known as the Self-Organizing Map (SOM). The technique is based on comparing the output speech signal to an artificial reference signal that is derived from a dataset of clean undistorted speech records. The performance of the proposed algorithm is tested with speech from a number of male subjects, distorted by a modulated noise reference unit (MNRU) under different conditions.

II. SELF-ORGANIZING MAP

The self-organising map (SOM) [1] is one of the most well-known neural network models, which has proven to be a powerful tool for clustering of data, correlation hunting and novelty detection due to its unsupervised learning and topology preserving properties. The model implements a nonlinear topology preserving mapping from a high dimensional input data space onto a low dimensional network or grid of neurons (usually 1D or 2D). Each neuron i of the SOM is an n-dimensional prototype vector $\mathbf{m}_i = [m_{i1},\ldots,m_{in}]$ where n represents the input space dimension. On each training step, a data sample $\mathbf{x}$ is chosen and the unit $\mathbf{m}_c$ closest to it (the best matching unit, BMU) is identified from the map. The prototype vectors of the BMU and its neighbours on the grid are moved towards the sample vector. The new position is than given by:

$$\mathbf{m}_i = \mathbf{m}_i + \alpha(t)\, h_{wi}(t)\, (\mathbf{x} - \mathbf{m}_i) \qquad (1)$$

with $\alpha(t)$ representing the learning rate at the time t and $h_{wi}(t)$ is a neighborhood kernel centered around the winner unit w. Both the learning rate and neighborhood kernel radius decrease monotonically with time. During the step-by-step training, the SOM behaves like elastic net that folds onto the "cloud" created by input data.

Due to its high efficiency and robustness, the SOM method has been used in the proposed measure to achieve the required matching process.

III. Objective Speech Quality Measures

Over the last decade, researchers and engineers in the field of objective measures of speech quality have developed different techniques based on various speech analysis models. Currently, the most popular techniques are those based on psychoacoustics models, referred to as perceptual domain measures [2]. In these measures, speech signals are transformed into a perceptually related domain using human auditory models. Most available objective assessment techniques are based on an input-to-output approach. In input-to-output objective assessment methods, as depicted in Fig.1, the speech quality is estimated by measuring the distortion between an "input" or a reference signal and an "output" or received signal. Using a regression technique, the distortion values are then mapped into estimated quality.

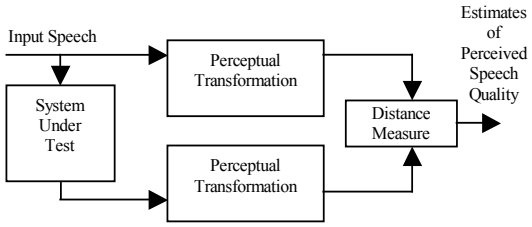


Fig. 1: Perception-based Approach to Quality Estimation

Currently there are a number of techniques that can be classified as perceptual domain measures. Examples of these include the Perceptual Analysis Measurement System (PAMS) and the ITU-T Perceptual Evaluation of Speech Quality (PESQ) measure [3],

There are three problems with the input-to-output speech quality measures. First, it is very difficult to achieve accurate synchronization between the input and the output signals. Secondly, the measurements can be seriously affected by background noise, as in the case of mobile networks, and hence would not provide true measure of the network's quality of service. Thirdly, in some situations the original speech is not available, as in case of mobile communications or satellite communications. Output-based measures, which do not need the input, are thus highly desirable.

IV. New Output-Based Approach

A new approach for a robust output-based objective speech quality measure, which correlates well with predicted subjective test, is detailed here. The approach, which is similar to that reported in [4], is based on comparing the output speech to an artificial reference signal representing the closest match from a database derived from undegraded speech material. The approach, which is depicted in Fig. 2, uses two different perception-based parametric representations of speech that have been shown effective in suppressing speaker-dependent details: the Bark Spectrum analysis [5] and Mel-Frequency Cepstral coefficients (MFCC) [6]

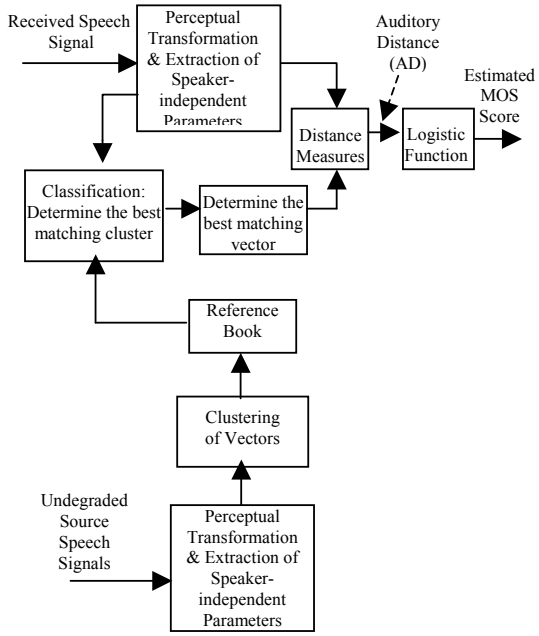


Fig 2: Block diagram of the new output-based approach

The general processing steps for the proposed output-based assessment approach are outlined below:

a) Establishment of datasets of high quality undegraded source and distorted speech records. The speech data are subjectively rated in terms of Mean Opinion Score (MOS).

b) Segmentation of the source (reference) and received (output) speech records into appropriately overlapped frames.

c) Derivation of an appropriate reference signal: this process involves the derivation of perceptually based speaker-independent speech parameter vectors from the distorted test (received) signal using two techniques: the Bark spectrum analysis and the Mel-Frequency Cepstral coefficients. Similar parameter vectors are also derived from a large data set of undegraded source speech records.

d) Application of clustering and classification techniques: this process involves three tasks. First the derived parameter vectors from the undegraded speech are clustered to produce a reference codebook corresponding to high quality speech. Secondly, the test vector is correlated with the clustered vectors stored in reference codebook in order to determine the best matching unit. Thirdly, by tracking the composition of the selected cluster, a best matching vector to the test vector is identified and an objective-auditory distance measure between the two vectors is computed. For the clustering, a dynamic and improved algorithm has been used (see

section IV sub-section A). The SOM has been used to perform the classification and determination of the best matching cluster and reference vector.

e) Distortion measure: due to the absence of the input speech, high quality clean speech records are used to formulate an artificial reference. The proposed objective measure is based on measuring the degree of mismatch between the distorted speech vectors and its best matching vector from the reference codebook. This has been affected by computing the median minimum distance (MMD), as described in Section IV subsection B.

f) Mapping the measured auditory distances into predicted subjective scores: finally, linear regression is used to map the measured distortion indicator, described in (e) above, into corresponding subjective quality score such as the Mean Opinion Score (MOS).

A. Determination of Number of Clusters

The k-means algorithm aims to minimize the sum of squared distances between all the data points and the cluster centre. The main inconvenience of this procedure is the determination of the best value of k that provides the optimum clustering for a given application. To alleviate this problem, the proposed objective quality measure uses a dynamic k-means method to determine the optimum number of clusters. The method starts by choosing K initial clusters centres z_1, z_2, … z_K. The coefficients of the reference vectors are distributed among the K clusters. To achieve the best clustering arrangement which results in a compact number of well separated clusters, two measurements are performed: the intra-cluster distance which is simply the average distance between a point and its cluster centre, and the inter cluster distance or the distance between the cluster centres, defines as:

$$\text{intra - cluster} = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}\left\|x - z_i\right\|^2 \qquad (2)$$

$$\text{inter - cluster} = \min\left(\left\|z_i - z_j\right\|^2\right),\ i = 1,2,\ldots K\text{-}1\ ;\ j = i+1,\ldots,K \qquad (3)$$

where x represents a given coefficient (point), N the number of points in a cluster, K the number of clusters centres, z_i is the cluster centre of cluster C_i and $\|.\|$ denotes an Euclidean distance operation. In order to determine the best clustering, the above two measurements are combined to give a 'validity' factor defined by:

$$\text{validity} = \frac{\text{intra - cluster}}{\text{inter - cluster}} \qquad (4)$$

Since we want to minimise the intra-cluster distance and this measure is in the numerator, we consequently want to minimize the validity measure. We also want to maximize the inter-cluster distance measure, and since is in the denominator, we again want to minimize the validity measure. Therefore, the clustering which gives a minimum value for the validity measure will tell us what the ideal value of K is in the k-means procedure.

B. Computation of the MMD

The Euclidean distance from a test vector $\mathbf{x}_l$ of the lth frame of the received speech signal to a reference vector $\mathbf{y}_m$ of the mth frame, which has been identified as the BMU, is detailed as:

$$dis(\mathbf{x}_l, \mathbf{y}_m) = \left\|\mathbf{x}_l - \mathbf{y}_m\right\| = \sqrt{[\mathbf{x}_l - \mathbf{y}_m]^T[\mathbf{x}_l - \mathbf{y}_m]} \qquad (5)$$

where T denotes transpose operation. After the distances for all frames are found, the median minimum distance (MMD) index for the received signal is computed as:

$$D_{MM} = \text{median}_L\,[dis\,(\mathbf{x}_l, \mathbf{y}_m)] \qquad (6)$$

where L is the number of frames in the received signal. The above distance measure provides an objective indication of the degradation in the received speech signal. Larger distances imply lower speech quality and vice versa.

V. Results And Discussion

The proposed output-based measure has been tested with speech distorted by a modulated noise reference unit (MNRU) under seven different conditions as those used in [7]. The tests were conducted on seven different cases with three levels of difficulty, using around 10 seconds of test speech signals taken from male subjects only. For each case, two versions of the proposed output-based quality measure are applied: the first is based on the use of the Bark spectrum analysis, and the second is based on the use of the MFCC.

For the first level (test cases 1 and 2), the proposed method was tested and trained using speech records from the same male speaker. Accordingly this represents the easiest possible test case. The main difference between these cases and a standard input-to-output objective measurement is that there is no frame-level time alignment between the input and output speech. For the second level of difficulty (cases 3, 4 and 5) two different male speakers, M1 and M2, were used and the spoken text was different. The third level (cases 6 and 7) is when the spoken text of the test speech was different from that of the reference speech and the speakers were also different. Correlation coefficients between the estimated and the actual subjective MOS of the test speech records for all the above cases are shown in Table I.

Inspection of the Table. I indicates the followings:

- For the first five test cases, the speech quality prediction of both versions of the proposed output-based measure seems to correlate very well with the actual MOS scores. Modern input-to-output based speech quality measures can typically achieve correlation in the range from 0.8 to 0.9. In contrast, the correlation coefficients for these five cases represent the upper limit of performance for an output-based algorithm, which has limited access to information compared to the input-to-output based approach.
- For the last two test cases the correlations with the actual MOS scores were comparatively lower. In addition the version of the proposed measure that is based on the Bark spectrum analysis seems to perform relatively better than that which is based on the MFCC. The last two test cases were repeated using longer speech records with duration of 30-50 seconds. The correlation coefficients were 0.9143 for Bark Spectrum and 0.9175 for MFCC Coefficients.

Table. I: Correlations between objective and subjective scores

Test Case	**Training Datasets**	**Testing Datasets**	**CORRELATION COEFFICIENTS**	
			Bark Spectrum	*MFCC Coefficients*
1	M1	M1	0.9950	0.9762
2	M2	M2	0.9986	0.9410
3	M1, M2	M1	0.9953	0.9638
4	M1, M2	M2	0.9988	0.9410
5	M1, M2	M1, M2	0.9881	0.9653
6	M1	M2	0.8869	0.7145
7	M2	M1	0.8256	0.7121

Aldo the system, as proposed here, has been designed to assess speech/voice quality for telecommunications networks, it can easily be adapted for biomedical applications. This can be done by replacing the subjective listening scale described in the paper (i.e. MOS) by an appropriate medical-based scale such as GRBAS [8]. The authors are currently working on these types of applications in collaboration with two departments from University of Florence: Department of Electronics and Telecommunications, and Department of Physics.

VI. Conclusions

In this paper a new output-based speech quality measure, which uses Bark Spectrum analysis and Mel-Frequency Cepstral coefficients, was introduced. The measure is based on comparing the output speech to an artificial reference signal that is appropriately selected from optimally clustered reference codebook, using the SOM approach coupled with an enhanced k-means technique. The codebook is formulated from a number of undistorted clean speech records taken form a variety of speakers. As part of an-going evaluation work, performance of the proposed measure were tested with speech distorted by modulated noise reference unit under different conditions. Test results indicated that the proposed output-based is generally effective in predicting the corresponding subjective speech quality, and is fairly robust against speakers and content variations. Further study is well underway to investigate the optimal clustering process, length of the frame size used to process the speech and its associated overlap, as well as the use of the SOM model for both clustering and matching process.

It is also indicated that the current work can be easily modified to be suitable for biomedical applications.

VII. Acknowledgment

The authors would like to thank Dr. Leigh Thorpe from Nortel Networks, Ottawa, Canada for providing the speech database used in this work.

References

[1] J. Vesanto and E. Alhonieni, “Clustering of the self-organizing map,” *IEEE Trans on Neural Networks*, Vol., No. 3, pp. 586-600, 2000.

[2] S.Voran, “Objective estimation of perceived speech quality-Part I: development of the measuring normalizing block technique,” *IEEE Trans. on Speech and Audio Process*., Vol., No. 4, pp. 371-382, 1999,

[3] J. Anderson, “Methods for measuring perceptual speech quality,” *Agilent Technologies-White Paper*, USA, May 2001.

[4] C. Jin and R. Kubichek, “Vector quantization techniques for output-based objective speech quality,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Process*., ICASSP-96, Vol.1, pp. 491-494, Atlanta, May 1996.

[5] S. Wang, A. Sekey, A. Gersho. “An objective measure for predicting subjective quality of speech coders,” *J. on Selected Areas in Communications*, Vol.10, pp 819-829, 1992.

[6] D. Hyun and C. Lee, “Optimisation of Mel-Cepstrum for speech recognition” *IEEE International Conference on Systems, Man, Cybernetics, SMC’99* 1:500-503, 1999.

[7] L. Thorpe and W. Yang, “Performance of current perceptual objective speech quality measure,” *Proc. IEEE Workshop on Speech Coding*, pp.144 –146, Porvoo, Finland, 1999.

[8] C. Manfredi, A. Schindler and P. Bruscaglioni, “A subspace approach for voice quality enhancement”, *6th International Conference: Advances in Qualitative Laryngocology, Voice and Speech Research,* No. 13, Hamburg, Germany 2003.

Special session on
Singing voice

3rd International Workshop MAVEBA 2003, 175-178

SAMPLE-BASED SINGING VOICE SYNTHESIZER USING SPECTRAL MODELS AND SOURCE-FILTER DECOMPOSITION

J. Bonada[1], A. Loscos[1], O. Mayor[1], H. Kenmochi[2]
[1] Music Technology Group, Audiovisual Institute, Universitat Pompeu Fabra, Barcelona, Spain
[2]Advanced System Development Center, YAMAHA corporation, Hamamatsu, Japan

Abstract: **This paper is a review of the work contained in the insides of a sample-based virtual singing synthesizer. Starting with a narrative of the evolution of the techniques involved in it, the paper focuses mainly on the description of its current components and processes and its most relevant features: from the singer databases creation to the final synthesis concatenation step.**

I. INTRODUCTION

The voice generation is typically explained as a source/filter system, in which a voiced/unvoiced excitation is filtered by the vocal tract resonances. The voiced excitation corresponds to the glottal pulses that originate the vocal fold vibrations whether the unvoiced excitation corresponds to the turbulent airflow that arises from the lungs. The voice filter is characterized by a set of resonances called formants that have their origin in the voice organs lengths and shapes (trachea, esophagus, larynx, …). This filter modulates the timbre of the sound by dynamically changing the amplitude, center frequencies and bandwidths of the resonances by moving the voice organs.

Some of the singing synthesizers developed since the beginnings of such discipline have focused in the source/filter decomposition (physical models based); others, rather than focusing on how the sound is produced, have focused on the perception of the sound (spectral models based); and others, such as the synthesizer we present in this paper, have tried to combine both models.

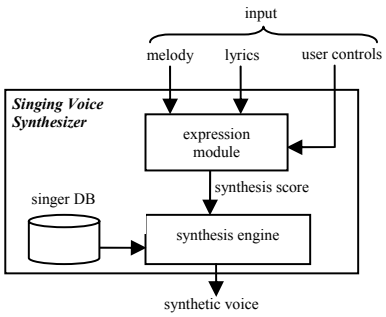


Figure 1: *General system diagram*

The system can be roughly described by a singer database, an input, an expression module and a synthesis engine (see Fig. 1). The input contains the melody and the lyrics of the song plus some expression controls. The expression module converts this input into an internal low-level synthesis score, and the synthesis engine reads this synthesis score, fetches the required samples from the singer database and transforms and concatenates them to obtain the synthetic output signal.

II. VOICE AND SPECTRAL MODELING

Since our system is a sample based synthesizer in which samples of a singer database are transformed and concatenated along time to compose the resulting audio, we have always considered the task of finding the most appropriate and the highest quality transformation techniques a crucial issue.

We initially used SMS [1] as the basic transformation technique with the addition of a time domain delta-based excitation to mimic the singer's voiced excitation [2]. SMS had the advantage of decomposing the voice into harmonics and residual. Both components were independently transformed, so the system yielded a great flexibility. But although the results were quite encouraging in voiced sustained parts, in transitory parts and consonants, especially in voiced fricatives, harmonic and residual components were not perceived as one.

Intending to improve our results, we moved to a spectral technique based on the phase-locked vocoder [3] where the magnitude spectrum is segmented into regions, each of which contains a spectral peak and its surroundings. These regions can be then freely shifted in amplitude and frequency. Regarding the phase spectrum, the relation between harmonics found at the beginning of each glottal period is kept after transformations [4]. On top of this technique we developed a frequency domain voice model that consists of an excitation curve, a set of resonances and a residual envelope. We call it EpR (Excitation plus Resonances) [2]. The excitation curve models the voiced source using an exponential decay function and a low frequency resonance. The vocal tract is modeled using the rest of the resonances and the

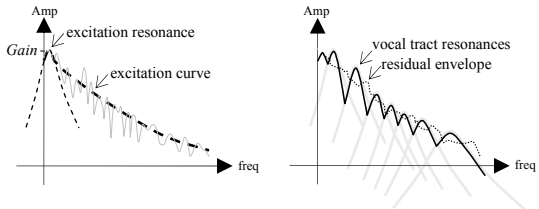


Figure 2: *The EpR model*

residual envelope stores the differences between the model and the spectral shape defined by the harmonics (see Fig. 2).

III. SINGER DATABASE

About two hours of dry singer performance recordings are required to build a database. The singer is asked to follow a detailed recording script that covers most possible phonetic contexts and several expression aspects [5]. These recordings are then segmented and analyzed using the spectral analysis algorithms. In order to speed up this process two free software toolkits [6, 7] are used as phonetic aligners between the audio files and the recording scripts. The resulting data fills the phonetic and the expression libraries and is stored in a set of files organized in tree structured folders.

The phonetic database contains timbres, steady-states and articulations. The timbre section stores the voice model (EpR) of different vowels at different pitches and dynamics, the steady-state section contains long sustained vowels at different pitches, and the articulation section contains an organized list of diphonemes samples at different pitches.

The expression database contains note and vibrato templates intended to keep some basic expression aspects of the singer's voice and therefore increase the naturalness of the synthesis. Note templates model singer's attacks, releases and transition behaviors in different musical and intentional contexts. These contexts are described by a set of meaningful labels, like sharp attack, legato transition or sexy release. Each template stores a set of controls (pitch, loudness, EpR excitation curve, breathiness, roughness) obtained from the analysis of the sample, each of which can be later used in synthesis to reproduce the voice excitation changes for each expressive context. Vibrato templates store the singer's excitation behavior for different types of vibrato and tremolos; basically they keep the pitch and the EpR excitation curve. Each template is segmented into attack, body and release parts. The body segment is mirror-looped at synthesis if needed.

IV. INPUT SCORE

The input score is an ASCII text file based on the METRIX control language [8] that contains the score of the song. Not only lyrics and notes can be specified, but also high level controls and all the possible music information that the system is capable to interpret. To achieve naturalness in the synthetic voice, the system defines some musically meaningful controls [5]. The idea is to cover the maximum situations that can appear in a real singing performance in order to avoid a lack of expression control that could bring about non-natural results.

The input score contain the so-called note parameters and control parameters. The note parameters refer to a specific note of the score and describe note attributes such as pitch, duration, loudness, lyrics, dynamics, vibrato, attack / release types, roughness, etc., while the control parameters refer to the whole song and describe song attributes such as singer, tempo, etc. Below you can see an example of input score where the lyrics are *fly me*.

```
Score_Info
{
      Tempo: 90
      Meter: 4/4
}
InstrumentInfo { Robert }

begin
  t1      Robert    NoteNumber: Ab2
                    Duration: t0.5
                    Lyrics: "f l al"
                    Loudness: 0.6
                    AttackType: "soft"
  t1.5    Robert    NoteNumber: G2
                    Duration: t1
                    Lyrics: "m l"
                    Loudness: 0.3
                    VibratoType: "wet"
                    VibratoRate: [(0,1)(1,0.6)]
                    VibratoDepth: [(0,0) (0.5,1)(1,0.7)]
                    ReleaseType: "long"
end
```

V. BUILDING THE SYNTHESIS SCORE

The expression module generates an internal low-level score (*synthesis score*) out of the input METRIX. This score is structured into several tracks and control envelopes, some of which are shown in Fig. 3. The phonetic track shows the articulations and steady-states to be fetched from the DB and their corresponding durations, which are calculated trying to make them as close to the original database sample durations as possible. The note and vibrato tracks contain information on the note and vibrato templates that must be applied at synthesis and their corresponding durations. The envelope controls (*vibrato depth and rate, pitch, pitch var, loudness, etc*) express their behavior along the performance with a time-varying function.

In addition to the note and vibrato templates, several models have been created to cover a wide variety of possibilities. However, templates extracted from real recordings are preferable to get a more authentic expressivity, although they may not sound natural when the synthesis context in which they are applied is far from the template context.

The phonetic track is filled out taking into account that the vowel onset should match the begin time of the note. Besides, as already mentioned, taking the original sample duration is preferable since this way we avoid time-scaling transformations, but this is not always possible because all required articulations must fit into the note segment. On the other hand, whenever the added duration of the articulations is less than the target note segment duration, a steady-state is added to fill out what is left.

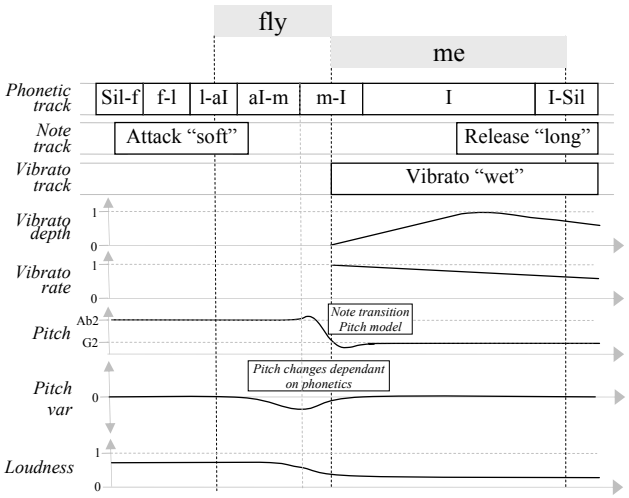


Figure 3: *Synthesis score*

In the synthesis score there are two envelope controls that specify the output synthesis pitch. The first envelope (*Pitch*) stores the absolute pitch values that come out from the notes specified by the input score. On the other hand *Pitch var* stores relative pitch variations due to changes originated by some phonetic combinations, such as certain voiced consonant - vowel combinations (b-a) in which the pitch decreases during the consonant sound.

In synthesis, the relative values of the *pitch var* envelope and the expression templates are added together to the absolute pitch values. In the case that an attack or release template is specified, the pitch variations of this template are applied when synthesizing to obtain a pitch curve similar to the one in the template. In the case of note transitions, the process is the same but whenever no template is specified, a pitch model is applied that overwrites the absolute pitch track of the score, like shown in Fig. 3, so to avoid pitch discontinuities. This pitch model has to be carefully generated to obtain a natural sounding pitch curve in the output synthesis. A mathematical model has been designed to produce

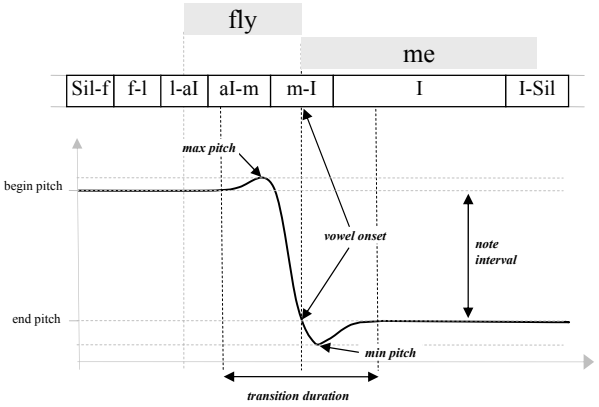


Figure 4**:** *Pitch model for note transitions*

smooth pitch transitions between notes and allow the control of some parameters like duration, shape and synchronization to phonetics and musical rhythm. This synchronization is basically attained by reaching the target pitch at the onset of the vowel of each syllable. In Fig. 4 we can see a more detailed drawing of this pitch model. The distance between *begin pitch* and *max pitch*, as well as between *min pitch* and *end pitch*, depends on the note interval (the bigger the interval, the bigger the distance, but with some limitations for big intervals). On the other hand, the transition curvature depends on both the note interval and the transition duration and its slope is restricted to a maximum value in order to guarantee smooth pitch variations in short transitions.

V. SYNTHESIS ENGINE

A. Sample transformations

The synthesis engine reads the synthesis score and retrieves the required samples and templates from the singer database selecting those units that are closer to the synthesis context (mainly pitch is considered). Once we have retrieved the samples, some transformations [4] are applied to match the synthesis score: transposition, equalization, time-scaling, loudness modification, vibrato and voice excitation based transformations. Finally, the transformed samples are concatenated to compose the resulting synthetic performance**.**

Transposition is applied to match the synthesis score pitch. Therefore, the transposition factor is calculated as the synthesis pitch divided by the sample pitch. This factor is calculated frame by frame. In terms of the spectral technique, harmonic peak's regions are shifted in frequency and harmonic peak's phases are corrected without altering the phase synchronization between harmonics.

Equalization is used to obtain transformations on timbre. When transposing, it is used to keep the original timbre but it can be applied as well to get generic timbre transformations. Equalization is achieved by shifting in amplitude the harmonic peak's regions so to match the desired timbre envelope.

Time-Scaling is applied to samples in order to match their durations with the synthesis score durations. The time-scale ratio is sometimes applied in a non-uniform way so that the synchronization between control parameters, phonetic and note tracks is not altered. For example, the phonetic articulation that contains the vowel onset should not change the timing of the vowel onset. Besides, in the case of abrupt phonetic changes, these should not be smoothed so not to degrade the intelligibility. The transformation is obtained by repeating or dropping some frames and interpolating them [9].

For loudness modification, database samples are considered to be sung at normal loudness, unless

otherwise specified. Thus, sample loudness is changed to match the synthesis score value. The transformation can be achieved by applying an equalization filter obtained from a recorded template where the singer sang a crescendo or a decrescendo. This filter represents the timbre envelope differences between the sample estimated loudness and the target loudness.

For vibrato transformation, the pitch and EpR excitation changes enclosed in the vibrato template are applied to the audio samples. The little nuances of the singer's vibrato are kept even after altering its depth and rate, and the EpR voice model allows the harmonics to follow the resonances as their frequency is modified, thus emulating the real situation.

Besides, some voice excitation based transformations can be applied to improve the naturalness and expressiveness of the synthetic voice, such as roughness, whisper and breathiness. Roughness is obtained by adding sinusoids to the spectrum in a way that glottal periods become irregular. Whisper comes out of equalizing an unvoiced excitation with the timbre envelope. Finally, breathiness is succeeded by adding together whisper and equalization effects and lowering the harmonic's peak adjoining bins.

B. Sample concatenation

The last step in the synthesis engine is the concatenation of samples. Once we have transformed the samples, we have to deal with the spectral shape and phase discontinuities that appear when connecting them. With the aim of minimizing such discontinuities, amplitude and phase corrections are spread out along a set of transition frames that surround the boundary [4]. The results are quite smooth and good enough in most cases. Sometimes, however, a gap in brightness can be heard, especially when connecting samples that have been transposed with rather different factors, due to the fact that although there are no harmonic peak's amplitude or phase discontinuities, there do exist harmonic peak's regions amplitude and phase shape gaps. This problem is inherent to only consider harmonic peak's discontinuities when connecting samples, thus our algorithm should be expanded to consider inside region characteristics.

VI. CONCLUSION

The system we present is able to generate synthetic performances with quite successful results. However, the more different from the database the synthesizer is asked to sing, the more artificial synthesis gets (it is difficult to make the system sing hip-hop using an opera singer database). Some of this difficulty arises from the fact that the synthesizer has been thought to preserve not only the timbre personality of the singer from which the database is created but also his/her expressivity.

In this sense, work has to be done to improve transformations naturalness, especially when the synthesis context is far from the original context in which the sample that is being transformed was recorded.

Other improvements directions include working on expression dependent timbre transformations and getting into a higher level transformation description in which the system could generate an expressive performance automatically out of the melody, the lyrics, the singer, and an expressive label such as sweet or aggressive.

REFERENCES

[1] Serra, X, "A system for sound analysis-transformation-synthesis based on a deterministic plus stochastic decomposition" *PhD thesis,* CCRMA, Dept. of Music, Stanford University, 1989.

[2] Bonada, J., Loscos, A., Cano, P., and Serra, X, "Spectral Approach to the Modeling of the Singing Voice" *Proceedings of the 111th AES Convention*, New York, USA, 2001.

[3] Laroche, J. and Dolson, M., "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 1999.

[4] Bonada, J., Loscos, A., and Kenmochi, K, "Sample-based singing voice synthesizer by spectral concatenation" *Proceedings of the Stockholm Music Acoustics Conference SMAC03*, Stockholm, Sweden, 2003.

[5] Bonada, J., Celma, O., Loscos, A., Ortolà, J., and Serra, X., "Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models" *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[6] Akinobu Lee et al, "Julius - An Open Source Real-Time Large Vocabulary Recognition Engine" *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1691-1694, 2001.

[7] Huang, X., Alleva, F., Hon, H., Hwang, M., and Rosenfeld, R., "The SPHINX-II speech recognition system: an overview" *Computer Speech and Language*, vol. 7, no. 2, pp. 137-148, 1993.

[8] Amatriain, X, "METRIX: A Musical Data Definition Language and Data Structure for a Spectral Modeling Based Synthesizer" *Proceedings of 98th Digital Audio Effects Workshop DAFX98*, Barcelona, Spain, 1998.

[9] Bonada, J., "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio" *Proceedings of the International Computer Music Conference*, Berlin, Germany, 2000.

3rd International Workshop MAVEBA 2003, 179-182

TOWARDS A NOVEL REAL-TIME VISUAL DISPLAY FOR SINGING TRAINING

David M Howard[1], Graham F Welch[2], Jude Brereton[1], and Evangelos Himonides[2]

[1]Media Engineering Research Group, Department of Electronics, University of York, Heslington, York YO10 5DD, UK
[2]School of Arts and Humanities, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK

Abstract: **Real-time visual displays have found application to be tested as part of a recently funded pilot project to investigate the usefulness or otherwise of computer displays in the singing studio. Following previous work that suggests that simple displays of a small number of analysis parameters are generally the most effective, the system makes available analyses plotted against time that relate to: pitch, spectral ratio, larynx closed quotient and vocal tract area. These can be viewed singly, multiply or in combination. The algorithms used will be described as well as previous analysis experiments that indicate their potential usefulness. A number of example output screens will be illustrated to indicate how users interact with the system. The on-going testing paradigm will also be described which is designed to establish whether or not displays such as these can be used in the singing studio to any useful advantage.**

.

Keywords : **visual displays, singing, vocal tract display**

I. INTRODUCTION

This paper describes the technology to be emplyed in a project during which the application of real-time visual feedback technology in the singing studio will be investigated, both during lessons and outside during private practice. In general, science and artistic musical performance tend to use different language codes and symbolisation for knowledge, and often, their ontological standpoints are different. Whilst it is not known to what extent these two language codes are reconcilable, the benefits from the application of technology have been demonstrated in many other fields, including the arts. There is no longer a widespread culture of technology phobia in non-scientific fields of human endeavour.

The standard pedagogical model employed in the conservatoire studio typically involves weekly/twice weekly lessons with an expert, supported by private practice and performance. The teacher is engaged in a psychological translation of the student's performance, for example by turning musical gestures into language, and the student is engaged in a further translation of the teacher's verbal and visual feedback into adapted singing performance. A dual possibility thereby exists for the misinterpretation of information. Anything that can provide more robust and easily understandable feedback to both teacher and student would seem to be worthwhile, and this forms the basic premise to investigate the use of technology in the signing studio.

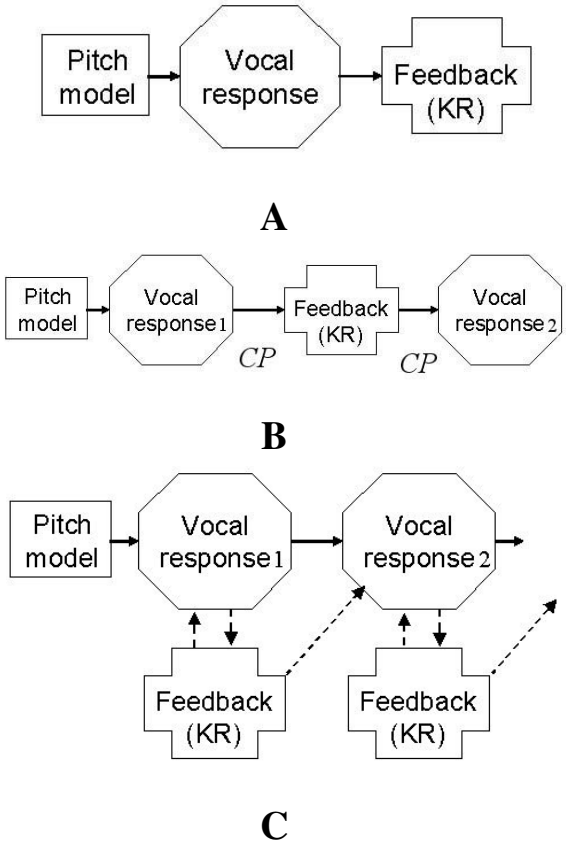


Figure 1: An illustration of the learning process for pitch in singing based on [1, 2]. Time is from left to right in these plots.
KEY: (A) the basic interaction between teacher and learner; (B) the on-going traditional learning process, and (C) the way in which real-time visual feedback can impact the learning process. KR = knowledge of results from an external source; CP = critical learning period.

Welch [1, 2] develops a model to characterise the learning process, taking pitch as an example, and this is illustrated in Fig. 1. During the traditional interaction between teacher and student, a model is provided, the

student makes an attempt vocally, and the teacher provides feedback to the student. A key issue in relation to this feedback is the gain the student makes in regard to knowing what s/he is supposed to be achieving in terms of a result, an external assessment being referred to as "knowledge of results" or "KR" as indicated in Fig. 1-A. Understanding what is required and how to recognise it is a vital aspect of the learning process.

Following feedback on a vocal response, the student subsequently will make another attempt as illustrated in Fig. 1-B. This is the nature of the traditional singing pedagogical process. The use of real-time visual feedback enables feedback to be provided *during* the student's vocal response, enabling modifications to be made immediately and their concurrent effect to be observed (see Fig. 1-C). Apart from the more obvious advantage of removing the time lag between a vocal response and the feedback that is inevitable without real-time provision, the student is able to make another attempt immediately based on observations of the feedback provided during the previous attempt as appropriate.

Quantifiable parameters have been identified that vary with training and experience for: (a) actors [3], (b) adult singers [4, 5], as well as (c) girl and boy cathedral choristers [6]. Real-time visual feedback has been previously used successfully with primary school children [7, 8] and adult singers [9, 10]. Our experience suggests that technological applications are only of potential benefit if they are easy to use by non-specialists and provide information that is meaningful, valid and useful. Such robust information can then underpin feedback to provide more accurate formative and summative assessments.

II. DISPLAYS TO BE EMPLOYED

A. Consultation with the community

A one day workshop was held with a group of singing teachers, the authors, and interested colleagues who research in the areas of speech and/or singing. The purpose of this event was to review existing displays that might be useful in the context of the singing studio, and to produce a specification for the software to be employed in the project. Colleagues were reminded that the project is not about testing the effectiveness of the technology itself, but to establish its potential usefulness or otherwise. Specific research questions include:

- the extent to which teachers and students will accept and make use of technology in the studio
- the ease-of-use of the technology, both in the studio and elsewhere for private practice
- the nature of the data offered by the technology
- how the data can be integrated into singing teaching and learning
- the readiness with which the data can be interpreted and utilised
- whether the technology overly intrudes into the learning and teaching experience
- any potential perceived threat posed to the teacher and/or the student by the use of technology.

In order to make the technology be potentially widely applicable, a windows-based PC implementation was targetted. Existing possibilities for real-time displays were demonstrated, and the following were identified as being appropriate as tools for use in the singing studio for this project:

- fundamental frequency against time
- spectral ratio against time
- vocal tract area
- summary vocal tract area measures against time
- side view camera.

Each of these is described and illustrated below.

B. Fundamental frequency against time

The measurement of fundamental frequency (f0) has been the subject of considerable research [e.g. 10]. No one technique exists that is accurate for all subjects, covering the complete human pitch range uttered in any acoustic. The choice of a technique should be matched to the situation where it is to be used. A real-time display must not exhibit any delay to the user, it should be accurate operating over a wide f0 range for singers, of the order of C2(65Hz) to C6(1047Hz). A peak-picking system was employed that was originally developed in analogue form for use in cochlear implants [12], and subsequently applied in the SINGAD system [7, 8]. Each of the elements of its circuit has been implemented in C++, and an example plot of f0 against time is shown in Fig. 2.

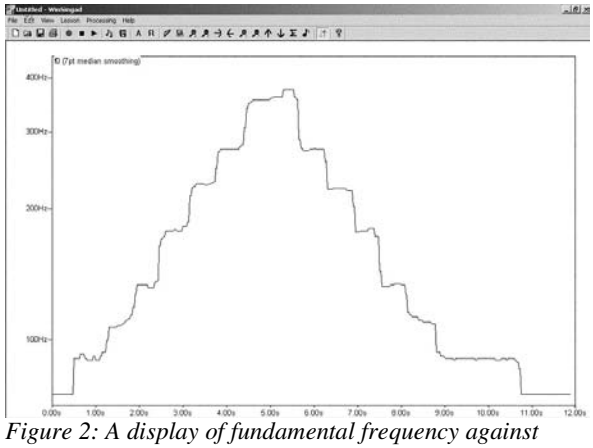


Figure 2: A display of fundamental frequency against time for a sung ascending and descending two octave arpeggio.

C Spectral ratio against time

A key element in singing training is that of voice projection, and one acoustic consequence of this is the appearance of a peak in the output spectrum in the region 2.5kHz to 4kHz, known as the singer's formant [e.g. 13]. The ratio of the energy in this band to the energy in the total signal is calculated. This measurement is constrained between 0 and 1 providing the full band extremes encompass the singer's formant band. In this implementation, these are set to (100Hz to 4000Hz) and (2500Hz to 4000Hz) respectively. These values can be changed by the user. Fig. 3 shows an example plot of this ratio against time for the vowel /a:/ sung in a projected and non-projected style.

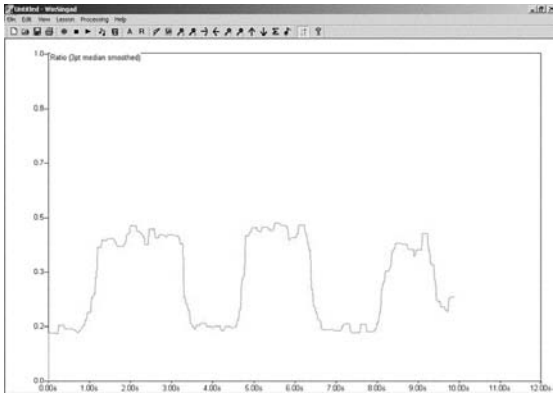


Figure 3: Example ratio against time display for/a:/ sung alternating between a non-projected.(lower ratio values) and a projected style (higher ratio values).

D Vocal tract area

A display of the vocal tract area can be obtained via a lattice filter model derived from a linear predictive analysis of the vocal output [14]. This models the vocal tract in terms of the areas (or diameters/radii) of a set of equal length tubes between the glottis (space between the vocal folds) and the lips. Fig. 4 shows an example vocal tract area display for a sung /a:/ vowel, where the glottis and lips are at the left and right edges of the display respectively.

There are, however, limitations associated with this representation. Firstly, it strictly only models non-nasal voiced sounds, due to the assumptions employed in linear prediction. Secondly, the output area values have no absolute area reference, and therefore they are arbitrary. They are usually therefore normalized either to a fixed glottis width (this is adopted in Fig. 4), or to a fixed maximum value. Finally, there are situations where more than one set of tube areas provides a solution, and results can be presented that could not be articulated by a human vocal tract. Due to the integrated nature of the solution process, it is not obvious how it might be constrained, for example, to vocal tract configurations that are physically possible.

It is for this reason that summary plots of the average, minimum or maximum vocal tract area against time will be incorporated.

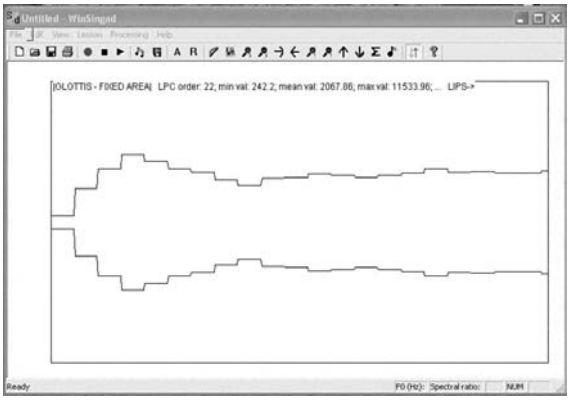


Figure 4: Example vocal tract area display for a sung /a:/ vowel . The glottis and the lips are at the left and right hand side of the plot respectively.

E Summary vocal tract area against time

The mean, minimum and maximum vocal tract area is calculated for each frame of input data, and these can be plotted against time. A plot of the mean area against time is shown in Fig. 5. An important aspect of singing training relates to the degree of perceived openness of the vocal tract, or the degree of constriction, and it is suggested that some indication of this may be given through reference to minimum vocal tract area against time.

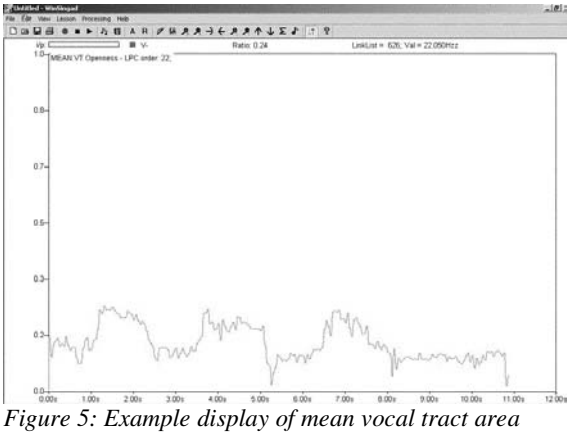


Figure 5: Example display of mean vocal tract area against time for /a:/ sung alternating between a non-projected.(lower values) and a projected style (higher values).

F Side view camera

Singers often make use of a mirror during training for feedback on their posture. With a computer display, it is

possible to make use of a camera with the result displayed on screen. We shall employ a camera to enable singers to view their posture from the side to enable the straightness of their spine to be observed. The screen will be placed at head height to encourage a vertical head position.

III. Discussion and conclusions

A set of displays to be employed in real-time in singing studios has been described. These are being integrated by a professional programmer into a complete system with the side view camera output, in which the user is given control over which single or arbitrary set of displays s/he wishes to use. Appropriate control over processing and display parameters will be provided to the user via standard menus and dialog boxes. In this way, attention can be drawn to individual parameters displayed alone, or to multiple parameters as familiarity and confidence grows, and areas of interest can be zoomed in on as desired. This system will provide the computer-based display system to enable the usefulness or otherwise of technology in the singing studio to be assessed.

An action research methodology is to be employed for this assessment, in which the teachers, students and the research assistants, acting as observers, keep diaries of progress and activities during lessons. Two teachers will be involved, each with an experimental and control group with two students in each.

The system will also allow both the audio signal (microphone) and video signal (side-view camera) to be recorded to enable vocal responses to be reviewed and/or archived for progress tracking.

IV. Acknowledgements

This project is supported by the Arts and Humanities Research Board (AHRB) in the UK under an innovation award numbered B/IA/AN8885/APN15651. The authors thank the singing teachers and other professional colleagues who contributed to the initial workshop.

References

[1] Welch, G.F. (1985a). A schema theory of how children learn to sing in tune. Psychology of Music, **13**, (1), 3-18.

[2] Welch, G.F. (1985b). Variability of practice and knowledge of results as factors in learning to sing in tune. Bulletin of the Council for Research in Music Education, **85**, 238-247

[3] Rossiter, D.P., Howard, D.M., and Comins, R. (1995). Objective measurement of voice source and acoustic output change with a short period of vocal tuition, *Voice*, **4**, (1), 16-31.

[4] Rossiter, D.P., and Howard, D.M. (1998). Observed change in mean speaking voice fundamental frequency of two subjects undergoing voice training, *Logopedics Phoniatrics Vocology*, **22**, (4), 187-189

[5] Howard, D.M. (1995). Variation of electrolaryngographically derived closed quotient for trained and untrained adult singers, *Journal of Voice*, **9**, (2), 163-172.

[6] Welch, G.F., and Howard, D.M. (2002). Gendered voice in the Cathedral choir, *Psychology of Music*, **30**, (1), 102-120

[7] Howard, D.M., and Welch, G.F. (1993). Visual displays for the assessment of vocal pitch matching development, *Applied Acoustics*, **39**, (3), 235-252.

[8] Welch, G.F., Howard, D.M., and Rush, C. (1989). Real-time visual feedback in the development of vocal pitch accuracy in singing, *Psychology of Music*, **17**, 146-157.

[9] Rossiter, D.P., Howard, D.M., and DeCosta, M. (1996). Voice development under training with and without the influence of real-time visually presented biofeedback, *Journal of the Acoustical Society of America*, **99**, (5), 3253-3256.

[10] Thorpe, C.W., Callghan, J., and van Doorn, J. (1999). Visual feedback of acoustic voice features for the teaching of singing, *Australian Voice*, 5, 32-39.

[11]. Hess, W. (1983). *Pitch determination of speech signals*, Berlin: Springer Verlag.

[12] Howard, D.M. (1989). Peak-picking fundamental period estimation for hearing prostheses, *Journal of the Acoustical Society of America*, **86**, 902-910.

[13] Sundberg, J. (1987). *The science of the singing voice*, Dekalb: Northern Illinois University Press.

[14] Rossiter, D.P., Howard, D.M., and Downes, M. (1995). A real-time LPC-based vocal tract area display for voice development, *Journal of Voice*, **8**, 4, 314-319.

3rd International Workshop MAVEBA 2003, 183-186

AUDITORY AND KINESTHETIC FEEDBACK IN SINGING – SIGNIFICAN AND EFFECTS OF TRAINING ON PITCH CONTROL*

D. Mürbe[1, 2], G. Hofmann[1 ,2], F. Pabst[2], J. Sundberg[3]

[1]Department of Otorhinolaryngology, Technical University of Dresden, Dresden, Germany
[2]Voice Research Laboratory, University of Music Carl Maria von Weber, Dresden, Germany
[3]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

* Parts of this investigation were presented at the 31th Annual Symposium Care of the Professional Voice, Philadelphi June 2002 and are accepted for publication in the Journal of Voice

Abstract: **An accurate control of fundamental frequency is one of the essential demands in professional singing. This control relies on auditory and kinesthetic feedback. However, a loud accompaniment may mask the auditory feedback, leaving the singers to rely on kinesthetic feedback. The object of the present study was to estimate the significance of auditory and kinesthetic feedback to pitch control in 28 students beginning a professional solo singer education. Since it seems reasonable to assume that pitch control can be improved by training, the same students were reinvestigated after 3 years of professional singing education. In both parts of the study the singers sang an ascending and descending triad pattern with and without masking noise in legato and staccato and in a slow and a fast tempo. Fundamental frequency and interval sizes between adjacent tones were determined and compared to their equivalents in the equally tempered tuning. The average deviations from these values were used as estimates of intonation accuracy. For both parts of the study, intonation accuracy was reduced by masking noise, by staccato as opposed to legato singing and by fast as opposed to slow performance. After education, the contribution of the auditory feedback to pitch control was not significantly improved while the kinesthetic feedback circuit was improved in slow legato and slow staccato tasks. The results support the assumption that the kinesthetic feedback contributes substantially to intonation accuracy and might be improved by training.**

Keywords : **singing, pitch control, training, auditory feedback, kinesthetic feedback**

I. INTRODUCTION

The high demands on intonation in professional singing require precisely acting pitch control systems. Auditory and kinesthetic feedback of the phonatory system have been described to contribute to singers' pitch control [1, 2].

Auditory cues are commonly regarded as the obvious tool for pitch control in singing under normal circumstances. However, auditory feedback cannot explain the fact that singers are able to continue phonating accurately even when they cannot hear their own voices. This situation is typically experienced in solo singing when the orchestral accompaniment is loud; SPL values as high as 110 dB have been observed on orchestral podia [3]. Under such conditions, singers have to rely on the performance of a second intraphonatory feedback circuit, based on kinesthetic discharges.

The aim of the present study was to estimate the importance of auditory and kinesthetic feedback to pitch control in students beginning their professional solo singer education. The effect on pitch control was investigated in tasks differing in complexity, such as legato or staccato, or slow and fast singing.

The effects of a professional training of the singing voice should include a sufficient accuracy of intonation. A longitudinal approach, in which the singer is used as his/her own control would represent a promising opportunity to test the effects of training. Therefore, the singing students were reinvestigated after 3 years of education to assess the effect of training on pitch control in singing.

II. METHODOLOGY

In the initial investigation 28 singing students were examined at the beginning of their professional solo singer education at the University of Music Carl Maria von Weber, Dresden [4]. After 3 years of professional solo singer education, 22 students, 13 female and 9 male students, mean age 24,0 ± 1, 6 years, still continued their studies and could be re-investigated [5].

Subjects were asked to sing an ascending and descending triad pattern up to the twelfth and back on the vowel [a:] at a moderate degree of vocal loudness. The starting pitch, chosen so as to fit comfortably the pitch range of the individual subject, was given by means of a synthesizer. Each subject sang the sequence twice, first without masking noise, and immediately

afterwards with a masking noise presented via headphones. The masker was a white noise band-pass filtered (24 dB/octave) at 50 Hz and 2000 Hz. The SPL of the noise was 105 dB_{As}. The masking efficiently eliminated the auditory feedback.

The sequences without and with masking noise were recorded in different conditions: a) legato slow, b) legato fast, c) staccato slow, d) staccato fast. The slow and fast tempi corresponded to metronome settings of 40 and 160 beats per minute, respectively. The output from a portable electroglottograph (EGG) (Laryngograph, London, UK), and the audio signal as picked up by a microphone (distance to mouth 0.3 m) (ECM-959DT SONY, Japan) were recorded on a digital audio tape (TCD-D10, SONY, Japan). The identical test program was recorded again after training.

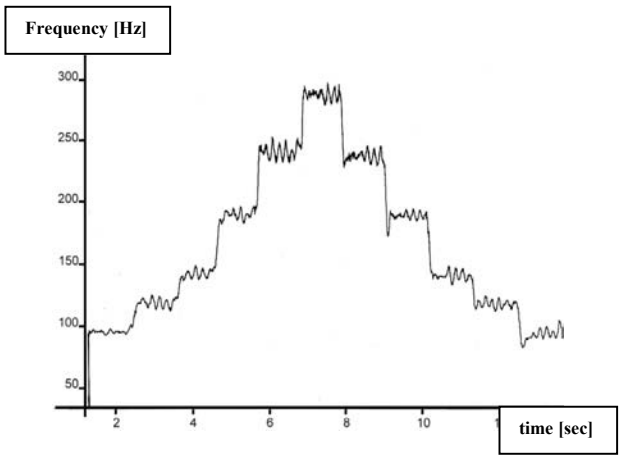


Figure 1. F0 contour of a recorded sequence.

Fundamental frequency (F0) was mostly estimated from the EGG signal using the Soundswell workstation program package which also displayed the resulting F0 contour on the computer screen (Fig.1) (Soundswell, Solna, Sweden) [6]. In some of the female subjects the EGG signal produced errors in the F0 measurement at high pitches. In such cases F0 was measured from the audio signal. For determining the mean F0 for each pitch, a set of complete vibrato cycles was selected from the quasi-steady state section, thus excluding onset and offset transients. The frequency distribution of this selection was analyzed, using the histogram module in the Soundswell package, which also displays the mean F0. The mean F0 of each tone was measured.

The sizes of the 10 intervals included in each triad sequence were determined by calculation of the F0 interval between adjacent tones, expressed in the logarithmic cent unit. The absolute values of the deviations of these intervals from their equivalents in the equally tempered tuning, henceforth the interval deviations, were determined and regarded as a measure of the accuracy of intonation. The averaged interval deviation of the 10 intervals contained in a complete triad sequence was defined as the mean interval deviation.

Interval deviation data were referred to a statistical analysis carried out by means of a repeated measures design (ANOVA), with time (before/after), masking (without/with masking), technique (legato/staccato) and tempo (slow/fast) as within subject factors.

III. Results

The measurements before the professional singing education showed a significant difference between the unmasked and masked conditions ($p<0.001$), mean interval deviations across all subjects amounting to 33.3 and 47.3 cent, respectively. The effect of masking appeared to be independent of technique and tempo. Figure 2 illustrates these results for the different conditions in terms of the distribution of individual mean interval deviations. Further, a significant difference was found between legato and staccato performances ($p< 0.001$) as well as between slow and fast performances ($p< 0.001$) [4].

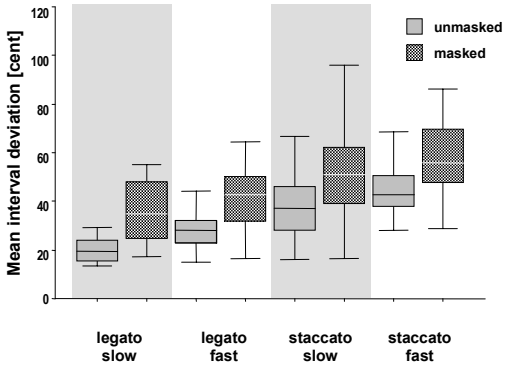


Figure 2. Box plot diagram showing the distributions of mean interval deviations (cent) for the different test conditions (subjects n=28). All data refer to the measurements before the singing education.

Comparison of the before and after education measurements did not show a general difference between these conditions. For the after education measurements, the masking increased the mean interval deviation across

all subjects from 35.3 cent to 45.1 cent [5]. Statistically, this effect of masking on pitch accuracy did not differ significantly between the before and after education measurements (p= 0.15). However, according to the ANOVA, there was a significant interaction effect of "time" and "tempo" (p= 0.001), reflecting different effects of education for the slow and fast performances. Intonation accuracy improved for the slow performances, the mean interval deviation across all subjects dropping from 37.7 cent before education to 32.7 cent after education. Fig. 3 shows the distribution of individual mean interval deviations for all slow performances, before and after education. The strongest effects appear for the masked test conditions, both for legato and staccato performances. No improvement of intonation accuracy was found for the fast performances after education.

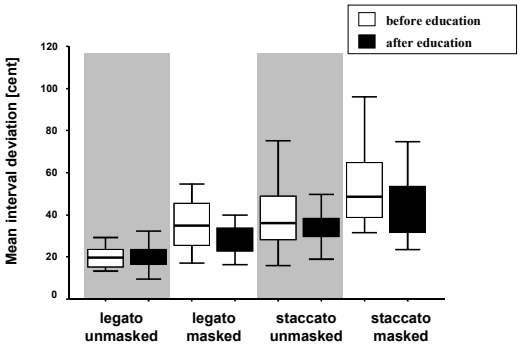


Figure 3. Comparison between before and after education data in terms of a box plot diagram showing the distribution of mean interval deviations (cent) for all slow tempo data (subjects n=22).

IV. Discussion

The present study was carried out to assess the significance of auditory and kinesthetic feedback on pitch control in singing and to investigate effects of training on both feedback circuits. The slow and fast as well as the legato and staccato conditions were included in our experimental design since they raise different demands on pitch control.

Intonation accuracy was found to be reduced by masking noise, by staccato as opposed to legato singing and by fast as opposed to slow performance. The masked and unmasked conditions allow an insight regarding the roles of the auditory and kinesthetic feedback systems in pitch control. Auditory feedback is commonly regarded as the main tool for pitch control in singing [7, 8]. However, under certain circumstances singers cannot hear their own voices, because the auditory feedback temporarily might be masked by the choral sound of the fellow singers or a loud orchestral accompaniment [3, 9] A significant effect of masking was observed, amounting to a mean deterioration of pitch accuracy by 14 cent at the beginning of the students' professional solo singer education [4]. This effect was only slightly smaller (10 cent) after education, a statistically non-significant difference. This suggests that the auditory feedback contributed to pitch control to a similar degree before and after education. The effect of masking was similar for the various tempo and technique conditions, see Figure 2. Therefore, the differences in intonation accuracy associated with these conditions should reflect the importance of the kinesthetic feedback.

The kinesthetic feedback circuit, a complex neuromuscular reflex system, depends on discharges of mechanoreceptors, mainly located in the intrinsic laryngeal muscles, the subglottic mucosa and the laryngeal joints [10, 11]. The afferent discharges from these receptors are fed back to the motoneurone pools in the brain stem operating as individual controllers for laryngeal action and to the overriding subcortical system [1]. Within the masked condition, intonation accuracy differed between the various tempo and technique conditions; a greater mean interval deviation was observed for the staccato than for the legato condition and also for the fast as compared to the slow conditions (see Fig.2). In a staccato performance singers would need to rely on an absolute neuromuscular memory of pitch while in a legato performance they could recruit also a relative neuromuscular memory [12]. The difference observed between staccato and legato performances suggests that the former memory is less precise than the latter.

Comparing data recorded before and after education, a significant improvement of pitch accuracy was found after education for the slow performances. For instance, for the masked slow staccato condition a mean pitch accuracy improvement of 9 cent was found after education. For the same condition, a study carried out by Ward and Burns showed a 17 cent better pitch accuracy in singers than in untrained subjects [2]. The difference between their results and our findings appear expected, given the fact that they compared singers and nonsingers. The improvement of intonation accuracy observed for the masked slow staccato task indicates that the accuracy of the absolute neuromuscular memory of pitch increased after education. Incidentally, this 'absolute kinesthesis' is important not only to staccato performances, where adjacent tones are separated by a pause. It is also essential for intonation at the beginning of a phrase, if no rehearsal of target pitch is allowed. In fast singing our study showed no improvement or even a modest impairment

was observed. Probably, a period of 3 years of professional training might not be long enough to improve pitch control in demanding vocal tasks such as fast singing. Also, the accuracy of measurement is smaller for short than for long tones; the shorter the tone sequence, the more difficult the pitch extraction.

It is interesting that our study showed no training effect for the basic, most easy condition – the unmasked slow legato. This task – singing slowly a triad or scale with normal auditory feedback – may reflect the limit of intonation accuracy, which would be reached early in any singing education. Finally, it is worthwhile to emphasize that, on average, the intonation errors were only slightly (10 cent) greater when the auditory feedback was eliminated. This implies that the kinesthetic feedback, contributes substantially to intonation accuracy.

V. Conclusion

The present investigation has shown that singers' intonation accuracy is reduced in the absence of auditory feedback. Under such conditions, the singers have to rely on kinesthetic feedback circuits. The performance of this feedback is significantly affected by the task that the singer performs. Thus, the mean intonation error was greater in fast than in slow singing. It was also greater in staccato than in legato singing. Professional solo singer education did not significantly affect the contribution of the auditory feedback to pitch control in singing. Such education seems mainly to affect intonation accuracy in terms of an improved accuracy of the kinesthetic feedback circuit.

References

[1] Wyke BD. Laryngeal Neuromuscular Control Systems in Singing. *Folia phoniat* 1974; 26: 295-306

[2] Ward WD, Burns EM. Singing without auditory feedback. *J Research in Singing* 1978; 1:24-44

[3] Jansson E, Axelsson A, Lindgren F, Karlsson K, Olaussen T. Do musicians of the symphony orchestra become deaf? In: *Acoustics of Choir and Orchestra.* Stockholm: Royal Swedish Academy of Music, publ no 52, 62-74, 1986

[4] Mürbe D, Pabst F, Hofmann G, Sundberg, J. Significance of auditory and kinesthetic feedback to singers' pitch control. *J Voice* 2002; 16: 44-51

[5] Mürbe D, Pabst F, Hofmann G, Sundberg, J. Effects of a professional solo singer education on auditory and kinesthetic feedback - a longitudinal study of singers' pitch control. *J Voice* accepted for publication

[6] Ternström S. *Sound swell manual*. Solna, Sweden: Sound Swell, 1991

[7] Elliot L, Niemoeller A. The role of hearing in controlling voice fundamental frequency. *Int Audiol* 1970; 9: 47-52

[8] Burnett TA, Senner JE, Larson CR. Voice F0 responses to Pitch-Shifted Auditory Feedback: A Preliminary Study. *J Voice* 1997; 11: 202-11

[9] Ternström S, Sundberg J. Intonation precision of choir singers. *J Acoust Soc Am* 1988; 84:59-69

[10] Abo-El-Enein MA, Wyke BD. Laryngeal myotatic reflexes. *Nature* 1966; 209:682-6

[11] Wyke BD. Laryngeal Myotatic Reflexes and Phonation. *Folia phoniat* 1974; 26: 249-64

[12] Sundberg J. *The Science of the Singing Voice*. Dekalb, IL: Northern Illinois Univ. Press. 1987

3rd International Workshop MAVEBA 2003, 187-190

ACOUSTIC ANALYSIS OF OVERTONE SINGING

M. Kob and C. Neuschaefer-Rube
Dep. of Phoniatrics, Pedaudiology, and Communication Disorders
Aachen University (RWTH), D-52074 Aachen

Abstract: **The articulatory configuration of an overtone singer is analysed with frequency analysis of the voice signal, sonographic visualisation of the tongue position, and analysis of the vocal tract impedance at the mouth. The biphonic character of the signal is observed in the spectrum plot. The sonographic analysis reveals a highly variable tongue position during production of a rising overtone. The high pitch of the produced biphonic sound is further analysed using the impedance technique. The extraordinary amplification of the melody pitch seems to be caused by the coincidence in frequency of two resonances. This findings support the theory that the overtone sound in sygyt style is a result of the filter effect of the vocal tract.**
Keywords: **Overtone singing, articulation, sonography, acoustic impedance**

I Introduction

The production of overtone singing has been a fascinating field of research since many decades. Trân Quang Hai gives an overview of the broad variety of different overtone styles [1, 2, 3]. The study of S. Adachi and M. Yamada [4] presents measurements and simulation of Xöömij singing in the sygyt style where a low pitch sound (drone) is accompaigned by a high melody pitch. Adachi supports the "resonance" theory, which considers the source for the melody tone to be a separated harmonic of the lower tone. F. Klingholz describes aspects of the voice source in [5]. Recent work of K. Sakakibara focuses on synthesis and analysis of the kargyraa style that is characterised by a very low fundamental frequency, probably due to vibrations of the ventricular folds, and a melody pitch [6].

Measured data of overtone singers are relatively rare. This might be caused by the fact that insight into the function of biphonic singing is of minor interest to most artists. Furthermore, the determination of voice physiology is rather invasive or very costly (laryngoscopy, MRI). However, completely noninvasive sonographic and acoustic measurements are possible [7].[1].

This contribution shall contribute to the understanding of the physical principle of the biphonic sound generation in the sygyt style. In a recent work [8, 9] a new method for analysis of the vocal tract configuration during overtone singing in the sygyt style has been developed. The method determines the acoustic impedance of the vocal tract at the mouth. These measurements are complemented by sonographic measurements of the tongue position and spectrum analyses of the voice signal.

II Methodology

2.1 Voice signal analysis

The voice of an overtone singer has been recorded during sustained phonation of a distinct overtone sound in sygyt style. In Fig. 1 the spectra of the voice signal at the mouth is shown for two cases. In the first (black line) the overtone singer did not yet amplify the melody pitch, in the second (gray/green line) the melody pitch was "switched" on. For a comparison to western phonation, in Fig. 2 the spectrum of the voice signal at the mouth is shown for the vowel /a/.

From a comparison between Fig. 1 and Fig. 2 it is obvious that the production of the overtone is different from the production of regular vowels. The amplification of the melody pitch over the amplitude of the fundamental is surprising since the vowel acoustics describes the vocal tract function mostly as a damping transmission line. Since the partials between the lowest few partials and the amplified partial are strongly damped, the latter is perceived as a separate sound.

[1] Sound examples from overtone recordings recorded at various occasions can be found at the Internet address URL: `www.akustik.rwth-aachen.de/~malte/overtone`

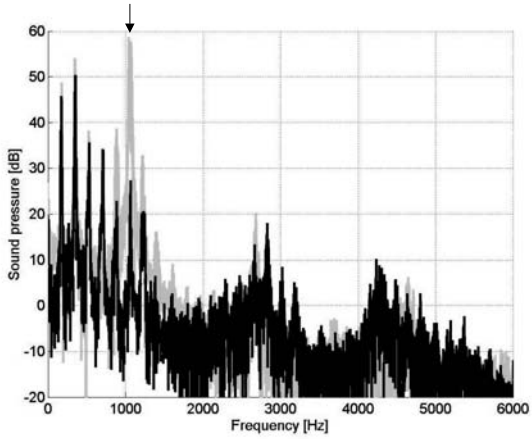


Figure 1: Voice spectra before the overtone is amplified (black) and after (gray/green)

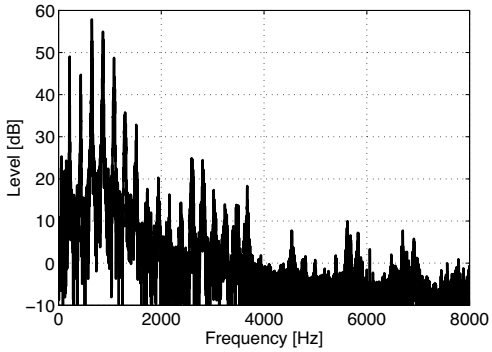


Figure 2: Voice spectrum of the vowel /a/

2.2 Sonographic analysis

The tongue movement of an overtone singer singing a rising sequence in sygyt style has been analysed with a sonograph using a 90°–3.5/5 MHz ultrasound probe. Within the same plane, the central submental position of the probe was not varied during the recorded performance of the overtone sequence.

In Fig. 3 and Fig. 4 the tongue position is shown as a sonographic image in the coronal respective mediosagittal plane. The image has been delineated by a marking procedure (white lines) that represents the interface between the dorsal tongue tissue and the oral air within the selected plane.

With rising pitch both, the images in the mediosagittal and in the coronal plane, exhibit a continuous change of the tongue position. In the mediosagittal plane the increasing backwards location of the dorsal tongue tissue can be observed, which

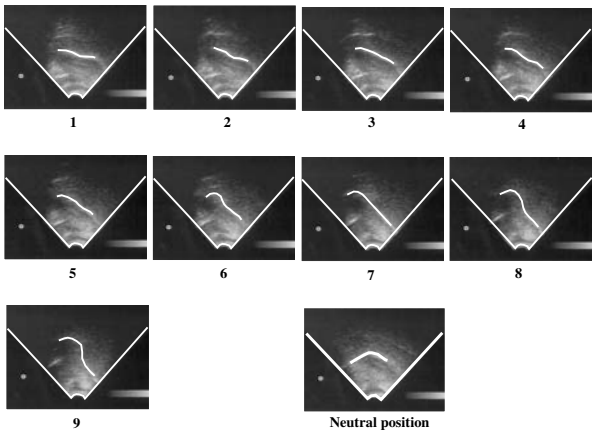


Figure 3: Sonographic mediosagittal view of the tongue during performance of a rising overtone sequence

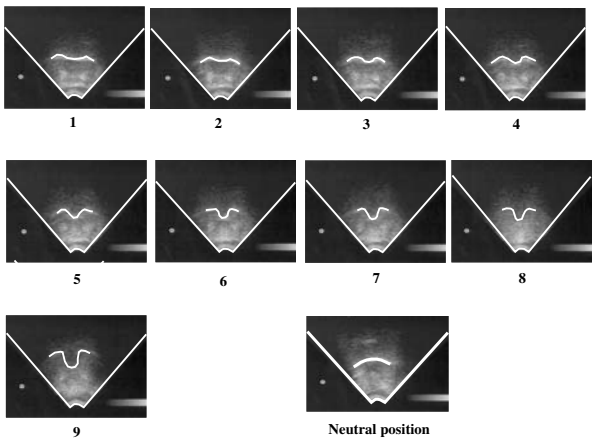


Figure 4: Sonographic view of the tongue in the coronal plane – during performance of a rising overtone sequence

forms a constriction in the vocal tract. In the coronal plane the forming of a channel with increasing depth can be observed.

2.3 Impedance analysis

The impedance analysis uses a method that determines the impedance spectrum of the vocal tract resonances. A sweep signal is generated, amplified, and emitted at the end of a horn. The horn is placed in such a way that the sound is emitted into the vocal tract. At the horn exit two sensors record the sound pressure p and the sound velocity v simultane-

ously. After a reference procedure and windowing the spectrum of the acoustic impedance Z is calculated from the Fourier spectra of both signals (Equation 1).

$$Z = \frac{\mathrm{FFT}(p)}{\mathrm{FFT}(v)} \quad (1)$$

The prototype of the measurement set-up is shown in Fig. 5. The signal flow is described in detail in [9].

Due to the sensor and loudspeaker specifications used in this set-up a frequency range from 500 Hz to 5 kHz could be evaluated.

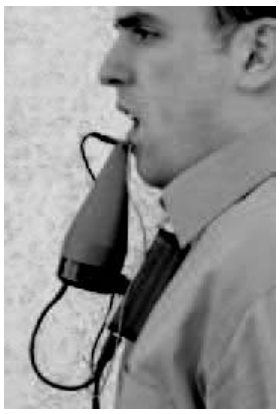


Figure 5: Prototype of the impedance measurement set-up

In Fig. 6 the impedance spectrum of the voice signal at the mouth divided by the free-field impedance Z_0 is shown for an overtone sequence similar to that described in section 2.2. The curves are shifted (from bottom to top) to visualise the course of time during the phonation of the rising overtone. With rising sequence the resonance structure of the vocal tract exhibits a strong resonance between 500 Hz and 2 kHz. In some cases, at higher frequencies of the melody pitch, a double resonance can be observed. Resonances apart from the one that corresponds to the melody pitch are not present.

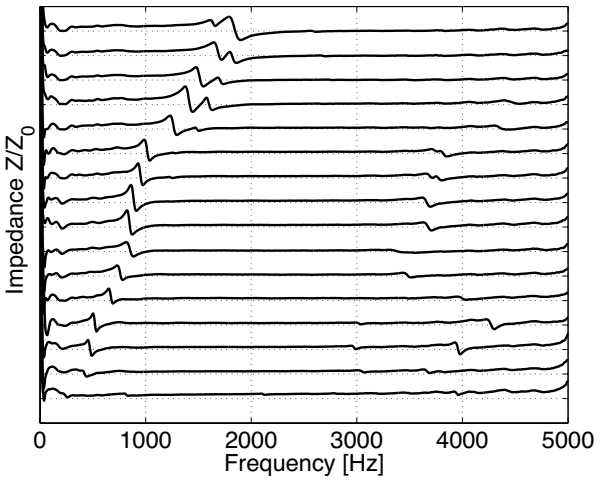


Figure 6: Impedance spectrum of an overtone sequence

In Figure 7 another impedance analysis is shown: the singer was asked to articulate the sound /a/ and then successively change the articulation towards an overtone sound. The sequence of shifted curves demonstrates the "morphing" from vowel /a:/ (bottom) to the configuration of an overtone (top).

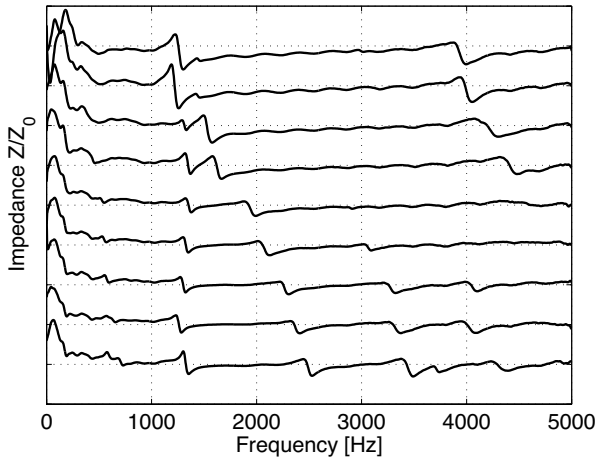


Figure 7: Impedance, morphing of [a:] to an overtone sound (bottom → top)

III Discussion

The sonographic analysis of the vocal tract configuration change with rising melody pitch indicates a change of the resonator structure.

The impedance plot in Figure 6 illustrates that, apart from the overtone resonances, only relatively weak resonances are excited between 3 kHz and 4 kHz. At higher resonances in the upper part of the plot a double resonance can be observed. This indicates that the overtone singer does not form a single resonance at the frequency of the melody tone but rather two closely neighboured resonances. This finding seems to be supported by the result from the morphing experiment shown in Fig. 7.

It is interesting to note that the second formant around 1300 Hz does not move significantly during the course of the sequence whereas the 3^{rd} formant moves from 2500 Hz downwards until it merges with the second one. The first formant of /a:/ cannot be resolved either because the lower frequency limit of the measurement set-up does not allow the visualisation or because first and second formant have the same frequency. All other frequencies are increasingly damped

towards the overtone configuration. However, a weak resonance can be observed at 4 kHz.

Even if two formants can be observed in the impedance spectrum, it is not clear how they are generated. Due to damping mechanisms in the vocal tract the longitudinal vocal tract resonances are not capable of producing very high quality formants.

One approach is to look for a different mechanism for the focalisation effect. The Helmholtz resonator is a well known resonator type that works as a main resonator in numerous musical instruments and — in the human voice organ — during whistling [10]. It could be possible that a combined longitudinal resonator and Helmholtz resonator could achieve a high quality formant when the resonance frequencies of both resonators coincide.

A numerical approach to verify this hypothesis is described in [8]. The calculation was based upon the equivalent area data published in [4]. A longitudinal resonator was assumed between glottis and constriction, and a Helmholtz resonator was supposed for the mouth cavity between constriction and mouth opening. The calculations confirm that the resonance frequencies of both resonators are of the same order of magnitude and that they are quite close for some overtones.

IV Conclusion

Within this contribution we could demonstrate that the simultaneous application of ultrasonography of the tongue, spectrum analysis of the overtone sound and impedance analysis of the vocal tract resonances during overtone singing support the filter theory. It could further be shown that in the case of sygyt style two resonances coincide at the frequency of the melody pitch.

In future investigations, the same procedure could be applied to the investigation of other singing styles, of both western and eastern cultures. Another current application of the impedance technique is the analysis of articulation disorders. With the help of this technique a mapping of acoustic resonances and dysfunction of the articulatory organs should be established.

Acknowledgments

Wolfgang Saus is thanked for participation in the measurements. The impedance measurement system has been developed within the Ph.D. work of the first author at the Institute of Technical Acoustics, Aachen University.

References

[1] Trân Quang Hai. Overtones — Recherches introspectives sur le chant diphonique et leurs applications (in french), 2001. URL: `http://tranquanghai.phapviet.com/english/introspectives.htm`.

[2] Trân Quang Hai. New Experiments on Overtone Singing. In *Dokumentation 3. Internationale Stuttgarter Stimmtage 2000*, pages 13–14. Akademie für Gesprochenes Wort, Staatliche Hochschule für Musik und Darstellende Kunst Stuttgart, 2000.

[3] Trân Quang Hai and D. Guillou. *Original research and acoustical analysis in connection with the Xöömij style of biphonic singing*, pages 162–173. Heibonsha, Tokyo, 1980.

[4] Seiji Adachi and Masashi Yamada. An acoustical study of sound production in biphonic singing, Xöömij. *J. Acoust. Soc. Am.*, 105(5):2920–2932, 1999.

[5] F. Klingholz. Obertonsingen. *Sprache — Stimme — Gehör*, 16:168–170, 1992.

[6] Ken-Ichi Sakakibara, Hiroshi Imagawa, Tomoko Konishi, Kazumasa Kondo, Emi Zuiki Murano, Masanobu Kumada, and Seiji Niimi. Vocal Fold and False Vocal Fold Vibrations in Throat Singing and Synthesis of Khöömei. In *Proc. ICMC*, pages 135–138. ICMA, 2001.

[7] C. Neuschaefer-Rube, W. Saus, G. Matern, M. Kob, and S. Klajman. Sonographische und endoskopische Untersuchungen beim Obertonsingen. In Hellmut K. Geissner, editor, *Stimmen hören – 3. Stuttgarter Stimmtage*, St. Ingbert, 2001. Röhrig Universitätsverlag. in press.

[8] Malte Kob. *Physical modeling of the singing voice*. PhD thesis, Aachen University (RWTH), 2002.

[9] Malte Kob and Christiane Neuschaefer-Rube. A method for measurement of the vocal tract impedance at the mouth. *Medical Engineering & Physics*, 24:467–471, 2002.

[10] A. Hirschberg, J. Kergomard, and G. Weinreich, editors. *Mechanics of musical instruments*, chapter Aero-acoustics of wind instruments, pages 291–369. Springer, 1995.

3rd International Workshop MAVEBA 2003, 191-193

LUCRETIUS, SONG AND MUSIC: A HISTORICAL APPROACH

Chantal Gabrielli
Department of Antiquities 'G. Pasquali', University of Florence, Florence, Italy

***Abstract*: The Latin poet Titus Lucretius Caro (I century B.C.), speaking of the origins of music in his work *De rerum natura*, expresses an interesting opinion on the scientific and technological progress that man has attained over the course of time.**

***Keywords*: Interdisciplinary paper, History of music, Song.**

The channels of man's expressions are numerous and varied. Song and music are without a doubt, possibly some of the richest ones, because they allow us to express our emotions in ways that are sociably and culturally accessible and acceptable by all.
The origins of song in Western culture can be identified with the first literary expressions of the Greek world. The famous Greek poet Homer is said to have composed the Iliad and the Odyssey, after having skilfully combined stories and episodes on ancient heroes that the oral bards, the so-called singers of tales, recited during banquets, travelling from one Greek city to another. The same structure of the Greek tragedy foresaw several parts sung entirely by a chorus. All the Greek lyrical texts were composed to be sung in public with instrumental accompaniment. In reality, music was practically present in all moments of communal life in Greek society, in religious ceremonies, in the sporting arena, in symposiums, in solemn festivities, even during political disputes. The importance of song remains in Latin culture to such an extent that the Fathers of the Church turned to song, in order to render efficacious their attempts at evangelisation among the numerous peoples of the Roman Empire, so diversified with regard to culture and language.
We know nothing of the ancient Greek and Roman music, which was composed before the III century B.C. The few musical texts that have been handed down to us from the Hellenistic and Roman period do not furnish precise and exhaustive reasons for their scarcity. There are only several inscriptions and a few fragments of papyrus, of which the interpretation and transcription are problematic. We do know that the musical system was based on the so-called tetrachords, that is, on elementary musical schemes, formed by the succession of four notes that, in Greek music, had the same function as the octave scales in our music. In addition, depending on the length of the intervals that separated these four sounds, various tetrachords existed: energetic, sweet or plaintive, according to the ambiance into which the song was introduced.

The long journey from the first sounds that man produced and individuated in nature up to the more modern compositions is also the result of a process of rationalisation that has at its core the relationship man-environment-music. And it is exactly on the origin of such a dialectic relationship between man and nature that I would like to focus, speaking of one of the most famous Latin poets who lived in the I century B.C. in Rome, Titus Lucretius Caro, and of his interesting ideas on song and music [1] [2] [3].
In the Fifth book of his work, the *De rerum natura*, the poet speaks of the origin and formation of our world and of the origin and development of humanity, describing several important steps that marked the progress of civilisation; the working of metals, weaving, the creation of language, the cultivation of the land, and song and music as well. From Greek philosophic thought comes the idea that men have learned the arts and crafts from animals, like the weaving of the spider's web, the construction of the swallow's nest, and music from the imitation of singing birds like the swan and the nightingale. However, in reality, in the discovery of the arts and techniques, man was guided by nature and pushed, according to the circumstances, by "need", by "necessity" and by what was "useful". In fact, the observation of nature caused mankind to desire its imitation; need, instead, forced him to look for instruments to better the conditions of his own life; while the benefit, or profit that emerged from the discoveries, continued to stimulate him to search with a desire to perfect his techniques [4].
Not by chance, Lucretius speaks of the origin of music (vv. 1379-1435) after the origin and the progress of techniques in the field of agriculture, almost wanting to indicate a separation between the arts that aided in the acquisition of goods, that were the first to occupy mankind to satisfy their impellent needs, and the other arts that followed, like poetry, music and song, when material necessities were no longer pressing, and one sought the pleasure of the spirit. Art, and therefore music, did not represent a necessity for man, but only a complement of his life: the liberal arts originated from the useful or advantageous. That is, from the pleasure that dance, song and poetry brought to people in moments of tranquillity or festivities. Here, the reference to the Greek philosopher Epicurus (IV-III century B.C.) is clear (his doctrine was diffuse in the work of Lucretius, his fervent disciple) and to the distinction he made between natural and necessary desires, those that are natural but not necessary, and those

that are neither natural nor necessary (Epicurus, *ad Men.*, 127, 130-131; *K. Δ.,* 15, 18, 29) [5]. Human needs corresponded to these desires and therefore some were real needs, others not; some were necessarily to be satisfied, others not necessarily. Epicurus appreciated the pleasure that music generated not only in the common person but also in the learned. He considered, however, the joy of music an unnecessary pleasure, that required a continuous learning process and constant practice, and for this reason could be criticised because it distracted the scholarly from a more important study, that which led to real knowledge, the study of philosophy.

As for everything including music, nature has been the inspirational model for man: song originated with the imitation of singing birds, the sound of the wind that blew within the reeds, created musical instruments like the flute and bagpipes. The flute, in fact, was constituted by only one reed with openings that were covered by the fingers in order to make music while the bagpipes were formed by larger reeds tied together, of various widths and sizes. The sound was produced by passing the lips from one reed to another. Lucretius only mentions wind instruments when he speaks of music while the lira, the most famous musical instrument in the Greek world, is not mentioned. Such an absence can be explained if one thinks of the context in which this passage was introduced. Lucretius is speaking of the theory of humanity and of the discoveries that man made to better his own life. The discovery of the lira is attributed to Hermes (Mercury), a God, and for this reason is not cited. Thus, the choice of the poet is conditioned by the idea that in the development of man from a savage state to modern day society, there was no divine intervention. It was need and reason to stimulate man and to make him advance over time. The gods, that lived isolated and indifferent in the *intermundia*, did not instruct man in the fields of agriculture, metallurgy or in the arts; it was rather nature and ingenuity that compelled man to improve when driven by necessity [6].

Moreover, the description of an idyllic scene of primitive life in the midst of music and dance offers Lucretius the occasion to reflect on the important differences that existed between the rustic music of the past and the refined music of his times and on the sense and value of these changes that the course of modern civilisation had imposed. Modern music, so perfected and refined, did not produce a greater pleasure than that which one's ancestors had experienced, who instead, with simplicity, used music and song to express sentiments of joy, pain, exaltation or depression. The same Greek philosophers gave great importance, in their meditations on culture and on the formation of man, to music and its relation to morality. They considered modern music to be in decline and felt that the refinements, which had been brought to it, were the means of its perversion. At the basis of such a moral consideration was the idea that true pleasure (*voluptas*), the goal of Epicurean thought, notwithstanding progress, must have a limit, or man, who is intent on its attainment, is destined to unhappiness. If in our lives we have not experienced something sweeter, we like what is at our disposal and this idea seems to prevail in whatever situation. If we then find something better, we immediately forget the previous pleasure and change our opinion of what we liked first. Progress, in fact, does nothing but manifest our restlessness, which then forces us to change. Moreover, it is a change, which can be compared to the individual who goes continuously from the house in the city to the house in the country and visa versa in the vain attempt to escape one's inner emptiness and to find happiness in a more pleasant place. For this reason, we must not ask progress to fill our emptiness but we should reserve our strengths for our inner perfection, following the precepts and teachings of the philosopher Epicurus [7]. These advise us to liberate ourselves from every ambition, every desire, every superstition, and every fear, to reach a state of perfect serenity similar to the beatitude of the gods. Man's happiness therefore could be identified with the healthy body and serene soul, and the pleasure (*voluptas*) which is merely the absence of pain for the body and anguish for the soul.

How can one not hear the modernity and relevance to the present in the words of Lucretius, when he reaffirms the moral damage that is caused to man by his search for continuously new objects and renewed pleasures, the same technology, if poorly used in war or in the production of superfluous consumer goods, conspires towards man's destruction and unhappiness, whose end is to then lead a chaotic and turbulent life like a violent and stormy sea: "And therefore the human race constantly suffers for nothing and consumes life in useless strife, because it doesn't know what limits possession has, and from where true pleasure is derived. That pushed life into the high seas little by little, and from the deep, unleashed the great waves of war." (vv. 1430-1435) [1] [2] [3].

REFERENCES

[1] *Lucrèce. De Rerum Natura*, A. Ernout - L. Robin eds., vol. III, Paris : Les Belles Lettres, 1928.

[2] *Titi Lucreti Cari De Rerum Natura Libri Sex,* C. Bailey ed., 2nd ed., vol. III, Oxford : Clarendon, 1950.

[3] *Titus Lucretius Carus, De rerum natura*, E. Flores ed., vol. I, Napoli : Bibliopolis 2002.

[4] A. Barigazzi, "Sulla chiusa del libro V di Lucrezio", *Prometheus*, vol. 15, pp. 67-79, 1989.

[5] G. Arrighetti, *Opere*. *Epicuro,* Torino: Einaudi, 1973.

[6] Ch. R. Beye, "Lucretius and Progress", *The Classical Journal*, vol. 58, pp., 160-169, 1962-1963.

[7] D. Furley, "Lucretius the Epicurean. On the History of Man", in *Entretiens Fondation Hardt sur l'antiquité classique*, vol. 24, pp. 1-27, 1977.

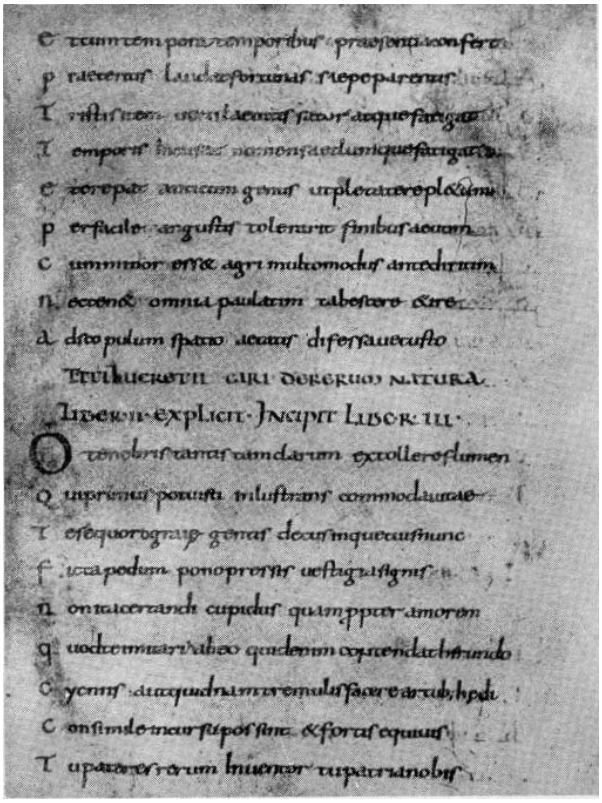


Figure 1 – Page of one of the oldest codex of *De rerum natura*, named *Oblongus* from its shape, and dated back to the 9th century A.D. It is preserved in the Library of Leiden, The Netherlands. The manuscript is also named *Vossianus*, from the name of its owner, J.Voss, a Dutch philologist. J.Voss owned another famous codex of *De rerum natura*, the *Quadratus*, which is now preserved in the Leiden Library as well.

Devices

3rd International Workshop MAVEBA 2003, 197-200

AN ASSESSMENT OF FLUENCY ENHANCEMENT TECHNIQUES FOR A TELEPHONE DEVICE FOR STUTTERERS

C. Bickley[1], M. Birnbaum[2], J. MacAuslan[3]
[1]Gallaudet University, Washington, DC, and Speech Technology and Applied Research, Lexington, MA, USA
[2]Martha Birnbaum Consulting, Cambridge, MA and Speech Technology and Applied Research, Lexington, MA
[3]Speech Technology and Applied Research, Lexington, MA

Abstract:
Telephone use is one of the most stressful communication situations for stutterers. We investigated a device to modify stuttered speech spoken into a telephone with the goal of ameliorating the stress and providing greater fluency. The device uses signal processing techniques to detect and correct certain types of dysfluencies. To assess dysfluent telephone input, stuttered speech exhibiting repetitions, prolongations, and blocks were recorded and then processed using phonetic classification technology to detect certain types of dysfluencies, and time-scale modification to correct them. In a series of experiments, listeners assessed the quality and intelligibility of the dysfluent (unprocessed) speech vs. the fluency-enhanced (processed) speech. Listeners assessed the processed speech as both more acceptable and more intelligible than the unprocessed speech.

Keywords : **acoustic analysis, landmarks, dysfluency, time-scale modification**

I. INTRODUCTION

The clinical literature on stuttering therapy frequently mentions that use of a telephone is one of the most commonly encountered stressful situations in daily living for persons who stutter (Zimmerman et al.,[7]; Adult Stuttering Therapy Tapes [1]; Bloodstein [4]).

It has been reported that stuttering affects approximately 1% of the population of the United States (Bloodstein [4]). As such, stuttering is a disabling speech impairment that is frequently a cause of significant stress to both the person who stutters and to communication partners. In a study that tested 19 stutterers and 19 matched normal controls, the frequency and severity of dysfluency were found to covary with physiological correlates of stress (Weber and Smith [6]).

The current work pursues the goal of offering stutterers an automated method for producing more fluent speech over the telephone. The study focused on three types of stuttering dysfluencies: **repetitions** of a syllable or sound, **prolongations** of a syllable, and **blocks**, or extended periods of silence. A multi-step approach was developed to identify and classify the three types of dysfluent events in acoustic terms, to recognize them algorithmically, and to correct them with a suite of signal processing techniques.

Our goals were threefold:
1) to determine how well certain acoustic events, identified as clusters of "landmarks", correspond to stuttered events;
2) to incorporate two established technologies with novel processing into a method that modifies stuttered speech by removing select dysfluencies; and
3) to determine which modifications are effective for improving listener judgments of speech acceptability.

II. METHODOLOGY

A. Overview and Techniques

The method for modifying stuttered speech incorporates two established techniques: landmark classification using a speech-event classifier, and time-scale modification of speech.

Landmarks are points in an utterance which mark perceptual foci and articulatory targets, and around which one may extract information about the underlying distinctive features (Stevens et al. [5]). Bitar and Espy-Wilson [3] have extended Stevens' theory to develop a knowledge-based signal representation based on phonetic features and associated acoustic events (the Event-Based Classifier, or EBS). EBS uses landmarks to classify acoustic events as one of several kinds of speech sounds. Some of the acoustic events, such as the ones associated with the phonetic feature `sonorant`, segment the speech signal into regions. Others, such as those associated with `nonsyllabic`, mark particular instants in time. The robustness of the acoustic events has been illustrated in a series of recognition experiments (Bitar and Espy-Wilson [2]).

Time-scale modification (TSM) of speech is a process of compressing or expanding the time-scale of an audio segment. A signal which is time-scale compressed has a shorter duration, while a time-scale expanded signal is longer in duration. Time-scale modification processing

has the important property of preserving the pitch, speaker identity, and intelligibility of the speech over a range of playback rates. In this way, the processed signal sounds like the same person speaking more slowly or quickly. This feature of TSM is particularly useful in the proposed elimination of stuttering dysfluencies because it is essential that acoustic characteristics relating to the perceived identity of the speaker be unaffected by fluency enhancement processing.

B. Data Collection and Transcription

Recordings were obtained from 3 subjects (two male, one female) who were representative of producing moderate to severe stuttered speech. The recordings contain examples of repetitions, prolongations, and tense blocks (as judged by a fluency therapist); some of the recordings contained fluent productions. The recordings were made during entry interviews with a fluency therapist. From an original set of 58 recordings, 43 were selected as the development set. The remaining recordings were used as reference data.

Two trained phoneticians manually edited each of the recordings in order to remove sounds not produced by the speaker. Disagreements between these two researchers were negotiated to establish a final agreed-upon set of utterances identified numerically on spectrographic output.

For each utterance, a trained speech therapist judged each stuttered episode as "repetition", "prolongation", "tense block", or "other". Utterances judged as containing at least one episode of type "repetition", "prolongation", or "tense block" formed the database of speech samples for this study. The distribution of these three types in the development set is shown in Table 1.

Table 1. Distribution of dysfluencies

TYPE	Number
Repetitions	24
Prolongations	7
Blocks	5
Other	7

C. Data Analysis

The development set of utterances was digitized and analyzed manually to determine patterns of acoustic landmarks that differentiate stuttered sequences from fluent productions. Spectrograms, formant tracks, and pitch tracks were computed using Sensimetrics SpeechStation™ and ESPS/Waves software. Spectrograms were hand-marked by the phoneticians for landmarks and associated features, based on acoustic parameters established in the literature. After a training period of hand-marking utterances together, inter-judge reliability was evaluated between the two judges on 10% of all utterances. Reliability for this task was defined as:

number of landmarks agreed upon by both judges / (number of disagreements + agreements)

Inter-judge reliability was 92%.

The development set was processed using the EBS software. The output of EBS was compared to the manually identified landmarks, as seen in spectrographic and waveform displays, to identify patterns of acoustic events in the stuttered speech. The goal of the analyses in this step was to identify the kinds of stuttered episodes that can be identified from the combined information of sound classes and patterns in landmark sequences.

D. Time-scale Modification of Data

The development set of stuttered utterances was manually altered to delete repetitions and audible block events of the type that would be detected automatically and to reduce the duration of prolongations within certain constraints of TSM processing.

An algorithm for editing episodes of stuttered speech was developed to meet conversational constraints. For example, ½ sec latency was maintained to preserve the real-time experience of a telephone conversation. In addition, maximum TSM speedups and slowdowns were established empirically to meet the ½ sec latency constraint.

These rules were applied to the development set of data to produce a set of 20 examples for evaluation. An original and the corresponding processed utterance appear in Figure 1.

Figure 1. Top: original "um..I hea- I hea- I hea- I hea- I hea- I heard about it, like um.". Bottom: processed "um.. I hea- n.. I hea- I heard about it, like um..".

E. Testing

Sets of recordings of paired original stuttered speech and modified speech were prepared for listener evaluation. The stimuli for this perceptual test comprised twenty phrases taken from entry interviews between a fluency therapist and her clients. The stimulus pairs were randomized with respect to speaker and phrase to form a set of 20 pairs. The order of presentation of the processed/original and original/processed pairs was also randomized, so that the listeners did not know which of a pair of stimuli was the processed file and which was the original speech.

Fourteen listeners evaluated the samples by listening over loudspeakers in an office-environment room. Each listener rated each phrase pair on a 5-point preference scale of 1 to 5, based on which phrase in the pair was more pleasant or more fluent. A "3" indicated "no preference" or that the difference between the sentences of the pair was not perceptible.

III. Results

Scores assigned by listeners were analyzed in order to determine intelligibility of fluency-enhanced utterances compared to the original stuttered ones in light of the requirement that any alteration to the speech signal not degrade the intelligibility of the message.

The scores from the listening test were recoded such that a 1 or 2 indicated strong or weak preference for the **processed** utterance and a 5 or 4 indicated strong or weak preference for the **unprocessed** utterance. Thus, scores below 3 denote preferences for the processed version over the unprocessed.

The average judgment score of the 14 listeners evaluating 20 stimulus pairs was 1.76, indicating a substantial preference for the processed speech. This result indicates that listeners found the processed speech in which dysfluencies had been removed or modified to be more pleasant and more fluent than the unprocessed speech. Overall, the preference for the processed utterances was 209 listener opinions vs. 31 for the unprocessed utterances: all listeners, all utterances. (p: infinitesimal). For even the least positive listener, the preference across all utterances was 8 (processed) to 4 (unprocessed) (p=.002, Fisher Exact Test). For even the least strong preferred utterance, the preference was 11 (processed) to 6 (unprocessed) (p<.0001, Fisher Exact Test).

Casual conversation normally contains occasional dysfluencies. If specific acoustic characteristics correlate with listener perceptions of stuttering, and if these characteristics can be detected and processed, the question remains, "Which acoustic characteristics cause listeners to perceive a speech event as 'stuttered', rather than 'occasionally dysfluent?'". A companion study was conducted to ascertain which automatically detectable stuttering events cause listeners to judge an event as "stuttered" and hence should be the focus of an automatic fluency enhancement device.

Listeners judged as "stuttered" those speech utterances with the following characteristics: irregular fundamental frequency, word-initial stop and fricative repetition, syllable repetition, lack of spectral and temporal variation, pauses, and whole-word repetition.

Those dysfluency types which are candidates for automatic algorithmic detection and correction are irregular fundamental frequency and stop and fricative repetition. Whole word repetition should not be a candidate for alternation because fluent speakers often repeat words in conversational speech. Similarly, pauses can be used intentionally and should not be removed without detailed analysis.

IV. Conclusion

Our initial goal was to show that we could improve at least half of the stuttered productions, and to not degrade the fluent productions of the speakers in the opinion of listeners. In fact, we were able to improve 90% of the stuttered productions and not degrade *any* of the fluent productions of the speakers, in the opinion of listeners.

Listener judgments indicate that several types of speech deemed dysfluent are good candidates for the automatic processing methods developed for making speech "socially acceptable" over the telephone.

References

[1] "Adult Stuttering Therapy Tapes," Stuttering Foundation of America, 1984.

[2] N. Bitar and C. Espy-Wilson, "A knowledge-based signal representation for speech recognition", *Proc. ICASSP '96,* Atlanta, 1996

[3] N. Bitar and C. Espy-Wilson, "A signal representation of speech based on phonetic features", *Proc. IEEE Dual-Use Technology and Applications Conf*, 310-316, 1995.

[4] O. Bloodstein, *A Handbook on Stuttering*, San Diego, Singular Publishing Group, 1995

[5] K. Stevens, S. Manual, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a model for lexical access based on features", *Proc. Internat. Conf. Spoken Lang. Proc.,* Banff, **1**, 499-502, 1992

[6] C.M. Weber and A. Smith, "Automatic correlates of stuttering and speech assessed in a range of experimental tasks", *J. Speech Lang. Hearing Res.,* **33,** 690-706, 1990

[7] S. Zimmerman, J. Kalinowski, A. Stuart, and M. Rastatter, "Effect of Altered Auditory Feedback on People Who Stutter During Scripted Telephone Conversations, *J. Speech Lang. Hearing Res.,* **40,** 5, 1130-1134, 1997.

3rd International Workshop MAVEBA 2003, 201-204

SPEAKING VALVES: INFLUENCE OF THE FATIGUE ON THE FLOW CHARACTERISTICS

G. Belforte, M. Carello, A. Dileno, M. Morero
Dipartimento di Meccanica, Politecnico di Torino
C.so Duca degli Abruzzi 24, 10129 Torino, Italy, e-mail: massimiliana.carello@polito.it

Abstract: **The *in vivo* operation of a speaking valve consists of two stages: 1) air passes through the razor-thin slit, the dome opens and the patient can speak; 2) the dome is closed and the patient cannot speak. The valve is thus subject to fatigue, as its service life is made up of a certain number of opening/closing cycles.**
Two types of valve were investigated: the Staffieri valve and a new valve prototype featuring a different angular extension of the razor-thin slit.
The investigation assessed fatigue degradation in valve flow characteristics; for this purpose a special test rig has been constructed.
Fatigue tests have been performed in four steps and the airflow resistance has been determined experimentally at the end of each step.
The experimental data have been used to make a statistical analysis to evaluate the effects of razor thin slit, type of valve, number of cycles and their interactions.
Keywords : **speaking valve, voice button, fatigue, flow characteristics.**

I. INTRODUCTION

Speaking valve is used for the rehabilitation of patients who have lost vocal function due to total laryngectomy. It is one-way valve, which thus permit expiratory air to pass from the trachea to the hypopharynx-oesophagus (direct flow) with as little resistance as possible, and prevent the passage of liquids in the opposite direction (reverse flow).

Previous papers discussed the experimental results obtained with two types of valves: the Staffieri and the new prototype [1, 2].

The Staffieri valve (a) and the new valve (b) are shown in Fig. 1. As can be seen from the figures, the most important differences between the two types of valve are the shape of the tracheal flange and the shape of the dome.

The authors established that the aerodynamic characteristics of the two types of valves are influenced by two important parameters: the type of dome and the razor-thin slit.

A properly test rig has been made, which reproduces valve opening/closing, to make the fatigue tests.

All valves were subjected to 50000 cycles in four steps. After each step the airflow resistance has been determined experimentally to establish the effect of fatigue on valve characteristics.

Fig. 1a – Staffieri valve

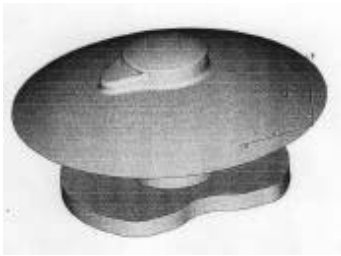


Fig. 1b – New valve

II. METHODOLOGY

The experimental plan involved 24 Staffieri valves and 24 new valves.

The hypopharynx-oesophagus exit (or razor-thin slit) is located at the base of the dome, and six different angular extensions α were considered (180°, 210°, 240°, 270°, 290°, 310°).

To determine the repeatability, four nominally identical valves were tested for each type and for each angular extension α.

The purpose of this investigation is to assess fatigue degradation in valve performance, in particular the airflow resistance.

Analysing the in vivo operation of the voice button has been observed that it consists of two stages:

1) The dome of the valve is lifted and opened by air. During this period the airflow passes through the razor-thin slit of the valve and the voice production is possible. The patient does not breathe but he can speak.

2) The valve is closed. In this second stage, the dome can be observed to move quickly until it is almost fully closed. This is followed by a slow final closing movement resulting from the material's elasticity, which positions the dome in contact with the oesophageal flange.

The reverse flow of food or saliva into the trachea must be prevented.

The authors have assumed a time of 6 seconds for each stage that has been considered a trade-off between actual in vivo conditions and the need to speed up fatigue tests, as total cycle time is thus equal to 12 s.

As valves carry out a certain number of cycles corresponding to the opened/closed stages, they are subject to a fatigue phenomenon.

The patient does not speak 24 hours a day then it is possible to assume around 200 cycles/day during in vivo operation.

The valves have been submitted to opening/closing cycles with direct flow, using airflow equal to the physiological rate (0.15 dm^3/s ANR).

The fatigue test rig is shown in Fig. 2. Valve PV supplies two timers T_1 and T_2, which regulate voice button V opening and closing time respectively. The additional counter C shows the number of cycles (open/close) logged. Resistance R is used to regulate airflow. Support S makes it possible to test 16 valves V simultaneously. Three identical test rigs were constructed so that all 48 valves could be tested simultaneously.

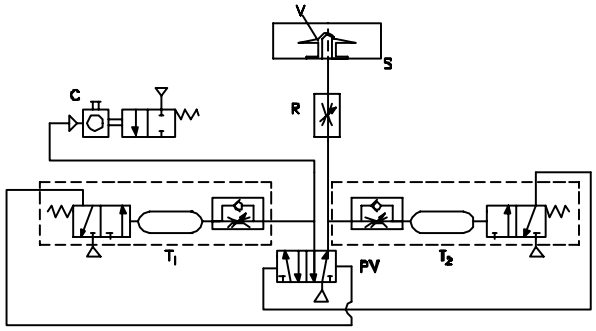


Fig. 2 – Test rig for fatigue tests

The proposed fatigue test method was optimised, carrying out four steps (10800, 16000, 26000 and 50000 cycles). At the end of each step the airflow characteristics have been experimentally obtained, in term of pressure and flow-rate, and the resistance of each valve has been calculated.

In general, it was observed that resistance drops as the number of cycles is increased, especially for small razor-thin slit angles α. This phenomenon was probably due to fatigue effects, which cause deterioration with a slight increase in α near the valve's oesophageal flange. For valves with larger α values, the influence of the number of cycles on resistance is comparable to experimental error. This applies to both Staffieri valves and the new prototype.

III. RESULTS

Figs. 3, 4, 5 and 6 show resistance versus flow rate for the Staffieri valve and the new valve prototype, with razor thin slit α=240°, after 10800 and 50000 cycles respectively.

The four dashed curves were obtained with four nominally identical valves, while the continuous line represents the average value. Standard deviation $\pm\sigma$ is also shown.

Average curves were taken into account in order to compare valves with different razor-slit extension α and domes.

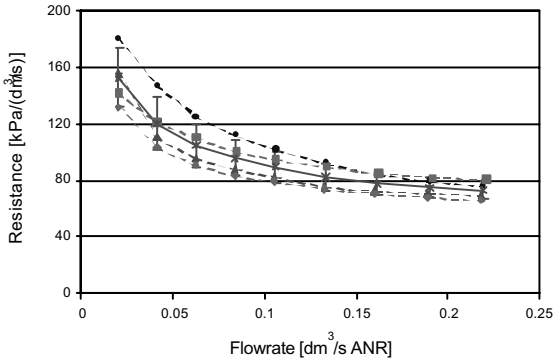


Fig. 3 – Resistance vs. flow-rate for Staffieri valves, α=240°, 10800 cycles

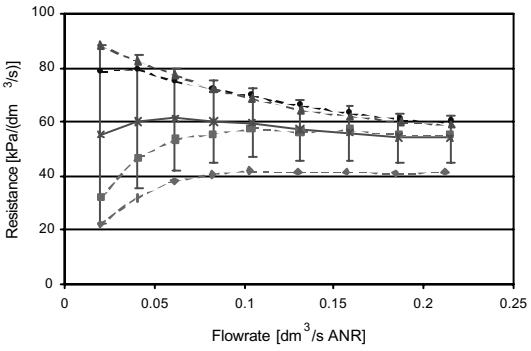


Fig. 4 – Resistance vs. flow-rate for Staffieri valves, α=240°, 50000 cycles

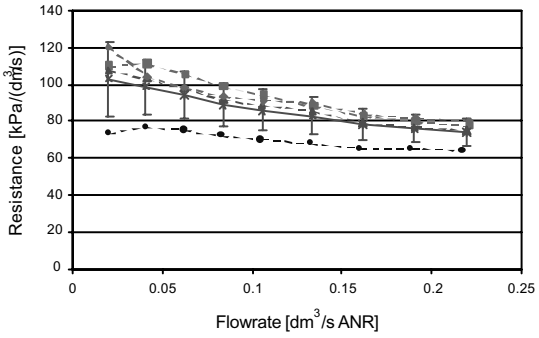


Fig. 5 – Resistance vs. flow-rate for New valves, α=240°, 10800 cycles

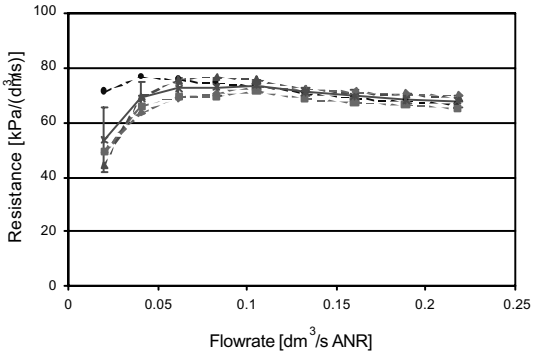


Fig. 6 – Resistance vs. flow-rate for New valves, α=240°, 50000 cycles

The influence of fatigue can be observed by considering the four steps. Figs. 7 and 8 show the average curves pressure P versus flow rate (continuous lines) and resistance R versus flow rate (dashed lines) for Staffieri and new valves respectively, both with α=240°.

Similar behaviours have been obtained for all the other value of α considered.

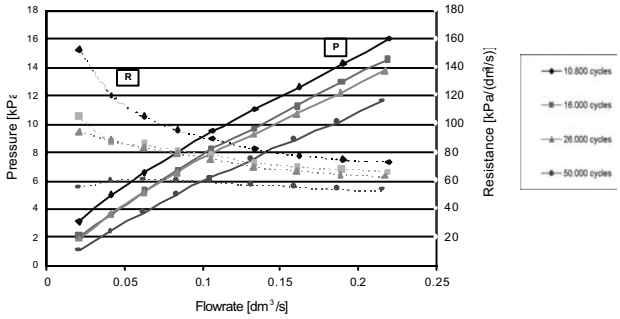


Fig. 7 – Pressure and resistance vs. flow-rate for Staffieri valves, α=240°

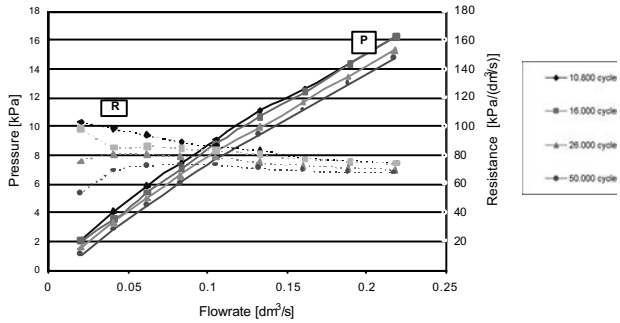


Fig. 8 – Pressure and resistance vs. flow-rate for New valves, α=240°

IV. DISCUSSION

An overall comparison can be made, for example, by considering all the Staffieri valves or all the new valves, and varying the angular extension of the razor-thin slit α while maintaining the same flow rate.

For the physiological flow rate (0.15 dm³/s ANR) in particular, resistance versus the number of cycles for the two type of valves are shown in Figs. 9 and 10.

Same behaviour has been obtained for different values of flowrate.

As can be seen, there is a general decrease in resistance as the number of cycle's increases.

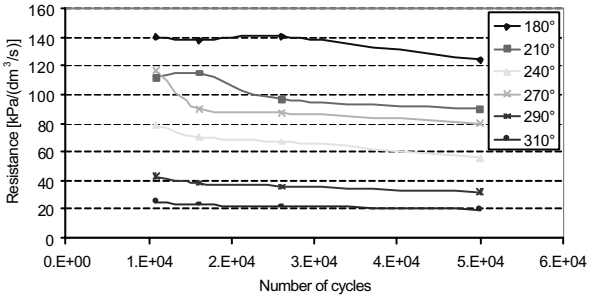


Fig. 9 – Resistance vs. number of cycles for Staffieri valves

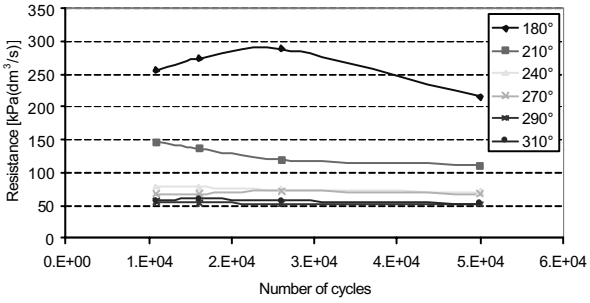


Fig. 10 – Resistance vs. number of cycles for New valves

The Staffieri valves have lower resistance than the new prototype for any given value of α.

The experimental results have been used to make the statistical analysis ANOVA. The ANalysis Of VAriance is a procedure that can be used to evaluate the effects of a group of experimental factors. It is then possible to assert with a certain significance level whether the data are influenced by these factors.

In this investigation, a sample of 48 valves was used, while the factors taken into account were:

- razor-thin slit extension α (factor A) with six levels (a=6) corresponding to the six angles (180°, 210°, 240°, 270°, 310°);
- number of cycles (factor B) with four levels (b=4) corresponding to the four steps (10800, 16000, 26000 and 50000 cycles);
- type of dome (factor C) with two levels (c=2) corresponding to the Staffieri valve and the new valve.

It was noted that there is a certain scatter in valve characteristics relative to the average value, thus repetitions were necessary to account for the experimental error. Specifically, four valves (n=4) were taken into account for every combination of razor-thin slit, number of cycles and valve type.

The analysis was performed using the airflow resistance obtained from the experimental data at constant flow-rate. Flow rates of 0.06 dm^3/s (low), 0.15 dm^3/s (physiological) and 0.2 dm^3/s (high) were taken into account. The result allows us to establish that for physiological flow the airflow resistance is influenced by the factors razor-thin slit, type of dome (and thus type of valve) and the interactions between razor-thin slit and dome; for these factors, in fact, the significance level β is very small and, therefore, the significance is very high. For the other factors/interactions, the value of β is large meaning that the effects of these factors cannot be regarded as significant.

Regarding the number of cycles, it should be noted that the risk of error in rejecting the null hypothesis ranges from 25% to 41% approximately. This is quite larger than the value usually adopted to consider an effect as significant (5% or less), however such factor could be not completely negligible.

Fig. 11 shows the significance level β of different factors-interactions, for the three flow-rates considered, and the β limit value (or critical β equal to 0.05). It can be noted that for the factors razor-thin slit, type of dome and the razor thin-slit/dome interaction the limit value of β is not exceeded and the significance behaviour moves away.

For the other double interactions, as well as for the triple interaction, the significance β is almost equal to unity, and then the factors have a smaller influence.

For the number of cycles it is possible to observe that β is larger than the critical β, but at the same time it is not negligible the effect on the valves performance.

V. CONCLUSION

The paper presented the results of fatigue testing on two types of tracheo-oesophageal valve: a Staffieri and a new prototype.

The valves was subjected to a certain number of cycles corresponding to the opened/closed stages, using a total cycle time equal to 12 s.

The airflow resistance has been checked at four cycling steps.

Generally has been observed a little reduction for the resistance increasing the number of cycles, for every razor thin slit and every type of valve.

Analysis of variance conducted with experimental airflow resistance values has allowed evaluating the influence of: angular extension of the razor thin slit, type of dome (or valve), number of cycles and their interactions.

The first two factors and the razor thin slit/type of dome interaction have a larger influence on flow characteristics (low significance). The number of cycles has a not negligible influence.

The other factors/interactions have a negligible effect (significance almost equal to 1).

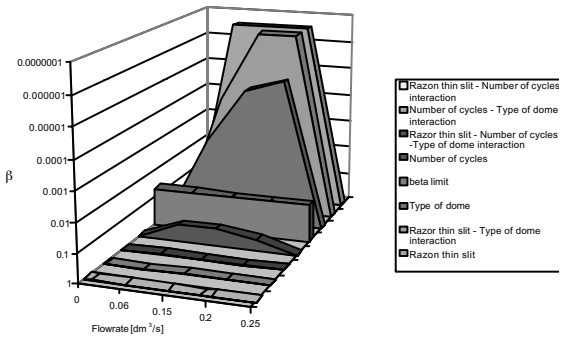


Fig. 11 – Significance level β of different factors-interaction

REFERENCES

[1] G. Belforte and M. Carello, "Experimental comparison between Staffieri and new tracheo-oesophageal prostheses", *2nd Int. Workshop MAVEBA*, Firenze (Italy), 13-15 September 2001.

[2] G. Belforte, M. Carello, C. Miani and A. Staffieri, "Staffieri tracheo-oesophageal prosthesis for voice rehabilitation after laryngectomy: an evaluation of characteristics*", Medical Biological Engineering and Computing*, 36, pp.754-760, 1998.

[3] C. Miani, A.M. Bergamin, A. Staffieri, S. Filippi, F. Miani and M. Zanzero, "Experiences in rapid prototyping: voice devices for patients who have undergone total laringectomy", *Fifth Int. Conf. on Advanced Manufacturing System and Technology*, pp.573-582, 1999.

[4] J.M. Heaton and A.J. Parker, "In vitro comparison of the Groningen high resistance, Groningen low resistance and Provox speaking valves", *The Journal of Laryngology and Otology*, 108, pp. 321-324, 1994.

[5] A.B. Jebria, M. Gioux, C. Henry, F. Devars and L. Traissac, "New prosthesis with low airflow resistance for voice restoration following total laryngectomy", *Medical & Biological Engineering & Computing*; 27, n. 2, pp. 204-206, 1989.

[6] H.F. Mahieu*, Voice and speech rehabilitation following laryngectomy,* Rijksuniversiteit Groningen, Doctoral dissertation, 1988.

[7] B. Weinberg, J.B and Moon, "Airway resistance of Blom-Singer and Panje low pressure tracheoesophageal puncture prostheses*", Journal of Speech and Hearing Disorders*,, 51, pp. 169-172, 1986.

3rd International Workshop MAVEBA 2003, 205-208

VLSI IMPLEMENTATION OF A TSM/FSM ALGORITHM

D. Breen [1], R. O'Neill [1], T. D. Smith[1], A. Th. Schwarzbacher[1].

[1]Dublin Institute of Technology, School of Electronics and Communications Engineering, Dublin 8, Ireland

Abstract: **The time scale modification (TSM) of speech is concerned with the compressing or expanding of audio signals in the time domain without affecting the signals pitch or naturalness. Conversely, the frequency scale modification (FSM) of speech is concerned with altering the pitch and formants of a signal without changing the signal duration.**

This paper describes a hardware implemented and optimized TSM/FSM system. Biomedical speech related applications for such a system include accelerated aural reading for the blind and improved speech recognition – In a voice controlled robotic system for the disabled, the speech can be effectively "slowed down" to improve the recognition rate. Other applications of the system include speech synthesis, foreign language learning, audio typing, and voice transformation.

Keywords: **TSM, FSM, VLSI**

I. INTRODUCTION

Time-Scale Modification (TSM) of speech consists of modifying the speed of the speech segment without affecting its naturalness or pitch. Conversely, Frequency-Scale Modification (FSM) of speech consists of modifying the pitch of the speech without changing the duration of the speech segment. Much research has been done in this type of speech processing since the early twentieth century and so a variety of algorithms exist. It is recognized however that some types of speech are more easily modified than others. Voiced speech segments are quasi-periodic in the time domain and in the frequency domain possess clearly defined pitch and harmonics. This is due to the vibration of the vocal cords while air is forced through the glottis. Typical voiced sounds are vowel sounds and broad consonant sounds such as 'y'. In contrast, unvoiced speech is spectrally noisy since there is no vocal cord vibration and the sound is instead produced in the oral cavity with the aid of the teeth and lips. Examples of unvoiced sounds are 's' sounds and 't' sounds. TSM/FSM algorithms exist to preserve the periodicity (continuity) and hence quality of voiced speech types and indeed music. The noisy nature of unvoiced speech means it is therefore unnecessary to employ algorithms for time-scale or frequency-scale modification. Distinctions between voiced speech and unvoiced speech may be based upon signal energy content and upon the signal's zero-crossings rate (the number of sign changes in a given period).

Algorithms for TSM and FSM fall broadly into three categories: time-domain techniques; frequency-domain techniques; parametric techniques. The level of output quality across the three categories is similar, however the time domain category is the most efficient in terms of computational burden [1]. By far the most widely used algorithm within this category is *synchronized overlap-add* (SOLA)[2] and its close relation, *pitch synchronized overlap-add* (PSOLA) [3]. However, the *adaptive overlap-add* (AOLA) algorithm due to Lawlor achieves similar quality with a saving in computational burden of an order of magnitude less [1]. Hence, this algorithm was selected over the others for implementation, since power consumption in a CMOS device is a strong function of switching activity and as such, the number of operations should be kept to minimum.

TSM and FSM are intrinsically related. If, for example, a speech segment is time-scale modified by a factor of two, the resultant speech segment is twice as long as the original segment. Playing this segment at double speed results in a speech segment that is the same duration as the original segment but its frequency content has doubled.

The possible applications for TSM/FSM algorithms are broad ranging. Possible speech related applications include speech synthesis, foreign language learning, audio-typing, accelerated aural reading for the blind, voice conversion, improved speech recognition, film/speech synchronisation, audio compression and noise reduction.

II. METHODOLOGY

For the modification of voiced speech the AOLA algorithm is used. The algorithm uses a fixed length rectangular stepping window and a simple peak alignment criterion to perform the overlap-add. Adjusting the overlap distance has the effect of increasing or decreasing the amount of expansion or compression. Overlap-adding in this way results in a local *natural expansion factor* or *natural scaling factor*. This factor is given by the ratio of the lengths of the original waveform and the newly formed synthetic segment and shall be denoted α_{ne}.

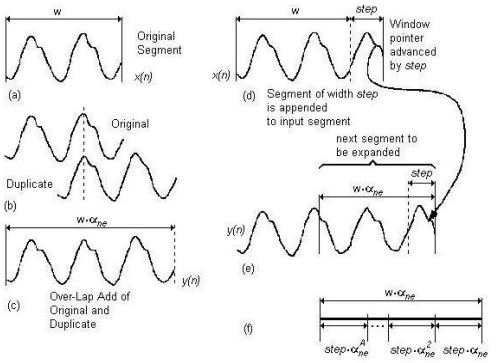


Figure 1: Steps in the AOLA algorithm.

Figure 1 [1] shows the alignment and output waveform synthesis procedures of AOLA for time-scale expansion. In the figure, the frame boundaries are marked by the dashed lines, *x(n)* is the input waveform and *y(n)* is the output waveform. Figure 1 (a) is the original segment to be expanded and is windowed with a rectangular window length *w*. In Figure 1 (b) the original segment is duplicated and the peak alignment procedure described earlier is performed about the dashed line. The result is shown in Figure 1 (c). It should be noted that this segment has been expanded by the natural expansion factor α_{ne}, and the length of the segment is now $w.\alpha_{ne}$. In Figure 1 (d) the input window is now advanced by a time *step*. Where this step ends coincides with the end of the next window to be expanded as indicated in Figure 1 (e). The segment preceding this new window is considered as expanded already and can be output. In (f), the expanded window $w.\alpha_{ne}$ is shown to be the accumulation of *A* expanded *step*s. From this the following equation is derived:

$$step\alpha_{ne}+step\alpha_{ne}^2+\ldots+step\alpha_{ne}^A \approx w\alpha_{ne} \quad (1)$$

$$\Rightarrow step \approx w.\frac{1-\alpha_{ne}}{1-\alpha_{ne}^A} = w.\frac{1-\alpha_{ne}}{1-\alpha_{de}} \quad (2)$$

There may be a discrepancy between the natural scaling factor α_{ne} and the desired scaling factor α_{de}. Therefore, *step* has to be updated for every advance step of the analysis window. The whole process repeats iteratively until the desired scaling factor is met. The AOLA algorithm accurately adapts to the local signal characteristics and ensures the signal is expanded by the desired scaling factor, α_{de}.

For time-scale compression the approach is similar. In this case the peaks or troughs are aligned as before but the signal to the left and right of the central overlapping region are discarded leaving a compressed segment. If the input segment has a natural compression factor of α_{nc} and the desired compression factor of α_{dc}, (5) becomes:

$$step = w.\frac{1-\alpha_{nc}}{1-\alpha_{dc}} \quad (3)$$

The algorithm can be recapitulated in the following three steps: 1. Isolate appropriate peaks; 2. Perform the overlap and determine the natural scaling factor; 3. Adapt as necessary and repeat.

The modification of unvoiced speech is a far simpler task. To achieve compression, the speech segment can simply be truncated as desired. As the frame boundaries are noisy, there will be no loss of continuity. In the case of expansion, a window of suitable length may be copied and appended to the end of the frame. As before, the integrity of the frame boundaries is preserved.

To ensure accuracy and efficiency, the system must discern between unvoiced speech and voiced speech. This distinction is based upon the short-term energy content and the zero-crossings rate mentioned earlier. In the case of short-term energy, a calculation is made of the energy content within a signal. Generally this energy content will be greater for a voiced speech segment than for an unvoiced segment of similar length. The total energy in a frame is given by the equation:

$$\sum_{n=1}^{N} s(n)^2 \quad (4)$$

Where N is the number of samples in the frame. Once the energy in a frame is known, it is compared with a reference value to decide if the energy present is indicative of voiced or unvoiced speech.

Since there will also be more energy in a voiced phrase that is louder than in the same phrase uttered softly, the zero-crossings decision mechanism is necessary. Unvoiced speech is spectrally noisy and will cross the time-domain origin a far greater number of times than voiced speech for a given segment. For a 20ms clean speech segment the crossing rate was found to be approximately 26 for voiced speech and more than 100 for unvoiced. These figures are used to determine whether the segment is voiced or not. Using both the methods outlined above, a more accurate decision is made.

III. IMPLEMENTATION

The system was coded and tested using VHDL. All VHDL code was synthesized and tested in the Synopsys Design environment. The system can be broken down into three major blocks of circuitry: 1. AOLA circuit; 2. unvoiced modification circuit; 3.decision circuit.

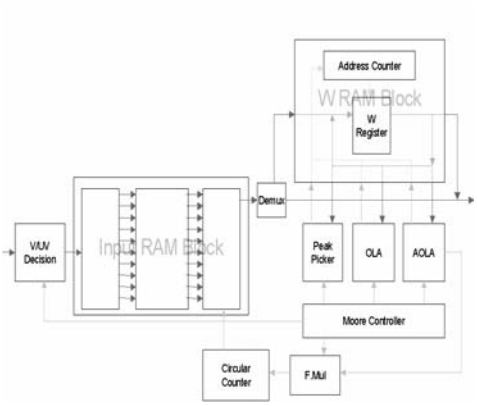


Figure 2: System block diagram.

In addition there is an input RAM structure, a RAM structure to hold the window of operation (W RAM) and a Controller module which synchronizes the system and controls system resets. The RAM structures are implemented as latch-multiplexer structures. These structures are more easily customizable, and are preferred to the RAM structures available from the existing libraries.

The AOLA algorithm is implemented in three modules corresponding to the three steps outlined earlier. The modules move samples as appropriate within the W RAM structure to perform the overlap-add, aswell as performing the step calculations of the algorithm. These latter operations include a number of multiplications and divisions. The divider employed operates on a subtract-shift-divide basis. The multiplier used is small, and operates within a single clock cycle.

The unvoiced modification circuit consists of a multiplier to establish α_{de} (desired scaling factor) in terms of the amount of samples (framesize x α_{de}), and a circular counter device which iteratively counts out the stored frame, α_{de} number of times.

The decision circuit consists of three modules, one for each of the decision mechanisms outlined earlier, and one to examine the results and make the decision. The modules operate on a running calculation basis. This allows a decision to be made at the input section as the input buffer is being filled with a reservoir of samples for working on. The input itself is a serial 8kHz sampled speech signal.

IV. Results

The system was tested with 8kHz quantised 8-bit speech samples. It was synthesized using the European Silicon Structures 0.7µm technology. The silicon area is shown in Figure 3. The total silicon area was 7518380µm^2 or 7.5mm^2, small enough for handheld devices.

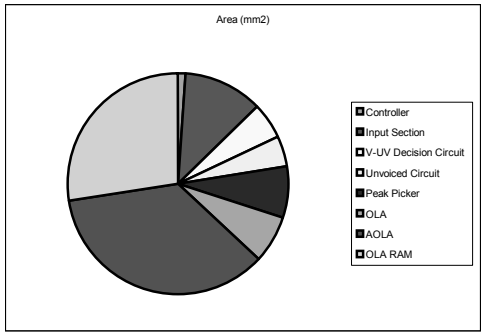


Figure 3: Silicon area of individual modules.

The following selected results show input and output waveforms for compression and expansion of both unvoiced and voiced speech. All inputs shown have a signal-to-noise ratio of 10. In the figure captions α_d is the desired scaling factor.

Figure 4: Unvoiced compression input (47.5ms) and output (35.625 ms), α_d = 0.75.

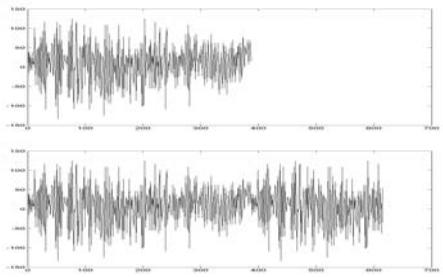


Figure 5: Unvoiced expansion input (47.5ms) and output (76 ms), α_d = 1.6.

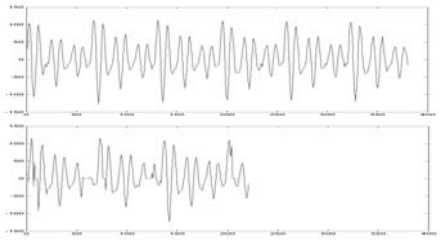


Figure 6: Voiced compression input (47.5ms) and output (29.25 ms), α_d = 0.6.

Figure 7: Voiced Expansion input (47.5 ms) and output (564 samples / 70.5 ms), α_d = 1.5.

The Circuit timing of the individual modules is shown in terms of propagation delays from input to output in the following table.

Module	*Delay (ns)*
Controller	2.08
Input RAM	2.10
Address counter	2.07
Decision circuit	2.08
Truncatenator	2.08
Peakpicker	1.98
OLA	2.09
AOLA	2.12
OLA RAM	2.14

Table 1: Propagation delays of individual modules.

The minimum operating clock frequency of the system is derived from the worst-case scenario time. This is when the maximum or minimum desired scaling factor is required and the minimum natural scaling factor occurs and the speech type is voiced. The time taken for the system to perform under these circumstances is approximately 10,000 clock cycles. Based on an 8kHz input signal the minimum operating frequency is therefore 80MHz. From the table above and based on the shortest route from input to output, the maximum allowable operating frequency was found to be approximately 159.75MHz.

V. Conclusion

The AOLA TSM/FSM algorithm was successfully implemented into hardware using high-level VLSI techniques and VHDL. In addition, a voiced/unvoiced decision circuit and an unvoiced speech modification circuit were also successfully implemented. Upon testing the system with various synthetic speech signals of varying signal-to-noise ratios (SNR), the circuit performed as expected. However for poorer SNR signals, the decision circuit occasionally made incorrect decisions. This problem may be overcome easily with a suitable adjustment of the reference threshold values for noisy environments.

The total silicon area was found to be 7.5mm^2 (based on the 0.7 μm library). This area is suitably small enough for handheld equipment such as mobile telephones, dictaphones or other portable speech processing equipment. However, it should be possible through additional optimization techniques to reduce this area further.

References

[1] Dr. Bob Lawlor, PhD Thesis, "Audio Time-Scale and Frequency-Scale modification", University College Dublin.

[2] Roucos, Salim and Wilgus, Alexander, M., "High-quality Time-Scale Modification for Speech", IEEE proceedings on acoustics, speech and signal processing, March 1985.

[3] Charpentier, Francis and Moulines, Eric, "Text-to-speech Algorithms based on FFT synthesis", ICASSP '88, pp. 667-670.

.

Voice/hearing impairment

3rd International Workshop MAVEBA 2003, 211-213

SOME EXPERIMENTS IN THE CZECH SPONTANEOUS SPEECH RECOGNITION DOMAIN

J.Kleckova, J.Krutisova
Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Czech Republic

***Abstract*— A spoken/dialog interpretation system is proposed, using prosodic information systematically at all processing stages. A prosody modul is used for parsing, dialog understanding, translation, generation and speech synthesis.** [1]
Keywords: Dialog system, spontaneous speech, prosody

I. Introduction

Instead of trying to match each unique sensory input directly onto a huge number of lexical entries, listener first recode the variable input into a smaller number of abstract units like phonemes or syllables that in turn serve to contact the lexicon. The intermediate representation based upon units can potentially guide the segmentation process. For instance, the onset of prosodic units or strong syllables could be used as starting points for the lexical matching process. To the extent that these segmentation points are likely to correspond to word boundaries, such heuristics would be helpful in reducing wasteful attempts to match the input with misaligned lexical candidates. In the framework that attributes a central role to intermediate levels of representation, we are led to search for the nature of the units making up this representation. The experiments reported in this paper were performed on a subset of Czech sentences. The computer assisted acoustic analyses allows for studying even subtle phonetic differences of pitch or stress, so that possibility to investigate the functional roles of these differences becomes possible. The important segmental characteristics are position and movement of formants, spectral tilt. The most important suprasegmental (prosodic) characteristics are pitch contour, rhythm, amplitude, prosodic boundaries, length of pauses, accents. For segmental characteristics description we concentrated on vocalic and its possible modification. Perception tests showed that the measure of a vocalic quality keeping and a vocalic quantity keeping together with a correct consonantal groups realization represents the primary segmental attribute of the utterance style.

II. New Aspects Of Segmental Characteristics

The estimated values represent results of the detailed experimental analysis of one speaker's spontaneous non-official speech (to avoid individual variations). The speech was realized a professional speaker. The corpus includes:

- Total 3040 syllables were measured.
- Total 326 vowels were analysed.

These vowels were realized in a comparable phoneme environment.

Their frequency representation shows the following table:

TABLE I
The frequency vowels representation in Czech spontaneous speech.

Vocal	Number	%
e	78	24%
o	75	23%
a	56	17%
i	53	16%
u	22	7%
í	21	6%
á	14	4%
ú	5	2%
é	2	1%
ó	0	0%

In experiments with a duration primarily we concentrated on phonologically short vowels. The previous task hypothesis was substantiated, the method "analysis by synthesis" [1] was verified: the average duration phonologically short vowels ahead of a pause achieves twice the phonologically short vowels average duration. The average duration of particular vowels don't differ, only the vowel /i/ is rather longer, because it occurs at the final position, vocal is inherency longer.

TABLE II
The average short vowel length in the spontaneous speech [ms].

Short vocal	/a/	/e/	/i/	/o/	/u/
Total average length	71	60	77	63	71
Average length at the stress group end	142	114	130	159	81

It was showed, that the phoneme environment doesn't feature the vowels length significantly. A phonologically short vowel prolongs before a pause so that its length exceeds the average phonologically long vowel length by 100 ms.

The table VI summarizes the measured values :

In following experiments we determinated the average of the F1 and F2. A comparison with conclusions of previous paper

[1]The work presented in this paper was supported by the Grant Agency Research Project No. 201/02/1553.

TABLE III
THE VALUES OF FORMANTS F1 AND F2 (CZECH VOWELS IN THE SPONTANEOUS SPEECH).

Vowels/Formants	F1	F2
a	650	1460
á	740	1470
e	490	1630
é	600	1660
i	320	2100
í	400	1480
o	460	1150
u	340	990
ó	310	930

suggests that articulation of Czech vowels may be changing. The articulation shift seems to reveal variation in measured formants frequency values. The comparison of the current situation with the previous papers was summarized in the table V:

TABLE IV
FORMANT F1 VALUES COMPARISON WITH THE PREVIOUS PAPERS.

Vowel	F1	F1	F1	F1
	[5]	[2]	[3]	Experiment
a	750	850	660	670
á	795	870	740	745
e	572	520	490	500
é	510	500	600	580
i	355	250	390	380
í	326	200	320	325
o	580	510	460	480
ó	530	490	-	500
u	385	260	310	315
ú	350	230	340	345

TABLE V
FORMANT F2 VALUES COMPARISON WITH THE PREVIOUS PAPERS.

Vowel	F2	F2	F2	F2
	[5]	[2]	[3]	Experiment
a	1280	1390	1450	1440
á	1175	1350	1470	1470
e	1660	2020	1630	1620
é	1750	2090	1660	1690
i	2120	2460	1890	1900
í	2230	2620	2100	2150
o	982	990	1160	1070
ó	900	920	-	890
u	758	730	930	900
ú	680	670	990	820

The present tendency demonstrates more open vowels in general and some reduction of differences in vowel quality among different phonemes. Changes of formants values by the particular vowels aren't equivalent. The first formant (F1) values put near the average value, that decreases by 22% in comparison with the previous measuring. The shifts are relative symetric; the forward vowels set (range) and the back vowels set (range) even put near the average values. The standard pronunciation divergence of a consonant articulation restricts to their duration changes. The analysis by ear and by experiments demonstrated the consonant duration is markedly increased due to emphatic accent.

III. INTONATION

The analysis of spontaneous speech showed interesting results in connection intonation. The professional speaker's utterance was analysed but his utterance has been attached attributes of unready spontaneous speech (for example free syntax). The intonation analysis was achieved on the short sentences with the definite syntax structure and the intonation at the end of sentences.The average F0 in this type of sentences is 120 - 163Hz, the standard deviation 159 - 211Hz. Maximum value F0 was founded repeatly (regularly, at all cases) on the posttonic syllable also in cases of short neutral (indifferent) sentences. There is another interesting conclusion - a terminal intonation very often is missing. The course of F0 is standard, e.g. in the lower third of the used range, maximum is usually in the first third of sentence and then it decreases to intonation minimum at the last syllable. The exception from this rule was held: The melodic top is fixed at posttonic syllable.

A. Intonation Scheme

From the point of the analysis goal, an important observation is that steadily repeating intonation schemes can be identified at the functionally equivalent syntactic positions. For example: The sentences are intonationally terminated one syllable before the end of the sentence. Then the last syllable is the first intonation syllable of the next intonation unit.

B. Differences of reading text

The intonation implementation of reading text is based on the contrast principle(fundamental). A rising or falling intonation may be correlated with incompleteness, and a falling intonation indicating completeness may also permit other intonation patters. The continuous gradient on the smaller groups is kept in the range of the one intonation unit. There is the most emphatic melodic contrast of consecutive syllables in the whole sentence. From whence it follows that the contrast principle and the theory of the maximum F0 at the posttonic syllable are claimed. In some cases the perception margin isn't given due to intonation contour (vide the fundamental frequency F0) but it is shown due to the emphatic vocal retardation and prolongation at the last syllable.

IV. EXPERIMENTAL RESULTS

The main reason, why the use of prosody in recognition system is not easy, are:

1) segmental (i.e. word chain) and suprasegmental (i.e. prosodic) information influence each other,

2) the prosodic functions which are realized to a great extent with the same prosodic parameters interfere with each other,
3) the use of prosodic means is optional - a specific function can be expressed with prosody but it does not have to, e.g. when other grammatical means are already sufficient.

The ability of the listeners to identify correctly and almost instantly a word from among the tens of thousands of other words stored in their mental lexicon constitutes one of the most extraordinary human cognitive feats. The speech signal indeed presents a formidable challenge. Both the speech is variable (every word takes on a different phonetic shape each time it is produced - the existence of large numbers of a highly similar words in the lexicon makes this variability even more troublesome) and speech is continuous (unlike written text, it contains no systematic spaces or reliable markers to indicate where word or utterance ends and the next one begins). The intonation often serves an information of a broad meaning nature. The fact that rising or level intonations are correlated with incompleteness and falling intonation with completeness admits other utilizations of the intonation. One of them helps to make clear the interpretation of potentially ambiguous utterances. The prosody is a very complex subject. Besides the intonation the hierachy of pauses is very important. Pauses of standard length in the places of punctuation marks between syntactic units are felt as bizzare in the spontaneous speech. After several experiments have been treid out, a three-tier pause hierarchy seems acceptable in Czech.

TABLE VI
THREE-TIER PAUSE HIERARCHY.

Pause	Duration of pause [ms] for speech rate	Classification of punctuaction marks
P1	8 - 10	,
P2	80 - 100	- :
P3	200 - 240	; . ? !

To make finer distinction of pauses would require to respect semantic relations of units in the dialog.

To summarize the results of spontaneous speech analysis we can state that we are able to detect the types of the sentences.

The intonation analysis was achieved on the short sentences with the definite syntax structure and the intonation at the end of sentences. The results of spontaneous speech analysis is carried into effect in the several experiments.

V. CONCLUSION

The analysis of spontaneous speech showed interesting results. It is currently that prosodic features have a very high significance for the dialog system. In the first phase, the prosody modul was developed that does not use phoneme-based but only word-based information. In the second phase, recognition system uses segmental and suprasegmenal characteristics in the several different modules. To describe the functions of the relevant features of sentence prosody would mean a significant step on the way towards a unified description of the system of language as a whole, from the phonetic form of the sentences to their underlying structure.

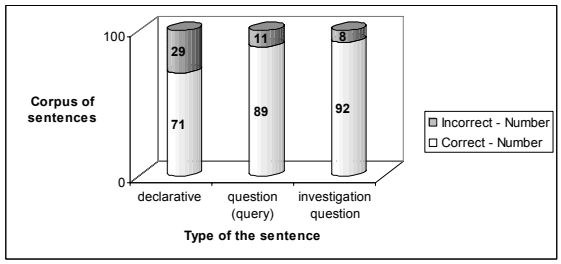


Fig. 1. Type of sentences.

REFERENCES

[1] Kleckova, J.,Krutisova, J., Matousek, V., Schwarz, J.: ∫Important Prosody Characteristics for Spontaneous Speech Recognition.∫ In: Proceedings of the 9th International Conference on Neural Information ICONIP 2002, Singapore, 2002.
[2] Dohalska, M., Dubeda, T., Mejvaldova, J.: ∫Dialogue of Analysis and Synthesis: Real Speech and Speech Synthesis.∫ In: Wodarz, H.-W., Palkova, Z.:Papers in Phonetics and Speech Processing, Forum Phoneticum, Hector,Frankfurt am Main,1999.
[3] Dohalska, M., Dubeda, T., Mejvaldova, J.: ∫Some Data on Spontaneous Czech: Prosody in Numbers.∫ 11th Czech-German Workshop - Speech Processing, Praha, 2001.
[4] Kolinsky,R.: ∫Spoken Word recognition: A Stage -processing Aproach to Language Differences.∫ In: European Journal of Cognitiove Psychology, vol.10, pp.1-40.
[5] Palkova, Z.: ∫Fonetika a fonologie cestiny∫. Univerzita Karlova Praha, 1994.
[6] Selkirk E.: ∫Sentence Prosody: Intonation, Stress, and Phrasing∫. In: Handbook of Phonological Theory. Ed. by J.A. Goldsmith, Oxford: Basil Blackwell, 1995, pp.550-569.

3rd International Workshop MAVEBA 2003, 215-218

TIME SYNCHRONIZATION OF SPEECH

Barbara Resch and W. Bastiaan Kleijn

Speech Processing Group, Dept. of Signals, Sensors and Systems
Royal Institute of Technology (KTH), Stockholm, Sweden
{barbara.resch,bastiaan.kleijn}@s3.kth.se

A time synchronization system is a helpful tool for different applications, such as language education and speech therapy. We present a system that performs temporal alignment of two utterances of the same phrase. The system consists of two parts. In the first part the time warping function is determined with Dynamic Time Warping (DTW). In the second part the time scale of one utterance is modified according to the time warping function. To obtain good performance, the dynamic time warping algorithm required significant modifications. Our listening test confirms that our time synchronization system has high precision and the resulting speech utterances are of natural quality.
Keywords: Time Synchronization, Time Scale Modification, DTW, WSOLA

I. Introduction

A system that time-aligns two utterances of speech can be used as a tool in language education and in the therapy of speech disorders. The acquisition of a good pronunciation is an important issue in language education. Time synchronization is valuable for this purpose. In speech therapy, synchronization can be applied in the therapy of voice problems, articulation problems, or in accent modification therapy.

In both, language education and speech therapy it is of great importance to observe certain differences in specific speech sounds. These differences get emphasized, if the temporal difference is removed The time synchronous utterances can be listened to either simultaneously or separately.

It is especially useful to change the speaking rate of the client or student if his or her utterance is nonuniform in speed. The parts of the sentence containing 'difficult' sounds, which need special attention and concentration to be pronounced properly, will often be spoken much slower than the remainder. Hearing the sentence in a natural speed, spoken with the own voice, encourages a natural way of speaking.

Another application of the time synchronization system can be found in the audio-for-video industries. Synchronization can be applied for dubbing material with another voice or post synchronization of outdoor recordings with studio recordings.

In the following, the signal that is to be modified in time is called the *source signal*, the resulting modified signal the *target signal*, and the signal that serves as the reference for the time scale the *reference signal*. Fig. 1 shows a block diagram of the described system. There are two parts: in the first part a relationship between the two utterances is established; in the second part the time scale of one utterance is modified to match the time scale of the other one.

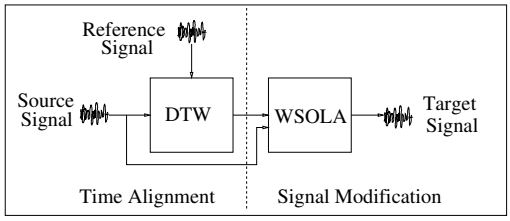


Fig. 1. Diagram of the time synchronization system.

The first part is realized with Dynamic Time Warping (DTW) [1]. DTW is mainly known from speech recognition, where it was featured in most systems in the 80's. Later on it was largely replaced by Hidden Markov Models (HMM's), as they proved to be advantageous for several reasons [2]. Although displaced from speech recognition, DTW has been in use in the 90's and later on for different applications such as speaker identification systems [3], signature verification systems [4], and in recent work for gesture recognition [5].

For the second part of the system Waveform Similarity Overlap and Add (WSOLA) synthesis [6] was selected. It falls in the class of time domain based Overlap and Add (OLA) methods. The idea behind the OLA synthesis methods is to create synthesized speech by concatenating small segments of speech. In doing so, the periodic structure of the speech signal has to be preserved. The different OLA methods such as the Synchronous Overlap and Add method (SOLA) [7] or Pitch Synchroneous Overlap and Add method (PSOLA) [8] offer various related solutions to this problem.

In [9] Verhelst presents an earlier system for time synchronization based on DTW and WSOLA. In his system DTW is applied without constraints or modifications. This leads to problems with the sound quality of the modified utterance if the reference and source signal are not sufficiently similar [9]. With our approach we account for acoustic and phonetic differences by introducing an accumulative local penalty constraint and a smoothing stage to the Dynamic Time Warping (sections 3.1 and 3.2).

II. Methodology

This section provides a short description of Dynamic Time Warping (DTW) and Waveform Similarity Overlap and Add (WSOLA) synthesis.

2.1. DTW algorithm

Dynamic Time Warping is a pattern matching algorithm with a non-linear time normalization effect. It is based on Bellman's prin-

ciple of optimality [10], which implies that, given an optimal path ϕ from A to B and a point C lying somewhere on this path, the path segments AC and CB are optimal paths from A to C and from C to B respectively.

The dynamic time warping algorithm [1] creates an alignment between two sequences of feature vectors, $(t_1, t_2, ...t_N)$ and $(s_1, s_2, ...s_M)$. A distance $d(i, j)$ can be evaluated between any two feature vectors t_i and s_j. This distance is referred to as the local distance. In DTW the global distance $D(i, j)$ of any two feature vectors t_i and s_j is computed recursively by adding its local distance $d(i, j)$ to the evaluated global distance for the best predecessor. The best predecessor is the one that gives the minimum global distance $D(i, j)$ at row i and column j:

$$D(i,j) = \min_{m \le i, k \le j} [D(m,k)] + d(i,j). \quad (1)$$

The computational complexity can be reduced by imposing constraints that prevent the selection of sequences that can not be optimal [1]. Global constraints affect the maximal overall stretching or compression. Local constraints affect the set of predecessors from which the best predecessor is chosen.

2.2. WSOLA algorithm

Waveform similarity overlap and add (WSOLA) is a time domain based algorithm for time scale modifications of speech [6] [11]. It gives high quality speech and allows scaling factors that may be specified in a time-varying fashion. One major advantage of the WSOLA method is that, in contrast to PSOLA [8], no pitch estimation is needed .

In OLA [12] (overlap and add) synthesis the modified signal is obtained by excising segments from the input signal, repositioning them along the time axis and performing a weighted overlap addition to construct the synthesized signal.

The basic idea of the WSOLA algorithm can be best explained graphically (see Fig. 2). The time warping function $\tau^{-1}(L_k)$ assigns one segment of the source signal to each synthesis instant L_k in the target signal. A timing offset Δ_k within a range of $2\Delta_{max}$ around the time warping function $\tau^{-1}(L_k)$ is needed to avoid pitch period discontinuities and phase jumps. In this way a proper segment synchronization in the synthesized signal is achieved. The timing offset Δ_k is determined such that the synthesized segment maintains maximal local similarity to the natural continuity existing in the original signal. Assume segment (1) in Fig. 2 was the last segment excised from the source signal and added to the target signal at L_{k-1}. Next, WSOLA tries to find a segment (2) lying in the region $[\tau^{-1}(L_k) - \Delta_{max}, \tau^{-1}(L_k) + \Delta_{max}]$ (shaded region), that is maximally similar to the natural continuation (segment (N1)).

III. Time alignment

Dynamic Time Warping (DTW) is used to establish a time scale alignment between source and reference signal. It results in a time warping vector Θ, describing the time alignment of segments of the two signals. Θ assigns a certain segment of the source signal to each of a set of regularly spaced synthesis instants in the target signal.

A preprocessing step is taken to remove silence in the beginning and the end of each utterance. This is done by applying a threshold on the energy of the signal evaluated in blocks of length 125 ms and overlap of 15 ms. The feature extraction is performed on the remaining signal. Attributes of speech relevant for differentiating phonemes are measured over short time intervals, within

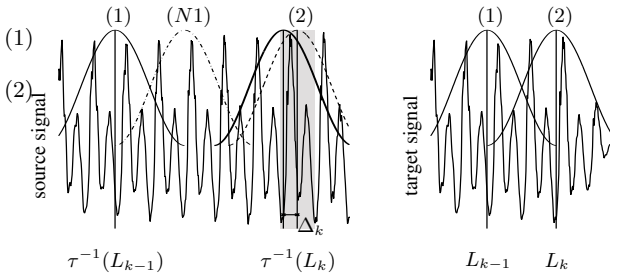


Fig. 2. Illustration of the WSOLA-algorithm. In each step, the synthesis segment (2) is selected within the tolerance region (shaded region centered around $\tau^{-1}(L_k)$) as the segment that is most similar to the natural continuation segment (N1) of the previous segment used for synthesis(1)).

which speech is considered to be quasi-stationary. The feature vectors are extracted from windowed segments of the signal of length 20 ms with 50 % overlap. The chosen features are 12 MelCepstrum coefficients [2] and the log energy.

the Euclidean distance (L_2) is applied to determine the distance between the feature vectors of the two sequences. As a global constraint, the search space of the DTW is limited to fall in a band of width G. This is illustrated in Fig. 3 a). G is determined by

$$G = 20 \cdot |\log_2 \frac{M}{N}| + 40, \quad (2)$$

where N is the number of feature vectors for the source signal and M for the reference signal. Thus, the bandwidth is dependent on M/N, the ratio of reference signal and source signal length. By using the base-2 logarithm an equal sized bandwidth is achieved for a time stretching by factor 2, as for time compression by factor 1/2. Fig. 3 b) shows the global constraint width G dependent on M/N. For the local distance, a modified version of the Sakoe-Chiba [13] local constraint has been used, as described in more detail in the next section.

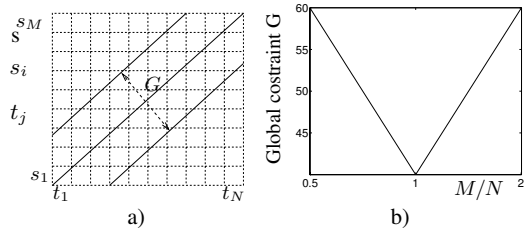


Fig. 3. a) Global constraint G in Dynamic Time Warping. b) Global constraint G as a function of the ratio. M/N.

3.1. Accumulated local penalty constraint

By choosing an appropriate local constraint, the first derivative of the time warping vector Θ may be limited in range. In [1] various local constraints are presented. One of the main criteria in the choice of the local constraint for our system is to preserve a certain flexibility in the alignment, needed to cope with local differences in the speaking rate within a wide range. For this reason we apply a symmetric Sakoe Chiba local constraint without slope constraint.

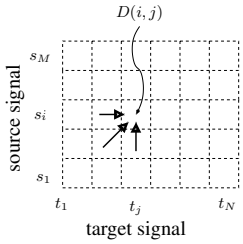


Fig. 4. Dynamic Time Warping (DTW). Calculating the global minimum distance using Sakoe Chiba [13] local constraint.

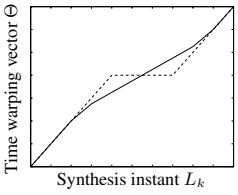


Fig. 5. Time warping vector before (dashed line) and after (solid line) the preprocessing stage.

Fig. 4 illustrates the symmetric Sakoe Chiba constraint. Only three possible predecessors are considered as candidates for the calculation of the global minimum distance:

$$D(i,j) = \min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j).$$

Despite of the desired flexibility, we need to control the warping curvature. Applying a local constraint without a slope constraint tends to result in a path with long horizontal and vertical subpaths. Horizontal subpaths result in a problem in the signal modification part, as they lead to the repetition of one segment several times. This yields a synthetic sound. Vertical subpaths denote the skipping of several segments of the original signal, resulting in an unnatural sound because of syllables or even whole words being omitted. In addition, vertical subpaths can result in pitch discontinuities in the target signal. Changes in pitch mostly occur during vowels in natural speech. In the case of modification from a long to a short vowel, the alignment may contain a vertical subpath, such that the modified vowel is composed by the on- and offset of the long vowel, omitting the middle part. If the long vowel contains a modulation in pitch this leads to a clearly audible pitch discontinuity in the target signal.

Vertical and horizontal subpaths are necessary in the alignment of two utterances containing pauses of different length. By a vertical subpath a pause can be cut off, by a horizontal subpath silent segments extended to a longer pause. Hence it seems reasonable to distinguish between segments containing speech and segments containing silence. A classification for each segment can be done by comparing its signal energy to a threshold obtained by statistics drawn from all segments. Sorting all occurring segment log energy values into 10 equally spaced groups, the threshold is heuristically appointed as the center of the third lowest group. The threshold is chosen so as to rather misclassify silence as speech sounds than vice versa.

We defined an accumulated local penalty constraint, such that the global distance is calculated as

$$D(i,j) = \min[D(i-1,j-1), \alpha D(i,j-1), \beta D(i-1,j)] + d(i,j),$$

where α and β are penalty factors. α and β are proportional to the number of contiguous preceding horizontal, resp. vertical moves of segments, that are classified as nonsilent. In doing so, sequences of horizontal and vertical moves in the time warping path become less likely.

3.2. Smoothing of the time warping vector

The output from DTW is a time warping vector Θ containing subpaths with slopes 0, 1 and infinity, corresponding to horizontal, diagonal and vertical moves. Longer horizontal subpaths are not suitable for time scaling with WSOLA. To ensure that the time warping vector Θ does not have a local slope smaller than a certain value σ for segments classified as speech, we use a smoothing stage. σ is chosen to be 0.5, corresponding to time stretching to the double length. The time warping vector can easily be modified in a left-to-right fashion. Horizontal subpaths are replaced by a subpath with slope σ and extended forwards and backwards in time with a minimum slope of σ until the new subpath meets the original subpath. This is illustrated in Fig. 5.

For segments classified as silence, horizontal subpaths remain to allow an extension of speaking pauses in the target signal. Repetition of the same segment inevitably results in synthetic sound, even if the segment contains no speech but background or breathing noise. To alleviate this we attenuate the repeated parts smoothed with a Hann window as illustrated in Fig. 6.

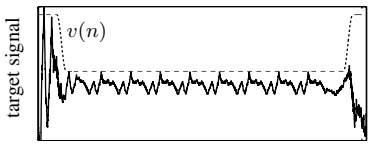


Fig. 6. The speech signal is faded out by multiplying with $v(n)$.

IV. Time scale modification

The time scale of the target signal is changed using the WSOLA algorithm. The time warping vector Θ determines a position of the segment to be excised from the source signal for each synthesis instant L_k as described in section 2.2.

The segment length is 20 ms, as used in the feature extraction and signal alignment procedure. Hann windowing with a 50 % overlap is applied. In WSOLA the segment for synthesis is picked from the tolerance region $[\tau^{-1}(L_k) - \Delta_{max}, \tau^{-1}(L_k) + \Delta_{max}]$ around the 'true' time instant $\tau^{-1}(L_k)$ as described in section 2.2. If Δ_{max} is chosen sufficiently large that the position of the natural continuation segment falls in the tolerance region, it leads to a repetition of the same segment even if the time warping function has been modified as described in section 3.2. With a slope limit σ of 0.5 Δ_{max} needs to be smaller than a quarter of the segment length. Hence the tolerance Δ_{max} is chosen to be 4.9 ms. For good performance, Δ_{max} must be selected to be larger then half a pitch period. Thus, our system functions well for a pitch down to 100 Hz. The cross-AMDF coefficient is used as measure of similarity [6].

Time stretching for unvoiced segments often leads to an audible periodicity using the basic WSOLA method. To reduce these artefacts, segments are classified as voiced or unvoiced, and every third consecutive unvoiced segment gets reversed in time. A similar procedure is known from PSOLA, where every other unvoiced segment is reversed [8]. Our classification is done by means of counting the zero crossings, where a high number of zero crossings indicates unvoiced sounds. With the described method a more natural sound is achieved.

V. Listening tests

We carried out a listening test on 17 listeners to evaluate the time synchronization system. We selected 10 different utterances from the TIMIT database, five spoken by male, five by female speakers.

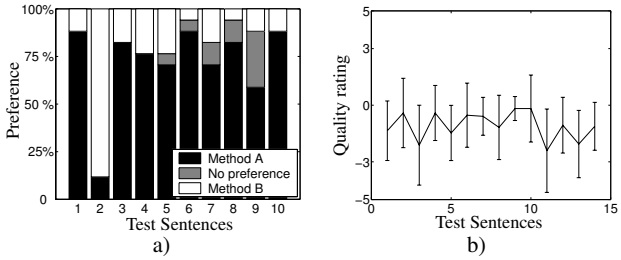


Fig. 7. Results from our listening test for different test sentences (horizontal axis). a) A-B preference test. Method A: Time synchronization system. Method B: Natural synchronization b) Quality rating from +5 (time synchronized version sounds much better) to -5 (time synchronized version sounds much worse)

To get a measure for synchronicity we recorded test utterances spoken by two male and two female speakers. The speakers were first asked to read a sentence independently, then in synchronization with a TIMIT sentence. They could practise as often as required to achieve satisfactory synchronization. The independently spoken utterances were processed by our time synchronization system. For the listening test, we selected 10 recorded utterances, that account for a wide range of time scaling factors. Four of these sentences were read by male speakers, six by female speakers.

The first part of the listening test was an A-B preference test, presenting the sentences synchronized by the speaker and by our system to the listeners. The 17 listeners were asked to judge the accuracy in synchronization. Fig. 7 a) shows the result of the preference test for the different test utterances. The height of the bars indicate the preference of the listeners. It can be seen that in most of the cases our time synchronization system is clearly preferred over to natural synchronization.

In the comparison one has to consider that the prosody of the independently spoken sentence might differ from the TIMIT reference sentence, whereas the speaker automatically will adjust the prosody speaking simultaneously with the reference. Therefore, the natural synchronization will sometimes be felt as more synchronous, even if it is not. This is the case for the second sentence from the test, which contains a large difference between the independent and reference prosody.

The aim of the second part of the listening test was to evaluate the quality of the processed files. To get a more balanced ratio of increased and decreased speaking rate, we added four additional test sentences where the TIMIT utterances were processed to make them synchronous with our recorded utterances. Thus, we obtained additional examples where the speaking rate is decreased, since the recorded test sentences are on average slower than the TIMIT utterances. The target signal and source signal were presented, and the same 17 listeners asked for a comparative qualitative rating between -5 (much worse) to +5 (much better). In Fig. 7 b) the mean rating and standard deviation for the test sentences over all listeners are depicted. The test results of the second part were inconsistent, showing that the judgement of speech quality for one sentence differs significantly for different persons. A reason for that might be that the judgment are influenced by the prosody, which is automatically changed by changing the time scale. Nevertheless, it can be concluded that the time modified sentences are experienced as being of good quality on average, in many cases rated better than the original.

VI. Conclusion

We presented a system that performs time synchronization between two different utterances of the same sentence based on DTW and WSOLA. In contrast to an earlier system (presented in [9]), our system can align utterances that differ severely (caused by a different speaker and speaking style), and makes the resulting time scaled utterances sound natural.

To obtain good time synchronization, major modifications are necessary to make the algorithms suitable for our application. We introduced an accumulated local penalty constraint in DTW to control the curvature of the time warping function. The constraint is made dependent on a classification of segments into speech or silence. A smoothing stage was added to handle the limitations of the WSOLA method in dealing with low slopes in the time warping function for speech segments. By that we achieved flexibility in the time warping function to cope with large local differences, as, for example, longer silence parts between words, while maintaining properties that guarantee good quality. Moreover, the speech quality was improved compared to the basic WSOLA algorithm by time reversing unvoiced segments.

References

[1] H. Silverman and D. Morgan, "The application of dynamic programming to connected speech recognition," *IEEE ASSP Magazine*, vol. 7, no. 3, pp. 6–25, 1990.

[2] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. New Jersey: Prentice Hall, 2001.

[3] I. Shahin and N. Botros, "Speaker identification using dynamic time warping with stress compensation technique," *Proc. IEEE*, pp. 65–68, April 1998.

[4] R. Martens and L. Claesen, "Dynamic programming optimisation for on-line signature verification," *Proc. Fourth Int. Conf. Document Anal. Recog.*, vol. 2, pp. 653–656, August 1997.

[5] A. Corradini, "Dynamic time warping for off-line recognition of a small gesture vocabulary," *Proc. IEEE ICCV Workshop Recog. Anal. Tracking Faces Gestures Real-Time Syst.*, pp. 82 –89, July 2001.

[6] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *IEEE Int. Conf. Acoust. Speech Sign. Proc. (ICASSP)*, vol. 2, pp. 554 –557, 1993.

[7] S. Roucos and A. Wilgus, "High quality time-scale modification of speech," *Proc. Int. Conf. Acoust. Speech Sign. Proc. (ICASSP)*, vol. 10, pp. 493 – 496, 1985.

[8] E. Moulines and F. Charpentier, "Pitch-synchroneous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Comm.*, vol. 9, no. 5, pp. 453–467, 1990.

[9] W. Verhelst, "Automatic post-synchronization of speech utterances," *Proc. Europ. Conf. Speech Comm. Techn.*, pp. 899–902, Sept. 1997.

[10] R. Bellman and S. Dreyfus, *Applied Dynamic Programming*. Princeton, NJ: Princeton University Press, 1962.

[11] E. Moulines and W. Verhelst, "An introduction to speech coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 519–555, Elsevier Science Publishers, 1995.

[12] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust. Speech Sign. Process.*, vol. 28, no. 1, pp. 99 –102, 1980.

[13] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech, Sign. Process.*, vol. 26, no. 1, pp. 43 –49, 1978.

3rd International Workshop MAVEBA 2003, 219-222

A CLASSIFICATION METHODOLOGY OF HEARING IMPAIRED PATHOLOGIES BASED ON DTW TECHNIQUE APPLIED ON VOCAL AUDIOMETRY

M. Kammoun, K.Ouni, A. Bouzid and N. Ellouze

Laboratory of Systems & Signal Processing, Ecole Nationale d'Ingénieurs de Tunis

BP.37 Le Belvédère, 1002, Tunis, Tunisia

kammoun_monia@yahoo.fr; (*kais.ouni; aicha.bouzid; N.ellouze*)*@enit.rnu.tn*

Abstract: **This paper describes a methodology based on DTW technique (Dynamic Time Warping) applied to vocal audiometry to classify the different pathologies of hearing impaired. This methodology is validated on a population of ten subjects composed of seven individuals suffering perception, transmission or mixed deafness and three nondisabled subjects. The obtained results show that this method can be used as a first step for classification of hearing impaired pathologies.**

Key words: **DTW, MFCC, Vocal Audiometry, Classification pathology.**

I. INTRODUCTION

Audiometry testing is used to identify and diagnose both hearing loss and hearing problems in children and adults [1]. With correct diagnosis of a person's specific pattern of hearing impairment, the right type of therapy, which might include hearing aids, corrective surgery, or speech therapy, can be prescribed [5].

There are two kinds of audiometry testing : the tonal and the vocal audiometry tests.

In the tonal audiometry test a trained audiologist uses an audiometer to conduct audiometry testing. This equipment emits sounds or tones, like musical notes, at various frequencies, or pitches, and at different volumes or levels in dBs of loudness. Testing is usually done in a soundproof testing room [3]. This diagnosis method is efficient since it localizes frequency areas non perceived by hearing impaired (H.I) persons [5].

Vocal or speech audiometry is another type of testing that uses a series of simple recorded words spoken at various volumes into headphones worn by the patient being tested. The patient repeats each word back to the audiologist as it is heard. An adult with normal hearing will be able to recognize and repeat 90-100% of the words.

However, diagnosis mechanism of hearing deficiencies by vocal audiometry is essentially manual and requires the introduction of new techniques of speech recognition. Also, because of the large variety of pathological cases, other possibilities and clinical tools for evaluation and supervision shall be useful [2][3].

For this purpose, the present work propose to the medical staff a classification methodology of deafness types using a technique based on DTW revealing score parameters [5]. This shall allow the measurement of hearing deficiency in order to determine the deafness type of tested subjects, thus they can benefit from a digital and automatic tool of rehabilitation necessary for speech recovery and practice [4].

In the following sections, we present the proposed diagnosis approach and its different stages. Then, we applied its validation on a population of some hearing impaired (H.I) subjects. Finally, we give some discussions and a conclusion.

II. DIAGNOSIS APPROACH

The proposed diagnosis approach is based on the scheme of the figure 1. A preprocessing stage is firstly applied, followed by an amplitude normalization operation on the input signals. Then a third stage is applied to extract the relevant parameters of the signal using a Mel-Cepstral analysis which will constitute the input values of the DTW stage. This last stage will generate a score which will be used for hearing pathologies classification.

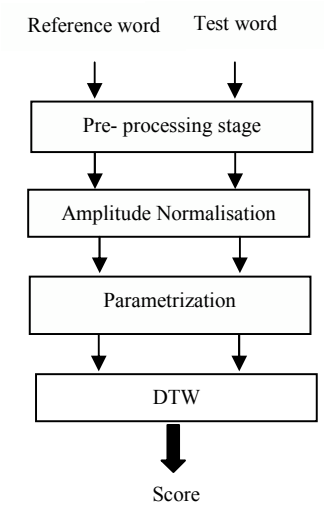


Fig. 1: The scheme of the proposed diagnosis approach.

A. Preprocessing Stage

Recorded words usually include different noise signals at their beginning and the end. This noise varies in each recording, increases processing requirements and slows down the comparison algorithm. This problem was resolved by removing noise sequences using sound processing software. Figure 2 represents the fixing procedure of the word.

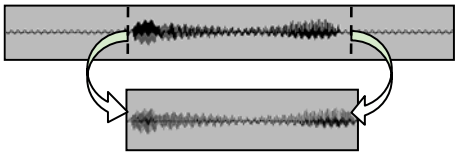


Fig. 2: Word start and ending points

B. Amplitude Normalization

Variation of the vocal signal level in elocution makes the comparison of the words more difficult in terms of decision on the resemblance of two words [1]. If two words are identical but have different levels, the comparison will lead to the conclusion that the two words are not similar which is a bad response. For this reason, an amplitude level normalization of the signal must be carried out, following the noise removal stage and prior to applying the comparison algorithm. This is achieved by dividing each word sample by the energy of the signal.

C. Parametrization Stage

In this stage each word will be processed and parameterized leading to a representation in the form of a matrix where the columns are the frames of the signal and the arrows are the parameters.

Parameterization is achieved by Mel-Cepstral analysis which produces a set of 12 coefficients called MFCC (Mel Frequency Cepstrum Coefficients) (Fig. 3). These parameters are the dominant features used for speech recognition [1]. In fact, several research studies showed the effectiveness of this method which allows a sound analysis equivalent to human perception mechanism. Figure 3 illustrates the corresponding parametrization scheme.

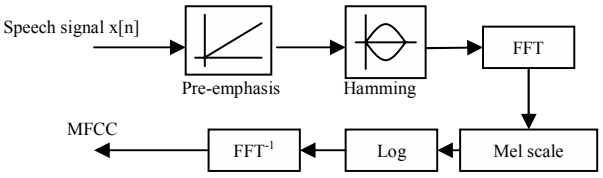


Fig. 3: MFCC Coefficients determination.

D. DTW Stage

In speech signals, the different acoustic tokens of a same speech utterance, are rarely realized at the same speaking rate across the entire utterance. This fact makes that when comparing different tokens of the same utterance, the speaking rate and the duration of the utterance should not contribute to the dissimilarity measurement. This leads to the need of normalizing the speaking rate fluctuations in order to compare the utterances in a coherent way. A solution to this problem can be achieved using dynamic programming techniques for time alignment as the well known DTW technique.

The obtained score from the DTW is the measurement of dissimilarity between two words. When the DTW is applied on the same word, a straight and linear path with a null distance or score are produced. In the case of two different words, the DTW produces a path which is deviated from the diagonal axe by an amount proportional to the difference between the two words. Also, the obtained score increases proportionally to the difference between the two words [5].

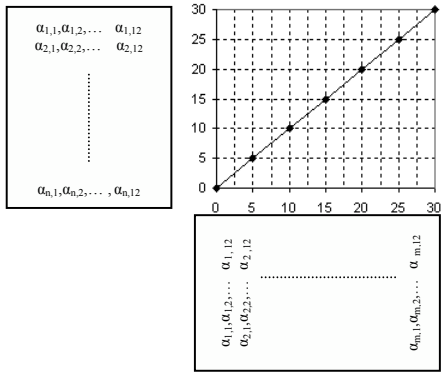


Fig. 4: Path representation of two identical words.

III. VALIDATION AND DISCUSSIONS

To validate the proposed approach a corpus of a series of simple recorded words spoken at various volumes into headphones worn by different subjects is built. Series of 100 phonetically balanced words are used. The subjects are composed of ten persons : three are non disabled (ND) subjects and seven are hearing impaired (HI) subjects. The pathology of the hearing impaired persons was previously provided by the medical staff as illustrated on table 1.

The obtained corpus contains a set of 620 words pronounced by the tested subjects representing the response of the HI subjects as well as the non-disabled persons (some lists were not used).

The application of DTW on the test word and the reference word at various decibel levels produces the score parameter which will be classified in tables corresponding to each type of deafness. The score obtained by the subject for each word of the list, for a given decibel level and the percentage of words repeated correctly by all tested subjects, are filled in columns. Average score (*AvrScore)* for each decibel level is calculated as follow :

$$AvrScore = \frac{\sum_{i=1}^{10} score(i)}{10} \qquad (1)$$

Score (i): represents the score of i th word in the same list.

Tables 2 to 5 (listed at the end of this paper) show the recognition rate of words classified in ten different lists at various decibel levels on one hand, and the average of the scores obtained from the recognition of ten different words on the same decibel level on the other hand. For example, the HI subject 3 obtained an average score of 6.1. This result reveals a recognition difficulty for this subject, since the score is relatively higher compared to a ND subject. Table 3 includes scores of the three ND subjects. The rate of recognition is 100% for all these subjects on any decibel level. The resulting score does not exceed the value 3.

Thus, this value can be referenced to classify the hearing impaired (HI) subjects according to their deafness. These scores are given in Figures 5 and 6 which trace the average score versus the decibel level for each subject.

According to figure 5, we notice that scores of ND subjects are not null because of variable factors such as voice timbre and speech velocity, but are quite lower than those obtained by HI subjects. This is due to better concentration on the way the reference word was pronounced, considering that these subjects do not need to recognize the word, but rather to repeat it with almost the same speed of elocution as the audiologist. The curves representing all subjects form two separate groups. A first group which includes ND subjects in the interval of 1.5 to 3 considered as a relatively low score. A second group includes HI subjects, limited to values between 4 and 7 revealing a possible hearing disorder of these subjects.

The scores obtained by HI subjects shall be isolated on a separate graph in order to better interpret their dispersion (Fig. 6).

This representation is significant as far as the arrangement of the average scores is concerned. The three subjects "3", "7" and "10" suffering from severe mixed deafness have scores varying from 5.5 to 6.5, these scores are relatively higher which is logical since this type of deafness is considered to be the most severe.

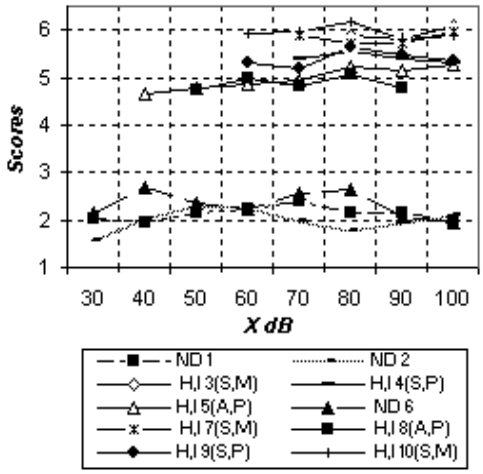


Fig. 5 : Average scores of all subjects versus dB's.
(S, P): severe perception, (A, P): average perception
(S, M): severe mixed

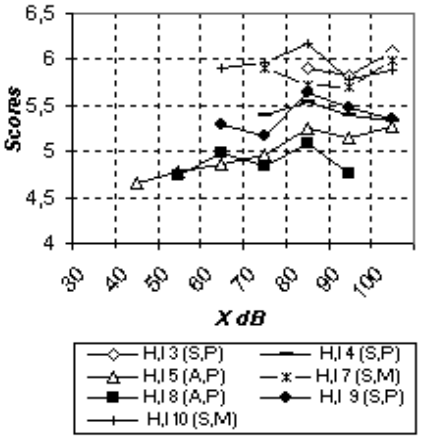


Fig. 6: Average scores of hearing impaired subjects versus dB's.

Subjects "4" and "9" diagnosed with severe perception deafness have scores ranging between 5 and 5.8, their deafness is classified severe since they perceive the word but distinguish the syllables with difficulty. Subjects "5" and "8" diagnosed with average perception deafness have scores ranging between 4 and 5.3; they represent subjects with the lowest scores.

IV. CONCLUSION

This work described a diagnosis approach based on DTW technique applied to the vocal audiometry. The purpose is to obtain a diagnosis and classification pathology method of hearing impairment.

This approach generates a unique score used to identify the kind of pathology of hearing impairment subjects.

The reference score value, obtained from non-disabled subjects, was around the score of 2. The scores obtained by subjects suffering severe mixed deafness averaged the value of 6; this value is relatively high since this type of deafness is considered extreme. Scores of subjects suffering severe perception deafness ranged from 5 to 5.8. The interval 4 to 5.3 includes scores of subjects with average perception deafness, representing the least impaired subjects. This proposed method establishes the use of scores to confirm the diagnosis of the type of deafness and classify hearing impaired individuals according to their pathology.

For future work, we suggest integrating a technique of speaker's normalization to surmount the problem due to the presence of several patients' types (a child, a woman or a male)

REFERENCES

[1] A. Bouzid, K. Ouni, N. Ellouze, "Dynamic Time Warping Applied to Vocal Audiometry", Tunisian-German Conference on Smart Systems and Devices, (SSD), Hammamet, 27-30 march 2001, Tunisia, pp. 463-466.

[2] D. M. Chabries, D. V. Anderson, T. G. Stockham, Jr. and R. W. Christiansen, "Application of a Human Auditory Model to Loudness Perception and Hearing Compensation." Proc. IEEE ICASSP'95, Vol. 5, pp. 3527--3530, 1995.

[3] J. Chalupper, H. Fastl, "Simulation of Hearing Impairment based on the Fourier Time Transformation", Proc. IEEE ICASSP'00. Vol. 2, pp. 857-860, 2000.

[4] J. M. Kates, "Signal Processing for hearing Aids," IEEE Signal Processing Magazine, pp. 41-43, September 1997.

[5] K. Ouni, N. Ellouze and N. Lakhoua, **"**Audiométrie numérique pour la mesure de déficiences auditives" Revue Les Annales Maghrébines de l'Ingénieur , Vol. 12, N° Hors Série, November 1998, pp. 365-370.

Table 1: Deafness type and level of HI subjects

Patient	Sex	Deafness Type	Deafness Level
Person 3	♂	Mixed	Severe
Person 4	♂	Perception	Severe
Person 5	♂	Perception	average
Person 7	♂	Mixed	Severe
Person 8	♂	Perception	Average
Person 9	♂	Perception	Severe
Person 10	♀	Mixed	Severe

Table 2: recapitulation of the various scores of subjects with severe mixed deafness

Subject: H.I. 10	*Subject: H.I. 7*	*Subject: H.I.3*	*Mixed severe*
5,88 100%	5,98 90%	6,1 90%	100 dB
5,79 90%	5,71 70%	5,82 70%	90 dB
6,18 90%	5,72 20%	5,9 10%	80 dB
5,96 80%	5,9 10%	- -	70 dB
5,91 60%	- -	- -	60 dB

Table 3: recapitulation of the various scores of nondisabled (ND) persons.

Normal hearing	*Subject: ND 1*	*Subject: ND 2*	*Subject: ND 6*
100 dB	1,98 100%	2,1 100%	1,96 100%
90 dB	2,13 100%	1,95 100%	2,05 100%
80 dB	2,14 100%	1,79 100%	2,64 100%
70 dB	2,4 100%	1,98 100%	2,57 100%
60 dB	2,24 100%	2,29 100%	2,25 100%
50 dB	2,13 100%	2,31 100%	2,34 100%
40 dB	1,95 100%	2,04 100%	2,68 100%
30 dB	2,03 100%	1,58 100%	2,15 100%

Table 4: recapitulation of the various scores of subjects with average perception deafness.

Average perception	*Subject: H.I. 5*	*Subject: H.I. 8*
100 dB	5,27 90%	- -
90 dB	5,14 100%	4,76 100%
80 dB	5,24 90%	5,08 100%
70 dB	4,96 90%	4,83 100%
60 dB	4,86 80%	4,98 100%
50 dB	4,77 70%	4,73 70%
40 dB	4,66 30%	- -

Table 5: recapitulation of the various scores of subjects with severe perception deafness.

Subject: H.I. 4	*Subject H.I. 9*	*Severe perception*
5,33 70%	5,36 80%	100 dB
5,39 80%	5,47 80%	90 dB
5,54 70%	5,64 90%	80 dB
5,4 30%	5,17 60%	70 dB
- -	5,29 40%	60 dB

3rd International Workshop MAVEBA 2003, 223-226

COMPARISON OF TWO FREQUENCY LOWERING ALGORITHMS FOR DIGITAL HEARING AIDS

Alan M. Marotta[1], Francisco J. Fraga[2]
[1] National Institute of Telecommunications, Santa Rita do Sapucaí - MG, Brazil
[2] National Institute of Telecommunications, Santa Rita do Sapucaí - MG, Brazil

Abstract: **A considerable percentage of listeners with severe hearing loss have audiograms where the losses are high for high frequencies and low for low frequencies. For these patients, lowering the speech spectrum to the frequencies where there is some residual hearing could be a good solution to be implemented for digital hearing aids. In this paper we have presented two different frequency–lowering algorithms: frequency compression and frequency shifting. Preliminary results have shown a slight better performance of the frequency shifting method relatively to the frequency compression method.**
Keywords: **digital hearing aids, frequency lowering, *spectral shaping***

I. INTRODUCTION

There are several kinds of hearing impairment. The origin of the sensorineural hearing losses can be due to defects in the cochlea, auditory nerve or both. These problems reduce the dynamic range of hearing. The threshold of hearing is elevated, but the threshold of discomfort (at which the loudness become uncomfortable) is almost the same as for normal–hearing listeners, or even may be lower. For some range of frequencies, the threshold of hearing is so high than it is equal to the threshold of discomfort, i.e., it is impossible for the listener hearing any sound at those frequencies.

Hearing loss is more common for high–frequency and mid–frequency sounds (1 to 3 kHz) than for low–frequency. Frequently, there are only small losses at low frequencies (below 1 kHz) but almost absolute deafness above 1.5 or 2 kHz.

These facts lead researchers to lower the spectrum of speech in order to match the residual low–frequency hearing of listeners with high–frequency impairments. Slow playback, vocoding, and zero–crossing rate division are some of the methods that have been employed in the last decades. All of these methods involve signal distortion, more or less noticeable, generally depending on the amount of the frequency shifting. Many of the lowering schemes have altered perceptually important characteristics of speech, such as temporal and rhythmic patterns, pitch and durations of segmental elements.

Hicks *et al.* [1] have done one of the most remarkable investigations about frequency lowering. Their technique involve pitch–synchronous, monotonic compression of the short–term spectral envelope, while at the same time avoiding some of the above-described problems observed in the other methods. Reed *et al.* [2] have conducted consonant discrimination experiments on normal hearing listeners. They have observed that Hick's frequency lowering scheme presented better performance for fricative and affricate sounds if compared with low pass filtering to an equivalent bandwidth. On the other hand, the performance of the low pass filtering was better for vowels, semivowels and nasal sounds. For plosive sounds, both methods have shown similar results. In general, the performance on the best frequency–lowering conditions was almost the same to that obtained on low pass filtering to an equivalent bandwidth. Further, Reed *et al.* [3] have extended the results of Hick's *et al.* system to listeners with high–frequency impairment. In general, the performance of the impaired subjects was inferior to that obtained by normal subjects.

Few years ago, Nelson and Revoile [4] have discovered that relative to the normal–hearing listeners, those with moderate to severe hearing loss required approximately double the peak–to–valley depth for detection of spectral peaks in bands of noise when signals have a high numbers of peaks per octave. Findings revealed that detection of spectral peaks in noise is significantly related to consonant identification abilities in listeners with moderate to severe hearing loss.

All previous mentioned frequency-lowering schemes compress the speech spectrum into a narrower band of frequencies, increasing the number of peaks per octave while maintaining the peak–to–valley depth. According to Nelson's and Revoile's investigation, applying sharpening processing to a frequency lowered speech may allow better detection of spectral peaks and better consonant identification.

Recently, Muñoz *et al.* [5] have combined sharpening (i.e., increasing the peak–to–valley depth) and frequency compression. They have demonstrated that the processed speech improved the understanding of fricative and affricate sounds, while providing no significant change in identification of vowels and other sounds by listeners with severe high–frequency hearing loss.

Based on Nelson's and Revoile's investigation, we hypothesize that the relatively poor performance of Hick's and Muñoz's frequency lowering schemes is due to the increasing of the numbers of peaks per octave, which is inherent to the frequency compression method used in these systems.

In this paper, we propose a new frequency-lowering algorithm that does not increase the number of peaks per octave because it uses *frequency shifting* instead of *frequency compression*. Furthermore, the frequency shifting is applied only for fricative and affricate sounds, leaving all others types of sounds untouched, because it is only for fricative sounds that the frequency lowering technique brings real benefits as have been demonstrated by all the previous mentioned works. We have also implemented a frequency compression algorithm based on Hick's [1] and Muñoz's [5] ideas. Preliminary results of subjective preference have confirmed our hypothesis about the better performance of the *frequency shifting* method compared to the *frequency compression* method.

II. METHODOLOGY

A. Audiometric data acquisition and processing

The first step of both frequency-lowering algorithms consists in audiometric data acquisition of the impaired subject. The audiometric exam is employed for measuring the degree of the hearing impairment of a given patient. In this exam, the listener is submitted to a perception test by continuously varying the sound pressure level (SPL) of a pure sinusoidal tone in a discrete frequency scale. The frequency values most frequently used are 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, 6 kHz and 8 kHz. For each of these frequencies, the minimum SPL in dB for which the patient is capable of perceiving the sound is registered in a graph. The audiogram is the result of the audiometric exam, which is presented by a graph with the values in dB SPL for each of the discrete frequencies. This graph is done separately for each ear of the subject. Since the level of 0 dB SPL is considered the minimum sound pressure level for normal hearing, the positive values in dB registered on the vertical axis of the audiogram can be considered as the hearing losses of the patient's ear.

If the losses are equal or inferior to 20 dB, the subject is considered as having normal hearing. From 21 to 40 dB, the losses are classified as mild. Moderate losses are those which are greater than 40 dB but until inferior to 70 dB. From 71 to 90 dB, we consider that the patient have severe hearing losses and more than 95 dB of loss is classified as profound [6].

The threshold of discomfort, for normal or impaired listeners, is no more than 120 dB SPL. Although less common, some audiograms bring both the threshold of discomfort and the threshold of hearing [7], as we can observe in Fig. 1. In this figure, the points of the audiogram corresponding to the right ear are signaled with a round mark and those corresponding to the left ear are signaled with an X mark. These marks are worldwide used in this way by audiologists [6]. The dynamic range of listening for each frequency is the threshold of discomfort minus the threshold of hearing.

Based on the acquired audiometric data, the algorithm analyses the range of frequencies where there is still some residual hearing. The criterion used is the following: first, it is verified if the patient have a ski–slope kind of losses, i.e., if the losses are increasing with frequency. Only patients with this type of impairment can be aided by any frequency lowering method. After that, the first frequency where there is a profound loss is determined. If this frequency is between 1.2 kHz and 3.4 kHz, a *destination frequency* to which the high–frequency spectrum will be shifted is calculated. Otherwise, no frequency shifting is needed (residual hearing above 3.4 kHZ) or profitable (residual hearing below 1.2 kHZ). This *destination frequency* is considered as the *geometrical* mean between 900 Hz and the highest frequency where there is still some residual hearing. The geometrical mean was empirically chosen because it provides a good tradeoff between minimum spectrum distortion and maximum residual hearing profit. In order to obtain more accuracy in the losses thresholds, the points of the audiogram are linearly interpolated.

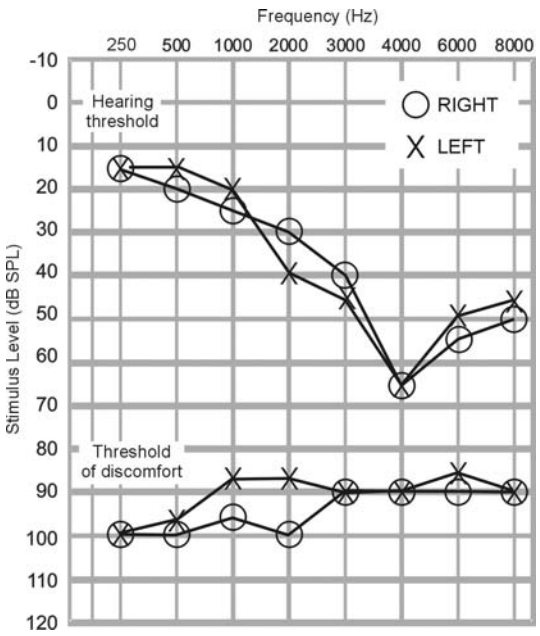


Fig. 1: 'Ski-slope' losses case

B. Speech data acquisition and processing

The speech signal is sampled at a 16 kHz rate and Hamming windowed with 25 msec windows. These windows are 50% overlapped, what means that the signal is analyzed at a frame rate that is the inverse of 12.5 msec. A 1024-point FFT is used for representing the high–resolution short–time speech spectrum in the frequency domain.

If in the previous audiometric data analysis a ski–slope kind of loss was detected and the frequency-shifting criterion was matched, a *destination frequency* have already been determined. Then, we have to find out (in a

frame–by–frame basis) if the short–time speech spectrum presents significant information at high frequencies that justify the frequency shifting operation. The criterion used for shifting or not the short–time spectrum of each speech frame is based in a threshold. When the signal has high energy in high frequencies the algorithm shifts this high frequency information to lower frequencies. The threshold is set for suppressing the processing of all vowels, nasals and the semivowels, while activating the frequency transposition for fricatives and affricates.

To decide which part of the spectrum will be shifted, the energy of 500 Hz bandwidth windows are calculated with 100 Hz spacing, from 1 kHz to 8 kHz. This is done with the aim of find out an *origin frequency*. The origin frequency is the frequency 100 Hz below the beginning of the 500 Hz bandwidth window that have maximum energy. The part of the spectrum that will be transposed corresponds to the range of all frequencies above the *origin frequency*. This empirical criterion guarantees that the unavoidable distortion due to the frequency lowering operation will be profitable. Because the most important part of the high–energy information will be shifted to the limited range of frequencies that are above 1 kHz (therefore maintaining untouched the low–frequency information) but still below the highest frequency where the patient presents residual hearing.

For comparison, the Hick's frequency compression scheme was already implemented, but now only when the same frequency lowering criterion (high/low frequency energy ratio) used for transposition was matched, i. e., only for fricatives and affricates. The frequency compression was done by means of an equation defined in [2]. But in practice, it is more useful to implement the inverse equation, which is

$$\frac{f_{IN}}{f_S} = \frac{1}{\pi}\tan^{-1}\left[\left(\frac{1-a}{1+a}\right)\tan\left(K\pi\frac{f_{OUT}}{f_S}\right)\right] \quad (1)$$

where f_{IN} is the original frequency, f_{OUT} is the corresponding compressed frequency, K is the frequency compression factor, a is the warping parameter and f_S is the sampling rate. For minimum distortion at low frequencies, the warping parameter must be chosen as being $a = (K–1)/(K+1)$.

The compression factor K was determined according to the degree of loss presented by the listener. Fig. 2 shows the curves of equation (1) for K = 2, 3 and 4. In this figure we can see that the low frequency information (below 1000 Hz) is barely compressed.

After frequency shifting or compressing (if it occurs), the FFT–spectrum of each speech frame is multiplied by the gain factor, which is calculated for each frequency in order to full compensate the hearing loss, unless the amplified sound pressure level exceed the threshold of discomfort. In this case, the gain factor is limited to the amount required for maintaining the loudness below the threshold of discomfort. The way we implemented this spectral shaping process is similar to that described in [8]. This last step was still under development in our digital hearing aids system.

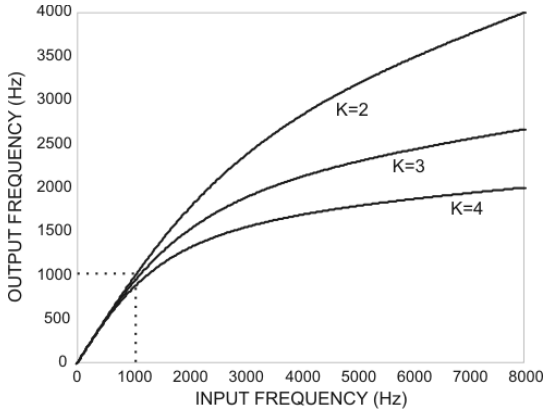


Fig. 2: Input vs. output frequency curves

Part (a) of Fig. 3 illustrates the original FFT–spectrum of a speech frame, in part (b) the same frame is shown compressed by a factor K = 4 and part (c) presents the frame after frequency shifting. It is important to observe that in the last case (frequency shifting) the shape of the spectrum is preserved, what does not occur in the case of frequency compression, where we can clearly note a great amount of shape distortion, but still preserving the low frequency information.

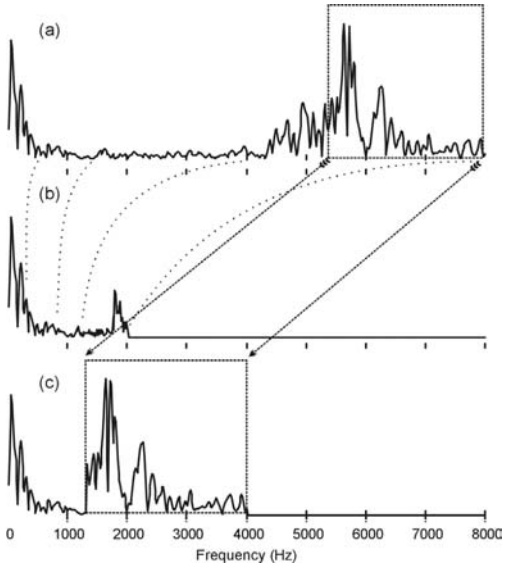


Fig. 3: Comparison of frequency lowering schemes

III. Preliminary Results

The two frequency lowering algorithms were not already tested with hearing impaired subjects because they final spectral shaping part are not completely developed, as mentioned in the last paragraph of section II. But we got some preliminary results with normal

listeners. In this case, a simple low pass filtering process simulates the losses above the frequency where there is no more residual hearing. This frequency was determined based on audiograms of real impaired subjects.

The experiment we have carried out consists of submitting the speech signal to the two frequency lowering algorithms. After that, the resulted signals were listened by two normal hearing subjects, one man and one woman. The listeners do not know anything about the origin of the signals and are asked for ranking the signals according to their intelligibility. In this preliminary test, only two speech signals were submitted to the algorithms. The original and processed spectrograms of one of these speech signals (pronunciation of the words 'loose management') are shown in Fig. 4, where we can appreciate again the visual difference between the two frequency lowering algorithms. According to the prevision, only the fricative speech sounds were frequency lowered in both algorithms. The unique exception is the phone [*l*], which is not fricative but lateral approximant. But in this case, its pronunciation had high frequency energy, as we can observe in the spectrogram of the original speech signal.

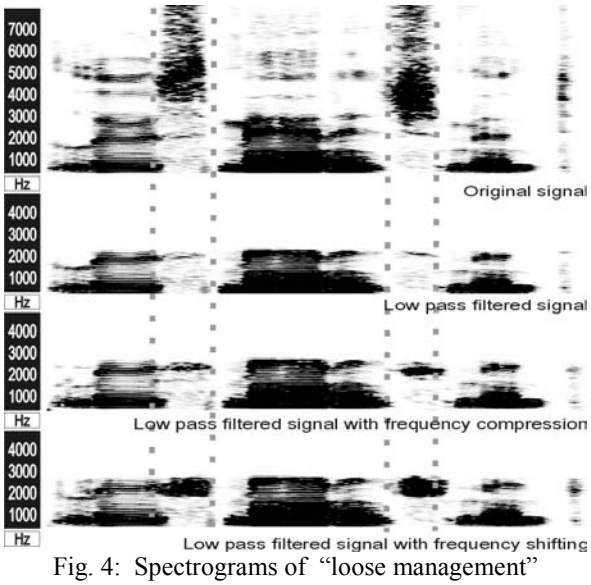


Fig. 4: Spectrograms of "loose management"

The preferences of the listeners were listed in Table 1. In this table, 'Signal 1' is the Portuguese word "pensando" (which means 'thinking') and 'Signal 2' is the English words 'loose management'.

Table 1: Listener's preferences

Speech signal	Man	Woman
Signal 1 low pass	1st	3rd
Signal 1 compr.	3rd	2nd
Signal 1 shifted	2nd	1st
Signal 2 low pass	2nd	2nd
Signal 2 compr.	3rd	3rd
Signal 2 shifted	1st	1st

IV. Discussion

These preliminary results indicate that the frequency shifting method was preferred by the listeners when compared with the frequency compression method. But it is important to remark that the subjective difference between the low pass filtered signal, the frequency–compressed signal and the frequency-shifted signal is very slight, as perceived by normal listeners.

V. Conclusion

It is necessary to finish the spectral shaping part of the system in order to submit the processed signals to hearing impaired listeners. The slight difference observed by the normal listeners may be due to the fact that the difference between the original signal (with frequencies up to 8 kHz) and the low pass filtered (2 kHz) signals is large. But for the impaired subject, that never had any perception of sounds with frequencies above 2 kHz, may be the difference between the processed signals was not so slight. Finally, it is important to remark that, with all the processing being done in the frequency domain, both algorithms have demonstrated to be fast enough for enabling the usage in real time applications.

References

[1] B. L. Hicks, L. D. Braida, and N. I. Durlach, "Pitch invariant frequency lowering with nonuniform spectral compression," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE New York)*, pp. 121-124, 1981.

[2] C. M. Reed, B. L. Hicks, L. D. Braida, and N. I. Durlach, "Discrimination of speech processed by low-pass filtering and pitch-invariant frequency lowering," *J. Acoust. Soc. Am,* vol. 74, pp. 409-419, 1983.

[3] C. M. Reed, K. I. Schultz, L. D. Braida, N. I. Durlach, "Discrimination and identification of frequency-lowered speech in listeners with high-frequency hearing impairment", *J. Acoust. Soc. Am,* vol. 78, pp. 2139-2141, 1985.

[4] P. Nelson, and S. Revoile, "Detection of spectral peaks in noise: Effects of hearing loss and frequency regions," *J. Acoust. Soc. Am*, 1998.

[5] C. M. Aguilera Muñoz, B. N. Peggy, J. C. Rutledge, A. Gago, "Frequency lowering processing for listeners with significant hearing loss", *IEEE*, pp. 741-744, 1999.

[6] S. Frota, *Fundamentos em Fonoaudiologia*, 1 st ed., vol. 1. Guanabara Koogan, 2001, pp. 40-59.

[7] Y. A. Alsaka, B. McLean, "Spectral Shaping for the Hearing Impaired", *IEEE*, pp. 103-106, 1996

[8] J. C. Tejero-Calado, B. N. Peggy, J. C. Rutledge, "Combination compression and linear gain processing for digital hearing aids", *IEEE*, pp. 3140-3143, 1998.

Numerical models

3rd International Workshop MAVEBA 2003, 229-232

VELOFARYNGEAL INSUFFICIENCY STUDIED USING FINITE ELEMEN MODELS OF MALE VOCAL TRACT WITH EXPERIMENTAL VERIFICATION

K. Dedouch[1], J. Horáček[2], T. Vampola[1], J.Vokřál[3]

[1]Institute of Mechanics, Faculty of Mechanical Engineering, Czech Technical University Prague, Czech Republic
[2]Institute of Thermomechanics, Academy of Science of the Czech Republic, Prague, Czech Republic
[3]Phoniatric Laboratory, 1st Faculty of Medicine, Charles University Prague, Czech Republic

Abstract: **Finite element (FE) models of acoustic spaces corresponding to the human nasal and vocal tract for vowel /*a*/ are used for numerical simulations. Simplified FE model of the vocal tract for English vowel was created from geometrical data published in literature and for the Czech vowel by transferring data directly from MRI images. The nasal cavities were added to the models manually according to anatomical literature. The acoustic signal for the vowel / *a* / is simulated using transient analysis of the FE models in time domain. The vocal tract is excited by time dependent displacement of a small circular plate moving at the position of the vocal folds. The time response and frequency response functions are calculated near the lips, nostrils and at the vocal folds. Effects of velofaryngeal insufficiency are simulated and compared to results from acoustic measurements.**
***Keywords*: Biomechanics of voice, acoustic transient and modal analysis, supraglottal spaces, cleft palate**

I. FINITE ELEMENT MODELS OF SUPRAGLOTTAL SPACES

In the previous papers of the authors [1,2] the acoustic frequency-modal characteristics of the human vocal tract were studied by FE modelling including the effects of cleft palate [3]. Here the study is extended to the time domain analysis using a real type of excitation of the acoustic spaces by pulses generated at the vocal folds. The simplified FE model of a male vocal tract for the English vowel /*a*/ was developed according to the MRI data published by Story et al. [4]. The FE model approximating the human supraglottal tract including the added nasal cavity spaces is presented in Fig. 1. The total length of the vocal tract from the vocal folds (on the right) to the lips (on the left) is 174.58 mm. The FE model used for simulation of phonation of the Czech vowel /*a*/ is shown in Fig. 2a [1].

A small connection (size of 20 finite elements) of the nasal and oral cavities was considered in the back area of the soft palate modelling the velofaryngeal insufficiency. The acoustic transient analysis was realised by the system ANSYS 5.7 using the acoustic finite elements FLUID30 considering the speed of sound $c_0 = 353$ ms^{-1} and the air density $\rho_0 = 1.2$ kgm^{-3}. Zero acoustic pressure ($p = 0$) was assumed at the lips and nostrils. Other boundary walls of the acoustic spaces were considered to be acoustically absorptive.

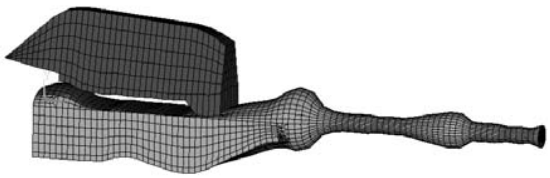


Fig. 1 FE model of male vocal tract for English vowel /*a*/ including the nasal cavity.

The acoustic damping, which is associated with the fluid-structure interface on the boundary between the air and the walls (tissues) of the vocal tract, was modelled by the boundary admittance coefficient $\mu = 0.006$ for supraglottal acoustic spaces and $\mu = 0.008$ for the nasal cavity. This coefficient defined as $\mu = x/\rho_0 c_0$ is a dimensionless quantity between 0 and 1 that is equal to the ratio of the real component of the specific acoustic impedance (resistance term x) associated with the sound absorbing material to the fluid characteristic impedance. Another frequently used characteristic of the sound absorption of the material is the dimensionless absorption coefficient α, which is related to the boundary admittance coefficient μ as

$$\alpha = \left[0{,}5 + 0{,}25\left(\mu + 1/\mu\right)\right]^{-1}.$$

The pulse excitation of the supraglottal spaces was realised by a small rigid circular plate (a piston) translating in the axial direction along the axes z. The plate was situated in the position of the vocal folds, and its diameter was equal to 1/3 of the diameter of the cross-section area of the FE model of the acoustic space at this point (see the detail in Fig. 2b). The translation motion of the plate in time was given by integration of the shape of volume velocity that approximately corresponds to the airflow through the vocal folds (see Fig. 3). Five subsequent excitation pulses with the period corresponding to the fundamental (pitch) frequency F0=100 Hz were considered in the transient

analysis. The interaction between the plate and the acoustic space was realised by the interactive acoustic finite elements.

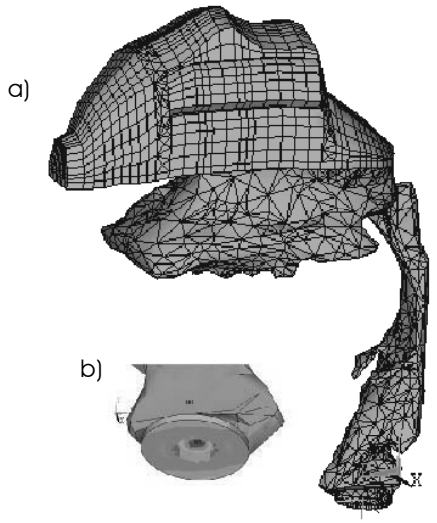


Fig. 2 a) FE model of male vocal tract for the Czech vowel /*a*/ obtained from MRI data file with the added nasal cavity, b) detail of the excitation location at the vocal folds.

In the model of the English vowel /*a*/ the effect of hard palate compliancy was included in the study. The material properties of the bone were assumed as follows: Young modulus $E_1 = 6.50 .10^9$Pa, Poisson ratio μ_1=0.21, density ρ_1=1.41 10^3 kg/m^3 and wall thickness h=0.6mm. The bone of hard palate was modeled using two separated parts. The first (lower) part of the finite elements SHELL63 was directly joined with the acoustic finite elements of the vocal tract using the material properties E_1, μ_1 and ρ_1. The second (upper) part of the finite elements SHELL63 was joined with the acoustic finite elements of the nasal tract on its lower boundary area. The material properties corresponding to the second part of the FE model of the bone were identical with the first part of the bone model except the Young modulus E_2=0.01E_1 respecting a much more compliant material. Each node of the lower part of the bone was connected with the corresponding node at the upper part of the hard palate FE model. This connection of corresponding nodes guarantees identical motion of the nodes in both parts of the FE model.

II. Mathematical formulation

Wave equation for the acoustic pressure can be written as:

$$\nabla^2 p = \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2}, \qquad (1)$$

where c_0 is the speed of sound, with the possible boundary conditions as follows

– on acoustically hard area and at the open end

$$\frac{\partial p}{\partial \mathbf{n}} = 0, \qquad p = 0, \qquad (2)$$

– between the flexible structure and the fluid elements:

$$\frac{\partial p}{\partial \mathbf{n}} = -\rho_0 \frac{\partial^2 \mathbf{w}_n}{\partial t^2}, \qquad (3)$$

where $\mathbf{n}$ is normal to the boundary area and $\mathbf{w}_n$ is the displacement of the structure in the normal direction to the vibrating surface.

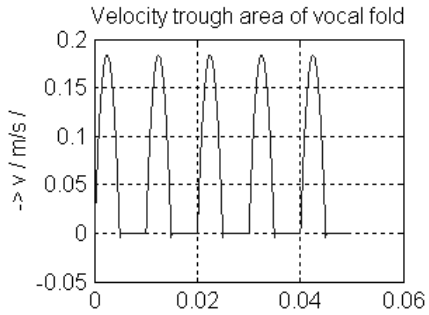


Fig. 3 Pulse excitation by airflow velocity in the glottis.

Equations of motion for the elasto-acoustic system after discretization can be written as

$$\begin{bmatrix} \mathbf{M}_s & \mathbf{0} \\ \rho_0 \mathbf{R}^T & \mathbf{M}_f \end{bmatrix} \begin{bmatrix} \ddot{\mathbf{u}} \\ \ddot{\mathbf{P}} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_f \end{bmatrix} \begin{bmatrix} \dot{\mathbf{u}} \\ \dot{\mathbf{P}} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_s & -\mathbf{R} \\ \mathbf{0} & \mathbf{K}_f \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{P} \end{bmatrix} = \mathbf{0} \qquad (4)$$

where $\mathbf{M}$, $\mathbf{B}$, $\mathbf{K}$ are the global mass, damping and stiffness matrices, $\mathbf{P}$ is the vector of nodal acoustic pressures, subscripts s or f denote the structure or fluid, $\mathbf{u}$ is the structural displacement, $\mathbf{R}$ is the coupling matrix and ρ_0 is the air density.

For the special case of **kinematic excitation** by the moving structure the following equations for the pressure describe the air vibration

$$\begin{aligned} \mathbf{RP} &= \mathbf{M}_s \ddot{\mathbf{u}} + \mathbf{K}_s \mathbf{u} \\ \mathbf{M}_f \ddot{\mathbf{P}} + \mathbf{B}_f \dot{\mathbf{P}} + \mathbf{K}_f \mathbf{P} &= \rho_0 \mathbf{R}^T \ddot{\mathbf{u}}, \end{aligned} \qquad (5)$$

where the structural motion $\mathbf{u}(t)$ is prescribed. The Newmark method of solution in time was used.

III. Numerical results

The results of the transient dynamic analysis of the FE models are the time responses of the acoustic pressure in selected points of supraglottal spaces near the vocal folds, the lips and the nostrils. The spectra of the exciting acoustic pressure pulses and the pressure time responses were calculated by MATLAB using FFT.

Fig. 3 presents excitation pulses of the airflow velocity through the glottis from where the corresponding displacement of the rigid plate was calculated and afterwards used for excitation of the vocal tract in the time domain.

The results of transient analysis of the FE models for English vowel /*a* / are presented in frequency domain in Fig. 4 showing the calculated acoustic pressure near the nostrils. The formant frequencies F1≈823 Hz, F2≈ 1164 Hz and F3≈2826 Hz calculated by modal analysis of the FE model can be detected in the spectrum. A nasopharynx (oro-nasal) resonant frequency f_{naso}≈2143 Hz is embodied in the frequency response function between the formants F2 and F3.

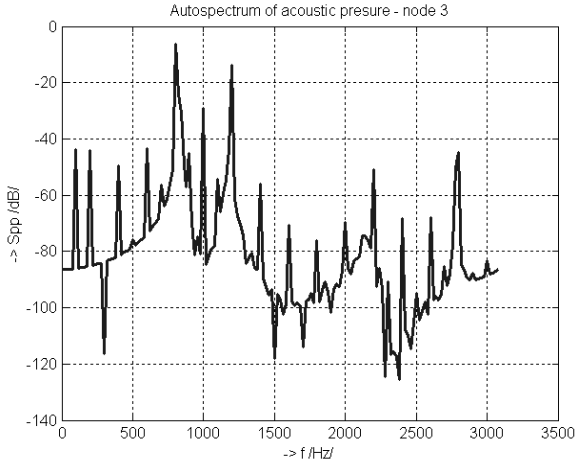


Fig. 4 Frequency response function for the acoustic pressure near the nostrils for the FE model of English vowel /*a*/.

Results of the transient analysis of the FE model for the Czech vowel /*a*/ are presented in Figs. 5 and 6 showing the spectra of the acoustic pressure calculated near the vocal folds and lips. The pressure levels near the vocal folds are much higher than the acoustic pressure near the lips. The formant frequencies F1≈623 Hz, F2≈890 Hz and F3≈2935 Hz can be found in the frequency response functions in Fig. 6. These formant frequencies are in good agreement with the data known from the Czech literature [6] as well as with calculations by modal analysis for the same FE models – see, e.g. [1,3]. Another resonant frequency f_{naso} ≈1707 Hz appears in the Fig. 6 due to the velofaryngeal insufficiency.

The differences between the Czech and English formants originate mainly in the fact that two very different types of the FE models were used, however, the results obtained are in a range of variability of the vocal /*a*/ production.

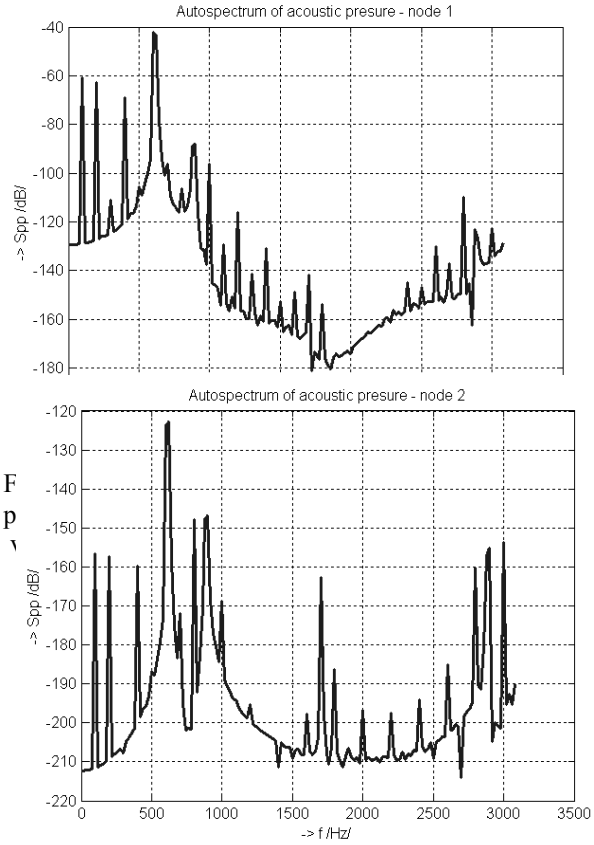


Fig.6 Frequency response function for the acoustic pressure near the lips for the FE model of Czech vowel /*a*/.

IV. Experimental verification of the model

The first experiment took into account the well-known phenomenon of connection vowel - nasal consonant - vowel. The passageway between the oral and nasal cavities of the first vowel is closed or almost closed in the Czech language for a clear sound to be pronounced. The velopharyngeal passageway must be opened when producing nasal consonants.

The vowels following the nasal consonants are nasalized because the passageway is still not closed. The differences in the velopharyngeal opening between the first and the second vowel should result in changes of the formant frequencies. Five normal subjects were asked to pronounce the interconnection /*ama*/ and the changes of the formants between the two vowels were studied.

The nasal and oral signals were picked up by microphones of the headset part of Nasometer 6200-3 (Kay Elemetrics Corp.) and analysed by Multi-Speech (Kay Elemetrics Corp.) programme.

V. Experimental Results

Examples of results from the practical experiments are shown in Fig. 7. The spectrogram of the interconnection /*ama*/, where the second vowel is more nasal than the first one, is shown in Fig. 7a. The signal was picked up in front of the nose. The position of the formants F1=800 Hz and F2=1100 Hz and F3= 3700 Hz is stable. The position of the oro-nasal formant changes from $f_{naso} \approx 2600$ Hz to 2950 Hz as approximately predicted by the FE models. Effects of increasing the cleft area of the hard palate were theoretically studied in detail in previous publications [2,3].

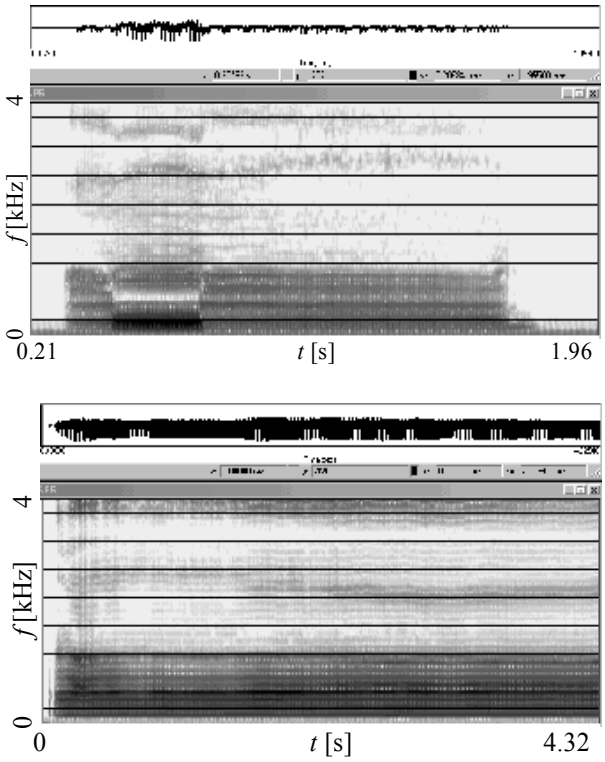


Fig. 7 Spectrograms: a) for interconnection /*ama*/; b) effect of continually closing the soft palate for vowel /*a*/.

The continual changes of the soft palate closing for vowel /*a*/ are demonstrated in Fig. 7b. The signal was picked up in front of the mouth and the measurement started from the soft palate opening. The formants F1=680 Hz, F2=1100 Hz, F3=3950 Hz remain practically unchanged. The oro-nasal formant changes its position from $f_{naso} \approx 2700$ Hz to 2350 Hz.

The second nasalized vowel /*a*/ in the interconnection /*ama*/ corresponded to an opening of the soft palate and simulated a velofaryngeal insufficiency.

VI. Conclusion

The transfer functions were obtained as the results of the transient analysis of the FE models of the vocal tract. The models were excited by a transient translation of a small rigid plate situated in the area of vocal folds and driven by a time signal which shape in the time domain approximately corresponds to a volume velocity of the air flowing through the vocal folds during phonation. The formant frequencies F1 – F3 evaluated from the resonances of the calculated frequency response functions for the pressure are in good agreement with the experimental data known for the formants from the literature [4-6] as well as with the results of the modal analysis performed [1-3]. The existence of calculated oro-nasal formants was verified by the measurements when the velofaryngeal insufficiency was simulated by the normal subjects.

Acknowledgements

The authors are very grateful to Doc. MUDr. Petr Krupa from the Hospital U svaté Anny in Brno for making possible the special measurements and providing the original MRI data sets for human vocal tract during phonation of the Czech vowels.

The research is supported by the Grant Agency of the Czech Republic by the project No 106/98/K019 "*Mathematical – Physical Modeling of Vibroacoustic Systems Important on Biomechanics of Voice and Hearing*."

References

[1] Dedouch K., Horáček J., Vampola T, Kršek P., Švec J.G., "Mathematical modelling of male vocal tract". In: *Proc. of the 2nd Inter. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Sept. 13-15, 2001, Firenze, Italy, 6 p.

[2] Dedouch K., Horáček J., Vampola T., Vohradník M., "Finite element models of supraglottal tract". In*: 5th Int. Conf. Advances in Quantitative Laryngoscopy, Voice and Speech Research,* April 27-28, 2001, Groningen, the Netherlands (CD-ROM, comp. by H. K. Schutte, Dept. of BioMedical Eng., University of Groningen), 7 p.

[3] Dedouch K., Horáček J., "Finite element modelling of male vocal tract with consideration of cleft palate", In: *Forum Acusticum*, Sevilla 2002, 16-20 Sept. 2002, Sevilla, Spain (CD ROM - Special Issue of the Revista de Acustica, Vol. XXXIII, No 3-4, 2002), Section Physical and Mathematical Models of Speech Production, Ref. No SPE-01-006, 6 p.

[4] Story, B. H., Titze, I.R. and Hoffman, E. A., "Vocal tract area functions from magnetic resonance imaging". *J. Acoust. Soc. Am.,* 100 (1), 537 – 554, 1996.

[5] Titze, I., R.: *Principles of Voice Production,* Prentice – Hall, London, 1994.

[6] Novák, A., *Phoniatrics and pedaudiology. Voice disorders – principles of voice physiology, diagnostics, treatment, reeducation and rehabilitation*, Unitisk s.r.o., Prague, 1996, (in Czech).

3rd International Workshop MAVEBA 2003, 233-236

A FORMANT-TRAJECTORY MODEL AND ITS USAGE IN COMPARING COARTICULATORY EFFECTS IN DYSARTHRIC AND NORMAL SPEECH

Xiaochuan Niu, Jan P. H. van Santen
Center for Spoken Language Understanding, OGI School of Science & Engineering at OHSU
20000 NW Walker Road, Beaverton, Oregon 97006, USA

Abstract: **Dysarthria is a diverse group of motor speech disorders that typically are associated with impaired intelligibility. As part of a project to develop augmentative communication technologies for intelligibility enhancement of dysarthric speech, a quantitative method is proposed for measuring the relative contributions to impaired intelligibility of vowels of three factors: First, target shift: Dysarthric speakers may have spectral targets that differ from those of normal speakers. Second, coarticulation: The degree of contextual influence on articulation may be greater in dysarthric speech than in normal speech. Third, random variability: Dysarthric speakers may articulate the same phoneme in the same context with more variability. The method is based on a linear model of formant trajectories of vowels in consonant contexts. The results from analysis of a dysarthric and a normal speech sample showed surprisingly similar target values, but increased coarticulation and random variability for the dysarthric sample.**
Keywords : **Dysarthria, coarticulation, formant**

I. INTRODUCTION

Dysarthria is a diverse group of motor speech disorders that typically are associated with impaired intelligibility and are caused by damage to the motor system [1, 2]. Since in most cases dysarthria is not reversible, major efforts have been made to create assistive devices, including devices based on speech enhancement [3], speech recognition [4], or speech transformation [5].

A recent perceptual study by Hosom et al. [5] focused on the relative contributions of segmental and prosodic factors to intelligibility of dysarthric speech. Using a human-supervised copy prosody technique that allowed for the independent modification of prosodic and spectral information in dysarthric speech, it was shown that significant improvements of intelligibility can be achieved through replacing either the prosodic features or the spectral features of a dysarthric speaker's speech with those of a normal speaker's speech. However, an automated baseline transformation system, based on speech transformation techniques to map the spectral features between the two speakers on a frame-by-frame basis [6], failed to improve intelligibility. A further analysis of the vowel formants indicated that their average values differ sharply between the dysarthric and the normal speech samples, with a much-reduced area of the vowel quadrilateral in the former case [Figs. 1 and 2].

These findings show that successful intelligibility enhancement requires an underlying model of the spectral differences between dysarthric and normal speech. Towards such a model, we consider here three factors that may account for these differences. First, *target shift*: Dysarthric speakers may develop special spectral targets that differ from those of normal speakers. Second, *coarticulation*: The degree of contextual influence on articulation may be greater in dysarthric speech than in normal speech. Third, *random variability*: Dysarthric speakers may articulate the same phoneme in the same context with more variability.

This paper provides an analysis approach that decomposes the contributions of these factors, so that they can be treated separately in the future intelligibility enhancement systems. This analysis will be applied to speech samples from one dysarthric and one normal speaker for demonstration purposes only; no claims are made about dysarthric speech in general.

Since formants constitute a concise acoustic representation closely related to the vocal-tract configuration, we focus our investigation on formant trajectories. We use a linear superposition model similar to a model by Broad and Clermont [7] to describe the trajectories of the first three formants through inter-consonantal vowel portions. In our model, target formants and coarticulatory effects are unknown parameters and are estimated from speech data. Beyond the structure of the model, nothing is assumed about these parameters, so that their estimated values provide unbiased information about the differences between dysarthric and normal speech. The experiments on dysarthric and normal speech data show surprisingly similar target values, but increased coarticulation and random variability for the dysarthric sample. We expect that these results can be used to construct an augmentative communication system.

II. METHODOLOGY

A. Speech data

For the purpose of comparison, the same speech data as in the previous study [5] were used. The data are utterances of one dysarthric speaker (LL) and one normal speaker (JP) from the Nemours database (For diagnostic information, see [8]). Each speaker read 74 syntactically correct nonsense sentences. The speech was recorded and stored in 16k Hz, 16-bit PCM format.

Each sentence in the database has been transcribed into a sequence of phoneme labels. The start and end times of each phoneme in the speech signals were indicated via manual segmentation. The segments considered in this study were syllables with a vowel between two consonants (CVC) in American English. The vowels consisted of /i:/, /u/, /A/, and /@/, as pronounced in the words beat, boot, father, and bad. They are supposed to represent four extreme vocal-tract configurations among the vowels in American English. The consonants consisted of the six stops, the four unvoiced fricatives, and the four approximants in American English.

ESPS software [9] was used to extract formant trajectories from the speech signals. The signals were down-sampled from 16k Hz to 10k Hz, and analyzed with a 49-ms Hanning window that was shifted in a 10-ms step. For each frame of windowed signals, a 12-order LPC analysis was performed and then continuous formant trajectories were obtained. Formant values at vowel midpoints were inspected and, when necessary, corrected manually with optional LPC-poles.

B. Coarticulatory model

We adopted the following model, similar to which was used in [7], to describe coarticulatory effects on the formant frequencies of vowels within different consonant contexts:

$$\vec{F}(t) = \alpha(t)\cdot\left(\vec{T}_C - \vec{T}_V\right) + \vec{T}_V + \beta(t)\cdot\left(\vec{T}_{C'} - \vec{T}_V\right), \qquad (1)$$

where $\vec{F}(t)$ is the observed formant vector as a function of time t, $\vec{T}_V$ is the target formant-vector of the vowel, $\vec{T}_C$ and $\vec{T}_{C'}$ are the target formant-vectors of the initial and final consonants, respectively. All formant vectors in Eq. (1) are 3×1 in dimension with the first three formant frequencies as elements. The first term in the right side of Eq. (1) represents formant transitions from the consonant C to the vowel V. This coarticulatory effect is proportional to the target difference and scaled by a function of the coarticulatory factor $\alpha(t)$. The last term represents a similar effect of the consonant C' on the vowel V, and $\beta(t)$ is the corresponding function of the coarticulatory factor.

If we let $\gamma(t) = \left(1-\alpha(t)-\beta(t)\right)$, then Eq. (1) becomes:

$$\vec{F}(t) = \alpha(t)\cdot\vec{T}_C + \gamma(t)\cdot\vec{T}_V + \beta(t)\cdot\vec{T}_{C'}\ , \qquad (2)$$

which shows that the observed formant vector of the vowel at any time point is a linear combination of the target formant-vectors of the phonemes C, V and C'. Note that, although the model describes full trajectories of formants, we only applied it to the vowel midpoints.

The model represents the three factors (target shift, coarticulation, and random variability) as follows: Target shift is represented by the differences in target values between the dysarthric and normal speaker; coarticulation by the values of the coarticulatory factors; and random variability by the relative goodness of fit of the model.

C. Estimation method

N denotes the number of samples of observed formant vectors, $\vec{F}^{(i)}\ (i = 1,\cdots,N)$. If target formant vectors are known, a least-square-error solution exists for each of the following equations derived from Eq. (1):

$$\left[\vec{F}^{(i)} - \vec{T}_V^{(i)}\right] = \left[\vec{T}_C^{(i)} - \vec{T}_V^{(i)} \quad \vec{T}_{C'}^{(i)} - \vec{T}_V^{(i)}\right]\begin{bmatrix}\alpha^{(i)}\\ \beta^{(i)}\end{bmatrix} \qquad (i = 1\sim N). \qquad (3)$$

When $\alpha^{(i)}$ and $\beta^{(i)}$ are fixed, Eq. (2) can also be rewritten in the following matrix form:

$$\vec{F}^{(i)} = \left[\alpha^{(i)}\cdot I \quad \left(1-\alpha^{(i)}-\beta^{(i)}\right)\cdot I \quad \beta^{(i)}\cdot I\right]\begin{bmatrix}\vec{T}_C^{(i)}\\ \vec{T}_V^{(i)}\\ \vec{T}_{C'}^{(i)}\end{bmatrix} \quad (i = 1\sim N), \qquad (4)$$

where I is a 3×3 identical matrix. Since the phonemes with the same identity in the samples share a common target formant-vector, all equations in (4) can be jointly solved in a least-square-error sense as long as the number of data samples is large enough. Thus, the estimation algorithm can be generally described as follows:

1. Initialize target formant-vectors;
2. Set a small number ε as the convergence threshold;
3. Solve equations in (3), update $\alpha^{(i)}$ and $\beta^{(i)}$, and calculate the square error $E1$;
4. Solve equations in (4), update target formant-vectors, and calculate the square error $E2$;
5. If $|E1 - E2| > \varepsilon$, then go to 3; else, output target formant-vectors, $\alpha^{(i)}$ and $\beta^{(i)}$.

D. Practical issues

When using this method to analyze the real speech data, additional efforts are needed to avoid physically meaningless solutions. This is discussed next.

Formant target initialization. One scheme adopted the formant values from the Klatt synthesizer [10]. For the vowels, we also used the medians of observed formants at vowel midpoints. The vowel targets estimated with the two initializing schemes were quite close.

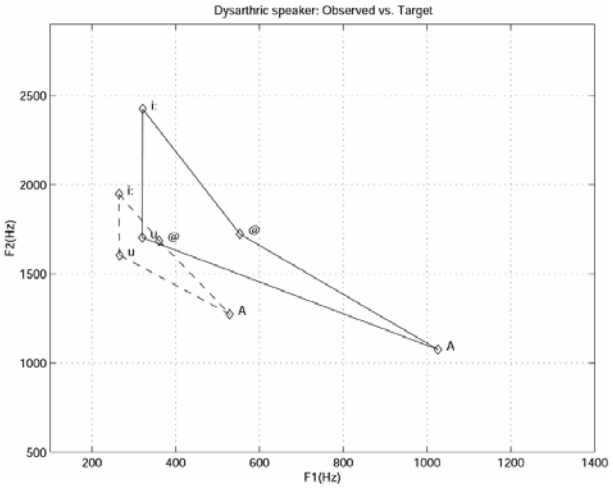


Figure 1. Observed and target vowel formants (dysarthric speech.) The medians of observed formants are linked with dashed lines; the estimated target formants are linked with solid lines.

Rank. When in Eq. (3) $\vec{T}_C^{(i)} = \vec{T}_{C'}^{(i)}$, the matrix on the right is not full-ranked, so that $\alpha^{(i)}$ and $\beta^{(i)}$ cannot be estimated. Thus, C and C' are required to be different.

Constraints. Constraints of the linear weights included: $\alpha^{(i)} \geq 0.025$, $\beta^{(i)} \geq 0.025$, and $\alpha^{(i)} + \beta^{(i)} \leq 0.95$. For a target formant vector, $\vec{T} = [f_1\ f_2\ f_3]'$, constraints were $90 \leq f_1 \leq 1300$, $500 \leq f_2 \leq 2800$, $1300 \leq f_3 \leq 3700$, and $f_1 < f_2 < f_3$, because the formants should be in reasonable ranges.

Normalization. Note that $\alpha^{(i)}$ and $\beta^{(i)}$ are scalar values, while formants are vectors. Hence, $\alpha^{(i)}$ and $\beta^{(i)}$ reflect only the average coarticulatory effects of the three formant frequencies. To balance the contributions of each formant to the fitting errors, each dimension of a formant vector was normalized by dividing it by the formant medians.

III. Results

A. Goodness of fit

The normalized sums of least squares deviations were 0.238 and 0.032 for the dysarthric and normal samples, respectively, indicating greater variability for the dysarthric speech.

B. Vowel space

Figs. 1 and 2 show the observed vowel space and the estimated target vowel space of the dysarthric and normal speaker, respectively. In both figures, the first and second formants (F1, F2) of four extreme vowels (/i:, u, A, @/) are plotted as points in the F1-F2 plane. The medians of observed formants are linked with dashed lines to represent an observed vowel space, and the estimated values of target formants are linked with solid lines to represent a target vowel space. As can be seen, the formant quadrilateral for each speaker shifts from the observed position to the target position, expanding the area of the vowel space. This expansion trend implies a potential way to increase the spectral separability of these vowels, which may be critical for intelligibility enhancement, by considering the target formants rather than the observed formants. Of critical importance is that, except for /@/, the dysarthric target values are surprisingly close to the normal target values.

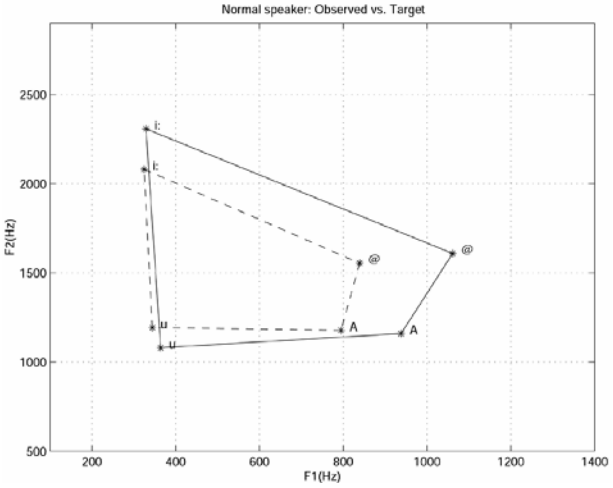


Figure 2. Observed and target vowel formants (normal speech.) The medians of observed formants are linked by the dashed lines; the estimated target formants are linked with solid lines.

C. Coarticulatory effects

Fig. 3 and 4 show the histograms of the value $(1-\alpha-\beta)$ for the dysarthric and normal speaker, respectively. Since the estimated values of the parameters α and β reflect the coarticulatory effects of the consonants C and C' on the vowel V, the value $(1-\alpha-\beta)$ can be interpreted as the weight of the vowel's contribution to the formant trajectory and hence can be used as an indicator of the degree of coarticulatory effects. The figures show that the distribution of $(1-\alpha-\beta)$ concentrates around 1 for the normal speaker, and has a wide spread for the dysarthric speaker. This shows that the speech of the dysarthric speaker is more coarticulated than the normal speech.

IV. Discussion and Conclusion

In summary, an approach to formant analysis was presented that decomposes the contributions of target

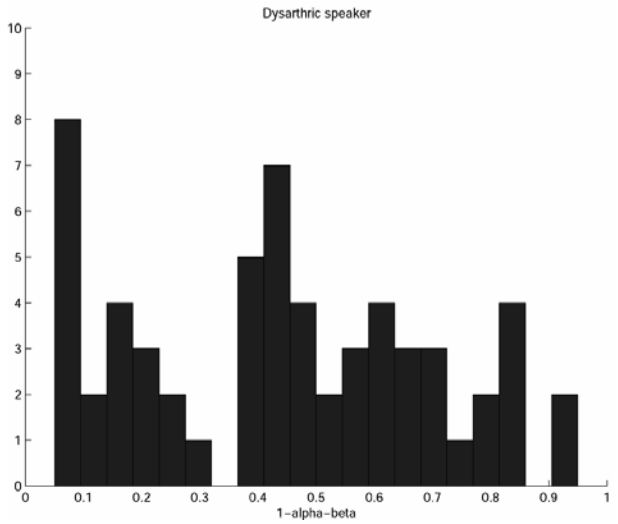


Figure 3. Distribution of coarticulatory factor values (dysarthric speech.)

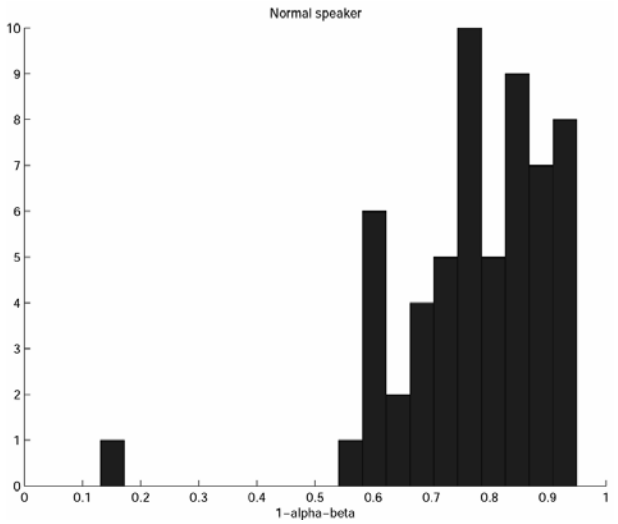


Figure 4. Distribution of coarticulatory factor values (normal speech.)

formants and coarticulatory effects on the formant trajectories of *CVC* syllables. The approach adopts a linear superposition model to describe formant trajectories. Using the method, target formants and coarticulatory factors can be estimated from speech data.

Using the method, we analyzed the speech data of a dysarthric speaker and a normal speaker to gain insight in the relative contributions of three factors that may be responsible for reduced intelligibility of dysarthric speech: target shift, coarticulation, and random variability. The results from this preliminary experiment revealed systematic differences between the two speakers. The target vowel space of the dysarthric speaker exhibits a specific distortion pattern of vowel production, but was surprisingly similar to the target space of the normal speaker. The analysis results also show a larger degree of coarticulatory effects in the speech of this dysarthric speaker, and more random variability.

The analysis results show that intelligibility enhancement may critically need algorithms for the "de-coarticulation" of dysarthric speech. In principle, if the system can recognize aspects of vowel environments, such as the place of articulation of surrounding consonants, this could be accomplished by applying Eq. (2) in reverse to recovered the true vowel formants from the observed formants and the inferred consonant targets.

We note that the model is extremely simple. For example, it assumes the same coarticulatory factor for the three formants at a certain time. This assumption does not necessarily hold.

ACKNOWLEDGEMENT

This research was partially supported by Grants 0117911 and 0082718 from the National Science Foundation. We thank Alexander Kain and John-Paul Hosom for inspiring discussions and valuable comments.

REFERENCES

[1] D. B. Freed, *Motor speech disorders : diagnosis and treatment,* San Diego: Singular Pub. Group, 2000.

[2] R. D. Kent, J. F. Kent, J. Duffy, and G. Weismer, "The dysarthrias: Speech-voice profiles, related dysfunctions, and neuropathology," *Journal of Medical Speech-language pathology,* vol. 6-4, pp. 165-211, 1998.

[3] Electronic Speech Enhancement, Inc., "The Speech Enhancer,", http://www.speechenhancer.com/.

[4] J. R. Deller, D. Hsu, L. J. Ferrier, "On the use of Hidden Markov Modeling for recognition of Dysarthric speech," *Computer Methods and Programs in Biomedicine,* vol. 35-2, pp.125-139, 1991.

[5] J. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," *Proc. of ICASSP 2003*, vol. I, pp. 876-879, Apr. 2003.

[6] A. B. Kain, and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," *Proc. of ICASSP 2001*.

[7] D. J. Broad and F. Clermont, "A methodology for modeling vowel formant contours in CVC context," *Journal of the Acoustical Society of America*, vol. 81-1, pp. 155-165, 1987.

[8] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech", *Proc. of ICSLP 1996*, pp. 1962-1965, Oct. 1996.

[9] D. Talkin, *ESPS*, Entropic Research Lab. Inc., 1993.

[10] J. Allen, M. S. Hunnicutt, and D. Klatt, *From text to speech: the MITalk system*, Cambridge [Cambridgeshire]; New York : Cambridge University Press, 1987.

3rd International Workshop MAVEBA 2003, 237-240

PRELIMINARY GLOTTAL SOURCE MODELING FOR PATHOLOGIC VOICES

Eoin O'Leidhin and Peter Murphy
Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland
Eoin.OLeidhin@ul.ie Peter.Murphy@ul.ie

Abstract: **A first attempt at implementing a flexible model for the glottal source waveform of pathologic voices is described. The LF (Liljencrants & Fant) model is the source model used. We also add various noise types, shimmer and jitter to the excitation source in order to replicate more closely the pathologic glottal waveform. Various vocal characteristics are then modeled in order to evaluate the performance of the glottal source model.**
Keywords: **Glottal source modeling, LF model, pathologic voice.**

I. INTRODUCTION

It has long been accepted that in-order to achieve natural sounding synthetic speech an accurate and versatile model of the voice source is needed. To this end a considerable amount of research has gone into trying to achieve a suitable glottal flow model for normal quality voice. However besides a few exceptions [1,2,3], there have been relatively few attempts reported to synthesise pathologic voice and to create an adequate glottal flow model for this purpose. As pointed out in [4], synthetic pathological voices could be useful for the introduction of a standard protocol for pathological voice quality assessment. The voice source that will be used for this study will be based on the LF model. This widely used model is chosen as it has relatively few controlling parameters yet is flexible enough to model a lot of different phonations. It was found in [3] that the LF model was quite capable of modeling variations that occur in speech pathology and that a more complex model was not needed.

Our purpose in this study is to implement a flexible glottal source model that may then be built upon to achieve an accurate speech synthesiser for pathologic voice.

II. GLOTTAL SOURCE MODEL

Traditionally there have been basically three types of excitation sources used in order to synthesise speech [5]. Impulse excitation with a glottal shaping filter is relatively simple however the vocal quality is poor and often fails to adequately reproduce characteristics of natural voicing. The second type is glottal waveforms calculated by inverse filtering. Although this type of excitation is of better quality than the first type, it is unsuitable to some situations as existing speech must be available and also the recorded data must be of a very high quality. The third type of excitation source is excitation waveform models, which most of the recent research in this area have used. This type is flexible, easy to use and can produce high quality natural sounding speech.

In [5] it was found that four factors were important for the characterisation of different voice production types. These were: the glottal pulse width, the glottal pulse skewness (the ratio of the glottal opening phase to the glottal closing phase), the abruptness of glottal closure and the turbulent noise component.

The LF model was therefore chosen as it has the ability to easily control the first three of these factors, and the fourth factor could be modeled by an added noise source. The LF Model is a four-parameter model that models the differentiated glottal flow [6]. The glottal flow derivative is chosen, as it is easier to identify points of interest of the glottal flow on its derivative. For instance it is easier to identify the moments of glottal onset and closure on the differentiated glottal flow than on the glottal flow waveform. The LF model is a combination of a growing sinusoid and an exponential.

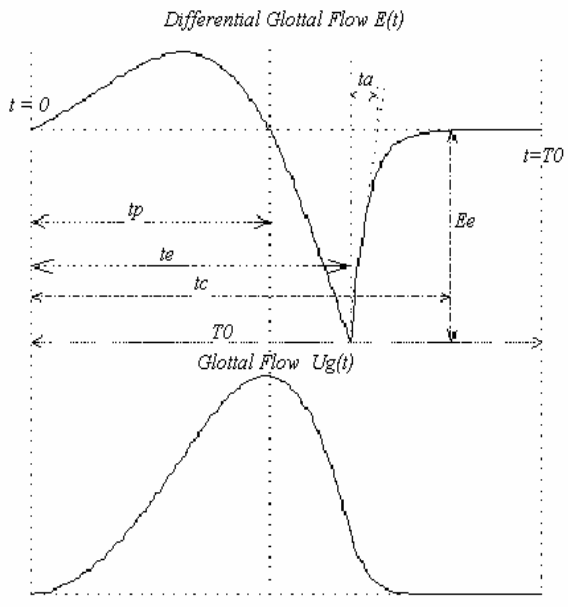


Fig 1: LF Model

The glottal waveform is given by:

$$E(t) = E_0 e^{\alpha t} \sin \omega_g t \qquad (0 \le t \le t_e) \quad (1)$$

$$= \frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{\varepsilon(t_c-t_e)}\right] \quad (t_e < t \le tc)$$

The four timing parameters are t_c, t_p, t_e and t_a. The parameter t_c is the instance of the complete glottal closure, t_p is the instance of maximum glottal flow (corresponds to the zero-crossing in the differentiated glottal flow). The parameter t_e is the maximum negative of the differentiated glottal flow waveform, while t_a which doesn't have an exact physical correspondence as such is found by projecting a tangent of the return phase back to the time axis. This parameter determines how quickly the exponential phase returns to zero. The following conditions hold for the LF model:

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)} \qquad (2)$$

$$\int_0^{t_c} E(t) = 0 \qquad (3)$$

$$E_0 = -\frac{E_e}{e^{\alpha t_e}\sin(\omega_g t_e)} \qquad (4)$$

$$\omega_g = \frac{\pi}{t_p}$$

where α is the parameter used to ensure that the glottal pulse returns to zero at closure and ε is the decay constant of the exponential phase.

The open quotient of a glottal pulse is defined as the open glottal period divided by the total pitch period. The open quotient is primarily determined by the glottal pulse width. For the LF model the open quotient can be defined as:

$$OQ = \frac{(t_e + kt_a)}{T_0} \qquad (5)$$

where k is a function of t_a, and will have a range of 2 to 3 when t_a is between 0 and 10% of the pitch period.

The speed quotient is defined as a measure of the glottal pulse skewness, and is found for the LF model by:

$$SQ = \frac{t_p}{t_e + kt_a - t_p} \qquad (6)$$

Often $t_e + kt_a$ is almost equal to t_c and may be used in the above equations. Finally the abruptness of the glottal closure is controlled by the parameter t_a. With a small t_a value causing a abrupt glottal closure and visa-versa.

Aspiration noise is the term given to noise that is generated at the glottis during voiced speech; it occurs especially when glottal closure is incomplete and when there is high airflow velocity. It is an important component in the perception of breathy vocal quality.

In order to synthesise the voice source of a pathologic voice an adequate noise source will be needed. Because of the different patterns of noise that is associated with pathologic voice, there will be some choice in the type of noise that can be added to the source.

- Additive random noise: This noise source will be superimposed on to the voice source. This is the type of noise that is used in most synthesizers for normal voice. This noise is white Gaussian noise. It is possible to add this to the whole background of the glottal flow as well as just in noise bursts in specific parts. In [5] it is stated that the energy of the turbulent noise is distributed over a wide range of frequencies (2-8 kHz). Therefore the noise is filtered through a high-pass first order FIR filter with a cutoff frequency of 2 kHz. There are also two gain parameters (one to control the noise over the whole speech segment and one to control the noise for noise bursts) that control the amplitude of the noise added. These gain parameters are calculated as a percentage of the energy of the glottal pulse.
- Multiplicative noise: This noise is calculated as a percentage of the amplitude of the glottal flow, therefore most noise will occur at the moment of maximum glottal flow.

Additive random noise is normally introduced in three different ways in order to model three distinct conditions of turbulent noise production. The noise may be added so that the peak noise occurs at the peak of the glottal flow in order to synthesis breathiness. Also the peak noise may be introduced at the glottal flow closure in order to synthesise roughness. Finally the noise may be introduced in a non-signal dependent way over the whole of the glottal cycle, this could be used in-order to model paralysis in one of the vocal folds. In this study each of these types of noise will be able to be modeled.

If the fundamental frequency is just held constant for the duration of the synthesised speech segment, a mechanical sound quality would be the result. Therefore perturbations of the fundamental frequency are introduced in the form of jitter and shimmer.

Jitter is defined as the cycle-to-cycle perturbation in the fundamental frequency of a signal. For modal voice a typical jitter value would be less than 1%. Obviously for breathy and pathologic voice the jitter can be considerably higher. In this study the jitter that will be added to the source, will be calculated using a random number generator.

Shimmer can be defined as the cycle-to-cycle variability in amplitude of the glottal flow waveform. For modal voice the shimmer level would be less than 0.7dB. In a number of the studies done on glottal pulse modeling for normal voice, shimmer is excluded, however in modeling the glottal pulses of pathologic voice the shimmer level could be quite significant and therefore needs to be modeled.

III. Vocal Tract Representation

In the source-filter model the source and filter are assumed to be non-interactive and linear. For this experiment we simply construct the formant synthesizer using 6 formant frequencies and bandwidths in order to model the vocal tract transfer function, the impulse response of which is then convolved with the LF source function.

IV. Implementation

For the LF model, ε is solved iteratively using equation (2) using $\varepsilon = 1/t_a$ as an initial estimate. Then the area under the return phase of the differential glottal flow may be calculated. Since according to equation (3) the area under the positive half of the curve must be equal to the area in the negative part of the curve, after making an initial estimate of α, *E0* and α may be also solved iteratively.

In order to implement Shimmer the calculated random shimmer levels are added to the glottal source model that is calculated without shimmer.

Jitter is a little more complicated to add in the LF model. Since all the timing parameters of the LF model are relative to the pitch period, if a change is made to the pitch period it would alter all the timing parameters. This would have a side effect of altering the amplitude of the glottal pulse (i.e. introducing shimmer). Since it is important that the exact amount of shimmer that is introduced is known, this is unsatisfactory. In order to solve this, the maximum amplitude of the glottal flow before the jitter (or shimmer) is introduced, is found. The jitter (calculated using a uniformly distributed random number generator) is then added to the pitch period and the glottal pulse is calculated for this new pitch period. Since now this glottal pulse contains shimmer as well, the maximum amplitude of this new glottal flow is calculated and the amount of shimmer that the change in the pitch period has introduced can be found and compensated for.

Noise is added to the glottal flow waveform, thus the LF model output is integrated, the noise then added and finally the resulting waveform derived again, which can then be convolved with the vocal-tract function. First of all in-order to implement the additive random noise; gain factors will control the level of the noise. These gain factors can be used to control the signal to noise ratio (SNR), which is described below. As mentioned already the additive noise may be added to the whole background of the glottal waveform, or in certain segments in order to simulate a glottal noise burst. The multiplicative noise is also calculated with regard to the glottal flow waveform. All of these noises can be used together.

In order to calculate the SNR, the noise that is added to the signal is calculated (by subtracting the clean derived glottal flow waveform from the noisy one). The SNR is then calculated as $10\times\log10[S/N]$, where N and S are the noise energy and the clean derived glottal flow energy respectively.

IV. Experiments

Experiments were then carried out in-order to evaluate the performance of the glottal pulse model. At first this evaluation was done using the LF model values used in [1]. In that paper, LF model values (plus some perturbation measurements) are given that model different glottal source pulses for various vocal characteristics, such as modal, creaky, breathy, rough and hoarse voices. In-order to model these vocal characteristics the sustained vowel /a/ was synthesised. The values of the glottal source pulses were held constant for the duration of the token.

Next the differentiated glottal flow waveform was calculated for some sustained vowel speech samples of various types of pathologic voice taken from the website [7]. Inverse-filtering software was used to find the glottal flow waveform of these speech samples [8]. It should be noted that this software wasn't designed specifically for pathologic voice so that when severe pathologic voice is analysed (as in Fig 3), the calculated glottal waveform could be quite different from the true glottal waveform. As in [3], for some of the voices analysed, high frequency formants were prone to being miscalculated, causing incorrect ripples on the flow derivative.

The inverse-filtered differentiated glottal waveform was then matched with the LF model waveform using both the best visual fit and also using an automatic LF fitting method similar to the one described in [9]. In order to achieve this automatic LF fitting, initial estimates of the LF parameters were found. Initial estimates for t_c, t_p, t_e, and E_e were relatively easy to obtain, (it was assumed that the pitch period was already accurately calculated), however an accurate estimate for t_a was more difficult to achieve. It was found, similarly to the study in [9] that the normalised maximum magnitude of the spectrum of the return phase gave a reasonable estimate of t_a. Then using the Least-Squares (LS) fit the LF parameters were optimised first using the Nelder-Mead simplex search method and then with a steepest descent algorithm.

It was found that in most cases the automatic fitting technique gave a lower LS error than the best visual fit method.

Fig 3 shows the comparison of the differentiated glottal waveform of a rough breathy male voice and the corresponding best-fit LF model waveform. As can be seen the LF model gives a reasonable model of the derived differentiated glottal waveform.

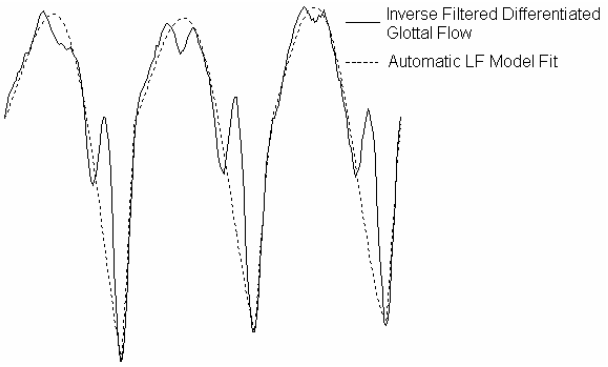


Fig. 3 - Differentiated Glottal Flow: Inverse Filtered vs. Automatic LF Model Fit

The LF was applied to various speech pathology samples. In most cases when an appropriate glottal waveform was calculated, the LF model provided a reasonable approximation. However in some cases the LF model is unable to provide an accurate model of the derived differentiated glottal waveform

With each of the speech files that was analysed, an attempt was also made to synthesise the speech, however even though for most cases the LF model gave a reasonably accurate model of the source, there was a considerable difference between the actual speech and the synthesised speech.

IV. Discussion

In the study [3], 3 different ways (least-squares fit, best visual fit and best perceptual fit) of matching LF model waveforms to filtered sources were examined. For the perceptual fit, an attempt was made to produce the best perceptual match of the target voice without regard to the calculated glottal flow derivative. It was shown that the best perceptual fit even though in some cases a good deal different from the calculated glottal flow derivative gave the best match to the original voice. The conclusion that was made was that there isn't always enough information about the source in the output of the inverse filter to reconstruct vocal quality adequately. This would seem to reflect the fact that in this study when attempts were made to re-synthesise the speech with the LF model Fit the results were disappointing.

This is only a preliminary study and a lot more research is needed to achieve an adequate glottal source model for various voice disorders. More research will be needed on disorder specific modeling. It is intended to perform EGG and video-stroboscopy recordings on various subjects with voice disorders, which will then be used to create closer models of these specific voice disorders.

V. Conclusion

The glottal source model that is implemented here performed reasonably well in synthesising sustained vowels for a variety of voice types. Along with the LF model parameters, the perturbations effects that may be added to the glottal source allow for a large variety of vocal characteristics to be modeled. However when attempts were made to re-synthesise existing sustained vowel segments of pathological voices, although the glottal waveform matched reasonably well, the results were rather disappointing when the two speech files were compared. Therefore more research is needed in order to examine the accuracy of using the output of the inverse filter as a sole model of the vocal source. Also in order to achieve a better glottal source model more research is needed into the modeling of in-complete closures of the glottis.

VI. Acknowledgements

The authors wish to thank Mr. Olatunji Akande from University of Limerick for the software used to perform the inverse filtering. This work is supported by Enterprise Ireland, Research Innovation Fund 2002/037

References

[1] A.L. Lalwani and D.G. Childers, "Modeling Vocal Disorders via Formant Synthesis", ICASSP 1991, pp.505-508, 1991.

[2] P. Bangayan, C. Long, A.A Alwan, J. Kreiman and B.R. Gerratt, "Analysis by synthesis of pathological voices using the Klatt synthesizer", Speech Communications 22, pp.343-368, 1997.

[3] J. Kreiman and B.R. Gerratt, "The perceptual structure of pathologic voice quality," *J. Acoust. Soc. Am.,* vol. 100, pp.1787-1795, 1996.

[4] M. Epstein, B. Gabelman, N. Antonanzas-Barroso, B. Gerratt and J. Kreiman, "Source Model Adequacy for Pathological Voice Synthesis", International Congress of Phonetics Science, 1999.

[5] D.G. Childers and C.K. Lee, "Voice quality factors: Analysis, synthesis and perception", *J. Acoust. Soc. Am.,* vol. 90, pp.1787-1795, 1991.

[6] G. Fant, J. Liljencrants and Q. G. Lin, "A four-parameter model of glottal flow", *STL-QPSR 4/1985,* pp. 1-13, 1985.

[7] http://www.icsl.ucla.edu/~spapl/

[8] O.O. Akande and P.J Murphy, "Split Band Inverse filtering of Speech with Application for Accurate Vocal Tract filter Estimation", AQL Conference, Hamburg, Germany, 2003.

[9] Helmer Strik, Bert Cranen and Louis Boves, "Fitting a LF-model to Inverse Filter Signals", EUROSPEECH-93, Berlin, Vo. 1,pp.103-106, 1993.

[10] D.G. Childers, "Speech Processing and Synthesis Toolboxes", Wiley, New York, 2000.

3rd International Workshop MAVEBA 2003, 241-244

MODELLING THE CREATION OF CZECH VOWELS BY MEANS OF THE VOCAL FOLDS MODEL AND THE MODELS OF VOCAL TRACTS

K. Prikryl
Institute of Mechanics, Faculty of Mechanical Engineering, Brno University of Technology, Brno, Czech Republic

Abstract **The key elements in generating speech are the vocal folds and the vocal tract. This paper deals with the modelling of the creation of Czech vowels by means of vocal folds model and the models of the vocal tracts. The folds model was devised by using the finite elements method and vocal tract models were designed by means of magnetic resonance. Source sound created by means of the "air bubbles" method was modified by the transfer function of vocal tracts. Models were applied both in time and frequency domains. Spectral analysis of the signal was carried out and completed in the area of the mouth and it was crowned by the spectra of vowels /a/,/i/,/o/ with marked formants.**
Keywords : **vowels, transfer function, spectrum, formants**

I. INTRODUCTION

Aural perception enables us to distinguish the vowels whose creation can be explained by the theory of the source of the sound and the filter. Under the source we understand sound spectrum produced as a result of periodic movement of the folds in interaction with the air. Under the filter we understand vocal tract the shape of which is changing depending on the requirements of the person producing sounds. People perceive vowels based on the two lowest natural frequencies of the vocal tract. These natural frequencies are called formants. Therefore, the key elements in generating speech are vocal folds and the vocal tract that, through the change of its shape, modifies its own natural frequencies. The author of the paper [1] has described the new method of producing the source sound by means of air „bubbles". Our aim is to prove that the model is functional and that using it enables modelling of Czech vowels when speaking aloud.

II. METHODOLOGY

The vowels possess the highest energy of a signal (sound, speech) and have their specific properties. When we utilise modelling for the analysis of the signal, i.e. via analysis of the signal in the mouth area of the modelled vocal tract, we are able to distinguish the formants of spoken vowels.

The vocal tract is modelled on the basis of shapes of the vocal tracts generated by the method of magnetic resonance. By utilising such procedure, finite-element models of the vocal tracts for pronouncing vowels /a/,/i/,/o/ were created. For adults, the range of formants of specific vowels is always similar.

The movement of the vocal folds has been identified by means of a defined subglottal pressure loop as presented in Fig. 1, in relation to a minimum gap between the vocal folds (glottis). Based on this relation, with the known integration time step of the transition analysis, the time dependence of the subglottal pressure (see Fig. 2) and a minimum gap between the vocal folds (see Fig. 3) were determined. The latter curve corresponds to a flow Ug. If the flow is derived the volume acceleration is obtained (Fig. 4). The models of the vocal tracts were driven by this volume acceleration dUg.

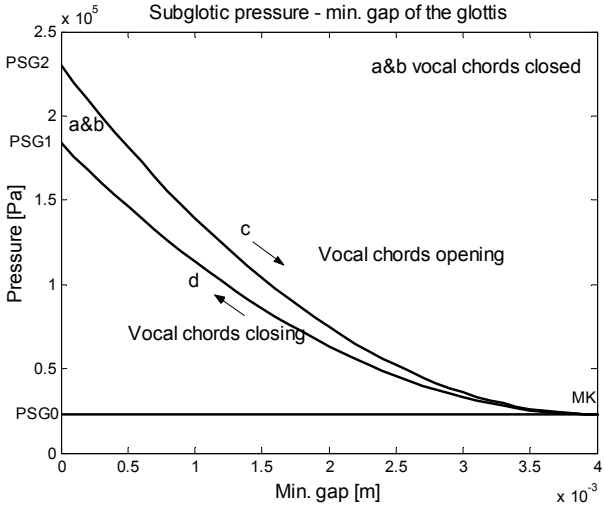


Fig. 1 Subglottal pressure in relation to the glottis

As mentioned before, our aim is to demonstrate that the vocal folds drive considered here is able to create the spectrum of the voice source and further, that the vocal tract (filter) can reshape a spectrum on the outlet in the mouth area. The spectrum must correspond to the vowel pronounced.

III. RESULTS

The experiment is carried out both in time and in the

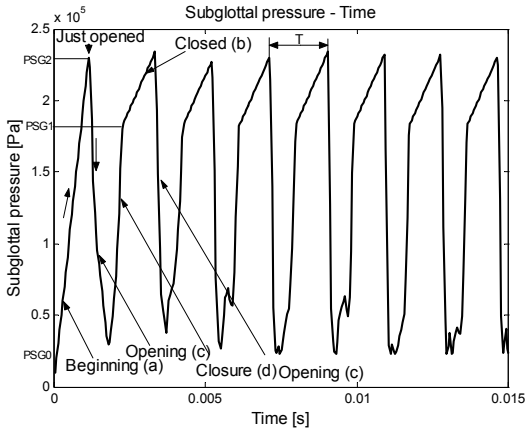


Fig. 2 Subglottal pressure in relation to time

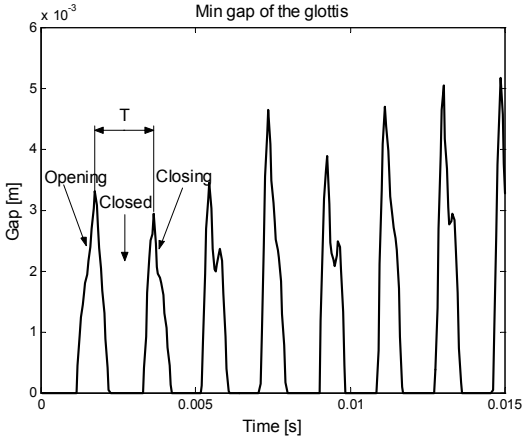


Fig. 3 Calculated minimum gap Ug between the folds

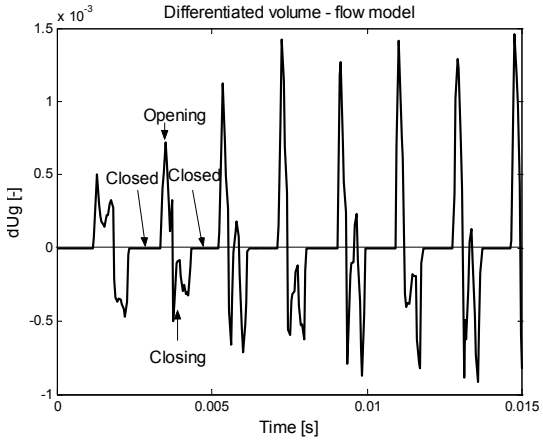


Fig. 4 Derivation of minimum gap dUg

frequency domains. The spectrum of the source should by reshaped by the filter. This means that if any of the natural frequencies of the vocal tract is identical with any of the harmonic of the source, this will affect its amplification. The vowel sound, if it appears in the speech is distinguished by the relative sizes of its harmonic. The individual vowels have harmonics, which have higher amplitudes close to the formants (resonance frequencies), or some of them are amplified directly by the filter.

Along with the change of tone height, the basic frequency of the folds and the distance of the harmonic elements [3] are changed, but the formants remain in the same places. Fig. 5 shows the spectrum of the source, determined by means of Fourier´s analysis of the course of volume acceleration on the figure Fig.4. The first harmonic is 546 Hz. Fig. 6 shows the transfer function (filter) of vowel /a/.

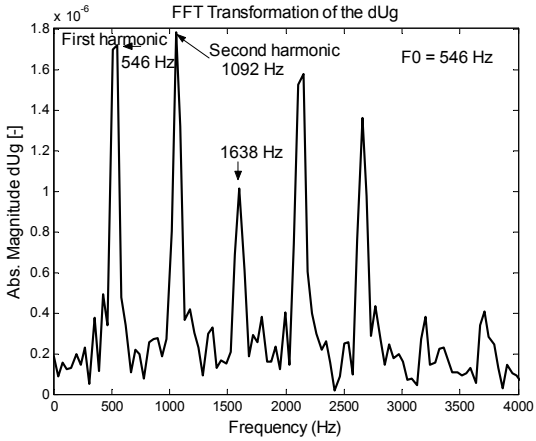


Fig. 5 Spectrum of the source, basic frequency F0=546 Hz

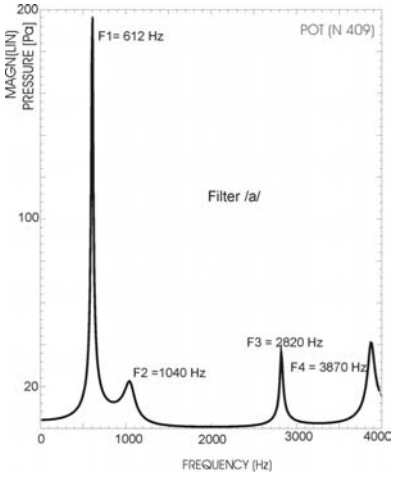


Fig. 6 Transfer function of vowel /a/

During the modelling of the vocal tracts, radiation impedance was included in the mouth area according to „Levin & Schwinger“ formula

$$Z = \left(0.24(ka)^2 + j*0.56(ka)\right)\rho c$$

Where ρ is mass density, c sound speed, k wave number, a opening of mouth. That allowed calculation of the acoustic pressures in the mouth area during phonation.
The results of the simulation are shown in Fig. 7 – the case of pronouncing of vowel /a/.

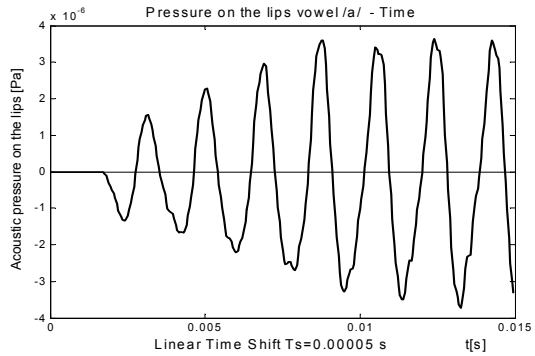


Fig. 7 Acoustic pressure in the course of pronouncing vowel /a/

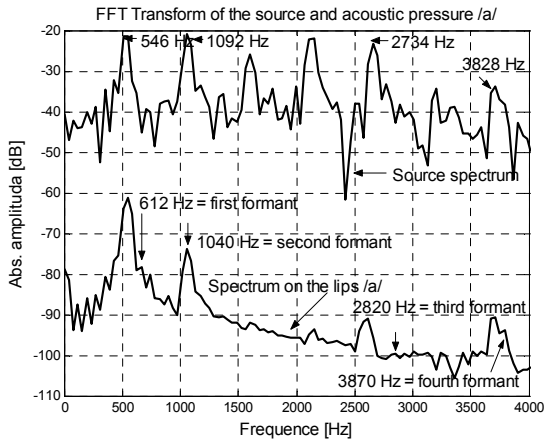


Fig. 8 Reshaping the source spectrum by the filter of the vocal tract /a/

It is obvious from Fig. 8 that, in accordance with the theory, those harmonic components of the source that are near the resonance frequencies of the filter of the vocal tract are amplified. E.g. according to transfer function, the first formant of the vowel /a/ is F1=612 Hz and the basic harmonic component of the source that is equal to 546 Hz is amplified, and the second formant is F2=1040 Hz and the second harmonic component of the source 1093 Hz is amplified. It is clear that the filter parameters cannot be constant, but they change as a result of movements of the articulators (tongue, jawbones etc.). Reshaping by means of the filter takes place in the frequency domain. Filter characteristics differ in the peaks of the transfer functions. These correspond to the natural frequencies of the vocal tracts for different vowels. The first two formants have a high level of correspondence, while the higher frequencies are comparatively more distinctly shifted.

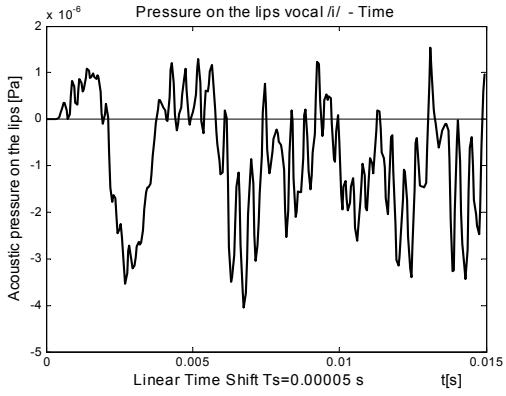


Fig. 9 Acoustic pressure in the course of pronouncing vowel /i/

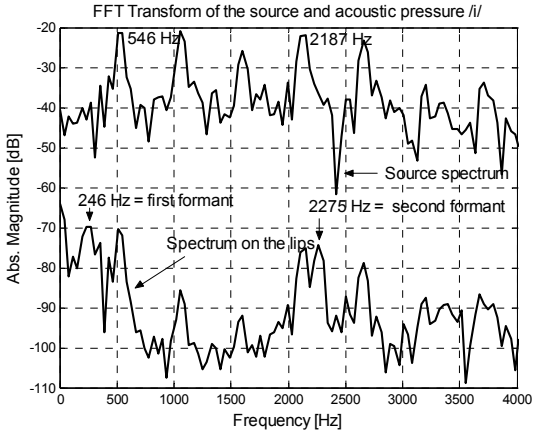


Fig. 10 Reshaping the source spectrum by means of the vocal tract of the vowel /i/

The first formant of the vowel /i/ is F1=246 Hz, the second one is F2=2275 Hz. These two formants are wide apart as shown in Fig. 10. Formants F3=3431 Hz and F4=3844 Hz are close to each other and they can be mutually affected. The first harmonic of the source cannot be amplified by the filter /i/ since it is very far from the first formant F1. The difference is only reached at the level of 221 Hz according to Fig. 10.

Based on the harmonic analysis the following formants were identified for vowel /o/: F1=516 Hz, F2=798 Hz, F3=2721 Hz and F4=3437 Hz. The first harmonic of the source is 546 Hz. The first formant of vowel /o/ is 516 Hz. The two frequencies are very close to each other and therefore this harmonic component is significantly amplified as shown in Fig. 12. Also the fifth and seventh harmonic elements of the source are amplified.

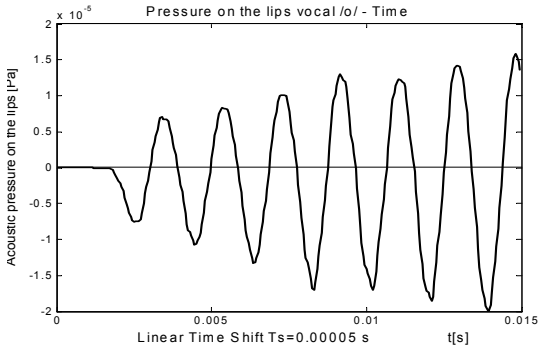


Fig. 11 Acoustic pressure when pronouncing vowel /o/

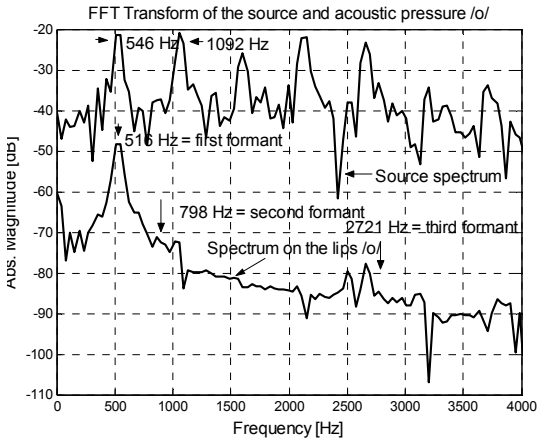


Fig. 12 Reforming of the source spectrum by the filter of the vocal tract /o/

IV. Discussion

Although all the vowels are excited by the same basic frequency of F0=546 Hz, their spectra differ on the outlet as a result of the different shapes of the vocal tract. The spectrum on the outlet is primarily determined by the resonance frequencies of the cavities and therefore by their shapes during articulation.

The harmonic components of the source are amplified near the mouth by the resonance frequencies and those components that are farther from the resonance frequencies lose energy. As an example, the first harmonic component of the source is 546 Hz (Fig.12) and the first natural frequency (first formant) of the vocal tract for vowel /o/ is 516 Hz. As they are very close to each other, the amplification is significant.

V. Conclusion

The primary source for generating vowels spoken aloud are the periodic movements of the vocal folds.

The model which was created by means of finite element method, was examined by loading it with subglottal pressure in relation to a minimum gap between the moving folds as shown in Fig. 1, where pressures are : PG1 just opened, PG2 just closed, PG0 beginning subglottal pressure. By means of this procedure we were able to achieve a substantial similarity of the phases of the vocal folds movements.

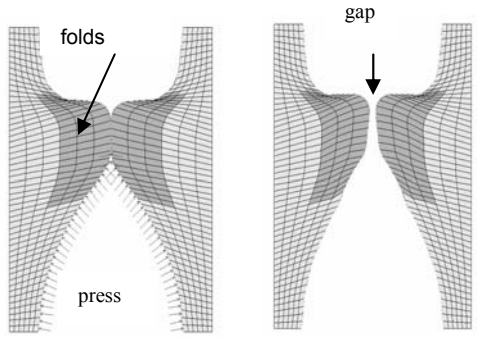


Fig. 13 Calculation model of the vocal folds

The time course of the gap between the folds corresponds to the course of the volume velocity and its derivative. Pressure wave with the spectrum corresponding to single vowel arises around mouth area. We can conclude that the modelling of the creation of the Czech vowels by means of the model defined about produces spectra that correspond to those identified by the measurements.

Acknowledgements

The research is supported by the Grant Agency of the Czech Republic by the project N0 106/98/K019 Mathematical-Physical Modelling of Vibroacoustic Systems Important in Biomechanics of Voice and Hearing, and by project No. MSM 262100001, Computational and physical modelling of engineering problems of thermo-fluid mechanics, mechanics of solids and phase changes.

References

[1] V. Misun, The principle of pressure air bubble within the phonation, Proc.Conf. Interaction and Feedbacks 2001, Prague, 27-28. Nov.2001, pp.115-122

[2] K.Prikryl , Numeric modelling of production of Czech vowels, 4th International conf. Mechatronic, Robotic and Biomechanics 2003, Hrotovice, March 2003

[3] Story B.H.: An overview of the physiology, physic and modelling of the sound source for vowels. Acoustic. Sci & Tech.23, 4 (2000)

Pathology detection

3rd International Workshop MAVEBA 2003, 247-250

THE EFFECT OF NORMALIZATION ON PARAMETERS IN DISCRIMINATION OF PATHOLOGICAL VOICE USING ARTIFICIAL NEURAL NETWORK

Tao Li, Il-suh Bak, Cheolwoo Jo
SASPL, School of Mechatronics, Changwon National University, Korea

Abstract: **In this paper we tried to examine the effect of normalization on discriminating the pathological voice into normal and abnormal classes using artificial neural network. Average values per each parameter were used to normalize each set of parameter values. Artificial neural network was used as a classifier. And the effect of normalization was evaluated by comparing the discrimination results between original and normalized parameter sets.**
Keywords: **Normalization, pathological, discrimination, neural network**

I. INTRODUCTION

These days there are many attempts to analyze and discriminate the pathological and normal voice by the original parameters (Jitter, Shimmer, NHR, SPI, etc.). The major purpose of such researches is to obtain some good standards and methods to classify and diagnose the patients who have diseases on their vocal folds. [1][2][3][4][5][6]

Even though there are some previous researches about discrimination of pathological voice, those only utilize original parameters' values as the data. Also artificial neural network has been widely used as a classifier because of random and complex characteristics of the pathological voice parameters. But the differences of the ranges of values among these parameters' magnitudes are very large. When bigger values and relatively much smaller values are input into the network for training at the same time, the effect of the parameters with the different magnitudes is not checked yet.

In this paper we suggest a normalization method to scale each parameter group's values and measure the effect of normalization by the classification rate from the artificial neural network.

II. DATA COLLECTION

To collect original voice data, collection system was installed in a room of the ENT department of hospital. The recording process was performed semi-automatically with the intervention of operator to control the quality and procedure. Also the voice materials from the same speaker were collected using DAT and CSL. [7][8] The sampling rate was 50 KHz and the resolution 16 bits. The collection was conducted in a hospital soundproof room. All the subjects were asked to pronounce /a/. Patient ages ranged between 23 and 75. Total voice data included 41 normal cases (33 males and 8 females), 59 pathological cases (43 males and 16 females) after removing invalid data from the raw data sets. The vocal diseases considered consisted of Vocal Polyposis, Hyperadduction, Vocal Cord Palsy, Vocal Nodule and Glottic Cancer etc. The parameters used are Jitter, Shimmer, NHR (Noise-to-Harmonic Ratio), SPI (Soft Phonation Index), APQ (Amplitude Perturbation Quotient) and RAP (Relative Average Perturbation). They were the 6 different kinds of parameters. [3]

III. NORMALIZATION

It is known that the units and magnitude ranges of the parameters Jitter, Shimmer, NHR, SPI, APQ, RAP, STD etc. are different. For example, Jitter is a percentage value but STD's unit is in Hz. And Shimmer's magnitude is much bigger than that of NHR. In the above, the measured parameter is 30.659 in one case of Shimmer, but 0.1296 in another case of NHR. As seen, there is great difference between these two parameters' magnitudes. When using an artificial neural network as a classifier with different parameters input, parameters with bigger value range may affect the classification rate more than those with smaller one.

Now in order to let these different parameters have the similar magnitude range, we normalized the 100 original measured values (41 are the normal data and 59 the pathological ones) for each parameter respectively (there are 6 kinds of parameters). Then we tried to observe the improvement of the classification rate with the normalized values comparing to that with the original values. By doing so, the effect of normalization can be measured and also we can measure how much each parameter affects the classification result under normalization.

Equation (1) and (2) shows how it is performed.

$$M_q = \frac{\sum_{i=1}^{K} P_i + \sum_{j=1}^{L} P_j}{K + L} \tag{1}$$

where P_i, $1 \le i \le K$, P_j, $1 \le j \le L$ are the original measured values of the normal and the pathological cases for parameter q respectively, K and L are the number of normal and pathological parameters respectively

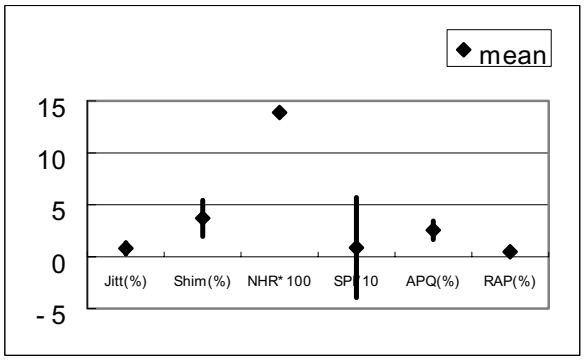
(K =41 and L=59 in this paper), and M_q is the mean value of the parameter q .

$$P_{nq} = \frac{P_o}{M_q} \tag{2}$$

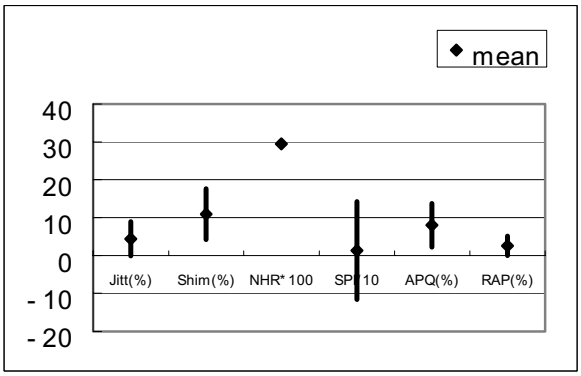
where P_o and P_{nq} are the original measured value and the corresponding normalized value for the specific parameter respectively. Then the normalized value for each parameter can be obtained by the equations (1) and (2).

After analyzing the collected voice materials using the analyzer and the above normalization method, we obtained the 6 different kinds of parameters (Jitter, Shimmer, NHR, SPI, APQ and RAP) which had the original measured values and especially the corresponding normalized values in this paper. And there were 100 original measured values and 100 corresponding normalized values for any kind of parameter. Also the 6 different kinds of parameters were divided into 3 categories according to their characteristics. They were pitch related (Jitter, RAP), amplitude related (Shimmer, APQ) and noise related (NHR, SPI). [6][7]

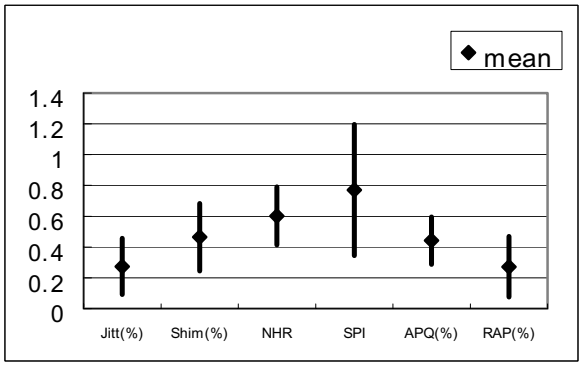
Fig.1 shows the comparisons of the original and normalized parameters from the normal and pathological voices. The graphs show the relative change of each parameter when DAT parameters are considered 1. [7]



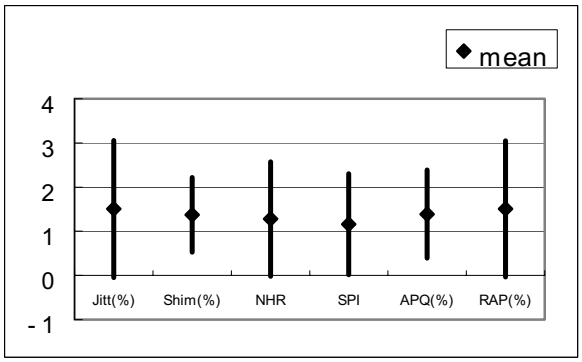
(a) Original normal parameter data



(b) Original pathological parameter data



(c) Normalized normal parameter data



(d) Normalized pathological parameter data

Fig .1 The comparisons of the original and normalized parameters

IV. EXPERIMENTS AND RESULTS

As a classifier of this experiment, three layered artificial neural network was used. The number of input layers was varied from 3 to 6 to find the optimal set of parameters. The number of output layers was fixed to 2. The number of hidden layers was set 3, 6, and 9 for the 3 inputs case, and 6, 9, and 12 for the 6 inputs case. [6] Because the total number of data was small, we tried to train and test the neural network by splitting total data sets into two parts. Two thirds of the data were used for training. The remaining one third was used for test. In each training session, the neural network was trained and tested separately using different combination of data sets. This was to compensate the small size of the data sets.

The original and normalized parameters were used to train and test the same structure neural network respectively. In order to accurately compare the difference between the classification results from the neural network using the original parameter input and that using the normalized parameter input, we must keep the same order when the two sets of different parameters were inputted into the different neural networks. That was, the inputting order of normalized parameters was corresponding to the original ones. So we could get the corresponding classification results.

Table.1 shows the classification rate from neural network training and testing with 6 parameters. Experiments were performed using 3 and 6 parameters respectively to see the different effects of the original and normalized parameters on discriminating the pathological voice into normal and abnormal classes. In case of 3 parameters, Jitter, Shimmer and NHR were used. And additional 3 parameters (SPI, APQ and RAP) were used for 6 parameters. There were 24 sets of result data in total.

V. DISCUSSION

From the experimental result we couldn't observe the significant difference among the corresponding classification rates when the original and normalized parameters were inputted into the different neural networks respectively as shown from table1. The results looked very similar. In order to obtained the observation results directly, we chose the best classification rate from every 24 sets of parameters to plot the changing trend curve of the classification rate as shown in Fig.2. In the Fig.2, any two sets of corresponding curves were very close and there was not big distance between them. But slightly better results were obtained for some the neural net configurations.

In the original parameter set, the value range of NHR was about 100 times bigger than other parameters. After normalization, the range became similar to those of others. But the classification result didn't change much, so NHR didn't play a significant role at the classification. In other parameters, the difference of relative change of values between before and after normalization was not so big. And normalization process didn't affect the performance of the network.

VI. CONCLUSION

In this paper we collected pathological voice materials using DAT. And the normalization method of the original parameters was introduced. Artificial neural network was used to classify the voice into normal and abnormal states by original and normalized parameters.

From the experiments we couldn't observe a significant improvement or decrease of performance by normalizing parameters with suggested way. And we can conclude that normalization process is not necessary for the classification of pathological voice when using artificial neural network.

But the total amount of voice data is still not enough to generalize the performance and more data collection is required.

ACKNOWLEDGEMENT

This study is supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea. (No. 02-PJ1-PG10-31401-0005)

REFERENCES

[1] Godino-Llorente, Santiago Aguilera-Navarro*, Carlos Hernandez-Espinosa**, Merced-es Fernandez-Redondo * *, Pedro Gomez-vilda+, "On the selection of meaningful speech parameters used by a pathologic /non pathologic voice register classifier," *Eurospeech'99, Budapest, Hungary,* vol. 1, pp. 563-566, 1999.

[2] Cheolwoo Jo, Daehyun Kim, Soogeon Wang, "Classification of pathological voice into normal /benign/malignant state," *Eurospeech'99, Budapest, Hungary,* pp. 399-402, 1999.

[3] Daehyun Kim, Cheolwoo Jo, "A study on the classification of pathological voice," *15th Workshop on Speech Communication and Signal Processing,* pp. 388-391, 1998.

[4] Kwangin Kim, Cheolwoo Jo, Daehyun Kim, Soogeon Wang, Gyerock Jeon, Sihun Ahn, Kiryon Kim, Yongju Kim, "Collection, analysis and classification of pathological voice using ARS and neural network," *13rd Korean Conference on Signal Processing,* pp. 955-958, 2000.

[5] Cheolwoo Jo, Daehyun Kim, Kwangin Kim, Soogeon Wang, Gyerock Jeon, Sihun Ahn, Kiryon Kim, Yongju Kim, "Implementation of analysis tools for pathological voice," *17th Workshop on Speech Communication and Signal Processing,* pp. 211-214, 2000.

[6] Cheolwoo Jo, Kwangin Kim, Daehyun Kim, Soogeon Wang, "Screening of pathological voice from ARS using neural networks," *International Workshop on Maveba, Firenze, Italy,* September 13-15, 2001.

[7] Operations Manual, "Multi-Dimensional Voice Program (MDVP)," Model 4305, Kay Elemetrics Corp, 1993.

[8] Cheolwoo Jo, Kwangin Kim, Daehyun Kim, Soogeon Wang, Gyerok Jeon, "Comparisons of acoustical characteristics between ARS and DAT voice," *Eurospeech'2001, Denmark,* 2001.

Table.1 The classification rate (%)

Neural Network Structure	Times	Original Data		Normalized Data	
		Test Data	Training Data	Test Data	Training Data
3 Inputs 3 Hidden layers 2 Outputs 3I3H	1st Run	87.5000	91.1765	84.3750	97.0588
	2nd Run	84.3750	86.7647	84.3750	97.0588
	3rd Run	84.3750	86.7647	84.3750	86.7647
	4th Run	81.2500	98.5294	78.1250	97.0588
	5th Run	81.2500	94.1176	78.1250	95.5882
3 Inputs 6 Hidden layers 2 Outputs 3I6H	1st Run	81.2500	92.6471	84.3750	100.0000
	2nd Run	81.2500	91.1765	81.2500	98.5294
	3rd Run	78.1250	95.5882	78.1250	100.0000
	4th Run	75.0000	100.0000	78.1250	100.0000
	5th Run	75.0000	97.0588	78.1250	100.0000
3 Inputs 9 Hidden layers 2 Outputs 3I9H	1st Run	87.5000	98.5294	84.3750	100.0000
	2nd Run	84.3750	100.0000	81.2500	100.0000
	3rd Run	84.3750	94.1176	81.2500	100.0000
	4th Run	78.1250	95.5882	81.2500	97.0588
	5th Run	75.0000	100.0000	78.1250	100.0000
6 Inputs 6 Hidden layers 2 Outputs 6I6H	1st Run	81.2500	100.0000	84.3750	100.0000
	2nd Run	78.1250	100.0000	84.3750	100.0000
	3rd Run	75.0000	98.5294	81.2500	100.0000
	4th Run	71.8750	100.0000	81.2500	100.0000
	5th Run	71.8750	100.0000	78.1250	100.0000
6 Inputs 9 Hidden layers 2 Outputs 6I9H	1st Run	81.2500	100.0000	87.5000	100.0000
	2nd Run	81.2500	100.0000	84.3750	100.0000
	3rd Run	78.1250	100.0000	81.2500	100.0000
	4th Run	78.1250	100.0000	78.1250	100.0000
	5th Run	75.0000	100.0000	75.0000	100.0000
6 Inputs 12 Hidden layers 2 Outputs 6I12H	1st Run	84.3750	97.0588	87.5000	100.0000
	2nd Run	78.1250	100.0000	81.2500	100.0000
	3rd Run	78.1250	100.0000	81.2500	100.0000
	4th Run	78.1250	100.0000	81.2500	100.0000
	5th Run	75.0000	100.0000	78.1250	100.0000

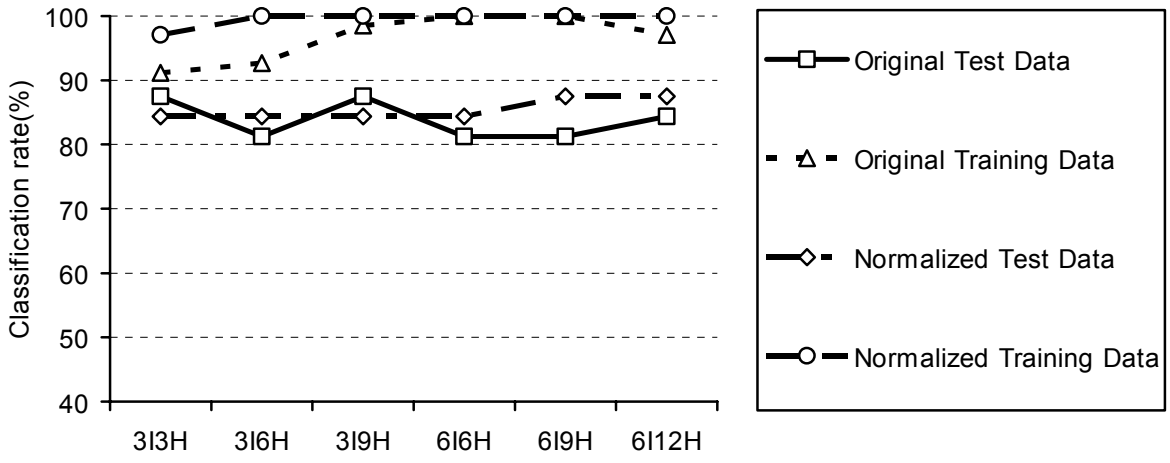


Fig.2 The changing trend curve of the classification rate

3rd International Workshop MAVEBA 2003, 251-254

AUTOMATIC DETECTION OF CHRONIC PIG COUGHING FROM CONTINUOUS REGISTRATION IN FIELD SITUATIONS

A. Van Hirtum[1], M. Guarino[2], A. Costa[2], P.Jans[1], K. Ghesquiere[1], J.-M. Aerts[1], P.L. Navarotto[2], D. Berckmans[1]

[1] Laboratory for Agricultural Buildings ResearchKatholieke Universiteit Leuven, Kasteelpark Arenberg 303001 Leuven, Belgium

[2]Department of Veterinary and Technological Sciences for Food safety, Faculty of Veterinary Medicine, via Celoria, 10, 20133 Milan, Italy

Abstract: **Cough is an important and present symptom in many respiratory diseases affecting the airways and lungs. Therefore it is interesting to monitor cough in a continuous, on-line way. The objective of this study was to test a cough recognition algorithm in real pig houses. Cough sounds were registered on 150 days old, 60 kg heavy Landrace x Large White x Duroc crosses with a microphone placed at 20-50cm from the animal. The analysis was done on a feature vector, containing energy, time-derivate energy and mean power spectral density. The feature vector was compared to the reference set using dynamic time warping. This resulted in a correct classification of 90%.**

Keywords: **Sound analysis, Cough, Diagnostic system, Health management.**

I. INTRODUCTION

Health care management is a critical and demanding issue in current livestock production. Discarding the economic cost related to large scale diseases early detection of diseases is important considering public health care issues like reducing antibiotics residuals. Also for reasons of animal welfare and monitoring and tracing of the food production chain, online disease monitoring is important. Therefore currently great effort is spent to the development and application of sensors and sensing techniques for diagnosis in the agricultural sector [1]. With respect to objective and automated detection of respiratory diseases in livestock, it has been shown that artificial intelligence is successfully applicable to obtain automated cough recognition from free field cough recognition. In [2,3,4] an accurate algorithm is presented to detect citric acid induced coughing originating from healthy individual piglets. In an intelligent free field recognizer is proposed to distinguish between coughing is evoked in absence or presence of a respiratory infection [5]. Although the mentioned references firmly emphasize the applicability of sound analysis in order to obtain an early, objective, contact less and continuous alarm system for coughing, the results are obtained on a database which is registered on individual subjects housed in a laboratory test-installation consisting of a laboratory inhalation-chamber. The test-installation, detailed in [5,6], allows to control environmental housing conditions, medical follow-up and to reduce environmental noises. So cough sounds are registered in optimal environmental sound conditions. Therefore the performance of the developed algorithms to recognize cough in field conditions needs to be assessed in order to validate the usage of sound analysis in livestock health management. The objective of this study was to test a cough recognition algorithm in real pig houses field conditions.

II. METHODOLOGY

Data capturing in field conditions: Experimental data were obtained in swine housing for finishing pigs assigned to the Parma ham production in Northern Italy.

Animals: The pigs (Landrace x Large White x Duroc crosses for Parma ham production) were in the first period of the finishing phase, their mean weight was around 60 kg and their mean age was 150 days. The farm was composed of three barns for piglets, sows, and finishing pigs. The barn for finishing pigs was an open-space 8,3 m x 83 m wide, it was subdivided in 16 boxes 6 x 5 m wide, and each boxes had a dunging area 1,3 m x 5 wide containing 50 pigs each. The boxes were delimited by a little wall in concrete, 1m high and 20 cm thick.

Sick pigs affected by cough, were confined in the six final boxes, in order to separate them from the healthy ones. A serological assay on blood sample to verify the presence of Pleuropneumonitis antibodies has been conducted on sick pigs to verify the source of coughing. After the slaughtering, Pleuropolmonitis was confirmed by the autopsy examine performed by the farm veterinarian. The average daily gain (ADG) in healthy pigs was 653 g/die, while the sick pigs showed a lower ADG calculated in 437 g/die.

Measurements: Pigs cough was recorded using a microphone linked to a portable computer. The operator, standing in the box, among pigs, recorded the coughs putting the microphone at 20-50 cm from the animal. In total, 44 cough attacks have been recorded from different animals in almost 4 hours.

Signal analysis: The applied signal analysis part consists of two main issues. Firstly individual sounds are objectively searched in the continuous sound registrations. Secondly suitable sound features are extracted to present to the classification algorithm given in the following subsection.

Individual sounds are retrieved by applying a threshold to the signal energy. The signal energy is calculated on signal-frames of 0.01s. The energy threshold is initiated with the energy level at the beginning at the signal, which is assumed to be silent. The threshold level is allowed to change smoothly in accordance with the variation in energy-level of the signal.

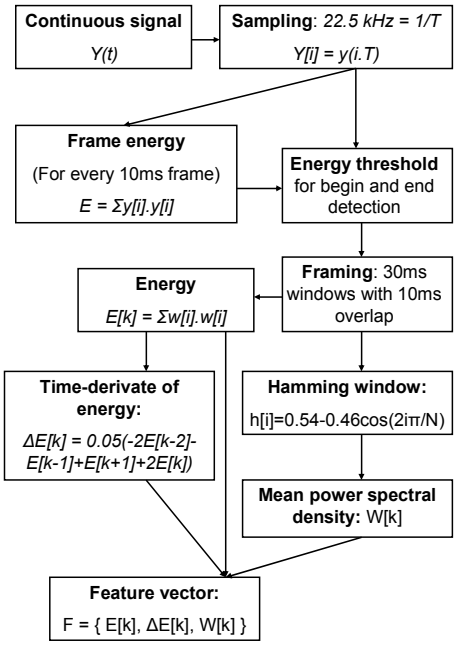


Fig. 1: Overview of the performed signal analysis.

In general, more objective, automated interpretation of respiratory related sounds is obtained considering the spectral power of the sound samples [7]. Spectral features are estimated by applying the discrete Fourier transform (DFT) or by averaging spectral estimates on the windowed N-element sound samples. In the last case N-element sample-parts are obtained by overlapping consecutive windows of width N. The relevance of frequency features towards the automated cough identification is shown in [2,5,8,9]. In [5,9] the averaged spectra are successfully used to distinguish between cough sounds originating from pathological and non-pathological subjects. In [2] averaged spectral features are employed to classify between citric acid induced cough sounds and other sounds.

Therefore in the current work spectral features are determined for the frequency range from 3 kHz up to 6 kHz on 30msec frames computed every 10msec. To consider the transient time behaviour in the cough, for each frame sound energy derivatives are added to the spectral features.

For clarity the consecutive steps in signal analysis are listed in the schematically in Fig. 1.

Classification: Cough sound recognition is assessed with dynamic programming i.e. 'dynamic time warping (DTW)' [10,11]. As indicated before, each sound is divided into frames of equal length and the features of each frame are stored in a feature vector. Thus, each sound is represented by a sequence of data feature vectors that form a sound template. The different duration of the cough sound results from non-uniform stretching and compression of the various portions in the cough sound. Consequently simple linear time alignment is not appropriate to compare two sounds of unequal duration. In order to compare two sound templates, the DTW algorithm uses one of them as a test pattern and the other one as a reference pattern. Taking frame by frame of the test sound template, DTW looks for the frame-path in the training template that results in the minimum distortion. For each test frame a set of specified frames in the training template is allowed for comparison. The set of allowed frames is determined by local continuity and monotonicity constraints. The constraints are imposed such that the temporal order in which frames occur is significant. Represent the sequence of data feature vectors from the test template by $X=(x_1,x_2,...,x_{T_X})$

and from the training template by $Y=(y_1,y_2,...,y_{T_Y})$.

Define two warping functions ϕ_x and ϕ_y which relates the indices of the test and training frames, respectively i_x= 1, 2, …, T_x and i_y= 1, 2, …, T_y, to a common time axis k=1, 2, ….,T so that $i_x=\phi_x(k)$, $i_y=\phi_y(k)$ and $\phi=(\phi_x,\phi_y)$ the function pair specifying the path. A global pattern dissimilarity measure between the test and training sequence $d_\phi(X,Y)$ is than defined as the accumulated distortion over the entire sound utterance or sequence as

$$d_\Phi(X,Y)=\sum_{k=1}^{T} d(x_{\Phi_x(k)}, y_{\Phi_y(k)})\frac{m(k)}{M_\Phi} \qquad (1)$$

with d the Euclidean distance, m(k)>0 a path weight coefficient and M_ϕ, a normalizing factor. In this paper local and global Itakura path constraints are applied allowing $\phi_x(k+1)$, $\phi_y(k+1)$ to take respectively the values $\{\phi_x(k)+1\}$ and $\{\phi_y(k)$,$\phi_y(k)+1$,$\phi_y(k)+2\}$ while $\phi_x(k+2)=\phi_x(k)$ is excluded. Depending on the value of

$\phi_y(k+1) \in \{\phi_y(k), \phi_y(k)+1, \phi_y(k)+2\}$ the weight coefficient m(k) takes respectively the value 1.5, 1 and 1.5. The global path constraints specifies the range of the points (i_x,i_y) which can be reached from the beginning point (1,1) via an allowable path according to the local constraints and the range of points that have a legal path to the ending point (T_x,T_y). This is expressed as follows:

$$1+\frac{[\Phi_x(k)-1]}{Q_{max}} \le \Phi_y(k) \le 1+Q_{max}[\Phi_x(k)-1]$$
$$T_y+Q_{max}[\Phi_x(k)-T_x] \le \Phi_y(k) \le T_y+\frac{[\Phi_x(k)-T_x]}{Q_{max}}. \quad (2)$$

Where Q_{max} and Q_{min} denote the values of maximum and minimum path expansion to $Q_{max}=2$ and $Q_{min}=0.5$.

During the recognition phase the template of the test sound X is compared to each template in the set of training templates using the DTW algorithm. The training template Y producing the minimum distortion, i.e.

$$d(X,Y) = \min_{\Phi}(d_{\Phi}(X,Y)) \quad (3),$$

determines the classification output.

III. Results

From Continuous Sound Registration To Individual Sounds:

Table 1: The continuous sound registration, described in the methodology subsection:

Number of on-line registered sound files	44 files
Duration	Min: 3.2 s / Max: 23.2 s / Average: 9.7 s
Number of individual sounds	592 sounds
Number of coughs	159 sounds (27 %)
Number of other sounds	433 sounds (73 %)

An exemplar of a continuous sound registration with duration 19.2s is given in Fig. 2. Since the sound registration implies no signal pre-processing at first, individual sounds are retrieved from the continuous registration.

Since the number of involved continuous registrations (44) is limited, all sound files are manually listened and visually inspected to validate the individual sound detection. In particular it is required that all cough events are detected as individual sound. If not, an error is introduced preceding the effectual classification approach outlined in the methodology subsection. In the continuous registration depicted in Fig. 1, 32 individual sounds are detected automatically of which 19 cough sounds. This number coincides with the auditively detected number of coughs. The detection of individual sound events is illustrated in Fig. 2. Individual sounds retrieved from the first 9.6s from the continuous sound registration shown on top of Fig. are indicated.

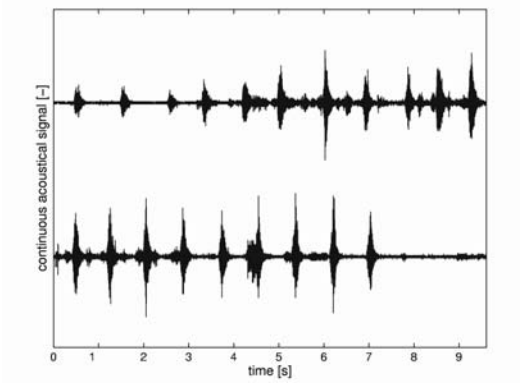


Fig. 2: Exemplar of 19.2s continuous sound registration

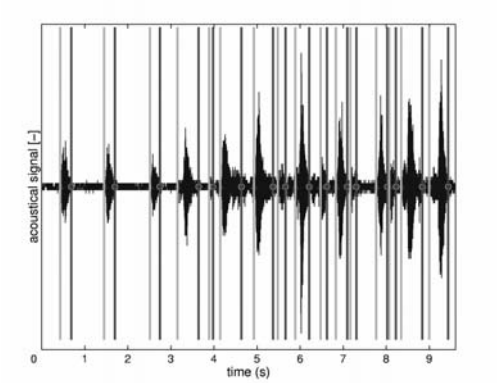


Fig. 3: Individual sounds retrieved from the first 9.6s from the continuous sound registration shown on top of Fig. 2. The beginning of an individual sound is indicated with a vertical line and a triangle (A), the end of an individual sound is indicated with a vertical line and a circle (o).

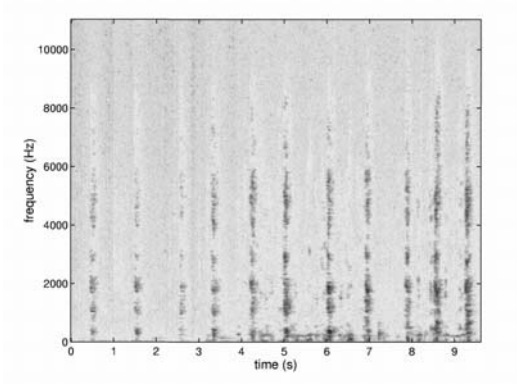


Fig. 4: Spectrogram for the first 9.6s from the continuous sound registration shown on top of Fig. 2.

The beginning of an individual sound event is pointed out with a vertical line and a triangle (Δ), the end of an individual sound is indicated with a vertical line and a circle (o).

In total a database of 592 individual sound events is objectively and automated obtained from the on-line continuous sound registrations under field conditions. From the total database 159 individual sounds or 27% involve cough events. The other 433 individual sounds or 73% comprise no cough events but contain mostly vocalisation sounds and background noises.

Scoring: The spectrogram for the first 9.6s from the continuous sound registration is shown on top of Fig. 2. Recognition performance is assessed applying the 'leave-10-out' method. The classifier is trained using all individual cough events except 10%. The remaining 10% is used for testing. The method is repeated until the entire database is tested. This method is known to provide a good estimation of the error in case of small databases.

To illustrate the features, the spectrogram for the first 9.6s from the continuous sound registration (shown on top of Fig. 2) is shown in Fig. 4. Limiting the spectral frequency to the range from 3kHz to 6kHz allows to eliminate low-frequency noises from mechanical origin while, as depicted in the spectrogram, the cough sound exhibits an important energy-peak in this range.

IV. Discussion

The accuracy of the cough recognition with the features and classification approach described in the methodology yields 90%. The recognition performance didn't depend on the test set and so reaches the same value for all repetitions. This is 4% lower than the recognition rate obtained in case of citric acid induced coughing [2, 3, 4]. Several factors contribute to the lower recognition rate. Firstly the data are registered in field conditions and not in a laboratory set-up as was the case in previous work. Secondly in contrast to the results presented in previous work, individual sounds are objectively and automated detected from the continuous and on-line sound registrations.

V. Conclusion

In this research it was demonstrated that the combination of on-line measured sound information by means of a cheap microphone with a cough sound recognition algorithm, can be used to monitor the health status of pigs in field conditions. The cough recognition algorithm was tested on 44 sound files recorded in field conditions. Cough could be classified successfully with an accuracy of 90%.

Acknowledgments

We thank the “Riva” farm (LO) for the availability.
Research funded by MIUR COFIN 40 %, 2001-2003

References

Nicks, M. Ansay, and P. Gustin, "Chronic exposure of pigs to airborne dust and endotoxins in an environmental chamber," *Veterinary Research Communications*, vol. 27, pp. 569-578, 1996.

[7] M. Oud, E. Dooijes, S. Taring, and S. van der Zee, "Asthmatic airways obstruction assessment based on detailed analysis of respiratory sound spectra", *IEEE transactions on biomedical engineering*, vol. 47(11), pp. 1450-1455, 2000.

[8] Y. Hiew, J. Smith, J. Earis, B. Cheetham, and A. Woodcock, "Dsp algorithm for cough identification and counting," *Proc. ICASSP'02,* Orlando, Florida, pp. 3888-3891, 2002.

[9] A. V. Hirtum and D. Berckmans, "Automated recognition of spontaneous versus voluntary cough,"

[1] I. Tothill, "Biosensors developments and potential applications in the agricultural diagnosis sector," *Computers and Electronics in Agriculture*, vol. 30, pp. 205-218, 2001.

[2] A. V. Hirtum and D. Berckmans, "Fuzzy approach for improved recognition of citric acid induced piglet coughing from continuous registration," *J. Sound and Vibration*, in press.

[3] A. V. Hirtum and D. Berckmans, "Objective recognition of cough-sound as biomarker for aerial pollutants," *International Journal of Indoor Air Quality and Health*, in press.

[4] D. Moshou, A. Chedad, A. V. Hirtum, J. D. Baerdemaeker, D. Berckmans, and H. Ramon, "Neural recognition system for swine cough," *Mathematics and Computers in Simulation*, vol. 56, pp. 475^87, 2001.

[5] A. V. Hirtum and D. Berckmans, "Intelligent free field cough sound recognition," *Proc. ICONS'03,* Faro, Portugal, pp. 453-58, 2003.

[6] B. Urbain, J. Provoust, D. Beerens, O. Michel, B. *Medical engineering & physics*, vol. 24, pp. 541-545, 2002.

[10] J. Deller, J. Proakis, and J. Hansen, *Discrete-time processing of speech signals*, Prentice Hall, New Jersey, 1993.

[11] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice Hall, New Jersey, 1993.

3rd International Workshop MAVEBA 2003, 255-258

SPECTRAL ENTROPY SIGNATURE OF SPEECH PERTURBATION IN ADULT ACQUIRED GROWTH HORMONE DEFICIENCY

C.J.Moore[1], K.Manickam[1], S.Shalet[2], T.Willard[3], S.Jones[4]

[1]North Western Medical Physics, Christie Hospital, Manchester, United Kingdom
[2]Department of Endocrinology, Christie Hospital, Manchester, United Kingdom
[3]North Western Medical Physics, Withington Hospital, Manchester, United Kingdom
[4]Department of Speech and Language Therapy, Wythenshawe Hospital, Manchester, United Kingdom

Abstract: **Approximate entropy (ApEn) adapted to quantify the pattern complexity across the electro-glottogram (EGG) spectral domain characterizes normal male vowel phonation in two groups, a majority group (G1) with high ApEn and a minority group (G2) with low ApEn. Using the ApEn measure of normality a sample of post-treatment male oncology patients with adult onset growth hormone deficiency (GHD) shows distinctive spectral entropy signatures. These are consistent with either disrupted larynx development in relative youth, with high normal-group G1 complexity and elevated pitch, or loss of conscious control in middle age, with low normal group G2 or worse complexity. This is at least initial evidence that speech perturbation may be of value in detecting the adult GHD in oncology.**

Keywords : **Speech, Complexity, Endocrine, Disruption, Oncology**

I. INTRODUCTION

It is well known that severe growth hormone deficiency (GHD) has a substantial impact on the growth of children and is subsequently reflected in a pronounced deficit of skeletal mass in adult life. It is perhaps less well known that GHD, acquired as an adult, strongly affects body composition as well as the continued health of the skeleton, cardiovascular risk and quality of life. In the context of quality of life in adult acquired GHD there is anecdotal evidence that subtle speech perturbation becomes noticeable but the quality of the perturbation depends on the timing of the onset of GHD. Thus we have conducted objective speech studies in four adult male oncology patients who acquired GHD in adult life. The group was measured against normal voicing adult male and female cohorts using a single-parameter measure of normality, which has recently been developed [1]. It was felt that such a study would provide useful information regarding the impact of GHD status on the development of the larynx and maintenance of speech pattern. In particular it is hoped that this will be the first stage in identifying the speech-perturbation signature of GHD.

By concentrating on vocal fold functionality a single parameter measure of normality has been developed by scientists at North Western Medical Physics in Manchester. It has been used to identify very tight bounds for 'normal' vowel production, or phonation, in both the healthy male and female populations. It is uniquely based on the quantification of the complexity of the entire spectral pattern from trans-larynx impedance measurements acquired during vowel phonation. The spectral pattern complexity for larynx cancer patients has been shown to be quite different and to change with the time elapsed following treatment [1]. Therefore, it is likely that complexity also has the potential to characterise subtle deviations from normal phonation resulting from GHD at different times of onset in adult life.

II. THEORY

The anterior pituitary hormone somatotrophin, commonly referred to as growth hormone, is involved in the stimulation of protein synthesis, amino-acid transport, fat and calcium uptake. Consequently, GHD in adults, regardless of gender, can result in decreased muscle mass, increased body fat and reduced bone density. In males the hormone testosterone is central to development of the adult male characteristics of musculature, bone mass, fat distribution, hair patterns, laryngeal enlargement, and vocal chord thickening. Once again pituitary disruption affects testosterone secretion. Consequently male oncology patients with GHD can be prescribed endocrine replacement therapies that include testosterone and the endocrine stimulant thyroxine.

The mass and composition of the folds, the integrity and mobility of the epithelium etc. are all reflected in the intricate pattern of fold vibration. Hence, it is possible that the vibration of the vocal folds is sensitive to minute

physiological changes [2] of the kind that might be expected in the development of GHD.

Fold vibration can be measured indirectly and non-invasively by exploiting trans-larynx electrical impedance variations [3]. A time series of impedance measurements is termed the electro-glottogram or EGG. A carefully gathered EGG is free from the complex resonant effects of the vocal tract, which can be variably configured by an individual. The relatively simple EGG time series is ideal for power spectral density (PSD) analysis, which is the natural choice for investigating vibration phenomena [4,5] including the EGG [6]. However, the usefulness of conventional PSD analysis is limited by the difficulty of interpreting the spectrum taken as a whole, especially where any perturbations are subtle.

In order to quantify subtle changes to fold functionality a measure of continuous spectral pattern complexity is required, which takes into account the entire spectral domain, rather than the selective analysis of a few discrete spectral peaks that are assumed to be of paramount importance. To progress towards characterization of spectral pattern, pre-normalisation can ameliorate the effects of pitch, $f_{o,}$ variation that would otherwise obscure any underlying spectral pattern in vowel phonation. Tracking f_o and expressing spectral components relative to f_o, combined with the normalisation of all component spectral powers, relative to the power of f_o , ensures that multiple spectral estimates can be averaged to reinforce common patterns. The authors term this 'Fundamental Harmonic Normalisation' [6]. For FHN-normalised spectra the influence of f_0 power is not lost, instead it is directly reflected across the scaled pattern of the normalised spectrum itself.

'Approximate entropy' is a measure of time series complexity that has been used in ECG studies of anaesthetised patients [7]. The measure is sensitive to noise in the time domain. However, in the *spectral* domain noise is relatively slowly varying, potentially flat, and does not directly distract from the complexity analysis of features of real interest. For this reason the authors have extended complexity analysis into the EGG, FHN-spectral domain for GHD patients. Since it is normalized, the pattern complexity of the FHN spectrum is concentrated from the first maximum of the harmonic peaks onwards. Therefore, for this study the FHN spectrum was truncated and the 7 harmonics following the first FHN spectral minimum selected for analysis. After taking the logarithm of the spectrum the standard deviation, σ,of the resultant spectral series is computed for use in approximate entropy calculations.

This study uses a specific formulation of approximate entropy ApEn described by Pincus [8]. Given N data points {u(i)}=u(1),u(2),….,u(N) and commencing with the i^{th} point, vector sequences $\mathbf{x}(1)$ to $\mathbf{x}(N–m+1)$ are formed consisting of m consecutive u $\mathbf{x}(i)=[u(i),\ldots,u(i+m-1)]$. Then the vector sequence, $\mathbf{x}(1),\mathbf{x}(2),\ldots,\mathbf{x}(N-m+1)$ is used to construct $C^{m,i}(r)$ values for each $i \leq (N-m+1)$ where;

$$C^{m,i}(r) = \text{number of } j \leq (N-m+1)$$

$$\text{such that} \quad d[\mathbf{x}(i),\mathbf{x}(j)] / (N-m+1) \leq r \qquad (1)$$

Where $d[\mathbf{x}(i),\mathbf{x}(j)]$ in (1) is the distance between vectors, defined as the maximum difference in their scalar components. The $C^{m,i}(r)$ values measure, within a tolerance r, the regularity or frequency of sequences occurring in the data set {u(i)}, which are similar to the given sequence, $\mathbf{x}(i)$ of length m. The Pincus approximate entropy statistic is then defined by;

$$A_pE_n = -(N-m)^{-1}\sum_{i=1}^{N-m} \ln [C^{m+1,i}(r)/C^{m,i}(r)] \qquad (2)$$

Equation (2) is interpreted heuristically as a measure of the average logarithmic likelihood, over all sequences $\mathbf{x}(1)$ to $\mathbf{x}(N-m+1)$, that any sequence in the data series {u(i)}, which is within a tolerance r of the given sequence $\mathbf{x}(i)$ of length m, remains within the same tolerance when the length of both sequences is increased by one data point. Tolerance r is proportional to the measured series standard deviation σ, i.e. $r=k\sigma$, where k is a constant. It is necessary to empirically determine k so that the widest range of complexity values is achieved.

III. METHODOLOGY

4 adult males (age range 23-47) who had acquired their GH deficiency in adult life, either as a consequence of tumour mass effect or radiation induced damage, were studied using a Laryngograph. Details are shown in Table-1. All had adult onset GHD for at least 2 years prior to the study and were assessed several years after diagnosis of the tumour mass or treatment. 89 healthy male volunteers were recruited to provide a 'normal' reference Laryngographic standard. Four volunteers were excluded because of errors during capture. For pathological and normal volunteers cases Laryngograph throat sensors were used to measure trans-larynx EGG signals for the sustained vowel /i/. Sampling was at 20 kHz for up to 4 seconds The resultant 4 pathological and 85 normal binary LX data-files were transmitted by FTP

to a COMPAQ Unix Alpha-server-2000 dual 4/275 processor system for storage. Visualisation, spectral and complexity analysis were then performed off-line on an AMD-Athlon, 1GHz processor, NT-PC equipped with 1Gbyte memory. All software utilities were written in Research Systems International IDL 5.5.

For sustained EGG signals multiple power spectral estimates were generated for individuals by segmenting the EGG data-stream into short frames of 1000 sample points. For each frame f_o was determined using the autocovariance function before power spectral density (PSD) computation by variance reduction and Fourier transformation. The frame PSD was then FHN normalised relative to the frequency and power of f_o for the frame itself. All frame FHN-spectra were then averaged to reinforce any shared pattern in each case.

Patient	Age	Pathology	Treatment
R	23	Pituitary stalk lesion	Nil
H	44	NF Pituitary adenoma	Pituitary Surgery & irradiation
W	47	Macroprolactinoma	Pituitary Surgery & irradiation
S	40	Craniopharyngioma	Pituitary Surgery

Table-1
Adult GHD cases, age at study, pathology and treatment (additional to endocrine replacement that includes testosterone for all patients and thyroxine for H, W & S).

The entire FHN, EGG spectral pattern for each individual was then characterised above fo using complexity analysis based on ApEn. Specifically the averaged FHN spectrum was truncated to produce a single new series extending from the first minimum to the 7th harmonic inclusive, taking logs and then computing the standard deviation σ of the result. In order to obtain the widest spread of ApEn, a k value of 0.6 was empirically determined for the computation of *r* in both normal and pathological cases.

IV. Results

Normal Population: Spectral Complexity

Normal spectral patterns clearly separate into two ApEn complexity groups (Students two tail t-test $p<<0.001$).

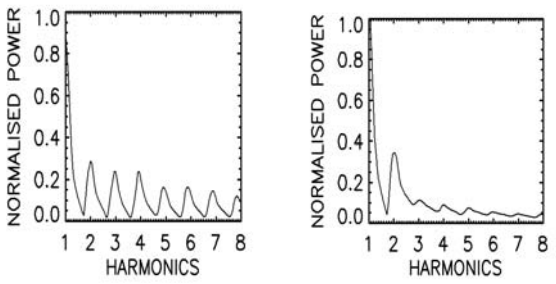


Figure-1
Two normal male population FHN spectral patterns; normalized power (ordinate) vs. harmonics (1 to 8).

Left: Normal G1 male, strong spectrum
Right: Normal G2 male, weak spectrum

The largest group, G1, had strong spectral features extending across the harmonic range within a slowly decaying spectral envelope. The smallest, G2, had weak spectral features that decayed rapidly towards the higher harmonics. Figure-1 shows the population averaged FHN for G1 and G2 populations. ApEn complexity analysis elegantly quantifies G1 and G2 differences. G1 has 55 individuals with high, mean complexity 0.34 (+/- 0.04). G2 has 30 individuals with low mean complexity 0.18 (+/- 0.05). Pitch analysis of EGG data showed no differences between G1 and G2, both having a mean f_o of 122-124 Hz (+/- 29 Hz)

Case	f_o (Hz)	Complexity
Normal Males G1	124 (29)	0.34 (0.04)
Normal Males G2	122 (29)	0.18 (0.05)
R	172	0.35
H	154	0.18
W	118	0.18
S	117	0.09

Table-2
Fundamental frequency fo (standard deviation) and complexity for normal and pathological cases.

Adult Acquired GHD Cases: Spectral Complexity

Table-2 shows the spectral complexities and fo values for the 4 adult acquired GHD cases. The first row shows case R in which fo is 172 Hz, intermediate between normal male and female pitch. It is the only case in which the spectral pattern is strong and well maintained to high harmonic levels. The spectral envelope is clearly 'bright' but erratic with spectral envelope decay reversed twice. The complexity at 0.35 is typical of the G1 population.
The remaining 3 adult acquired GHD cases have characteristically low, male fo. Their spectra clearly exhibit the G2 decaying envelope and the pulsatile

reduction with increasing harmonic level. Progressing from case H to S there is the characteristic obliteration of spectral features by noise. Complexity levels are low and comparable to the G2 normals, particularly for case S, where there are few spectral features and a pathologically low ApEn of 0.09.

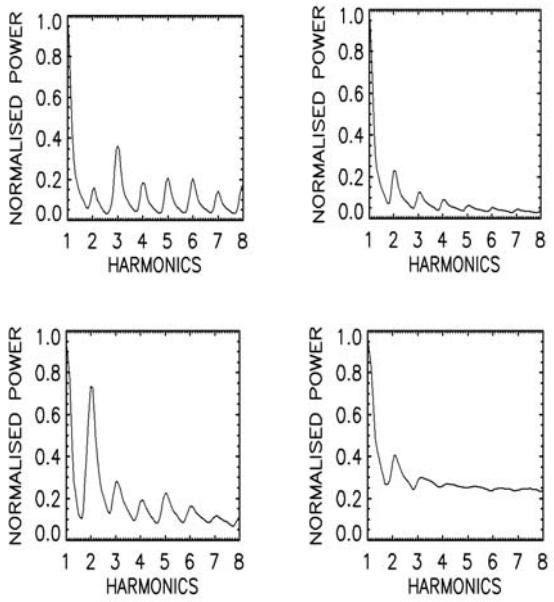


Figure-2
Adult acquired GHD cases FHN spectra; normalized power (ordinate) vs. harmonics (1 to 8). Only Case-R shows a non-exponential fall in spectral envelope due to multiple, high harmonic peaks.

Top Left: R ApEn 0.35 *Top Right*: H ApEn 0.18
Lower Left: W ApEn 0.18 *Lower Right*: S ApEn 0.09

V. Discussion

Three of four adult acquired GHD cases were middle aged. Hence larynx development was completed before onset of GHD. Their low complexity levels are comparable to the minority G2 male normal population. The membership ratio G1:G2 is 2:1 but this is more than reversed at 1:2, with one other case well below G2, in the adult acquired GHD cases and so is likely to be significant despite the small sample of cases. Life style, such as smoking etc., may be a possible cause. However, given the similarity of treatment regime in all three cases, these effects could conceivably have been produced by a reduction in conscious control over fold functionality. Conscious change of control has been demonstrated by Moore et al [9].

The fourth GHD case, R, has high, G1-level complexity and a f_o bordering on female levels. Apart from endocrine replacement, R received no treatment. The onset of acquired GHD occurred after reaching final height but before full adult development in respect of body composition and bone mass. Environment and life-style are unlikely causes for this effect. Hence, perturbation of larynx development due to GHD is a potential explanation for these characteristics.

VI. Conclusion

There is at least initial evidence that adult acquired GHD, occurring between late puberty and adulthood could disrupt larynx development and be detected by EGG complexity analysis. Furthermore, this can be differentiated from post pubertal adult cases of acquired GHD, in which the result of treatment may be reduced conscious control of fold functionality and produce a measured EGG complexity that is comparable to the lowest G2 level of normality or pathologically low.

References

[1] Manickam K, Moore C J, Willard T, Jones S and Slevin N, "Approximate Entropy In The Analysis And Monitoring Of Voice Quality Changes In Larynx Cancer Patients Following Radiotherapy", *Procs IEE Medical Applications of Signal processing*, London, October 2002 . UK ISSN 0963-3308 Ref No 2002/110.

[2] Titze I R, *Principles of Voice Production*, Prentice Hall, 1994

[3] Fourcin A J, "Electrolaryngographic Assessment of Vocal Fold Function", *Journal of Phonetics,* 14, 435-442, 1986.

[4] Priestley M B. *Spectral Analysis and Time Series,* Academic Press 1981.

[5] Rabiner,L.R. et al. ,*Digital Processing of Speech Signals,* Prentice Hall, 1978.

[6] Moore C J, Slevin N and Winstanley S, "Characterising Vowel Phonation By Fundamental Spectral Normalisation of Lx-Waveforms", *Procs Intl Workshop Models & Analysis of Vocal Emissions for Biomedical Appls*, Florence Italy, 1-6, 1999

[7] M. Nirmalan, T. Willard, P.Sinnot, J. E. Eddleston and R. A. Little,"Heart rate variability in patients with severe brain injury", *Br Jnl Anaesth*, 86: 312P, 2001

[8] Pincus S M, "Approximate Entropy as a Measure of System Complexity", *Proc. Natl. Acad. Sci. USA.* March 15; 88 (6): 2297-2301, 1991.

[9] Moore C J and Jones S, "Stimulated production of Vowel-like LX-Waveforms and Spectral Features in the Absence of Phonation", *Med Eng Phys*, 24,461-465, 2002

3rd International Workshop MAVEBA 2003, 259-262

IDENTIFITCATION OF VOICE PATHOLOGY USING AUTOMATED SPEECH ANALYSIS

C. Maguire[1], P. de Chazal[1], R.B. Reilly[1], P.D. Lacy[2]
[1]Department of Electronic and Electrical Engineering, University College Dublin, Ireland
[2]Royal Victoria Eye and Ear Hospital, Dublin, Ireland

Abstract: The classification performance of an automatic classifier of voice pathology for the detection of normal and pathologic voice types is presented. The proposed classification system is non-intrusive and fully automated. Speech files of sustained phonation of the vowel sound /a/ in the 'Disordered Voice Database Model 4337' provided 631 subjects of both genders (58 normal, 573 pathologic). This database includes features extracted by the Multi Dimensional Voice Program (MDVP). Mel frequency cepstral coefficients (MFCC) were extracted for all of the speech files. Discrete Fourier transform (DFT) features, Log DFT and Cepstral features were also extracted. Cross-fold validation was used to measure the classifier performance. Linear discriminant analysis was employed as the classifier model. The MDVP feature set of shimmer and signal-to-noise ratios are shown to have similar classification performance to the Log DFT and the MFCC features.
Keywords: **Voice Pathology, speech analysis, Linear Discriminant Analysis.**

I. INTRODUCTION

A wide variety of vocal fold pathologies are found in patients with vocal disorders. These pathologies can be found in varying degrees of severity and development. They can be classed as physical, neuromuscular, traumatic and psychogenic and all directly affect the quality of the voice. At present a number of diagnostic tools are available to the otolaryngologists and speech pathologists such as videostroboscopy [1] and videokymography. However these current methods are time and personnel intensive and lack objectivity.

Research has been reported on the development of reliable and simple methods to aid in early detection, diagnosis, assessment and treatment of laryngeal disorders. This research has lead to the development of feature extraction from acoustic signals to aid diagnosis. Much focus has been centred on perturbation analysis measures such as jitter and shimmer and on signal-to-noise ratios of voiced speech, which reflect the internal functioning of the voice. Through this research it has been shown that these features can discriminate between normal and pathologic speakers [2], [3], [4], [5].

The aim of this research was to investigate the performance of a voice pathology classifier categorising sustained phonation of the vowel sound /a/ from a large labelled database into either a normal or pathologic class. The goal of this project was to produce a stand-alone classifier that would be non-intrusive and objective.

II. METHODOLOGY

Each stage of the flow chart of a voice pathology classifier in Figure 1 is discussed below.

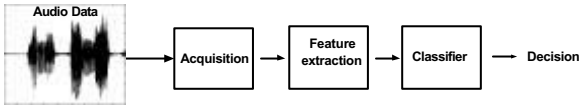


Figure 1. Flow Chart of the Processes involved in a Voice Pathology Classifier

Acquisition: The labelled voice pathology database "Disordered Voice Database Model 4337" [6] acquired at the Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory and distributed by Kay Elemetrics was used in this study. A detailed description of the database can be found at [6], [7].

Digitised voice recordings of the sustained phonation of the vowel sound /a/ were used for training and testing the classifier. The database contains 631 recordings of which gender information is available for 389 recordings. In this study we divided the available data in to three datasets in order to investigate the influence of gender on classification performance:

1. A mixed gender dataset containing 631 subjects (58 normal, 573 pathological)
2. A male dataset containing 164 subjects (22 normal, 142 pathological)
3. A female dataset comprising 225 subjects (36 normal, 189 pathological)

Feature Extraction: The Multi Dimensional Voice Program (MDVP) [8] was used as a feature extractor. For each sustained phonation of the vowel sound /a/ in the database there are 33 associated MDVP features. These 33 features can be divided into six subsets. Each subset is a grouping of features that describe specific properties of the phonation. Namely: 1) the fundamental frequency, F_0 2) jitter (short-term, cycle-to-cycle, perturbation in the fundamental frequency of the voice), 3) shimmer (short-term, cycle-to-cycle, perturbation in the amplitude of the voice), 4) Signal-to-noise ratios, S/N 5) count and 6)

duration features. Some recordings contained missing MDVP feature values. In these cases missing features were replaced by the average value of that MDVP feature. This ensured that the replaced features would not aid in the classification. The histogram was examined for each feature and where appropriate a log transformation was applied. This forced all the features to have an approximately Gaussian distribution.

The Mel Frequency Cepstral Coefficients (MFCC) features are commonly used in Automatic Speech Recognition (ASR) and also Automatic Speaker Recognition systems [9]. These coefficients are formed by taking the Discrete Fourier Transform (DFT) of the speech signal. Then a linearly spaced filterbank in the Mel frequency domain that translates to a log spaced filterbank in the Frequency domain is applied to the spectrum of the signal. The Mel scale is based on the non-linear human perception of sounds. Finally the signal is log transformed and the inverse discrete Fourier transform or the discrete cosine transform is applied.

The MFCC were extracted from the speech signal using the Hidden Markov Model toolkit (HTK) that is commonly used in speech research [10]. A first order pre-emphasis filter using a coefficient of 0.97 was utilised here so that the measured spectrum has a similar dynamic range across the entire frequency band. The signal was then separated into 20ms frames using a Hamming window with an overlap of 10ms between each frame. HTK employs the DCT to transform the outputs of the filterbanks to the cepstral domain. MFCC were calculated for each frame and then averaged across all frames in a recording. Thus each speech recording is represented by the averaged MFCC for that particular speech recording. These averaged MFCC were used as features for the classifier.

The Disordered Voice Database speech files are sampled at two different sampling frequencies 25 or 50 kHz. The location of the filterbank channels used in calculating the MFCC would differ for speech recordings that have different sampling frequencies. In order to standardise the recordings for all subsequent processing all the speech recordings sampled at 50 kHz were downsampled to 25 kHz.

The DFT, Log DFT and Cepstral coefficients were calculated in Matlab by applying similar methods to the speech signal as HTK, i.e. a pre-emphasis filter using a coefficient of 0.97 and segmenting the speech signal into 20ms frames using a Hamming window with an overlap of 10ms between each frame. The Cepstral coefficients were calculated in the same way as the MFCC except that the filter bank is not applied to the signal. For the DFT, Log DFT, Cepstral coefficients and MDVP each set of features were divided into subsets in order to investigate the system performance using subsets of features. Different MFCC feature sets were extracted from the speech recordings with a varying number of filter channels and also a varying number of MFCC.

Classifier: Linear discriminants (LD) [11] partition the feature space into the different classes using a set of hyper-planes. The parameters of this classifier model were fitted to the available training data by using the method of maximum likelihood. Using this method the calculation required for training is achieved by direct calculation and is extremely fast relative to other classifier building techniques such as neural networks. This model assumes that the feature data has a Gaussian distribution for each class. In response to input features, linear discriminants provide a probability estimate of each class. The final classification is obtained by choosing the class with the highest probability estimate.

Estimating the classifier performance: The cross-validation scheme [12] was used for estimating the classifier performance. The variance of the performance estimates was decreased by averaging results from multiple runs of cross validation where a different random split of the training data into folds is used for each run. In this study ten repetitions of ten-fold cross-validation were used to estimate classifier performance figures. For each run of cross fold validation the total normal population and a randomly selected group of abnormals equal in size to the normal population was utilised. This results in a more realistic reflection of the predictive ability of the system.

In this study the performance of the classifier is quoted using the class sensitivities, predictivities and the overall accuracy. The sensitivity of the classifier to a particular voice class is the fraction of speech files in the class that are correctly classified. The specificity is the sensitivity calculation applied to the normal class. The positive/negative predictivity is the fraction of speech files detected as abnormal/normal that are correctly classified. The overall accuracy is the fraction of the total number of subjects' voices that are classified correctly.

III. Results

All MDVP features were log-transformed so that the resulting histograms more closely approximated Gaussian distributions. Classification results were obtained for the MDVP, MFCC, DFT, Log DFT and Cepstral features as well as the combination of these features for mixed genders together and for each gender individually. The number of filterbank channels and coefficients used in the MFCC was examined. Through testing it was seen that utilisation of 15 filterbank channels and 15 coefficients resulted in satisfactory system performance.

Table 1: Classification results for feature sets that were examined. The test set accuracy, mean specificity, mean sensitivity, mean negative predictivity and mean positive predictivity cases are shown for the mixed gender classifier while only test set accuracy is shown for the male and female classifiers

Feature set	Gender	Test set (%) Acc	Sens	Spec	P.Pred	N.Pred	Gender	Test set (%) Acc	Gender	Test set (%) Acc
MDVP (F_0, Jitter, Shimmer, S/N)	Mixed	84.74	84.83	84.64	84.83	84.64	Male	75.97	Female	78.14
MDVP (Shimmer, S/N)	Mixed	87.16	83.28	80.8	81.45	82.68	Male	90.61	Female	82.51
MFCC (1:15)	Mixed	82.65	83.62	81.68	82.2	83.13	Male	88.4	Female	73.95
MFCC (1:5)	Mixed	83.35	83.45	83.25	83.45	83.25	Male	88.4	Female	76.32
DFT Magn (1:8)	Mixed	81.18	87.59	74.69	77.79	85.6	Male	83.15	Female	79.42
Log DFT (1:8)	Mixed	81.53	83.1	79.93	80.74	82.37	Male	84.81	Female	81.79
Cepstrum (1:8)	Mixed	77.19	74.66	79.76	78.87	75.66	Male	80.11	Female	71.95
MDVP (F_0, Jitter, Shimmer, S/N) **&** MFCC (1:15)	Mixed	85.69	86.03	85.34	85.59	85.79	Male	75.41	Female	67.76
Log DFT (1:8) **&** MDVP (Shimmer, S/N)	Mixed	88.55	88.28	88.83	88.89	88.21	Male	83.98	Female	81.42
Log DFT (1:8) **&** MFCC (1:5)	Mixed	85.86	86.55	85.17	85.52	86.22	Male	86.74	Female	78.14
DFT Magn (1:8) **&** MFCC (1:5)	Mixed	84.82	87.41	82.2	83.25	86.58	Male	84.81	Female	76.87
Cepstrum (1:8) **&** MFCC (1:5)	Mixed	82.57	82.59	82.55	82.73	82.4	Male	81.49	Female	75.77

The duration features of the MDVP were not included as intuitively there was no link between the duration of the recording and any pathology. The predictive ability of the count features was found to be poor and so this group was disregarded for the rest of the study. The classification performance of different feature sets is shown in Table 1. The feature set of shimmer and signal-to-noise ratio combined gives the highest classification performance among the MDVP feature subsets. The DFT magnitude, Log DFT and Cepstral coefficients achieve optimal classification performance via the first eight coefficients. In the frequency domain this corresponds to frequencies between 0 and 385 Hz.

IV. Discussion

The MDVP feature set performs well for the mixed gender classifier achieving a classification accuracy of 84.74%. However, its performance falls off when utilised in the individual gender classifiers, 75.97% and 78.14% respectively. The reduced set of MDVP features using shimmer and signal-to-noise ratios performs at a much more consistent level though all of the different gender classifiers with an accuracy of 87.10%, 90.61% and 82.51% respectively.

The reason why only the first eight coefficients are significant for the DFT, Log DFT and cepstral coefficients is due to the fact that it is a vowel sound /a/ that is being analysed and hence most of the fundamental frequency content will be contained in the lower frequencies. Utilisation of the DFT magnitude and Log DFT features with all three gender classification systems achieve consistently high results of 81.18, 83.15, 79.42% and 81.53, 84.81, 81.79% respectively.

The Cepstral feature set did not perform as well as the MFCC feature set resulting in an accuracy of 77.19%, 80.11% and 71.95% for the mixed, male and female gender classifiers. This illustrates that by incorporating the human auditory system's non-linear perception of the audio spectrum through application of the Mel scale improves the performance of the system.

Through the use of the first five MFCC it is possible to achieve the same classification rates as achieved using all 15 MFCC. This trend is consistent with research reported by [13] where the authors observed that only the first few MFCC were required for automatic speaker recognition systems. The test set accuracies for the system employing the MFCC perform well in the mixed gender and male gender classifiers, 82.65 and 88.40%, but the accuracy was lower for the female speech recordings, 73.95%. The MFCC are based on homomorphic analysis whose function is to deconvolute the speech signal, i.e. to separate the excitation and impulse response of a linear time-invariant system. The coefficients at the beginning of the MFCC and Cepstrum represent the impulse response of a linear system that combines the effects of the glottal wave shape, the vocal tract impulse response and the radiation impulse response [14]. For this reason these features should yield information about the health

of a person's vocal system. The DFT magnitude and Log DFT features contain information about the source and vocal tract simultaneously.

Various combinations of the feature sets were examined however we observed that the systems performance was not improved significantly.
A number of research groups [15], [16], [17] have reported results for detection rates for voice pathologies of 94.87%, 76% and 96.30% respectively. In [15] the Disordered voice database was employed and their results may be compared with the results obtained in this study. However results from [15] should be considered biased as the authors used the MDVP speech recording duration features "SEG" and "PER". In the database the normal recordings are three times longer in duration than the pathologic recordings and therefore the "SEG" and "PER" features are three times as large for normal recordings than for pathologic recordings. Hence the features based on the recording duration could be used to distinguish the normals from pathological cases with high success due to the different durations of normal and pathologic recordings.

In study [16] different databases were used and a direct comparison of results cannot be made. The database used in the present study provides a large amount of pathologic subjects that might not fairly represent the pathologies present in other studies conducted in this area or those encountered by the medical profession on a day to day basis. The predictive ability of this model could be confirmed through external validity. The latter study [17] utilises similar features to the ones used in this study however their classification performances were based on correct classification of individual frames from the speech files which implies that the training data used would consist of data very similar to the testing data.

V. CONCLUSION

The MDVP feature set containing the shimmer and signal-to-noise features offers the best classification results over each of the gender classifiers. The utilisation of the Log DFT and MFCC feature set in the classification system performs almost as well as the MDVP features. However the Log DFT and MFCC features are implemented with very little computational cost in comparison to the MDVP features.

In this study, the performance of the mixed-gender classifiers was similar to the classification performance of the single-gender classifiers. These results suggest that for this particular automatic dassification system there is no advantage to be gained by utilising single-gender classifiers to detect pathologic voice.

The Support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

REFERENCES

[1] B. Schneider, J. Wendler and W. Seidner "The relevance of stroboscopy in functional dysphonias", *Folia Phon.*, Vol 54, No. 1, pp 44-54, 2002

[2] P. Liebermann "Perturbations in vocal pitch" *J. Acoust. Soc. Am* , Vol. 33, No. 5, pp 597-603, 1961

[3] I.R. Titze, "Workshop on Acoustic Voice Analysis", National Centre for Voice and Speech, America, 1994

[4] G. de Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments", *J. Speech. Hear. Res*., Vol. 38, pp 794-811, 1995

[5] D. Michaelis, M. Frohlich, H.W. Strube, "Selection and combination of acoustic features for the description of pathologic voices", *J. Acoust. Soc. Am*., Vol. 103, No. 3, pp 1628-1639, 1998

[6]"Disorder Voice Database Model 4337" Massachusetts Eye and Ear Infirmary Voice and Speech Lab, Boston, MA, Jan. 1994. Kay Elemetrics Corporation.

[7] C. Maguire, P. de Chazal, R.B. Reilly, P. Lacy "Automatic Classification of voice pathology using speech analysis", *World Congress on Biomedical Engineering and Medical Physics* , Sydney, August 2003.

[8] "Multi Dimensional Voice Program" Kay Elemetrics Corporation.

[9] J.P. Campbell, "Speaker Identification: A tutorial", *Proc. of the IEEE*, Vol. 85, No. 9, pp 1437-1462, 1997.

[10] S.J. Young, "The HTK HMM toolkit: Design and philosophy", Cambridge Univ. Eng. Dept. Tech. Rpt. CUED/F-INFENG/TR.152, 1993.

[11] R. O. Duda, P. E. Hart, and H. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, NY, 2000.

[12] R. Kohavi, "A study of cross validation and bootstrap for accuracy estimation and model selection," *Proc. 14^h Int. Conf on Art. Intel.*, pp. 1137-1143, 1995

[13] H. Gish, "Text Independent Speaker Identification", *IEEE Signal Processing Magazine*, Vol. 11, No. 4, pp 18-32, 1994.

[14] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Sig*nals. Englewood Cliffs, NJ, 1978.

[15] G. Llorente, S. Navarro et al, "On The Selection of Meaningful Speech Parameters Used Pathologic/Nonpathologic Voice Register Classifier", *Eurospeech '99*, Volume 1, Page 563-566, 1997

[16] D. G. Childers, "Detection of Laryngeal Function using Speech and Electrographic Data" *IEEE Transactions on Biomedical Engineering*, Vol. 39, No. 1, pp 19-25, JAN 1992

[17] M. E. Cesar, R. L. Hugo, "Acoustic Analysis of Speech for Detection of Laryngeal Pathologies", Proc. 22nd Annual EMBS Int. Conf., pp 2369-2372, July 2000.

Voice analysis

3rd International Workshop MAVEBA 2003, 265-268

COMPARISON OF OBJECTIVE AND SUBJECTIVE CLASSIFICATION OF UNVOICED STOP CONSONANTS IN STOP-VOWEL SYLLABLES

T. Hirvonen[1], U. K. Laine [1]
[1]Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland

Abstract: **The objective and subjective classification of unvoiced stop consonants in varying vowel contexts were studied. The objective classification was based on auditory feature vectors obtained by warped linear prediction (WLP) and vector autoregressive (VAR) models for parameter trajectories. In the case of known vowel the unvoiced consonants were classified 98-100% correctly based on the auditory spectral features of the bursts whereas the VAR models for the parameter (formant) trajectories gave at best only 52-68% correct classification. The importance of the burst part also in the human perception was confirmed by a listening test.**
Keywords: **Speech, syllables, classification**

I. INTRODUCTION

The unvoiced stop consonants /k, p, t/ are clearly the most difficult sounds for a phonemic speech recognizer. The developed continuous speech triphone based speaker dependent HMM recognizer for Finnish produced the largest errors (about 25-30%) in recognition of unvoiced stops whereas the error rate for most phonemes was only 1-3% [1].

The human listener utilizes three different features in unvoiced stop-vowel syllable recognition: 1. The spectral structure of the burst. 2. The voice onset time (VOT). 3. The formant transitions. According to earlier studies, the priority of these three factors in the human perception corresponds the list above, the spectral structure of the burst being the most important cue [2].

The aim of this study is to objectively evaluate the importance of the burst spectral structure in comparison with the formant transitions of the voiced part of the syllable in /k, p, t/ classification. An additional goal was to compare objective results to subjective ones via a simple listening test. Samples of the same speech material were used in both classification tasks. The formant transitions were not explicitly modeled, but rather indirectly represented through an autoregressive prediction matrix. The study was limited to combinations of stops /k, p, t/ and four Finnish vowels /a, e, i, u/.

The study shows that with a proper design of the classifier, close to 100\% performance can be reached. This is comparable to the human perceptual ability. However, this result may occur only when the context, i.e., the vowel part of the syllable, is first correctly recognized.

II. MEATERIALS AND METHODS

A. Speech Material

The speech material used in this research consisted from 80 sentences in Finnish that were spoken by one male person. The sentences were in wav-format with a 22050 Hz sampling frequency. The material had been manually segmented to phonemic units. This segment information was used as a basis for these tests.

A total of 12 different stop-vowel syllable types shown in Table 1 were used for the objective classification. The material consisted of different amounts of different syllables since it was decided that all possible instances of each syllable from the original sentences should be included. Table 1 shows the 12 syllable classes and their corresponding number of occurrences.

Table 1: *The 12 stop-vowel syllables used for the objective classification and their corresponding quantities.*

syllable	quantity	syllable	quantity
/ka/	30	/ku/	9
/pa/	5	/pu/	11
/ta/	22	/tu/	18
/ki/	17	/ke/	12
/pi/	7	/pe/	8
/ti/	19	/te/	7

B. VQ Auditory Feature Vectors

The feature vectors used for the objective classification were obtained with warped linear prediction (WLP) [3]. WarpTB Matlab toolbox [4], along with some custom functions was used for samples processing.

The feature vectors were constructed by obtaining 12^{th} order warped linear prediction coefficients calculated from the time-domain signal. The warping factor was 0.676. The WLP vectors were calculated using 16-ms frames per one vector. The frame window hop was 1 ms. The WLP coefficients were further transformed into line

spectral frequencies (LSF). Thus each segment was described by a (12*60)-matrix.

Fig. 1 shows an example of a segment used in these tests. The upper picture illustrates the time-domain signal. A spectrogram of the signal is shown in the middle picture. Finally, the 60 LSF vectors transformed from the WLP coefficients are drawn in the undermost picture. It can be seen that the LFS vectors follow the structure of the spectrogram.

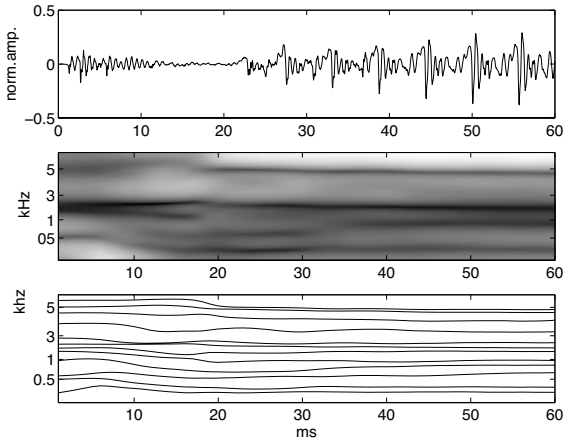


Figure 1: *Three representations of one /ka/ - segment used in the classification: the time-domain signal, the spectrogram and the LSF vectors transformed from the WLP coefficients (from up to down).*

C. Vector Auto Regression (VAR)

Vector auto regression is a method much similar to conventional linear prediction, with the difference that the prediction is done for vectors instead of scalars. A mathematical description of the method can be found in [5].

Fig. 2 shows a 1st order VAR procedure applied to a similar speech segment as seen in Fig. 1. A VAR model was calculated from the whole segment and then used to produce the prediction vectors. The model requires an initial state which has been chosen from the middle of the original segment (x = 0 ms). The prediction has been continued ``past'' the original segment limit (x > 30 ms). It can be seen that the prediction vectors model the original curves fairly well. In this study, the VAR models of the feature vectors are used to parameterize the trajectories in the feature space. The trajectories reflect the formant transitions caused by articulatory movements.

III. Classification by Trajectories

A 1ST order VAR prediction matrix was calculated for the LSF feature vectors of each speech segment.

Classification was based on comparing the eigenvectors of the VAR matrices. A descriptive vector **u** was calculated for each prediction matrix according to (1).

$$\mathbf{u} = (\mathbf{c};\mathbf{v1};\mathbf{v2}) \quad (1)$$

, where **v1** and **v2** are the eigenvectors associated with the two largest eigenvalues of the corresponding prediction matrix. Here the variable **c** is the scaling vector associated with the VAR prediction matrix [5]. A few other descriptive vectors were also tested but the above method produced the best results.

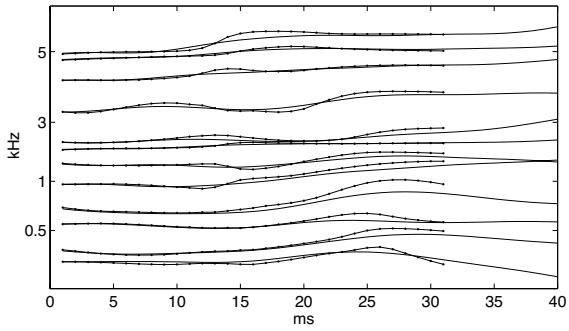


Figure 2: *Prediction of feature vectors with a VAR model: original vectors (thick line) and the predicted vectors (thin line).*

The actual classification was done for each of the four vowels separately so that there were three possible alternatives to which a syllable could be classified, according to the consonant. A median vector for each of the three consonant classes was calculated for each vowel. The Euclidean distance between the descriptive vector **u** of each segment and the median vectors of each of the three classes was calculated. The segment was assigned to the class to which this distance was the smallest.

Results for the classification based on trajectories represented by the prediction matrices are shown in Fig. 3. The x-axis indicates value the amount of feature vectors excluded from the beginning of the segment when calculating the VAR models. It can be seen that the classification percentage does not rise above 70%.

Other schemes for the construction of the VAR model were also tried. In addition to the previous method, feature vectors were also removed from the end of the segments before the modeling procedure. Also, both of these schemes were combined in various ways. The overall classification percentage did not increase as a result of these experiments.

IV. Classification by Brurst Structure

The classification in the previous section was based on the comparison of the VAR prediction matrices. The method modeled mainly the formant transition structure of the speech segments. In this section, the temporal spectral structure classification based on the optimal time window is investigated as well.

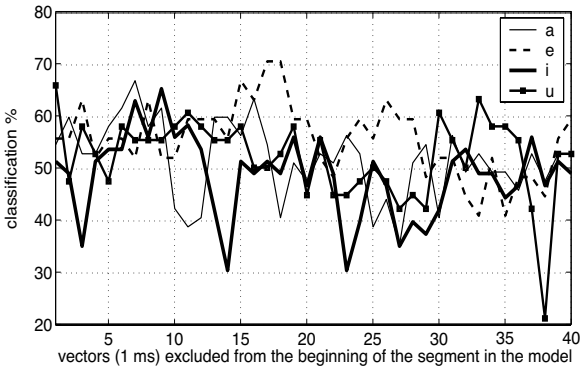


Figure 3: *Classification percentage based on VAR modeling as a function of feature vectors included to the model.*

The method used for the classification was to compare the auditory feature vectors (see Section 2.2) directly in terms of Euclidean distances. For each ($12*60$)-feature vector matrix representing a given speech segment, a mean of three adjacent vectors, starting from the beginning of the segment, was taken as a basis for the classification. The mean three vectors was obtained for all cases in the same syllable class and a median vector of these cases was calculated. This vector represented the average feature vector of a syllable class at given point of the segment.

As in the previous section, the actual classification was done for each of the four vowels separately, based on the Euclidean distance between the mean of three feature vectors of each segment and the median vectors of each of the three classes.

Each speech segment used in the classification was the same length, i.e. 60 ms starting from the beginning of the consonant burst. The previous procedure was repeated within the area of 1 - 50 ms from the beginning of the segments with a 1 ms hop. In this way, the optimal segment point where the classification yielded the best results could be found.

Fig. 4 illustrates the results of the feature vector-based classification. Location at x = 0 ms represents the beginning of the segment and the classification percentage is calculated at 1 ms intervals onward to the end of the segment. The best classification percentage is achieved in all four cases by comparing the feature vectors within the first 10 ms of the segments. By investigating for example the /ka/-segment in Fig. 1 it can be seen that the burst part of the syllable exceeds over this limit. The situation was similar to other segments as well.

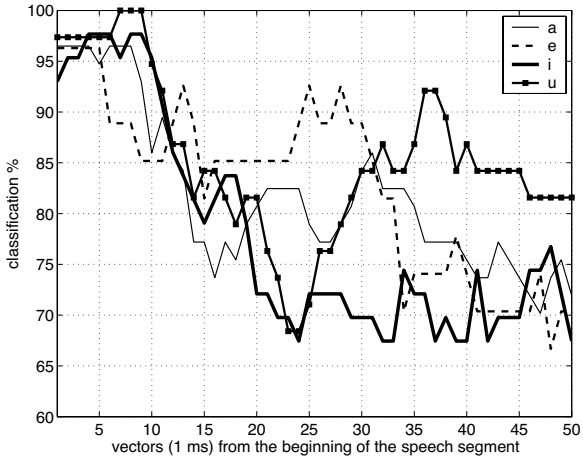


Figure 4: *Classification percentage based on feature vectors as a function of the location of the vectors in the speech segments.*

The overall classification percentage when investigating the feature vectors of the burst part was higher than in the previous section where the classification was based on VAR models. For this reason, the structure of the burst is determined to be more important classification cue than the formant transitions between the consonant and the vowel.

V. Listening Test

The results from the previous experiments were compared with those of a subjective listening test. The purpose was to establish which cue is more important in the subjective classification of unvoiced stop-vowel syllables; the burst spectral structure with right VOT or the formant transitions of the voiced part.

A. Method

The samples for the listening test were constructed by taking one speech sample from each of the 12 classes given in Table 1 and portioning the 12 segments to burst and voiced parts. For each vowel, there were three burst parts and three voiced parts. These were then combined to form a total of nine synthetic speech syllables per vowel. The test thus included 36 samples, i.e. three of each syllables shown in Table 1.

The voiced part of the samples was segmented so that it included 25 ms from the beginning of the first clear

pitch period, as well as 10 ms before this point. The average length of this part was 34 ms. The burst part started from the beginning of the burst and ended to the point overlapping 5 ms with the voiced part. The samples for the listening test were constructed by linearly cross-fading the burst and the vowel over this 5 ms part.

This method produced samples where the burst parts were combined with the voiced parts that included the formant transition information. If the synthetic samples could be correctly recognized by the subjects, the importance of the burst part compared to formant transitions as a classification cue would be established.

B. Test Procedure

The test was done in a quiet listening room, whose specifications can be found in [6]. The sound reproduction device was a pair of Sennheiser HD600 headphones. The subjects classified the samples with a graphical test interface. The task presented was to choose the stop consonant for all samples from the three options (/k, /p// or /t/). The subjects could listen the samples as many times as they wanted. A total of 11 subjects participated to the test.

B. Results

The overall classification percentage was 98.48 for the 11 subjects. In six cases, the syllable /ti/ was classified as /ki/. Each of the three different /ti/-samples were classified as being /ki/ two times. The subjects classified the samples perfectly in all other cases.

The results from the listening test indicate that the burst part of a stop-consonant syllable could be replaced with another burst part so that the subjects could distinguish the synthetic syllables correctly. Thus the importance of the burst part is emphasized.

VI. DISCUSSION

This study confirms that the primary strategy of the human perception in classification of unvoiced stop-vowel syllables relates to the spectral information of the burst part. The burst section was found to carry the most important cues necessary for the classification task by objective studies as well. However, the pronunciation of the syllables may vary and in some contexts the burst part may almost be missing. In these cases the time-frequency structure of the voiced part is the only a short-term cue. Thus the human perceptual system may utilize this time-frequency structure especially in noisy conditions and in such way increase the robustness. In the human perception the language model also has an important role when meaningful words are produced.

The outcomes of the two objective classification schemes, one based on the instantaneous spectral features and the second on the trajectory modeling, gave results which may need some further studies. The method based on the spectral features gives 68-93\% right classifications over the voiced parts whereas the prediction model results to only 50\% on the average, even though it is capable of combining and predicting the same features. This may be related to the problem on how to find the optimal time window position and size for the VAR model. Another problem is to find the most optimal metrics for the VAR model comparison.

Theoretically, the classifier based on spectral feature vectors is able to give very close to 100\% correct classification when the individual classifications are combined in time. This can be done by statistical models for the chains of the feature vectors. However, it has to be remembered that this is true only when the vowel is first correctly classified. In other words, the classification of /k, p, t/ depends strongly on the right classification of the following vowel. Thus the burst parts of these stop consonants are strongly context dependent.

VII. CONCLUSION

Our study on the objective and subjective classification of stop-vowel syllables showed that the human perception utilizes effectively the most important objective spectral information of the syllable located in the burst part. Thus an optimal objective classifier can be constructed based on the spectral features of the bursts. When the vowel context is known it is possible to reach close to 100\% right classification in the cases where the burst energy is high enough to allow for the bursts to be detected.

REFERENCES

[1] M. Ursin, *Triphone Clustering in Finnish Continuous Speech Recognition.* M.Sc. thesis. Helsinki University of Technology, 2002.

[2] D. Kewley-Port, "Measurement of Formant Transitions in Naturally Produced Stop Consonant-Vowel Syllables", *J. Acoust. Soc. Amer.*, Vol. 72(2), pp. 379-389. 1982.

[3] A. Härmä, M. Karjalainen, L. Savioja, V. Välimaki, U. K. Laine, and J. Huopaniemi, "Frequency-Warped Signal Processing for Audio Applications", *J. Aud. Eng. Soc*., November 2000.

[4] A. Härmä and M. Karjalainen, "WarpTB – Matlab Toolbox for Warped DSP", *http://www.acoustics.hut.fi/software/warp/*

[5] J. D. Hamilton, *Time Series Analysis, Princeton University Press*, pp. 257-291, 1994.

[6] A. Järvinen, *Kuunteluhuoneen Suunnittelu ja Mallinnus*, M.Sc. thesis, Helsinki University of Technology, 1999.

3rd International Workshop MAVEBA 2003, 269-272

APPLICATION OF ACOUSTIC ANALYSIS OF SPEECH SIGNAL FOR EVALUATION OF INTUBATION-RELATED DAMAGES OF THE SPEECH ORGAN

Joanna Szaleniec[1], Maciej Modrzejewski[1], Wieslaw Wszolek[2]

[1] Chair of Otolaryngology, Collegium Medicum UJ, 31-501 Krakow, Sniadeckich 2 asiat@agh.edu.pl
[2]AGH University of Science and Technology, 30-059 Krakow, Mickiewicza 30 wwszolek@agh.edu.pl

Abstract: **Endotracheal intubation is a method commonly applied nowadays in medicine, particularly in surgical procedures carried out under general anaesthetic. It is however an invasive method which may result in many complications, including mechanical injuries of the larynx [1]. The problem is not of purely medical nature. However from the research point of view the detailed analysis of vocal folds damages is necessary. In the present work the attention is mainly focused on the prospects of application of a dedicated acoustical analysis of the speech signal, based on professional methods of signal processing. In the field considered in the present work the objectives of the signal processing and classification are different from the usual ones (revealing the origins of the deformation and evaluation of the signal deformation level in relation to the standard. The acoustic and phonetic properties of the signal itself are essentially different from the widely known parameters of correct speech.**
Keywords: **speech analysis, pathological speech, surgical treatment, speech processing**

I. INTRODUCTION

In many problems of medical diagnosis, as well as in planning and monitoring of the therapy and rehabilitation of vocal organs, the evaluation of quality of the deformed speech signal is very important. The intubation-related damages of vocal folds are not a medical problem of great importance. However if the problem is occasionally observed it should be thoroughly investigated. Fortunately the number of observed cases is not very high, and the scale of the injuries is not very extensive. What's more rather fast and easy recovery process can be achieved, as a result of natural regeneration processes or intentional rehabilitation procedure. Still the research analysis of the intubation-related vocal folds injuries seems necessary, especially, when taking into account the fact, that the progress and achievements of anesthesia and surgery have lead to a situation, when more and more often the surgical procedure under general anesthetic becomes a therapy of choice, thus affecting greater and greater number of patients. Therefore the situation requires solution of a nontrivial task of elaboration of examination methodology able to reveal the intubations-related injuries of larynx (in particular vocal chords). The examination should allow the evaluation of the extent of the revealed injury, monitoring of the rehabilitation process in the cases of intubation-related injuries and possible supervision of the whole process in situations when phoniatric intervention is required. The examination can be also used as a basis for objective evaluation of risk factors, related to occurrences of intubation-related larynx injuries.
Therefore the in the present work attention has been focused on the possibilities created by a properly dedicated acoustic analysis of speech signal in the field related to the discussed problem.

II. RESEARCH MATERIAL AND METHOD

The studies of speech clarity have been carried out for patient surgically treated in the Otolaryngology Clinic of CM UJ, Cracow, after various types of operations not related to the vocal tract.
In the preliminary stage a group of 24 patients have been examined. The registration of acoustic signal has been carried out in an anechoic chamber, where a digital magnetic recorder has been used for registration of time dependencies of the acoustic pressure during the test utterance. The study was of prospective nature. The patient's voice was registered twice: before the operational treatment connected with the intubation and approx. 24 hours after the treatment. For some patients who were ready to co-operate an additional voice registration has been carried out during a check-up examination several months later and thus a reference material has been obtained, showing the effects of long-term rehabilitation.
From the point of view assumed in the present work it is particularly important that the available computer programs dedicated to analysis and processing of sound signals are able to extract and objectively evaluate even very subtle changes in the structure of the sounds examined. It is of critical importance, because the changes in the speech signal, observed as an aftermath of possible larynx injury during the intubation procedure, are minute and they are usually hardly detectable even for a experienced ear. This is, by the way, one of the reasons that up-to-date those changes have not become a subject of any extensive studies.
Table 1 presents detailed information about the examined patients.

Table 1. Detailed information about the examined group

Patient	Gender	Age	Diagnosis
1	Male	50	Tonsillitis chronica
2	Male	21	Cystis colli lateralis
3	Male	47	Polypi nasi
4	Female	56	Tumor nasi
5	Female	52	Polypi nasi
6	Male	19	Otitis media chronica
7	Female	56	Sinusitis maxillaris chronica
8	Female	21	Tonisillitis chronica
9	Male	54	Polypi nasi
10	Male	68	Polypi nasi
11	Female	46	Pansinusitis chronica
12	Female	47	Pansinusitis chronica
13	Male	36	Sinusitis chronica
14	Female	49	Sinusitis frontalis
15	Male	64	Polypi nasi
16	Female	10	Otitis media chronica
17	Female	9	Tumor orbitae
18	Male	49	Ca. baso. regionis retroauricularis
19	Female	59	Polypi nasi
20	Female	48	Polypi nasi
21	Male	44	Otitis media chronica
22	Male	33	Cystis sinus maxillaris
23	Male	17	Otitis media chronica
24	Male	45	Tumor glandulae parotis

The study has been based on the analysis of changes in signal characteristics observed in quasi-stationary states, because the main subject of the study was the functioning of the laryngeal pitch generator, not possible disfunctions of articulation organs. Therefore from the continuous speech signal, registered in the anechoic chamber conditions for all the patients before and after the operation (connected with the intubation) and for some patients also during a check-up examination after a long-term rehabilitation process, fragments containing quasi-stationary vowel sound have been extracted using computer procedures. That element of the applied research methodology has probably removed from the acoustic research material some potentially valuable information (related to the articulation of all transients, which can be also affected by the intubation procedure), however such a restriction of the analysis has considerably simplified the applied research techniques and resulted in better, clearer interpretation of the obtained results. The general scheme of data processing and analysis used in the research described below is shown on Fig. 1. The acoustic files obtained in this way were analysed with the Voice Analysis and Screening System (VASS) 3.0 [4]. The parameters measured in VASS which were taken into consideration in the research were:

- Pitch Perturbation Quotient (**jitter**) and Amplitude Perturbation Quotient (**shimmer**)
- TNI - Turbulence Noise Index
- HNR - ratio of harmonic energy to noise energy,
- NNE - Normalized Noise Energy
- HFHE - Normalized First Harmonic Energy - ratio of the amplitude of the first harmonic from the power spectrum to the total energy [8].

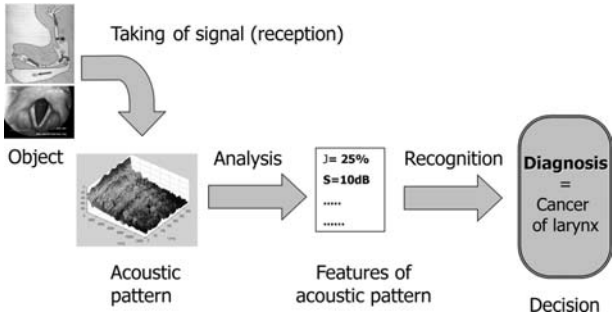


Fig.1. General scheme of data processing

The described method of acoustic evaluation of intubation-related changes, occurring in the larynx as a result of the intubation procedure, based on the acoustic analysis of the speech signal exhibits a number of advantages. Its main advantage is the fact that it is completely non-invasive and can be easily applied. Further advantage of the method is the simplicity of the required calculations, resulting from the fact that the presented parameters of the acoustic speech signal are among the signal parameters that are the easiest for evaluation.

III. RESULTS OF THE STUDY

It has been shown that after the intubation considerable disturbance of vocal chords' functioning occurs, manifested mainly by the changes in the relative energy of the first harmonic frequency of the laryngeal pitch. For some of the patients examined the considered parameter noticeably increases, what can be a direct indication, that as a result of mechanical stretching of vocal chords during the intubation procedure, temporary injury of the chords occurs, manifested mainly by decrease in their elasticity. However it seems reasonable to presume, that the examined parameters of the voice signal (relative energy of the first harmonic frequency, jitter and shimmer) exhibit natural variability specific for a given person, resulting from that fact that no man is able to speak exactly in the same way during two consecutive recording sessions carried out with time separation of several days. The range of voice changeability in the control group for jitter and NFHE in each vowel is presented in table 2.

Table 2. Voice changeability range for jitter and normalized first harmonic energy (NFHE) in the control group.

	mean	standard deviation	Physiological range of voice changeability
jitter /a/	0,0068	0,5591	-1,1114; 1,125
NFHE /a/	0,5572	1,7632	-2,9692; 4,0836
jitter /e/	0,5539	0,7018	-0,8497; 1,9575
NFHE /e/	1,3279	2,4888	-3,6496; 6,3055
jitter /i/	0,0401	0,5881	-1,1361; 1,22
NFHE /i/	1,0806	1,7212	-2,3618; 4,5231
jitter /u/	0,4113	0,8699	-1,3285; 2,151
NFHE /u/	0,9771	2,65158	-4,3260; 6,2803

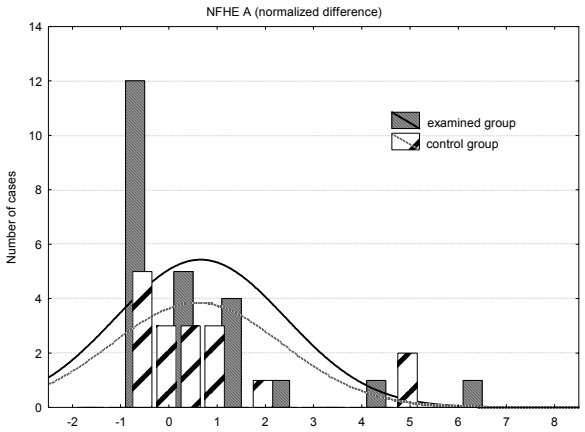
After analysis of all results we found 8 patients in the examined group that presented values of the normalized difference in jitter or NFHE that lay beyond the calculated physiological range (the values are shown in table 3). The deviations were present only in some parameters (jitter /a/, /e/, /i/, NFHE /a/, /i/, /u/) and only in one or two vowels for each patient.

Table3. Normalized difference between the postintubation and preintubation recordings that go beyond the physiological range.

Patient number	parameter	vowel	normalized difference
1	jitter	/a/	3,8346
	jitter	/i/	1,2490
2	NFHE	/a/	6, 6439
3	NFHE	/a/	2,8158
	jitter	/i/	2,2870
4	NFHE	/a/	1,9356
	NFHE	/i/	16,8855
5	jitter	/e/	4,5930
6	jitter	/i/	1,6160
7	NFHE	/i/	5,7044
8	NFHE	/i/	12,2911
	NFHE	/u/	55,6839

The figures 2. and 3, present histograms of the normalized difference for example parameters in which deviations have been observed.

As can be seen in the pictures 2 - 3, most of the patients present minor postintubation voice changes similar to those observed in the control group. In several patients however larger changes occur. In all the cases that lay beyond the physiological range the value of the normalized difference is positive, while for the other patients (and the control group) both positive and negative outcomes are observed. In addition, for one parameter (jitter /i/) a statistically significant difference between the normalized difference distribution between



the examined group and the control group was observed (unpaired one-side t - Student test, $p < 0,05$).

Fig. 2. Histograms of the normalized difference for jitter /a/ in the examined group and in the control group.

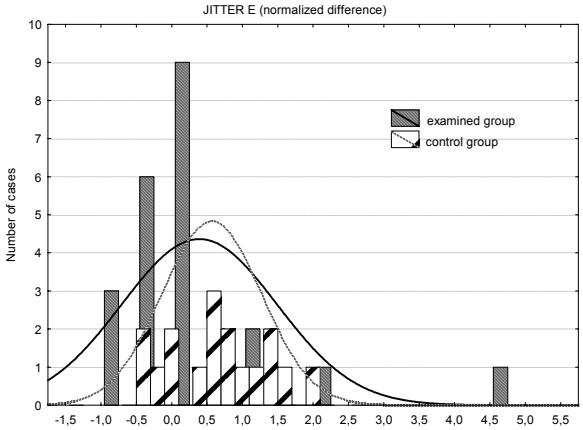


Fig.3. Histograms of the normalized difference for NFHE /a/ in the examined group and in the control group.

IV. Discussion

The presented results are preliminary ones and their full interpretation is not known yet. It can be assumed that no statistically significant difference between the preintubation and postintubation recording in the whole examined group was observed because the majority of the patients did not suffer intubation-related vocal fold damage. In those patients only minor positive and negative voice changes that lay in the physiological range were present. The changes may be attributed to atmospheric conditions, differences in voice effort before the recording, imperfectly identical pronunciation of the sustained vowels, etc. as it was observed in the control group.

Eight intubated patients presented much greater positive changes in jitter or normalized first harmonic energy that

the control group and the other patients. Currently it cannot be stated unambiguously whether the observed deviations are connected with intubation. It must be noted that no patient presented deviations in all vowels. Moreover, the deviations do not always occur in the same parameters and the same vowels. Therefore it is not possible to decide which of the patients actually suffered vocal fold damage during intubation or which of the parameters really reflects the intubation-related trauma.

V. CONCLUSIONS

In approximately 30% of patients changes in parameters: jitter /a/, jitter /e/, jitter /i/, NFHE /a/, NFHE /i/ and NFHE /u/ laying beyond the natural voice changeability range were observed. The distribution of the normalized difference of jitter for the vowel /i/ was significantly different in the examined group than in the control group.

The parameters presented above may be useful in identification and assessment of the of the intubation-related laryngeal trauma, but their usefullness in practice requires further verification.

Further research plans

Because the percentage of the patients who suffer vocal fold trauma during intubation seems to be low, the number of patients with postintubation complications in the examined group may be to small to draw correct conclusions.

The verification of the preliminary results requires:

- continuation of the research on a larger examined group and control group,
- taking into consideration patients after longer intubations (in the currently examined cases the intubation lasted approximately 30 minutes to 3 hours) which might cause more significant voice changes in a higher percentage of patients,
- a more accurate way of choosing the most stationary portion of the sustained vowel, e.g. according to the method proposed by Prosek [9],
- research on the influence of intubation on other acoustical measures (e.g. sound pressure level perturbation quotient [10]),
- a reference examination (preliminary identification of patients with more probable vocal fold trauma by a professional listener).

The further research goals are:

- choice of the most reliable parameter which reflects the intubation-related laryngeal trauma,
- choice of a phoneme in which the acoustic effect of the intubation trauma is most distinct,
- creating a standard method of preliminary signal processing,
- decision what values of the parameter change suggest laryngeal trauma,
- research on possible risk factors, especially in patients group with persistence pathology of the larynx.

REFERENCES

[1] Latkowski B.: *Otolaryngologia,* Wydawnictwo Lekarskie, Warszawa 1998

[2] Hadjitodorov S., Mitev P.: *A computer system for acoustic analysis of pathological voices and laryngeal diseases screening,* Medical Engineering & Physics, 24 (2002) 419–429

[3] Tadeusiewicz R.: *Sygnał mowy*, Wydawnictwa Komunikacji i Łączności, Warszawa 1988

[4] Reroń E., Tadeusiewicz R., Modrzejewski M., Wszołek W.: *Application of Neural Networks and Pattern Recognition Methods to the Evaluation of Speech Deformation Degree for Patients Surgically Treated for Larynx Cancer,* Neuroendocrinology Letters, vol. 19, No. 3, 1998, pp. 147-157

[5] Tadeusiewicz R., Modrzejewski M., Wszołek W.: *Una Aplicatión de Redes Neuronales y Métodas de Reconocimiento de Imágenes en la Evaluación del Grado de Deformación de la Voz en los Enfermos de Cancer de Laringine Sometidos a Tratamientos Quirúrgicos,* Simulación, Boletin de Investigación Instituto McLeod de Ciencias de Simulación, Universidad Panamericana, Vol. 3, No. 4, 1998, pp. 4 – 13

[6] Wszołek W., Modrzejewski M., Tadeusiewicz R., Izworski A.: *Acoustic Assessment of Voice Signal Deformation After Partial Surgery of the Larynx,* in: Medical & Biological Engineering & Computing (Journal of the International Federation for Medical & Biological Engineering), vol. 37, supl. 2, part I, 1999, pp. 530-531

[7] Wszołek W., Modrzejewski M., Tadeusiewicz R., Wszołek T.: *Methods of voice signal analysis after ENT surgery*. In: Hutten H., Kroesl P. (eds.): Advances of Medicine and Health Care through Technology – the Challenge to Biomedical Engineering in Europe, vol. 3 (1): Biosignal Processing, EMBEC Vienna 2002, pp. 536-537

[8] Wszołek W., Modrzejewski M., Tadeusiewicz, R., Wszołek T.: *Quality of voice signal after ENT surgery,* In Manfredi C. and Bruscaglioni P. (eds.): Proceedings of 2nd International Workshop on Models And Analysis of Vocal Emissions for Biomedical Applications, University of Firenze, 2001

[9] Prosek, R. A., Montgomery, A. A., Walden, B. E., Hawkins, D. B., 1987. An evaluatoin of residue features as correlates of voice disorders. *J. Commun. Disorders* 20, 105-107

[10] Yonick, T. A., Reich, A. R., Minifie, F. D., 1990. Acoustical effects of endotracheal intubation. *J Speech Hear Disord* 55, 427-33

3rd International Workshop MAVEBA 2003, 273-276

ANALYSIS AND EVALUATION OF NASALIZED [G] CONSONANT IN CONTINUOUS JAPANESE

Hisao Kuwabara

Teikyo University of Science & Technology, Uenohara, Kitatsuru-gun, Yamanashi 409-0193, Japan
Tel. (0554)63-4411, Fax (0554)63-4431, E-mail: kuwabara@ntu.ac.jp

***Abstract*: Nasalized veolar consonant [g] in continuous Japanese is often observed in some dialect and is said to decrease in frequency year by year. This paper deals with acoustic and perceptual analysis of this phenomenon. Test materials used in this experiment are read version of Japanese short sentences by NHK's (Japan Broadcasting Corporation) professional announcers. Each sentence includes at least one [g] consonant that would likely be pronounced as nasalized. An evaluation test reveals that less than 60% of nasalization has been found to occur for [g] consonants for which 100% nasalization bad been observed decades ago. Acoustic analysis for nasalized and non-nasalized [g] sounds has been performed mainly through waveform parameters. It has been found that power ratio between consonant and vowel is the most effective parameter for distinguishing nasals from non-nasals. But it is highly speaker dependent.**
***Keywords*: Nasals, veolar, perception, waveforms**

I. INTRODUCTION

It is well known that the [g] consonant, a velar voiced plosive, in Japanese continuous speech is often nasalized unless it appears at the word-initial position. Nasalized [g] consonant, which is expressed as [ŋ**],** takes place in dialects mainly spoken in northern districts including Tokyo area where the standard Japanese is spoken. There have been arguments among Japanese linguist whether [ŋ**]** consonant exists independently from [g] consonant or it is viewed as a phonetic variant of the consonant in standard Japanese [1]. Shiro Hattori, for example, took the former view on [g] and [ŋ] distinction in Tokyo dialect [2]. As a speech technology engineer, I myself would like to view the phenomenon an allophone of the phoneme.

In the so-called common Japanese, which is based on the Tokyo dialect, [ŋ] consonant sometimes takes place. This way of speaking used to be, and still to some extent is, regarded as a beautiful pronunciation of Japanese. TV/radio casters and announcers used to pronounce [ŋ**]** consonant as much as possible and this has been the norm of pronunciation for NHK (Japan Broadcasting Corporation) announcers for years. They used to be trained to pronounce [ŋ**]** sounds for proper portions while they are training to become a professional announcer. However, this trend declines gradually and less young people speak [ŋ**]** sound than ever before. The present study deals with perceptual evaluation and acoustic analysis of [g] and [ŋ**]** consonants in common Japanese.

II. PERCEPTUAL EVALUATION

Speech material to be examined has been offered by NHK Broadcasting Culture Research Institute. Several announcers, including those fresh announcers under training, participated in the recordings. The first evaluation has been performed at the NHK Institute using 32 short sentences, each includes at least one [g] consonant somewhere in-between, uttered by 24 speakers. The result is shown in Fig. 1 in which a percentage of [g] vs [ŋ] pronunciation among 32 utterances for each speaker. Speaker 23, for instance, shows a 100% [ŋ] pronunciation while speakers 8 and 11 exhibit very small percent of [ŋ]. As we can see at the right-most bar, less than 60% of [ŋ] pronunciation can be observed on the average. The above result was obtained based on the auditory perception by experienced persons including former professional announcers at the NHK Culture Research Laboratories and the speech materials they used are supposed to be 100% [ŋ] utterance for all speakers when they would have been used decades ago.

III. WAVEFORM ANALYSIS

Preliminary waveform analysis has been conducted for different speech materials from those used in the perceptual evaluation shown in Fig. 1.

3.1. Speech material

Materials used here is a set of 7 short sentences, each contains at least one [g] consonant, uttered by 10 announcers. In this speech material, two kinds of pronunciations, one is intentionally [ŋ]and the other intentionally [g] pronunciation for the same sentences by two speakers, are included. A sentence (No. 3) includes two [g] consonants while the rest has only one [g]

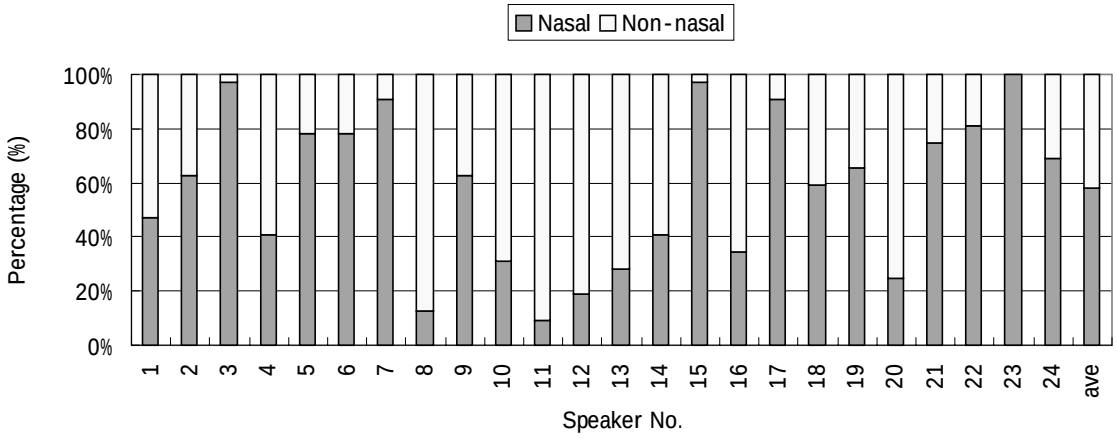


Fig. 1 Ratio of nasalized and non-nasalized perception of /g/ consonant in 32 sentences for individual 24 professional announcers

consonant each. There are 8 [g] consonants as a whole. Of the 8 [g] consonants examined, only one is a sample from word-initial position and the rest 7 samples are in a intervocalic position. Then, there will be little or no contextual effects for analyzing the data.

3.2. Waveform parameter

In order to find out waveform parameters that characterize the [ŋ], we have selected 10 acoustic parameters that are considered to reflect the waveform shape [3]. For power-related features, four kinds of parameters, and three kinds of parameters each for amplitude-related and duration-related features, as shown in Table 1. Among the ten parameters, the most important factor was found be the ratio between consonant and vowel parts followed by the energy level of consonant. At first, this experiment was intended to find an acoustic parameter that can automatically distinguish [ŋ] consonants from [g] without conducting human audition. Before trying to find acoustic parameters that can automatically distinguish [ŋ] consonant from [g] consonant, the above ten parameters have been examined.

3.3 Application of waveform parameters to [ng] and [g] consonant

The ten waveform parameters defined above have been applied to the 8 [g] consonants in the 7 short sentences. There is no clear-cut distinction in determining the phoneme boundaries especially between consonant and vowel when the consonant in question becomes nasalized. In such cases, facilities available such as waveforms, spectrum, hearing by ear, are incorporated to find an appropriate boundary. Hearing by ear has been found to be the most promising tool to decide the boundary. Table 2 shows the result for [g], [ŋ] and their ratio when applied to the test sentences.

In Table 2, the "ratio" stands for the value [ŋ] divided by [g]. This value is considered to

Table 1 Ten waveform parameters used in analyzing nasalized/non-nasalized [g] consonants.

Parameters	Definition	Name
(1) *Power Parameter*		
1) consonant and vowel ratio	power ratio of consonant part to vowel part	**prcv**
2) consonant and syllable ratio	power ratio of consonant part to the whole syllable	**prtc**
3) normalized consonant and vowel ratio	consonant power normalized by its length divided by normalized by vowel power	**prncv**
4) normalized consonant and syllable ratio	consonant power normalized by its length divided by normalized syllable power	**prntc**
(2) *Amplitude Parameter*		
5) consonant and vowel ratio	ratio of mean amplitude of consonant to mean amplitude of vowel	**lrcv**
6) consonant and syllable ratio	ratio of mean amplitude of consonant to mean amplitude of the whole syllable	**lrtc**
7) maximum consonant and syllable ratio	maximum consonant amplitude divided by maximum vowel amplitude	**mlrcv**
(3) *Duration Parameter*		
8) consonant duration	duration of consonant part	**cl**
9) vowel duration	duration of vowel part	**vl**
10) consonant and vowel ratio	duration ratio of consonant to vowel	**rlcv**

Table 2 Ten parameter values for /g/, /ŋ/ consonants and their ratio.

	/g/	/ŋ/	**ratio**
prcv	0.03	0.18	5.48
prtc	1.96	13.7	7.00
prncv	0.09	0.25	2.85
prntc	10.54	32.8	3.11
lrcv	0.24	0.48	2.00
lrtc	28.68	58.0	2.02
mlrcv	0.24	0.4	1.67
cl	33.81	68.6	2.03
vl	102.1	103.5	1.01
rclv	0.36	0.7	1.95

reflect the difference between [g] and [ŋ] sounds. The largest difference can be seen in the second parameter "prtc", the power ratio between consonant and syllable, followed by the first parameter "prcv." The third largest parameter is for "prntc," and the least one is naturally for "vl."

IV. Correlation between waveform parameters and perception of nasality

Another perceptual experiment, which is different from the one described in Section 2, has been performed about the nasality of [g] consonant using different speech materials uttered also by NHK announcers. Twenty five short Japanese sentences including the same ones used in the waveform analysis were used. Twelve NHK announcers read the sentences without any instructions about the nasalization of [g] consonants. They were allowed to read the sentences as exactly the same way as they used to pronounce when broadcasting news materials. Each sentence contains at least one [g] consonant and the number of [g] consonants to be examined is 31.

4.1. Perceptual experiment on nasalization

A perceptual experiment has been performed for [g] consonant whether it is nalasized or not. A whole CV-syllable that contained [g] consonant was excised from the running speech and the listeners were asked to judge if the portion in question was nasalized or not. Judgment of nasalization is hard to decide somehow. In fact, individual listeners responses show a great inconsistency from trial to trial. However, there are some speech samples, though very few, that can be served as objective data for which all listeners give consistent judgment towards [g]- [ŋ] distinction. Using these consistent judgment data for [ŋ], a correlation analysis between these perceptual data and the waveform parameters has been examined.

Before going into correlation analysis, let's take a brief look at the perceptual result. Fig. 2 shows the percentage of nasalization for 12 speakers. Ten listeners participated in this experiment. For each syllable excised from the running speech, they were asked to judge whether the consonant in question was nasalized or not.

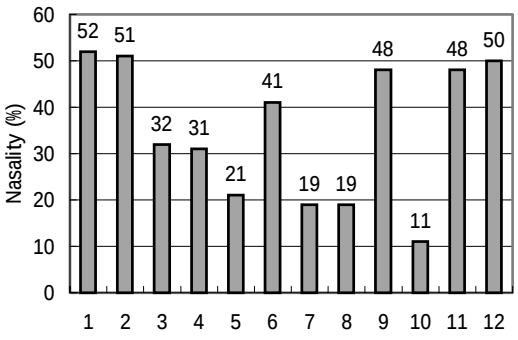


Fig. 2 Perceptual result on nasality for 12 speakers.

4.2. Correlation between perception and three acoustic parameters

In order to find out which waveform parameters relate closely to the perceptual results of [g]- [ŋ] distinction, we have performed correlation analysis between perceptual data and the waveform parameters. But this time, we have chosen three waveform parameters. They are 1) normalized consonant and vowel power ratio –**prncv**, 2) consonant and vowel amplitude ratio –**lrcv**, and 3) another parameter –**psmax**. The third parameter, **psmax** which is not included in the ten parameters listed in Table 2, is defined here. The parameter **psmax** is defined as a differential coefficient regarding the smoothed waveform-envelope as a function of time, which represents a kind of abruptness of the envelope change from [g] consonant to the following vowel.

There are 31 [g] consonants to be examined. The correlation has been taken between two vectors in the 31 dimensional space. For each speaker, let X be a vector in the 31 dimensional space,

$$X = (x_1, x_2, \cdots, x_{31}) \qquad (1)$$

where a component x_i stands for the percentage of [ŋ] response, averaged over ten listeners, for the i-th [g] consonant. Also, let Y be an another vector in the same 31 dimensional space,

$$Y = (y_1, y_2, \cdots, y_{31}) \qquad (2)$$

where the 31components represent acoustic values measured either one of the three waveform parameters described above and arranged in the same order as those of vector X . Then the correlation R between vectors X and Y is defined as,

$$R = \frac{(X, Y)}{\|X\| \cdot \|Y\|} \qquad (3)$$

where (X, Y) stands for the vector's inner product and $\|\cdot\|$ represents a vector's norm.

The result of correlation analysis is shown in Table 3 for speakers individually. From the result, it reveals that there are no specific waveform parameters that highly correlate with the perceptual data. In other words, there are no specific waveform parameters that characterize the nasalization regardless of speakers. If we look at the results more closely with speakers individually, we can find a few parameters for one speaker that show high correlate with the nasalization. Speakers 1 and 6, for example, show relatively high correlation with all the three

Table 3 Correlation coefficients between /ŋ/ perception and three waveform parameters for each speaker.

speaker	psmax	prncv	lrcv
1	0.76	0.72	0.79
2	0.74	0.42	0.61
3	0.64	0.74	0.76
4	0.68	0.69	0.78
5	0.39	0.41	0.51
6	0.79	0.71	0.80
7	0.72	0.72	0.77
8	0.72	0.63	0.76
9	0.53	0.55	0.57
10	0.62	0.48	0.57
11	0.62	0.65	0.75
12	0.45	0.54	0.55
average	**0.64**	**0.61**	**0.62**

waveform parameters, while speaker 5 does not. The result shows rather high speaker dependency. On average, it is around 0.6 for each acoustic parameter.

V. Spectrum analysis

It is likely that nasalization/non-nasalization distinction will appear most obviously in spectral envelope of the consonant in question. A spectral analysis has been made on the consonant. From the perceptual result described in Section 4.1, we have chosen speech data that clearly show [ŋ] pronunciation and [g] pronunciation. Spectral analysis has been made for these speech samples separately and the results are compared. Analysis has been made at the center of consonant part for each syllable. There are many ways to conduct spectral analysis but only the frequency analysis of the lowest three formants has been made here. Fig. 3 shows the result of frequency analysis. It is observed that there are small differences among the three formants between [ŋ] and [g] consonants.

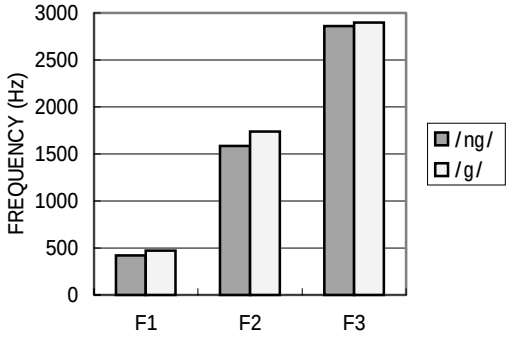


Fig. 3 Formant frequencies for /ŋ/ and /g/ consonants

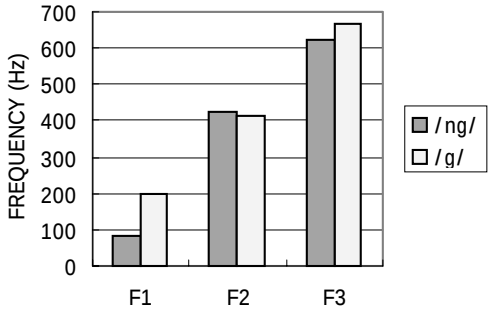


Fig. 4 Standard deviation of formants for /ŋ/ and /g/ consonants

For [ŋ] consonant, all formants, especially 2^{nd} formant, are slightly lower than those for [g] consonant but these differences are not statistically significant. Fig. 4 stands for the standard deviation of three formants for [ŋ] and [g] consonants. Again there are small differences between the two consonants. But in F1, relatively large difference can be observed; standard deviation for [ŋ] is significantly smaller than for [g] consonant.

VI. Discussion

As far as waveform parameters are concerned, we could not find a single parameter that can clearly separate nasalized [g]. Besides, waveform parameters largely differ from speaker to speaker and no specific speaker-independent parameter can be found so far. It is obvious that the distinction between nasalized and non-nasalized consonant appears dominant in spectral region. There are some zeros (anti-formants) in a nasal sound. Formant frequencies themselves appear not significant acoustic parameters that distinguish [g] and [ŋ] consonants. As far as spectral analysis is concerned, most dominant factor that seems to differ between the two consonants is the shape of spectral envelope. For [g] consonant, the spectral envelope is rather "flat" over the entire frequency region while that for [ŋ] is not. Quantitative analysis will be needed to express this "flatness" between the two consonants.

VII. Acknowledgement

Speech materials of NHK announcers were offered by NHK Broadcasting Culture Research Laboratories. The author would like to express his gratitude to Mr. Ohnishi and Mr. Shibata.

References

[1] H. Jo: Nihongo, Iwanami Shoten, pp.109-145, 1977
[2] S. Hattori: Onseigaku, Iwanami Zensho, p.102, 1968
[3] M. Nakamura and H. Kuwabara: "Waveform analysis of /ŋ/ consonant," *Proceedings for Spring Meeting of ASJ*, pp.363-364, 2001

3rd International Workshop MAVEBA 2003, 277-280

Topologically equivalent reconstruction of instationary, voiced speech

F.R. Drepper
Zentrallabor für Elektronik, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Abstract: **Voiced speech is characterized by qualitatively rich mode locking phenomena linking harmonically excited acoustic modes of the vocal tract. Due to the strong instationarity of speech, a differentiated analysis of these modes cannot be achieved with the help of a linear, time invariant source and filter model (based on stationary sources). As alternative, the characteristic mode locking is described as generalized synchronization in drive - response systems with an instationary, common (fundamental) drive. By introducing a combined harmonic and logarithmic (audiological) scale subband decomposition adapted to the frequency of the master oscillator of phonation, a self-consistently confirmed, topologically equivalent reconstruction of a number of acoustic modes of an acoustic object is generated. Whereas the invariant resonator properties (Lyapunov exponents) of the reconstructed response dynamics are characteristic for vowels, the generalized synchronization manifolds (lines or surfaces) in the combined state space of drive and respective response band can be used for the distinction of consonants. The topologically equivalent reconstruction of the phonation process is potentially useful for phoniatric diagnoses.**
Keywords : **Subband decomposition, drive – response reconstruction, transfer function model, voiced speech, generalized synchronization**

I. Introduction

The characteristic mode locking of voiced speech results from harmonic excitations, which are synchronized by glottal closure events [1]. In the context of generalized synchronization in drive – response systems it has been shown recently [2,3], that mode locking or synchronization is not an elementary phenomenon, but a header for a larger number of qualitatively different coordination possibilities, which are characterized by more or less smooth and/or continuous invariant manifolds in the combined state space of coupled drive - response oscillator pairs, the manifolds being defined by maps, which relate a state of the response uniquely to the simultaneous state of the drive [2-4]. In the context of speech recognition the topological equivalence between drive and response represents an important special case [5], which is characterized by a conjugation (a continuous and uniquely invertible map). Together with the more general concept of conditional asymptotic stability [6] these notions are useful for a differentiated analysis of the synchronization or coordination phenomena of voiced speech.

The ubiquitous instationarity of the amplitude and pitch of phonation is a second essential feature of voiced speech, the variation of the amplitude being relevant on time scales down to less than 50 ms. In this context it is important to note that the long time known phenomenon of synchronization is not limited to periodic or quasi periodic driving but may as well occur for stochastic [7] or deterministic chaotic driving [2]. So far the application of the source and filter model to the recognition of voiced speech is based on the assumption of a stationnary phonation process [8]. This assumption limits the source and filter model to the description of relatively short sections of speech (typically 20 ms). Such short sections, however, are insufficiently suited to detect the characteristic invariant manifolds of voiced phonemes. The ubiquitous instationarity of human speech motivates, to replace the assumption of stationnary phonation (implicitly implied when estimating spectra) by the assumption of generalized synchronization between the instationary and/or nonlinear drive and the acoustic response. Thus the atoms or objects (in particular the phonemes) of speech are no longer interpreted as stationary processes but as stationnary or invariant manifolds (lines or surfaces) in the combined state space of instationary drive and response oscillator pairs. However, neither the acoustic response modes within the vocal tract nor the excitation within the glottis can directly be observed in the situation of speech communication.

II. Subband reconstruction

As a characteristic feature the present approach uses suitably chosen bandpass filters to determine a fundamental driver mode as well as higher frequency subbands, which represent topologically equivalent reconstructions of corresponding acoustic modes of the vocal tract. The choice of the appropriate bandpass filters is based on the fact that voiced speech is characterized by a concentration of power in comparatively narrow frequency ranges and that due to the approximate periodicity of the voice source these frequency ranges show a comb like pattern, aligned to the fundamental frequency defined as (short time) average of the frequency of glottal closure events. The bandwidths of the bandpass filters should be chosen

sufficiently narrow, to resolve as many harmonics as possible, however also sufficiently broad, such that the relative bandwidth exceeds the one of the instationary frequency fluctuations of the fundamental drive process. Obviously the ERB bandwidths (according to the equivalent rectangular bandwidth model) [9,10], known from masking experiments in psychoacoustics, represent an evolutionarily successful compromise. This choice introduces an a priory limit on the harmonic number *h* of resolvable subbands, (*h* < 10). When generating a vowel, the vocal tract shows no branching and no additional constriction (apart from the glottis). In this situation the feasibility of a harmonic scale aligned subband decomposition is guaranteed, since the response processes of the different harmonic excitations superpose without perturbation and can thus be separated by appropriate bandpass filters due to their differing frequencies.

Even in the case of nasals or voiced approximants like /l/ or /v/ in veal and voiced sibilants like /th/ in thumb, the concentration of power of the primary voice source (in space and frequency) implies or supports a phonation dynamics, which features a causal pinhole expressed by a low dimensional, potentially instationary master oscillator, which "enslaves" [4] the faster state variables of sound production or at least their long distance effect on the acoustic field. According to the so far rather limited study there is no contradiction, that at least in the case of healthy phonation the voiced part of the excitation of the acoustic modes can be expressed as synchronization manifolds, which are driven by a pair of fundamental amplitude and phase. The complex wavelet transformation [11,12],

$$A_t\, e^{i\psi_t} = \sum_k X_{t+k} \,(e^{i\,k\,\omega} - e^{-0.5\omega/\sigma^2})\, e^{-0.5\,k^2/\sigma^2} \;,$$

turns out to be particularly suited for the extraction of the amplitude A_t and phase ψ_t of the master oscillator from the speech signal. The centre frequency ω is chosen as an appropriate multiple of the fundamental frequency F_0, which is obtained by a conventional method.

Following the well accepted linear source and filter model of speech production [1,8] it is plausible to represent the voiced part of each subband specific excitation as product of drive amplitude A_t and an oscillatory driver phase dependent excitation function $G_j(\psi_t)$, which thus takes the central role in the phenomenological description of complex voiced phones. The enslavement of the fast degrees of freedom of the excitation implies a periodicity of the excitation function. In the context of instationary phonation it is important to note that this periodicity does not refer to time but to the phase of the glottal drive. The period length $2\pi\, p_j$ of the excitation function is potentially speaker dependent and coincides usually with the fundamental period 2π. Due to the band limitation each excitation function can nicely be approximated by a finite Fourier series, the terms of which may be interpreted as purely harmonic, elementary excitations.

Following the linear source and filter model, subband *j* $\{X_{j,t} \mid t = 0,1,...\}$ is approximated as a finite dimensional, linear response to a drive synchronous excitation. Due to the described band limitation of the subbands as well as due to the band adapted time step length Δ (chosen as a quarter of the period length defined by the band specific central filter frequency) a two dimensional response dynamics turns out to be sufficient,

$$X_{j,(n+1)\Delta} = a_j\, X_{j,n\Delta} + b_j\, X_{j,(n-1)\Delta} + A_{n\Delta}\, G_j(\psi_{n\Delta})$$

with $G_j(\psi_{n\Delta}) = \sum_{k=0}^{K_j} c_{j,k} \cos(k\psi_{n\Delta}/p_j - \gamma_{j,k})$,

n = 0, 1, ... and $K_j \le 2\,h_j$, where h_j represents the band index dependent harmonic. The goal of the phonation process adapted bandpass decomposition is characterized by subbands, which can be approximated as two-dimensional response to a single, pure harmonic, elementary excitation. In the case of the higher harmonic subbands, in particular of consonants, the goal reduces to maximal diagonal dominance of the subband specific elementary harmonic excitation. The average distance of index *k* to the band specific harmonic h_j turns out to be a useful objective function,

$$\overline{\Delta k}_j \;=\; \frac{1}{\sum_{k=0}^{K_j} c_{j,k}^2} \sum_{k=0}^{K_j} c_{j,k}^2 \left|k - h_j\right| \;.$$

The central filter frequency of the fundamental subband filter represents the essential adaptation parameter to achieve the diagonal dominance of the elementary excitations.

III. TOPOLOGICAL EQUIVALENCE

The introduction of time dependent and time related (continuously extended, unwrapped) phases as state variables of the response dynamics opens the possibility to identify (1:n) or (n:m) mode- or phase locking as a (near linear, diffeomorphic) conjugation. Due to transitivity and invertibility of conjugations in a chain of conjugated oscillators, the evidence of a near linear conjugation between the subband oscillators of a voiced signal can be taken as a confirmation of the topological equivalence of all oscillators involved, including the equivalence between the respective harmonically excited acoustic mode within the vocal tract and the corresponding subband (figure 1). The confirmation of topological

equivalence of the subband dynamics can be used as a basis for the quantitative determination of the topological invariants [13] of the resonator dynamics.

According to the so far rather limited empirical basis (50 ms subsections of 5 vowels and 6 sustainable voiced consonants uttered by 4 male and 2 female subjects), the described subband decomposition of voiced phonemes generally offers the possibility to detect near linear conjugation between the lower harmonic subbands (figure 1). This way the phase and amplitude of the fundamental drive can generally be confirmed as topologically equivalent image of state variables of the fundamental glottal mode. The presented approach is thus well suited for a robust and precise determination of the momentary pitch of voiced speech and potentially also for phoniatric diagnoses.

For subbands within the harmonically resolvable range (harmonic number $h < 9$), a missing conjugation to the driver band can be attributed to a break up of the conjugation chain within the vocal tract and not to a break up on the way from the vocal tract to the ear or microphone (figure 1). In the case of voiced approximants and sibilants, in particular, the loss of conjugation between subbands does not indicate a loss of conditional asymptotic stability [6] of the higher harmonic subbands. A general definition of complex voiced phones of human speech can thus be given as existence of a bandpass filter based subband decomposition, which contains one fundamental drive oscillator and further conditionally stable response bands, where the conditioning is limited to the amplitude and phase of the drive and where the drive can be confirmed to be (1:1) equivalent to the fundamental glottal oscillator.

Strikingly many distinctive properties of voiced phonemes coincide with topological invariants of the response dynamics or with topologically invariant geometric properties of the related invariant manifolds. The most important topological invariants of the subband dynamics are the (conditional) Lyapunov exponents [6], since they express resonator properties of the vocal tract, like resonator quality and eigen-frequency, which are known to be strongly dependent on the geometry of the vocal tract and thus particularly suited for the distinction of vowels. The distinctive properties of consonants are predominantly related to geometric properties of invariant manifolds in the four-dimensional state space of drive - response oscillator pairs (like kinks or jumps in the case of nasals). Stop consonants are characterized by a pronounced visibility (audibility) of the amplitude – amplitude coupling between the drive and the respective response bands, whereas for sustainable voiced consonants the coupling of the response phases to the driver phase plays the more important distinctive role.

IV. Evolutionary aspects of voiced speech

As a striking feature of human speech, the confirmation of the topological equivalence can often be achieved for subbands with harmonic numbers higher than 10. (Due to resonant excitation, the detectability of higher harmonic, phase locked modes becomes extreme in the case of singing.) The surprisingly extended success of the described approach towards the determination of phonation and vocal tract equivalent excitation and response processes, can only be explained within the framework of evolutionary and ontogenetic adaptation, characterized by a near optimal fit between properties of human speech and the abilities of auditive perception. Thus voiced speech and singing have to be interpreted as results of adaptation processes, which favor easy detectability within a confusion of voices.

In view of the pronounced differentiation of the synchronisation phenomena of voiced speech, auditive perception of humans can be assumed to be able to perform and select the skilled bandpass decomposition, which uncovers the more or less smooth, stationary manifolds in the combined state space of the subbands - even in the case of instationary phonation. There are several empirical facts, which support a perception equivalent model of hearing, which is build on the described synchronization analysis of voiced speech. Firstly there is the central role of the pitch known to be relevant on different semantic layers of speech communication and to be perceived even in the case of imperfect harmonicity [14]. Further support can be seen in the astonishing monaural voice separation and speaker identification ability of the auditive perception of humans, which (in particular in the case of rough phonation) could so far neither be explained by perceptional models nor imitated by speech - or speaker recognition algorithms.

Based on highly developed abilities of higher vertebrates [15,16], the astonishing speaker identification ability indicates that the auditive perception of humans is in command of analysing abilities of the nonlinear dynamics of phonation, including recognition of subharmonics or of co-existing meta-stable periodic trajectories (unstable periodic orbits, UPO's) [3,17]. In order to avoid dangerously large bandwidths of the fundamental driver mode it is advantageous to represent the influence of the mentioned nonlinear phonation dynamics with the help of periodicity p_j of the driver phase dependent excitation function. The potential richness of the combination possibilities of periodicity p_j of an excitation manifold with the periodicity q_j of the resulting response manifold and the winding number w_j of the corresponding response phase offers a plausible explanation for the astonishing speaker recognition ability of auditive perception.

V. Conclusion

Contrary to the conventional approach towards speech analysis, which is based on the assumption of a stationary, high dimensional source and the use of a broad band version of the linear source and filter model, the newly proposed approach describes the source as a synchronized response to a low-dimensional instationary drive, which is determined self-consistently as a topologically equivalent image of the underlying fundamental glottal mode. The self-consistency is based on a skilled subband decomposition, the subbands of which can optimally be interpreted as linear response to harmonically distinct, voiced excitations. Apart from providing evidence of the topological equivalence of the common drive, the skilled subband decomposition discloses topologically equivalent images of the invariant manifolds, which characterize the synchronization of the higher harmonic acoustic modes of the vocal tract. The distinction of consonants is hypothesizes to rely largely on topologically invariant geometric properties of these manifolds. Since the parameters of the excitation manifolds can be estimated efficiently with the help of multiple linear regression, the outlined synchronization based analysis of voiced speech is expected to be feasible in real time.

The author would like to thank G. Langner, Darmstadt, V. Hohmann, B. Kollmeier and J. Nix, Oldenburg, C. Neuschaefer-Rube, C.Hoelper and P. Vary, Aachen, N. Stollenwerk, London, P. Grassberger, H. Halling, M. Schiek and P. Tass, Jülich for helpful discussions.

References

[1] Fant G., *Acoustic theory of speech production*, Mouton, 'S-Gravenhage (1960)
[1] Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I.
[2] Abarbanel, *Phys. Rev. E* 51, 980-994 (1995)
[3] Rulkov N.F. et al., *Phys. Rev. E* 64, 016217 (2001)
[4] Haken H., *Synergetics*, Springer Verlag, Berlin (1977)
[5] Kocarev L., U. Parlitz, *Phys. Rev. Lett*. 76, 1816 (1996)
[6] Pecora L.M. and T.L. Caroll, *Phys. Rev. Lett*. 64, 821 (1990)
[7] Afraimovich V.S., N.N. Verichev, M.I. Rabinovich, *Radiophys. Quantum Electron*. 29, 795 (1986)
[8] Vary P., U. Heute, W. Hess, *Digitale Sprachsignalverarbeitung*, B.G. Teubner Verlag, Stuttgart (1998)
[9] Moore B.C.J., *An introduction to the Psychology of hearing*, Academic Press (1989)
[10] Hohmann V., *Acta Acustica* 10, 433-442 (2002)
[11] Lachaux J., E. Rodriguez, J. Martinerie and F.Varela, *Hum. Brain. Mapp*. 8, 194 (1999)
[12] Quiroga R.Q., A. Kraskov, T. Kreuz and P. Grassberger, *Phys. Rev. E* 65, (2002)
[13] Kantz H., T. Schreiber, *Nonlinear time series analysis*, Cambridge Univ. Press (1997)
[14] Schouten J.F., R.J.Ritsma and B.L.Cardozo, J. Acoust. Soc. Am., 34, 1165-1168 (1962)
[15] Fitch T., J. Neubauer and H. Herzel, *Animal Behaviour* 63, 407-418 (2002)
[16] Langner G., C. Simonis and S. Braun, *Fortschritte der Akustik-DAGA'02*, (2002)
[17] Herzel H., D. Berry, I.R. Titze and I. Steinecke, *Chaos* 5, 30-34 (1995)

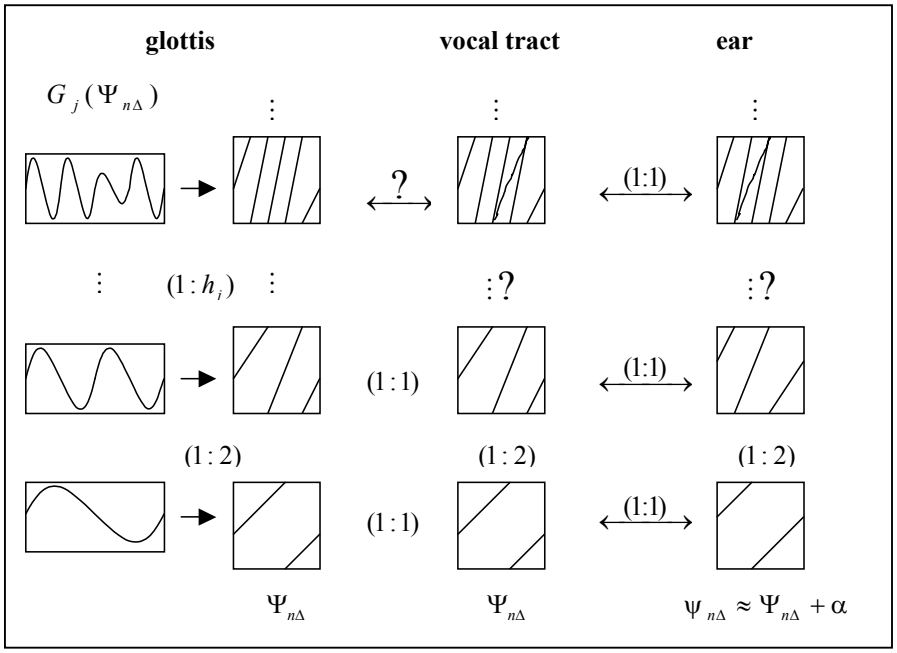


Figure 1: Voiced phones of human speech are characterized by stationary manifolds (lines or surfaces) in the combined state space of drive and response oscillator pairs, which differ with respect to the distance to sound generation inside the glottis as well as with respect to the respective oscillation or winding number h of the subband specific excitation. The dynamics of the excitations as well as of their resulting response processes are reduced to the corresponding phase dynamics, the respective driver phases ψ or Ψ being indicated horizontally and the corresponding response phases vertically.

3rd International Workshop MAVEBA 2003, 281-284

POINCARÉ SECTIONS FOR PITCH MARK DETERMINATION IN DYSPHONIC SPEECH

Martin Hagmüller, Gernot Kubin
Institute of Communications and Wave Propagation, Graz University of Technology, Graz, Austria

***Abstract*: In this paper a Poincaré approach to pitch mark determination is presented. While speech has been interpreted in terms of nonlinear systems theory for quite some time, not much effort has been made to exploit this knowledge in the problem of pitch mark detection. This algorithm uses nonlinear state space embedding and calculates the Poincaré section at a chosen point in state space, pitch-marks are then found at the crossing of the trajectories with the Poincaré plane. The procedure is performed frame-wise to account for the changing dynamics of the speech production system. First results show promising performance, comparable to the pitch marking algorithm used in 'Praat', and outperforming it in case of irregular voices.**
***Keywords*: Dysphonic speech, state-space-embedding, Poincaré section, pitch-marks.**

I. Introduction

For pitch–synchronous processing of speech, accurate pitch-marks are essential. A particular challenge is the correct determination of pitch-marks for dysphonic voices. On the other hand, having a reliable method for pitch marking available, this could be used for enhancement of rough pitch, by reducing the fluctuations of the fundamental period. Accurate and robust methods for pitch detection are of interest for the analysis of dysphonic voices [1] and, e.g., for the measurement of jitter, methods to reliably determine the instantaneous fundamental period are necessary.

The nonlinear nature of the speech signal has been of increasing interest for several years now, starting in the early nineties [2].

Conventional algorithms, such as correlation based methods, assume linear models of speech production, though even for normal voices those models cannot fully explain the properties of the signal. For dysphonic speech, those models more or less fail due to the higher dimensional non-linearity inherent in the system. Especially, for strongly irregular voices, conventional algorithms for pitch mark determination fail and, therefore, the need for new methods is at hand. Non-linear methods seem to be a promising way of overcoming the weaknesses of the currently used approaches.

State-space approaches for dysphonic voice analysis have been proposed recently [3], [4]. Voice irregularities have been treated with nonlinear methods before, e.g. by performing noise reduction in state space [5].

The paper is organized as follows. Section II will give some background and review existing algorithms for pitch determination in state space. In section III the proposed state-space approach for pitch marking will be introduced and the algorithm will be explained. Section IV will show some results and finally section V will conclude the paper with a summary and an outlook.

II. Background and Related Work

A non-linear dynamical system can be embedded in a reconstructed state-space by the methods of delays. According to Takens [6], the state space of a dynamical system can be topologically equivalently reconstructed from a single observed one-dimensional system variable. For a D-dimensional attactor it is sufficient to form a $M \geq 2D+1$ state space vector. The M-dimensional trajectory is formed from a speech signal vector $\mathbf{x}(n)$ by delayed versions of the signal $\mathbf{x}(n)$,

$$\mathbf{x}(n) = [x(n), x(n-\tau_d), ..., x(n-(N+1)\tau_d)], \quad (1)$$

where τ_d is the delay time, which has to be chosen to optimally unfold the attractor. If one chooses an arbitrary point on the attractor in an M-dimensional space then one can create a hyper-plane which is orthogonal to the flow of the trajectory at the chosen point. This is called the Poincaré plane. All trajectories, that return to a certain neighborhood of the initial point, cross the hyperplane and can be represented in dimension $M-1$ compared to the original trajectory.

In 1997 Kubin [7] first suggested to use those Poincaré sections for the determination of pitch-marks and mentioned special applications for signals with irregular pitch period. Experiments showed very promising results for an example with vocal fry, where the pitch period doubles for some time. The pitch period was followed correctly.

Later Mann and McLaughlin [8] further worked with Poincaré maps and applied them to epoch marking for speech signals. They again saw promising results, but reported inability to resynchronize after, e.g., stochastic portions of speech.

More recently Terez [9] introduced another state space approach to pitch detection, using space-time separation histograms. Each point on the trajectory in state space is separated by a spatial distance r and a time distance Δt. One can draw a scatter plot of Δt versus r or,

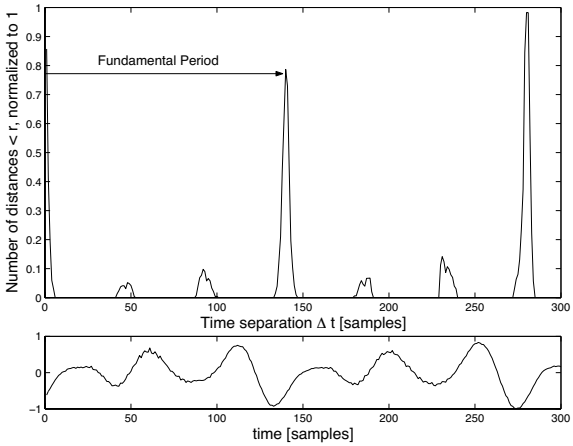


Fig. 1. *Histogram of space-time separation. The normalized number of state space-distances within a certain neighborhood r for every time distance Δt is plotted.*

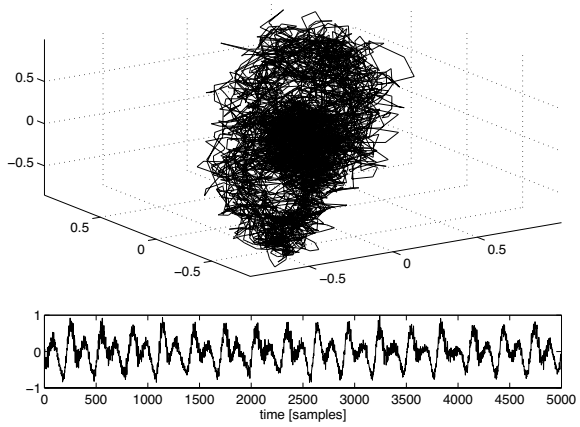


Fig. 2. *State-space embedding before low-pass filtering.*

for every time distance, count the pairs within a certain neighborhood r. This can then be normalized to yield a histogram (fig. 1). In case of periodicity in the signal, the histogram concentrates at certain Δt values, whereas others have rather low values. The first maximum of the histogram indicates the fundamental pitch period. Compared to the autocorrelation function the peak is much more significant and, therefore, offers improved performance. In case of noise-like signals the histogram is more evenly spread over all time distances. Since histograms are based on averaging statistics, localized pitchmarks cannot be determined reliably with this approach.

III. Description the Algorithm

Our algorithm builds on the before mentioned approaches. The algorithm works on a frame-by-frame basis to handle the changing system parameters.

For pitch mark detection the low-dimensional characteristics of the signal need to be observed. So the noise has to be removed, otherwise the attractor is hardly visible with 3-dimensional embedding (fig. 2). If the embedding dimension is high enough, intersections with the Poincaré plane would still be corrosponding to the pitch period, less reliable, though. For a noise reduced attractor a singular-value-decomposition (SVD) embedding approach has been proposed [8], but similar results can be achieved by a simple low-pass filter. The latter is computationally less demanding of course, so this is chosen for noise reduction.

Then the signal is upsampled to f_s = 96kHz to increase the resolution of the pitch marks, since at low sampling rates the pitchmarks would exhibit too much discretisation noise. The embedding in the state space is done by the method of delays, the embedding dimension was chosen be $M = 9$. The delay for the chosen sampling frequency is around $\tau = 50$.

Poincaré section

At the heart of the algorithm is the calculation of the Poincaré hyperplane (fig. 3). Around a chosen point $\mathbf{x}(n_0)$, the neighborhood within a certain radius r is searched for points. Then a mean flow direction $\mathbf{f}(n_0)$ of the trajectories in this neighborhood $\mathcal{N}(n_0)$ is calculated (considering only those trajectories, with a flow in the same direction as the initial point).

$$\mathbf{f}(n_0) = mean[\mathbf{x}(n+1) - \mathbf{x}(n)] \qquad \forall n \in \mathcal{N}(n_0) \quad (2)$$

So for every frame the Poincaré hyperplane is defined as the hyperplane through $\mathbf{x}(n_0)$, which is perpendicular to $\mathbf{f}(n_0)$ (fig. 3).

Mann et al. [8] reported the loss of synchronicity in case of unvoiced portions of the signal. Since in running speech this is usually the case, we decided to use the minimum of the low-pass filtered time-domain signal as an additional criterion for synchronization. So, in every frame we initialize the algorithm with $n_0 = min(x)$.

Points in the neighborhood $\mathcal{N}(n_0)$, within a certain distance r from the plane are considered as pitch mark candidates. Of these candidates, we select those which correspond to an absolute minimum in the time domain.

To remove the influence of a changing amplitude automatic gain control was applied for every frame. This moves the trajectories of quasi-periodic signals closer together, which means, that the attractor is contracted, if it was spread due to amplitude changes.

The length one frame has to be chosen so that at least two periods of the expected minimum frequency fit into the frame. If the signal is periodic, the trajectory returns at least once into the chosen neighborhood and intersects the Poincaré hyperplane and a pitchmark can be detected. The hopsize depends on the the last pitchmark in the current frame. The beginning of the following frame is set to the last pitchmark.

A proper voiced/unvoiced decision is not yet solved. Right now a frame is considered as unvoiced, if no

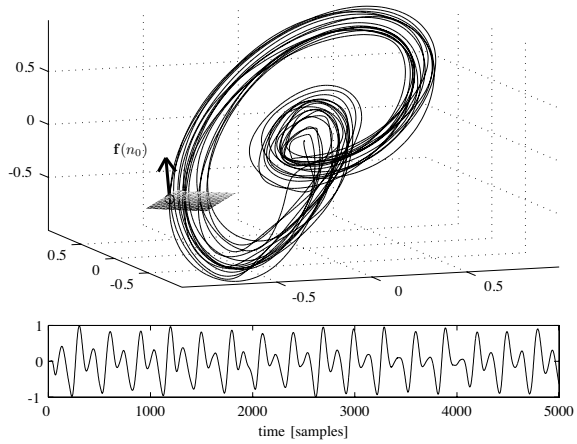


Fig. 3. *State-space embedding of low-pass filtered speech signal with mean flow vector and Poincaré plane.*

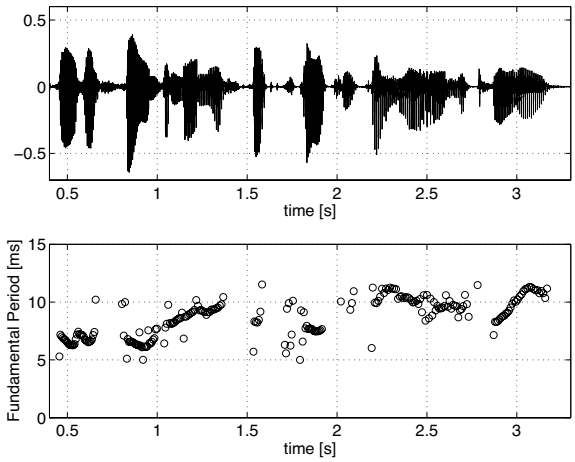


Fig. 4. *Top plot: Waveform plot of the sentence 'Judith found the manuscripts waiting for her on the piano' and the pitch marks obtained by Poincaré section. Bottom plot: Fundamental period*

neighbors can be found, because the trajecories do not come back to a chosen neighborhood anymore.

IV. Results

Formal evaluation of the pitch marking problem still has to be performed. Informal results using the pitch detection evaluation database by Paul Bagshaw [10] (http://www.cstr.ed.ac.uk/projects/fda/) and recordings of dysphonic voices from Graz University Hospital [11] are very promising.

In figure 4 the results of the algorithm on running speech can be seen. The sentence *'Judith found the manuscripts waiting for her on the piano'* is spoken by a male speaker with modal voice. Most of the pitch marks are correctly set.

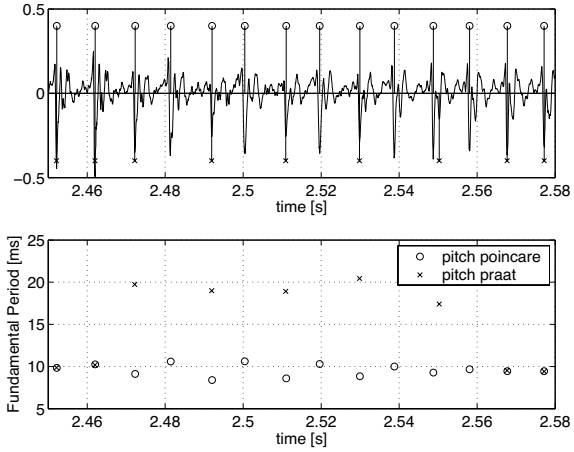


Fig. 5. *Top plot: Time-domain waveform plot with pitch-marks with Poincare section (positive peaks) and praat (negative peaks). Bottom plot: Fundamental period.*

In figure 5 a segment is taken out of the same speech file. There, a short period of diplophonic fundamental frequency is present (sentence *'rl040'* from the bagshaw database). Other algorithms like Praat [12] fail at this instance or detect a period doubling if the chosen minimum pitch value allows for such long pitch periods. The Poincaré method recognizes the rapidly alternating pitch period correctly. Though in this case it is a matter of definition whether the alternating period or the period doubling is the correct interpretation.

The state space plot of the same segment (fig. 6) shows an interesting property. There are two loops with different sizes in the plot. The interpretation is that depending on the period cycle length the state space vector follows either the larger or the smaller loop.

Fig. 7 shows a speech waveform, of a male speaker uttering the German phrase *'nie und nimmer'*. This utterance is described by speech therapists as hoarse, with strong diplophonia and some breathiness; his mean pitch is unusually high for a male person. Besides a few errors the pitch seems to be marked correctly.

Fig. 8 shows a segment of this phrase showing the irregular fundamental period. This case, of course calls for a comparison with a laryngograph signal, which is not available in the database [11].

V. Conclusion

An algorithm using Poincaré sections for pitch mark determination for dysphonic voices was presented. The algorithm works on running speech, overcoming the synchronization problem by sticking to the minimum of the time domain signal. A diplophonic case was presented where the alternating pitch period is correctly identified. The results are very promising, and will receive further evaluation.

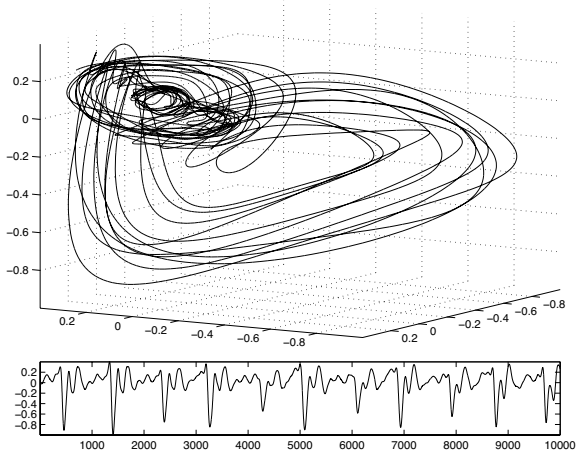


Fig. 6. *Top plot: State-space embedding of the diplophonic speech sample. One can interpret the two loops as the two different attractors for the two fundamental periods. Bottom plot: waveform plot*

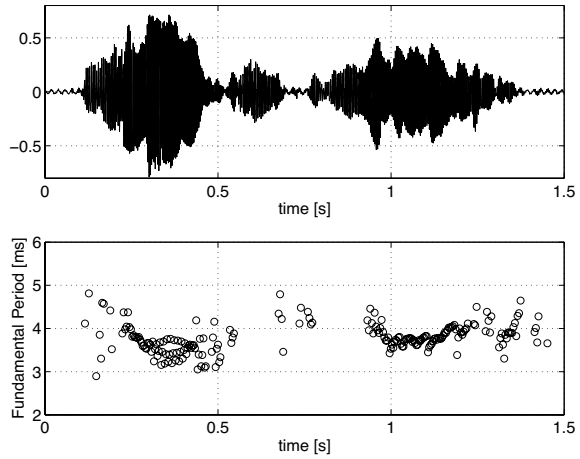


Fig. 7. *Top plot: Waveform plot of the German phrase 'nie und nimmer'. Bottom plot: fundamental period obtained with Poincare section.*

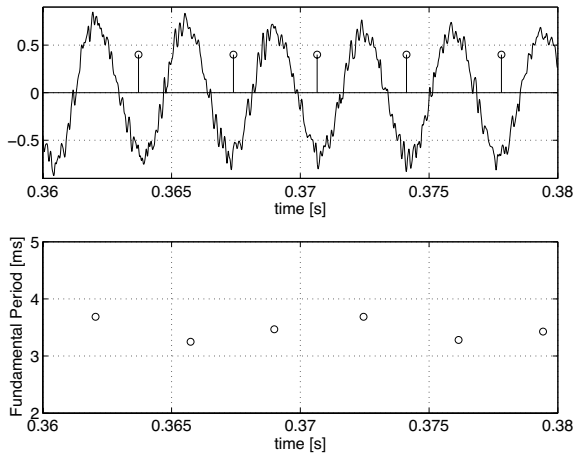


Fig. 8. *Top plot: Waveform plot of a segement of the German phrase 'nie und nimmer' and pitch marks. Bottom plot: fundamental period obtained with Poincare section.*

REFERENCES

[1] Ingo R. Tietze, "Workshop on acoustic voice analysis - summary statement," in *Proc. Workshop on Acoustic Voice Analysis*, Denver, Colorado, Feb. 1994.

[2] N. Tishby, "A dynamical systems approach to speech processing," in *IEEE Proc. ICASSP'90*, Albuquerque, NM, Apr. 1990, vol. 4, pp. 365–368.

[3] Lorenzo Matassini, Rainer Hegger, Holger Kantz, and Claudia Manfredi, "Analysis of vocal disorder in a feature space," *Med Eng Phys*, vol. 22, no. 6, pp. 413–418, 2000.

[4] Hanspeter Herzel, Joachim Holzfuss, Zbigniew J. Kowalik, Bernd Pompe, and Robert Reuter, "Detecting bifurcations in voice signals," in *Proc. Nonlinear Techniques in Physiological Time-Series-Analysis, Freital, '95*, H. Kantz, J. Kurths, and G. Mayr-Kress, Eds. Oct. 1996, pp. 23–27, Springer Verlag.

[5] Lorenzo Matassini and Claudia Manfredi, "Software correction of vocal disorders," *Computer Methods and Programs in Biomedicine*, vol. 68, no. 2, pp. 135–145, 2002.

[6] Floris Takens, *Detecting Strange Attractors in Turbulence*, vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Springer, New York, 1981.

[7] Gernot Kubin, "Poincaré section techniques for speech," in *Proc. of IEEE Workshop on Speech Coding for Telecommunication '97*, Pocono Manor, PA, Sept. 1997, pp. 7–8.

[8] Iain Mann and Steve McLaughlin, "A nonlinear algorithm for epoch marking in speech signals using poincare maps," in *Proceedings of the 9th European Signal Processing Conference*, Rhodes Greece, Sept. 1998, vol. 2, pp. 701–704.

[9] D.E. Terez, "Robust pitch determination using nonlinear state-space embedding," in *Proc IEEE ICASSP*, 2002, vol. 1, pp. 345–348.

[10] Paul Bagshaw, *Automatic prosodic analysis for computer aided pronunciation teaching*, Ph.D. thesis, University of Edinburgh, UK, 1994.

[11] Birgit Wrentschur, *Die Perzeptuelle Evaluation von Stimmstörungen*, Ph.D. thesis, University of Graz, 1998, Includes 2 audio CDs with Disordered Voice Samples.

[12] Paul Boersma and David Weenink, "Praat, software, downloaded from http://www.praat.org, 8/2003," .

3rd International Workshop MAVEBA 2003, 285-
© Firenze University Press 2003, ISBN 88-8453-154-3

AUTHOR INDEX

Stampato da
Grafiche Cappelli
Osmannoro, Firenze
Novembre 2003