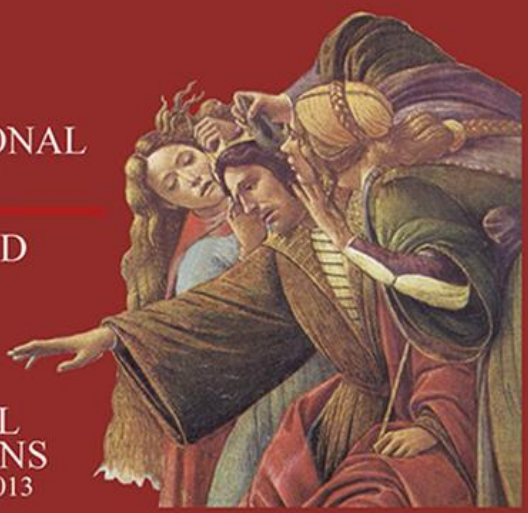
UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

8th
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

December 16-18, 2013
Firenze, Italy



PROCEEDINGS

FIRENZE
UNIVERSITY
PRESS

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

8th INTERNATIONAL WORKSHOP

December 16-18, 2013
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2013

Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2013.
(Proceedings and report ; 98)

http://digital.casalini.it/9788866554707

ISBN 978-88-6655-469-1 (print)
ISBN 978-88-6655-470-7 (online)

Cover: designed by CdC, Firenze, Italy.

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published in the online catalogue of the FUP (http://www.fupress.com).

Università degli Studi di Firenze
Firenze University Press
Borgo Albizi, 28, 50122 Firenze, Italy
http://www.fupress.com
Printed in Italy

MAVEBA 2013

Firenze, Italy

The MAVEBA 2013 Workshop is sponsored by:

Università degli Studi di Firenze
Department of Information Engineering - DINFO

and is supported by:

Biomedical Signal Processing and Control, Elsevier Ltd.
http://www.journals.elsevier.com/biomedical-signal-processing-and-control/

Italian Ministry of Health
http://www.salute.gov.it/
Project GR-2008-1143201 "Non-invasive tools for early detection of Autism Spectrum Disorders"

Ente Cassa Risparmio di Firenze
Via Bufalini 6, Firenze
http://www.entecarifirenze.it/

CONTENTS

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

MAVEBA
2013
Firenze, Italy

FOREWORD

As organizer and chairperson of this conference, I would like to express to all the participants my warmest welcome at the 8TH International Workshop MAVEBA2013, which takes place once again in Firenze, Italy, after 14 years since the first edition in 1999.

This event, never discontinued over the years, has now reached full maturity and fully expresses what was the original aim, namely to collect contributions of multidisciplinary research in the increasingly extensive field of the study of issues related to the human phonatory apparatus.

In fact, during these years there has been a continuous parallel expansion in clinical research and technology devoted to this field. This has led to an increasing need for interaction between researchers in technological and clinical disciplines, with extremely positive results as evidenced by the numerous papers presented at this Workshop that, for the first time, covers three whole days.

I am therefore confident that this cooperation will continue and grow in the future.

The eighth edition of MAVEBA is characterized by two Special Sessions, one devoted to the investigation of human neurological diseases of the vocal apparatus and related methods of classification (held by Prof. Shimon Sapir, Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel) and the other to the analysis of the complex mechanisms that regulate the neonatal cry as early indicator of autism spectrum disorders (held by Maria Luisa Scattoni, Department of Cell Biology and Neuroscience,Istituto Superiore di Sanità, Roma, and Dr. Silvia Orlandi, Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy, and presented by Prof. Philip Sanford Zeskind, Director of Neurodevelopmental Research Levine Children's Hospital, Carolinas Medical Center, Charlotte, North Carolina, U.S.A.).

Other equally important subjects are exploited in more sessions: two are dedicated to the basic theme of the workshop, namely modelling and analysis of the voice signal, one to the analysis of endoscopic video images at high speed, very timely topic due to technological advances in this area, two other sessions are devoted to the study of the characteristics of the singing voice, an emerging field of research with important implications in the field of occupational voice disorders, one to the development of devices for voice monitoring, nowadays of great socio-economical impact, and, last but not least, three shorter sessions devoted respectively to the impact on the quality of the voice of neurological diseases, psychiatric disorders and factors related to sex and hormonal therapies.

As always, the three intensive days of the workshop (16, 17 and 18 December 2013) will be also an opportunity for participants to visit places of Firenze not included in the traditional touristic routes and to attend musical events. At the Military Geographical Institute, the historic building housing the workshop, a welcome cocktail will be offered followed by a concert at the Conservatorio Luigi Cherubini, while the visit of the Casa Martelli museum, one of the oldest and most renowned Florentine families, will be followed by a gala dinner in its beautiful ballroom.

This year and for the first time the MAVEBA 2013 Workshop has instituted two awards: MAVEBA Best Paper Award and MAVEBA Best Poster Award. They are intended to encourage and reward the scholarly efforts of early-stage investigators in voice modelling and analysis. The awards are offered by the Journal: Biomedical Signal Processing and Control, Elsevier Ltd., and will be given to young presenters with accepted papers that are judged by a scientific committee to have the greatest positive impact on the mission and quality of MAVEBA 2013.

Hoping that this initiative will stimulate young researchers, I therefore wish to express my gratitude to this important international journal that, as for past editions, will publish a special issue collecting MAVEBA's most significant contributions.

My thanks also go to the anonymous reviewers of the papers and to the Committee for the selection of winners of the awards for young researchers, who have freely devoted part of their valuable time to the success of the Workshop.

But most of all I thank the participants that with the high level of their papers make this 8th MAVEBA Workshop an event of great scientific relevance worldwide.

Lastly, I want to devote special thanks to my co-workers Silvia, Andrea and Davide, and to my friends that since many years support and sustain me, now more than ever after the loss of who more than anyone believed in me until the end and that, I am sure, continues to encourage me wherever he is.

Claudia Manfredi
Conference Chair

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Special session: Early detection of neurologic diseases by acoustic speech analysis and machine learning and classification

Organizer: Prof. Shimon Sapir, Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel

Introduction: Prof. Shimon Sapir, Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel

EARLY MOTOR SIGNS OF PARKINSON'S DISEASE DETECTED BY ACOUSTIC SPEECH ANALYSIS AND CLASSIFICATION METHODS

S. Sapir[1], E. Sprecher[1], S. Skodda[2]

[1]Departments of Physiotherapy and Communication Sciences and Disorders, University of Haifa, Haifa, Israel, sapir@research.haifa.ac.il
[2]Department of Neurology, Knappschaftskrankenhaus, Ruhr-University of Bochum, Bochum, Germany, sabine.skodda@kk-bochum.de

***Abstract* - Purpose: Parkinson's disease (PD) is a slowly progressing and highly debilitating disease. By the time it is diagnosed there is already substantial damage to the central nervous system. There is no medical treatment yet to prevent or decelerate the disease process. Brain imaging and other technological methods can detect the disease earlier than by clinical examination, but such technology is extremely expensive. Speech abnormalities might be among the earlier manifestations of the disease. However, they might be too subtle to be detectable perceptually. Acoustic analysis of speech is objective, valid and inexpensive method. The purpose of this study was to find predictors of early motor signs of Parkinson's disease (EMSPD) by acoustic speech analysis and classification methods. Methods: Twenty seven individuals with EMSPD (mean age= 63.56±10.50; H&Y=1.59±0.42; UPDRS (motor)= 19.07±8.38; years since diagnosis= 1.48±0.51), all optimally medicated during the study, and 86 healthy, age-matched, controls participated in the study. They sustained vowel phonation and read a paragraph. Potential predictors of PD risk were age, gender, and acoustic measures of vowels, voice fundamental frequency, temporal aspects speech articulation, and measures of vocal stability. Results: A multivariate stepwise selection model process yielded four surviving predictors, all reflecting vocal and articulatory instability. ROC area under the curve (AUC) was 0.905. At logistic regression probability 38% or higher, sensitivity was 78.8%, specificity 88.1%, with overall 85.5% correct prediction. Conclusions: Detection of EMSPD by speech acoustic analysis and classification methods is feasible. Whether these methods can detect speech abnormalities in the prodromal /preclinical stage is yet to be explored.**

***Keywords*: voice analysis, classification, early detection, Parkinson's disease**

I. INTRODUCTION

Parkinson' disease (PD) is a slowly progressive and highly debilitating CNS disease. By the time it is firmly diagnosed via routine clinical neurological examination, there is already substantial damage to the CNS [1]. Dysarthria is present in 70-90% of individuals with PD [2]. It has been suggested that subtle signs of the dysarthria, detectable only by acoustic methods, might serve as biomarkers to help detect the presence of the disease in its early stages [3,4,5]. The purpose of this study was to determine whether acoustic analysis of speech and classification methods can help detect early motor signs of PD (EMSPD).

II. METHODS

Subjects. Twenty seven (27, M=16, F=11) individuals with EMSPD (mean age= 63.56±10.50; H&Y= 1.59±0.42, range:1.0-2.0; UPDRS (motor) =19.07±8.38, range:5-32); Years since diagnosis=1.48±0.51, range:1-3) and 86 healthy controls (M=42, F=44, mean age=64.78+8.45) participated in the study. The PD individuals were optimally medicated during the study.

Speech Tasks. Participants sustained vowel phonation and read aloud a paragraph in German, as described elsewhere [6].

Acoustic analyses. The speech was analyzed acoustically [6], with measures of vowels and consonants, voice quality and stability, prosodic pitch inflection, and pauses and rhythmic aspects of speech. The potential predictors of PD risk were age; gender; and the following acoustic measures: the first (F1) and second (F2) formants of the vowels /i/, /a/, and /u/; Vowel Articulation Index (VAI); the mean fundamental frequency (Fo) of each of the vowels; the mean Fo of words; standard deviation of Fo of words; temporal measures: Net Speech Rate (NSR), i.e., total speech time (TST) minus total pause time (TPT); Pause ratio (PR), i.e., % of TPT re: TST; and percent

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

pauses within multisyllabic words (Pinw); and measures of vocal instability: shimmer (SHIMM) and jitter (JIT).

Statistical analyses. Besides basic examinations of data and their distributions, preliminary univariate logistic regression analyses of potential predictors of subject status (EMSPD or healthy control) were performed. These basic analyses employed JMP (SAS Institute, Cary, NC).

The next statistical analysis consisted of a stepwise selection logistic regression procedure, to determine and optimize a set of vocal characteristic predictors for subject status. Age and gender were also included to adjust for their influence in this study group. PROC LOGISTIC of SAS (SAS Institute, Cary, NC) was employed for statistical analysis, with default options for stepwise selection. Model diagnostic procedures and ROC analyses were performed, and P values, odds ratios and associated 95% confidence intervals, along with optimized model cutoffs, were determined. Note that, as will all such selection procedures, probabilities associated with the derived performance parameters should be regarded as potentially inflated.

III. RESULTS

The multivariate stepwise selection model process yielded four surviving predictors: SHIMM, NSR, PR and Pinw. The overall model was significant on LR (Likelihood Ratio), Score and Wald statistics, P<0.0001 for all. The Hosmer and Lemeshow test was not significant, P=0.38, indicating adequate model fit. Wald Chi-square statistics and Odds Ratios and 95% CIs for individual predictors are provided in the table below.

Parameter	df	Estimate	Stand. Error	Wald Chi-Square	p	OR	95% Wald OR Confidence Limits	
Intercept	1	-6.9794	3.3432	4.3582	0.0368			
SHIMM	1	0.3613	0.1079	11.2061	0.0008	1.435	1.162	1.773
NSR	1	1.2982	0.5334	5.923	0.0149	3.663	1.288	10.419
PR	1	-0.1522	0.0603	6.3767	0.0116	0.859	0.763	0.966
Pinw	1	-0.0619	0.0223	7.6833	0.0056	0.94	0.9	0.982

Note: OR is based on a single whole unit change in the predictor, for predicting PD. As the units of measurement differ among the predictors, their relative magnitude does not necessarily indicate their relative influence in the same way that P value does. However, note that OR values above 1 indicate greater values of the predictor predispose toward PD, where OR values below 1 indicate a predisposition against PD. The ROC area under the curve (AUC) was 0.905, indicating good overall sensitivity and specificity for various model cutoff values. At the logistic regression default probability 50% or better cutoff assignment to PD default, sensitivity was 57.6%, specificity 91.7%, with overall 82.1% correct prediction. Optimization of the cutoff for maximum overall correct prediction (at logistic regression probability 38% or higher), sensitivity was 78.8%, specificity 88.1%, with overall 85.5% correct prediction.

IV. DISCUSSION

Early detection of PD is extremely important to prevent or halt this debilitating disease. At this point there is no medical intervention that effectively treats the disease. Until such treatment is available, there is a need to develop biomarkers that can be sensitive to the disease and its progression. According to the

model proposed by Braak et al. [1], the disease initially affects brain stem motor systems such as the glossopharyngeal and vagal nerves. These nerves are likely to affect phonatory and articulation movements. The present findings suggest that there are speech abnormalities at EMSPD and that these abnormalities are characterized by temporal instability, albeit subtle, of the phonatory (SHIMM) and articulatory (NSR, PR, Pinw) systems. Such abnormalities may be imperceptible to the ears of the patient and clinician, thus the need for acoustic speech analyses and other noninvasive, sensitive, valid, and reliable methods. The present findings are preliminary and their interpretation should considered tentative.

V. CONCLUSIONS

The present study indicates that acoustic speech analysis combined with statistical and classification methods can differentiate between individuals with EMSPD and healthy controls. Recently there have been other studies attempting to detect individuals with EMSPD [5,7,8]. The different studies have identified different acoustic parameters as biomarkers of EMSPD. These differences are most likely related to the different languages of the speakers, different speech tasks, different acoustic measures, and different phonetic inventories. Thus, there is a need to determine which acoustic metrics and methods of analyses best predict the presence of PD in individuals at risk for PD.

REFERENCES

[1] Braak, H., Ghebremedhin, E., Rüb, U., Bratzke, H., Del Tredici, K., et al (2004). Stages in the development of Parkinson's disease-related pathology. *Cell Tissue Res*, 318, 121-34

[2] Sapir, S., Ramig, L, & Fox, C. (2008). Speech and swallowing disorders in Parkinson's disease. *Curr Opin Otolaryngol Head Neck Surgery*, 16, 205-210.

[3] Adler, C. (2011). Premotor symptoms and early diagnosis of Parkinson's disease. *Int J Neurosci*, 121 Suppl 2, 3-8.

[4] Becker G, Müller A, Braune S, Büttner T, Benecke R, Greulich W, Klein W et al. (2002). Early diagnosis of Parkinson's disease. *J Neurol*. 249 Suppl 3:III/40-8.

[5] Hazan, H., Hilu, D., Manevitz, L., & Sapir, S., (2012). Early Diagnosis of Parkinson's Disease via Machine Learning on Speech Data, in *2012 IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, 2012, pp. 1 –4. [proceedings]

[6] Skodda, S., Flasskamp, A., & Schlegel, U. (2010). Instability of syllable repetition as a model for impaired motor processing: is Parkinson's disease a "rhythm disorder"? *J Neur Trans*, 117, 605-12.

[7] Rusz, J., Cmejla, R., Ruzickova, H., & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *J Acoust Soc Am*, 129, 350-67.

[8] Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., Picmausova, J., Roth, J., Ruzicka, E. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. *J Acoust Soc Am*, 134, 2171-81

STEADINESS OF SYLLABLE REPETITION IN EARLY MOTOR STAGES OF PARKINSON´S DISEASE

8TH INTERNATIONAL MAVEBA WORKSHOP

S. Skodda
Department of Neurology, Knappschaftskrankenhaus, Ruhr-University of Bochum,
sabine.skodda@kk-bochum.de

Abstract: **Patients with Parkinson´s disease (PD) show characteristic abnormalities in the performance of simple repetitive movements which can also be observed concerning speech rate and rhythm. The aim of the current study was to survey if patients with early PD already feature impairments of steady vocal pace performance based upon a simple syllable repetition paradigm. N = 31 patients with PD with mild to moderate motor impairment and n = 32 age-matched healthy controls were tested. Participants had to repeat a single syllable or a pair of alternating syllables in a self chosen steady pace or in a given pace of 80/min. As main result, vocal pace performance was observed to be irregular in all tasks, but showed a further decline when two equal demands (keeping the steady pace and alternating the syllables) were present indicating an additional impairment when the complexity of the task was increased as a possible hint for executive dysfunction already detectable in the early motor stages. Given that subsequent studies are able to confirm these preliminary results, analysis of steadiness of syllable repetition might be a promising non-invasive tool for detection of subtle abnormalities of motor speech performance even in the early motor stages of PD.**

Keywords: **Parkinson´s disease, dysarthria, motor speech performance, repetitive movements, steadiness of pace**

I. INTRODUCTION

Parkinson´s disease (PD) is a chronic progressive neurodegenerative neurological disease with a veraity of motor and non-motor symptoms. According to the prevailing concept of the Braak stages, Lewy bodies as the neuropathological hallmarks of neurodegeneration in PD can initially be found in the olfactory bulb and lower brain stem nuclei years before the involvement of the dopamine producing cells in the substantia nigra pars compacta [1]. However, first motor signs of PD do not occur until a substantial number of dopaminergic midbrain neurons damage are degenerated [2]. And even in these early motor stages of PD, the clinical signs as akinesia, tremor and rigidity are often subtle and inconclusive leading to a delay of the diagnosis and of the initiation of treatment [3]. Therefore, there is an urgent need for the establishment of meaningful and easy appliable clinical tests to facilitate an early diagnosis of PD.

Abnormalities of the steady performance of simple repetitive "automated" movements are well-known features of PD and can be identified in different motor modalities as hand and finger movements, gait and also in Parkinsonian hypokinetic dysarthria. This characteristic pattern of "motor instability" throughout the performance is thought to be induced by the complex dysfunction of planning, preparing, scaling and maintaining a once chosen simple motor program as a consequence of the underlying basal ganglia dysfunction [4]. Since speech can be subdivided down to the level of single utterances, one might expect abnormalities of vocal pace performance already on the level of very basic non-speech articulatory gestures. Indeed, in previous studies, our group had been able to show that patients in different stages of PD featured marked difficulties to steadily repeat a single syllable without changing the speed of the repetition [5,6].

The aim of the current study was to survey, if these abnormalities of steady vocal pace performance are already detectable in the early motor stages of PD.

II. METHODOS

N = 31 patients with PD (20 male) with mild to moderate motor impairment and n = 32 age-matched healthy controls (19 male) were tested. In the patients´ group Hoehn&Yahr /H&Y stages ranged from 1.5 to 2 (average H&Y 1.82, standard deviation/ SD 0.24) and the average Unified Parkinson´s Disease Motor Score /UPDRS III was 12.74 pts. (range from 5 to 21, SD 4.28). To the time of examintaion, all patients were under stable but uncontrolled regimen of dopaminergic medication for at least four weeks. Speech and motor examinations were

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

performed 60 to 90 minutes after the morning dose of medication to ensure the "on"-state.

Speech samples were digitally recorded using a commercial audio software and a head-set microphone The speech task consisted of four subtests which have been described in detail in one previous study of our group [6]. Test 0: Participants had to reiterate the syllables /pa/ and /pa-ti/ as fast as possible for at least 5 seconds for the description of maximum syllable repetition capacity (maxSylRep in syllables per second). Test 1: Repetition of the syllable /pa/ in a self chosen steady (isochronous) pace without acceleration or slowing articulatory velocity. Test 2: Repetition of the syllable /pa/ in a velocity of 80/min given by a metronome; participants had to listen to the pace first, then start with the syllable repetition; the metronome was stopped after four utterances, and participants had to keep the given pace. Test 3: Alternating repetition of the syllables /pa/ and /ti/ with the given metronome-based velocity of 80/min.

Each subtest was performed twice; the average values of first and second cycle were taken for the definite analyses. In each test the participants were asked to repeat the syllables at least 40 times. Only the first 30 utterances were taken for the definite analyses in order to avoid a modification of participants´ articulatory velocity by the expectance of the imminent end of the task. Based upon the oscillographic sound pressure signal of the recorded audio material, the period from onset of one vocalization until the following vocalization was defined as "interval"; interval duration (IntDur) was measured manually in milliseconds (ms). Stability of pace of the utterances was defined as relative coefficient of variation (COV_{5-30}) calculated for the intervals 5 to 30 in relation to the average interval length of the first 4 utterances ($avIntDur_{1-4}$) following the formula: $COV_{5-30} = SD_{5-30} / [(avIntDur_{1-4}) / \sqrt{26}] \times 100$. Furthermore, in test 3 (alternating repetition of a pair of syllables in a given pace), the average interval duration of the first syllable /pa/ was related to the average interval duration of the second syllable /ti/ ("/pa/-/ti/ ratio").

Winstat© was used for statistical analyses. T-test for independent groups was performed, since the variables were normally distributed (Shapiro-Wilk test). The adjusted level of significance was set at $p = 0.008$. Pearson correlation was used to test for significant correlations.

III. Results

Patients with PD showed significant aberrations of the steadiness of pace throughout the performance of different syllable repetition tasks (table 1). COV in the PD group was elevated in the tasks consisting of a single syllable repetition in self-chosen pace and in the task with repetition of single syllables in a given pace as well. A further deterioration of steadiness of pace was observed in the task with repetition of alternating syllables in a given pace. Furthermore, the /pa/-/ti/ ratio was significantly reduced in the PD group indicating that PD speakers connected the syllables /pa/ and /ti/ at the expense of keeping the steady pace.

No correlations were seen between the measures of the vocal pace performance and patients´ characteristics as age, disease duration, H&Y stage and UPDRS motor scores.

	control group	**PD group**	
max SylRep	3.72 ± 1.33	3.79 ± 0.98	n.s.
COV_{test1}	1.00 ± 0.28	1.62 ± 0.96	p=0.001
avIntDur	467 ± 176	507 ± 237	n.s.
COV_{test2}	0.96 ± 0.35	1.64 ± 1.03	p=0.001
COV_{test3}	1.05 ± 0.44	1.76 ± 0.81	p<0.001
pa-ti ratio	0.987 ± 0.06	0.936 ± 0.010	p=0.015

table 1: comparison of the results between control and PD

IV. Discussion

According to the current data, impairment of steady vocal pace performance can be already detected in the early motor stages of PD although the abnormalities are less pronounced than in previous investigations of our group performed in a mixed group of patients with a much higher disease duration and overall motor disability [5,6]. Interestingly, in the current study, performance of syllable repetition was similar in all three tests and did not significantly worsen with higher complexity (keeping a given pace, alternating syllables) which we had previously found and interpreted as a hint for disturbed executive function [6].

The present data seem to corroborate the hypothesis of an early disruption of basic speech motor performance caused by dysfunctional basal ganglia networks with instability of basic motor programs which normally run in quasi automated mode.

However, the present study has some important limitations since the patients were under uncontrolled regimen of dopaminergic medication and therefore, treatment effects on vocal pace performance cannot be ruled out. Furthermore, all PD patients were in H&Y stage 1.5 or 2 which means mild to moderate motor impairment, however, with already present axial symptoms (e.g. akinesia and/or rigidity of neck and facial muscles, disturbed posture and/or gait) which clinically indicates bilateral neurodegeneration of the substantia nigra as some supposed predisposition for motor speech

impairment. Since all patients had previously been diagnosed to suffer from PD according to established clinical criteria, our findings cannot answer the question, if abnormal vocal pace performance is already present in the "grey area" of very early motor PD when accepted clinical criteria are often not yet applicable for the diagnosis of PD.

V. Conclusion

Investigation of pace and steadiness of syllable repetition is a easily applicable, non-intrusive and cost effective method which reveals significant abnormalities of basic motor speech performance in patients with mild to moderate motor stages of PD. Further longitudinal studies are warranted in drug-naïve patients with supposed, but still not verifiable PD to survey, if the proposed method can become a helpful tool for the early diagnosis of PD.

References

[1] H. Braak, K. Del Tredici, U. Rüb, R.A. de Vos, E.N. Jansen Steur and E. Braak. "Staging of brain pathology related to sporadic Parkinson´s disease," *Neurobiol. Aging*, vol. 24, pp.197-211, 2003.

[2] M. Guttman, J. Burkholder, S.J. Kish, D. Hussey, A. Wilson, J. DaSilva and S. Houle. "[11C]RTI-32 PET studies of the dopamine transporter in early dopa-naive Parkinson's disease: implications for the symptomatic threshold," *Neurology,* vol. 48, pp.1578-1583, 1997.

[3] B. Pakkenberg, A. Moller, H.J. Gundersen, A. Mouritzen Dam and H. Pakkenberg. „The absolute number of nerve cells in substantia nigra in normal subjects and in patients with Parkinson's disease estimated with an unbiased stereological method," *J. Neurol. Neurosurg. Psychiatry*, vol. 54, pp. 30-33, 1991.

[4] R. Iansek, J.L. Bradshaw, J.G. Phillips, R. Cunnington and M.E. Morris. "Interaction of the basal ganglia and supplementary motor area in the elaboration of movements," In: *Motor control and sensory motor integration: Issues and directions*, D.J. Glencross, J.P. Piek (eds.) Elsevier Science BV, 1995, pp. 37-59.

[5] S. Skodda, A. Flasskamp, U. Schlegel. „Instability of syllable repetition as a model for impaired motor processing: is Parkinson's disease a "rhythm disorder"?", *J. Neural. Transm.,* vol. 117, pp. 605-612, 2010.

[6] S. Skodda, J. Lorenz, U. Schlegel. „Instability of syllable repetition in Parkinson's disease—Impairment of automated speech performance?" *Basal Ganglia*, vol.3, pp.33-37, 2013.

ACOUSTIC FINDINGS OF VOICE DISORDERS IN HUNTINGTON'S DISEASE COMPARED TO PARKINSON'S DISEASE

J. Rusz[1,2], J. Klempir[2], E. Baborova[2], T. Tykalova[1], V. Majerova[2], R. Cmejla[1], E. Ruzicka[2], J. Roth[2]

[1] Department of Circuit Theory, Czech Technical University in Prague, Faculty of Electrical Engineering, Prague, Czech Republic
[2] Department of Neurology and Centre of Clinical Neuroscience, Charles University in Prague, First Faculty of Medicine, Prague, Czech Republic
ruszjan@fel.cvut.cz

Abstract: **One common finding in Huntington's disease (HD) is related to phonatory disruptions that can be perceptually characterized by harshness, strained-strangled voice quality, and pitch fluctuations. These alterations of voice occur mainly as a consequence of underlying involuntary contractions, variable muscle tone, or even tremor of laryngeal musculature. Recently, several new acoustic analysis methods have been introduced to capture different aspects of these phonatory abnormalities. In this report, we summarize objective acoustic metrics suitable for assessment of phonatory dysfunction and provide their classification accuracy in separation between patients with HD and healthy controls. For this purpose, data consists of 272 phonations collected from 34 individuals with HD and 34 healthy controls. As impairment of phonatory function in HD was found across all investigated measurements, voice analysis may potentially serve as a marker of disease progression.**

Keywords: **Huntington's disease, hyperkinetic dysarthria, dysphonia, acoustic analysis, classification.**

I. INTRODUCTION

Huntington's disease (HD), which is caused by an expansion of the number of CAG repeats located on the short arm of chromosome 4 at 4p16.3 [1,2], is a chronic, degenerative, neuropsychiatric disorder, characterized by progressively increasing of choreiform movements,. In the course of the illness, the patients with HD typically develop a distinctive alteration of speech termed as hyperkinetic dysarthria [3]. Hyperkinetic dysarthria in HD is mainly affected by the involuntary contractions of speech mechanism musculature, occurring mainly as a consequence of underlying choreatic movements. Such involuntary contractions of vocal muscles can especially transcend during speaking task such as sustained vowel phonation which demands stable coordination of the jaw, tongue, palate, and facial movements. Recently, we have introduced several metrics that were sensitive to differentiate between healthy and HD voices [4]. The aim of the current study was to review the most successful algorithms to capture phonatory dysfunction in HD and investigate their ability to predict HD membership.

II. METHODS

A. Data

The data for this study were collected as the part of the previous study [4]. From 2011 to 2012, a total of 34 Czech native participants (15 men and 19 women) with genetically verified HD were recruited. Their mean age was 45.2 ± SD 13.3 (range 23–67) years, mean age at HD onset was 39.3 ± 13.5 (14–62) years, mean disease duration 5.9 ± 3.1 (2–16) years, and average number of CAG triplet repeats 46.4 ± 5.8 (40–70). As a control group, 34 persons (15 men and 19 women) of comparable age, mean age 45.5 ± 13.6 (range 24–68) years, with no history of neurological or communication disorders were included. None of the participants had undergone voice therapy and all gave their consent to the vocal tasks and recording procedure. Every subject was instructed to perform sustained phonation of the vowel /a/ and vowel /i/, each one repeated two times.

B. Acoustic measurements

Acoustic analyses were performed using several phonatory measurements in order to investigate different aspects of speech in HD patients and controls. To assess airflow insufficiency, we examined maximum phonation time (MPT) [5], and MPT until the occurrence of the first voice break (MPT_{VB}) [4]. To investigate aperiodicity, we evaluated number of voice breaks (NVB) and degree of voicelessness (DUV) [6]. With respect to irregular

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

vibrations of vocal folds, we extracted fundamental frequency variations (F0 SD) [7], recurrence period density entropy (RPDE) [8], and pitch period entropy (PPE) [9]. To examine signal perturbations, we investigated jitter and shimmer [6]. To capture problems with increased noise, we calculated harmonics-to-noise ratio (HNR) [6], and fluctuation analysis (DFA) [8]. Finally, we have also introduced new acoustic parameter related to articulation deficiency based upon mel-frequency cepstral coefficients (hereinafter, MFCC) [4], which was defined as the mean of the standard deviations of the 1st-12th MFCCs using the implementation of Brooke's Matlab toolbox [10].

C. Classification experiment

Each designed acoustic feature underwent classification experiment, where support vector machine (SVM) with Gaussian radial basis kernel was used to decide whether the speech performance belongs to HD or control speaker. The cross-validation scheme was applied where all data (136 phonations of HD patients and 136 phonations of controls) were randomly separated into training (80%) and testing (20%) subsets; the process of cross-validation was repeated 20 times for each parameter.

III. Results

According to the SVM classifier, four metrics including MPT, MPT_{VB}, F0 SD, and MFCC achieved greater classification accuracy exceeding 80% in differentiation between HD and control speakers (Table 1). The best single parameter reflecting phonatory dysfunction in HD was found to be MPT_{VB} with classification accuracy of 89.4 ± 3.9% (sensitivity: 91.8 ± 4.9%; specificity 87.9 ± 5.5%). This parameter represents sudden phonation interruptions and can be associated with motor impersistence, which is the inability to sustain certain simple voluntary act such as keeping the tongue protruded or maintaining a firm grip.

Table 1: List of classification results of acoustic phonatory measures with mean and standard deviation (SD) values for differentiation between patients with HD and healthy controls.

Parameter	Classification score % (Mean ± SD)			Rank
	Overall	Sensitivity	Specificity	
Airflow insufficiency				
MPT	85.5 ± 4.6	92.1 ± 5.5	81.3 ± 5.4	3rd
MPT_{VB}	89.4 ± 3.9	91.8 ± 4.9	87.9 ± 5.5	1st
Aperiodicity				
NVB	65.5 ± 6.1	80.9 ± 9.4	60.5 ± 4.0	9th
DUV	72.8 ± 6.1	93.8 ± 5.7	65.7 ± 4.3	6th
Irregular vibrations of vocal folds				
F0 SD	84.9 ± 4.3	92.3 ± 4.6	80.2 ± 5.1	4th
RPDE	79.9 ± 5.4	86.1 ± 6.9	76.0 ± 6.0	5th
PPE	68.5 ± 6.1	68.6 ± 6.7	69.2 ± 6.9	7th
Signal perturbations				
Jitter	63.8 ± 5.8	66.7 ± 6.7	62.1 ± 5.6	10th
Shimmer	62.5 ± 5.9	67.3 ± 9.0	60.3 ± 4.9	12th
Increased noise				
HNR	62.9 ± 5.8	67.4 ± 8.6	60.7 ± 4.7	11th
DFA	66.1 ± 5.1	69.8 ± 6.3	63.9 ± 4.6	8th
Articulation deficiency				
MFCC	88.8 ± 3.6	92.4 ± 4.6	86.2 ± 4.9	2nd

MPT = maximum phonation time, MPT_{VB} = maximum phonation time until first break, NVB = number of voice breaks, DUV = degree of voicelessness, F0 SD = variability of fundamental frequency, RPDE = recurrence period density entropy, PPE = pitch period entropy, HNR = harmonics-to-noise ratio, DFA = detrended fluctuation analysis, MFCC = mel-frequency cepstral coefficient.

IV. Discussion

The current study shows the potential of voice analysis in documentation the degree and patterns of hyperkinetic dysarthria in HD. The patients with HD showed deterioration in all measured parameters, however, the most prominent pattern of dysphonia was related to sudden phonation interruptions with classification accuracy up to 90% in prediction of HD group membership.

Our findings are in accordance with previous studies reporting voice in HD patients as harsh, breathy, strained-strangled with irregular pitch fluctuations and arrests [5,11-13]. Considering main phonatory deficits in patients with HD revealed in this study from physiological point of view, we can hypothesize that (a) airflow insufficiency and aperiodicity reflected by sudden phonation interruptions are a consequence of choreatic contractions, abnormal muscle tone, or hyper-adduction of vocal folds, (b) articulation deficiency is mainly caused by problems in coordination of articulators including misplacement of tongue, lips, jaw, and face, whereas (c) irregular vibrations of vocal folds manifested as pitch fluctuations occur as a consequence of inefficient nervous system control.

In fact, recognizing of specific signs of speech and voice disorders can provide important clues about the etiology of the disease, and may be useful in differential diagnosis [3,14,15]. Comparing the current finding of hyperkinetic dysarthria in HD patients to better described hypokinetic dysarthria in Parkinson's disease (PD) patients, we can note several differences. Both hyperkinetic and hypokinetic dysarthrias manifest decreased quality of voice (breathiness, harshness, hoarseness) [16]. In contrast, the higher incidence of voice breaks seems to be more specific for hyperkinetic dysarthria. Slight misplacement of articulators during phonation captured by MFCC has also been shown in PD

[17], whereas parkinsonian patients do not manifest such marked pitch fluctuations as observed in HD subjects [7]. Table 2 summarizes main results for HD group and compares it to previous findings in PD group.

V. Conclusion

A precise description of vocal patterns may significantly contribute to existing assessment batteries for monitoring disease onset and progression, and may be beneficial in the differential diagnosis of movement disorders. In addition, a qualitative description of voice dysfunction may be helpful to gain better insight into the pathophysiology of the vocal mechanism. In practice, the measurement of speech is non-invasive, fast, easy to apply, and inexpensive. Future studies combining various aspects of voice may extend our knowledge to identify longitudinal changes of phonatory dysfunction in HD patients as well as in subjects at risk for HD.

Acknowledgment

This research was supported by the Czech Science Foundation (GACR 102/12/2230) and Czech Ministry of Education (MSM 0021620849).

Table 2: Summary of results: comparison of voice features between patient and controls groups for HD and PD.

Parameter	Group	
	HD	**PD**
Airflow insufficiency		
MPT	↑↑↑	—
MPT_{VB}	↑↑↑	—
Aperiodicity		
NVB	↑	—
DUV	↑↑↑	—
Irregular vibrations of vocal folds		
F0 SD	↑↑↑	—
RPDE	↑↑↑	↑↑
PPE	↑↑↑	↑↑
Signal perturbations		
Jitter	↑↑	↑↑↑
Shimmer	↑↑	↑↑↑
Increased noise		
HNR	↑↑↑	↑↑↑
DFA	↑↑↑	—
Articulation deficiency		
MFCC	↑↑↑	↑

—: no difference, ↑ slightly affected ($0.01 \leq p < 0.05$), ↑↑ affected ($0.001 \leq p < 0.01$), ↑↑↑ markedly affected ($p < 0.001$).
¥ For the purposes of comparison, the data for PD group were adopted from our previous study [18]. Note that HD and PD groups have different characteristic related to duration and severity of disease.

References

[1] Kremer B, Goldberg P, Andrew SE, Theilmann J, Telenius H, et al. (1994) A worldwide study of the Huntington's disease mutation: The sensitivity and specificity of measuring CAG repeats. New Engl J Med 330: 1401–1406.
[2] Hayden MR (1981) Huntington's chorea. New York, Springer-Verlag, pp. 59–92.
[3] Duffy JR (2005) Motor Speech Disorders: Substrates, Differential Diagnosis and Management, 2nd ed., Mosby, New York, p. 592.
[4] Rusz J, Klempir J, Baborova E, Tykalova T, Majerova V, et al. (2013) Objective acoustic quantification of phonatory dysfunction in Huntington's disease. PLoS One 8: e65881.
[5] Ramig LA (1986) Acoustic analysis of phonation in patients with Huntington's disease. Preliminary report. Ann Otol Rhinol Laryngol 95: 288–293.
[6] Boersma P, Weenink D (2001) PRAAT, a system for doing phonetics by computer. Glot International 5: 341–345.
[7] Rusz J, Cmejla R, Ruzickova H, Ruzicka E (2011) Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J Acoust Soc Am 129: 350–369.
[8] Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM (2007) Exploiting Nonlinear recurrence and Fractal scaling properties for voice disorder detection. Biomedical Engineering Online 6: 23.
[9] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO (2009) Suitability of dysphonia measurement for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng 56: 1015–1022.
[10] Brookes M (2009) VOICEBOX, Speech Processing Toolbox for Matlab, Department of Electrical & Electronic Engineering, Imperial College.
[11] Zwirner P, Murry T, Woodson GE (1991) Phonatory function of neurologically impaired patients. J Commun Disord 24: 287–300.
[12] Hartelius L, Carlstedt A, Ytterberg M, Lillvik M, Laakso K (2003) Speech disorders in mild and moderate Huntington's disease: Results of dysarthria assessment of 19 individuals. J Med Speech-Lang Pa 1:1–14.
[13] Velasco Garcia MJ, Cobeta I, Martin G, Alonso-Navarro H, Jimenez-Jimenez FJ (2011) Acoustic analysis of voice in Huntington's disease. J Voice 25: 208–217.
[14] Skodda S (2012) Analysis of voice and speech performance in Parkinson's disease: a promising tool for the monitoring of disease progression and differential diagnosis. Neurodegen Dis Manage 2: 535–545.
[15] Kim Y, Kent RD, Weismer G (2011) An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. J Speech Lang Hear Res 54: 417-429.

[16] Sapir S, Ramig L, Fox C (2008) Speech and swallowing disorders in Parkinson's disease. Curr Opin Otolaryngol Head Neck Surg 16: 205-210.
[17] Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO (2012) Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE T Bio-Med Eng 5: 1264–1271.
[18] Rusz J, Cmejla R, Ruzickova H, Klemir J, Majerova V, et al. (2011) Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test. Movement Disord 26: 1951–1952.

Vocalization deficits in translational rodent models of Parkinson disease

M.R.Ciucci[1, 2, 3], L. M. Grant [1,2], C.A. Kelm-Nelson [1], L. Fulks[4], T. Kyser[4], K.B. Seroogy[4], S.M. Fleming[4]

[1]Department of Surgery, Division of Otolaryngology, University of Wisconsin, Madison, WI
[2]Department of Communication Sciences and Disorders, University of Wisconsin, Madison, WI
[3]Neuroscience Training Program, University of Wisconsin, Madison, WI
[4]Department of Neurology, University of Cincinnati, Cincinnati, OH

Abstract: Vocalization deficits are common in Parkinson disease (PD) and can significantly compromise the ability to communicate. These deficits are reported to occur early in the disease and are not responsive to dopaminergic therapies, leaving patients with few treatment options. Over the last several years our laboratory has developed a novel approach of measuring vocalization deficits relevant to PD in rodent models. In the present study we use these methods to measure ultrasonic vocalizations in two novel genetic rat models of PD, PINK1 knockout and DJ-1 knockout rats. We show that PINK1 rats develop vocalization impairments at an early age that persist over time similar to those observed in patients and in toxin models of PD. In contrast, DJ-1 knockout rats display more subtle alterations in ultrasonic vocalizations that differ from PINK1 knockout rats. These findings will be important for identifying pathological correlates related to vocalization deficits as well as potential therapeutic targets for PD.

Key words: Parkinson disease, ultrasonic vocalization, transgenic, rat

I. INTRODUCTION

In Parkinson disease (PD) the development of both sensorimotor and cranial sensorimotor deficits can severely reduce the quality of life for patients. While much is known about the underlying pathology associated with sensorimotor deficits such as bradykinesia, resting tremor, and rigidity, very little is known about the neural correlates of cranial sensorimotor impairments such as vocalization deficits. Vocalization deficits are common in PD and several studies suggest voice impairments may actually precede the cardinal motor signs [1-5], making it an attractive target for mechanistic and preclinical studies. We have shown that cranial sensorimotor function can be assessed in rodent models of PD by measuring the quality of their ultrasonic vocalizations (USVs) including call intensity, bandwidth, duration, and peak frequency [6]. In our studies using the classic unilateral 6-hydroxydopamine rat model of PD, we found call intensity and bandwidth were significantly reduced compared to control rats [6,7]. More recently we have started to study vocalizations in the novel genetic models of PD. The strength of these models is that they have excellent construct validity because they have a mutation known to cause PD in some families [8,9].

They are also amenable to testing the progression of deficits and pathology whereas the toxin models are more of an acute model of the disease [10,11]. In mice overexpressing human alpha-synuclein, a presynaptic protein found in Lewy bodies and involved in multiple inherited forms of PD [12], we found USV deficits early, prior to dopamine content reduction in the striatum [13]. In the present study we are investigating USVs in two novel genetic rat models of PD: PINK1 knockout (KO) and DJ-1 KO rats. PINK1 is associated with mitochondrial function and DJ-1 is involved the oxidative stress response [13, 14]. Both mitochondrial dysfunction and increased oxidative stress have long been implicated in PD pathology. We hypothesize that vocalization deficits will develop in these models and resemble vocalization deficits observed in PD as well as deficits seen in the 6-OHDA and alpha-synuclein models. In the PINK1 KO experiment, PINK1 KO homozygous (PINK1-HOM), PINK1-HET, and wild type (WT) rat USVs were analyzed at 2, 4, 6, and 8 months of age. For the DJ-1 experiment, homozygous DJ-1 KO and WT rat USVs were measured at 9 months of age.

II. METHODS

Animals: Male PINK1 KO and DJ-1 KO rats were generated and maintained on a Long-Evans background strain by SAGE laboratories (Sigma-Aldrich). For the PINK1 experiment the groups included PINK1 KO-HOM n=16, PINK1 KO-HET n=16, and WT n=16. USVs were measured at 2, 4, 6, and 8 months of age. For the DJ-1 experiment, DJ-1 KO (n=6) and WT (n=10) USVs were measured at 9 months of age. Animal care was conducted in accordance with the United States Public Health Service Guide for the Care and Use of Laboratory Animals, and procedures were approved by the Institutional Animal Care and Use Committee at the University of Wisconsin and the University of Cincinnati.

Ultrasonic vocalization: For all experiments, recordings were made using an ultrasonic microphone (CM16, Avisoft, Germany) with 16-bit depth and sampling at a rate of 250-kHz, mounted 15 cm above a standard polycarbonate rat cage. A receptive female rat was placed in a test enclosure containing a male rat. When the male demonstrated interest in the female (sniffing, mounting, chasing), the female was removed and recording captured only male vocalizations for 90

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

seconds. Offline acoustic analysis was performed with a customized automated program using SASLab Pro (Avisoft, Germany). Spectrograms were built from each waveform with the frequency resolution set to a FFT of 512 points, frame size of 100%, flat top window, and the temporal resolution set to display 75 % overlap. From these calls, bandwidth in Hertz (Hz), peak frequency in Hz, intensity in Decibels (dB), and percent complex calls were measured (see Ciucci, et al., 2009 [1] and Johnson et al., 2011 [3] for details). Data will be presented here for the intensity measures.

Statistical Analysis: For the PINK1 experiment, a mixed design 3x4 ANOVA design was employed to compare genotype (PINK1 HOM, PINK1 HET, WT) and age (2, 4, 6 8 months). Post-hoc analyses were performed with Fisher's LSD. For the DJ-1 experiment Student's t-test was used to compare WT and DJ-1 KO rats at 9 months of age. Critical level for significance is 0.05.

III. RESULTS

Results are presented for acoustic data (intensity only) for both PINK1 and DJ-1 experiments.

Average intensity of frequency modulated calls in PINK1 rats. There were significant effects for genotype (Table 1) $F(2, 45)=24.80$, $p<0.0001$) and age $F(3, 110)=4.62$, $p=0.0044$). Homozygous rats demonstrated lower intensity of calls as compared to WT ($p<0.0001$) and heterozygous rats ($p<0.0001$), regardless of age. There was not a significant difference between WT and heterozygous rats. For age effects, all rats at 4, 6, and 8 months of age had significantly louder calls compared to 2 months of age ($p<0.0001$, $p=0.0003$, respectively). In the DJ-1 knockout rats mean intensity was significantly increased compared to WT rats at 9 months of age [$t(14)= 2.29$, $p<0.05$].

Average Intensity of frequency modulated calls in DJ-1 KO rats: In contrast, as shown in Figure 3, DJ-1 KO rats at 9 months of age show an increase in mean call intensity compared to WT [$t(14)= 2.29$, $p<0.05$].

Table 2 shows alterations in intensity observed in the PINK1 and DJ-1 rats and how it compares with our previous studies in other models, including aged rats, unilateral 6-OHDA rats, and the alpha-synuclein overexpressing mouse.

IV. DISCUSSION

Overall, our preliminary results demonstrate homozygous *PINK1* knockout rats show early vocalization (intensity) deficits beginning at two months of age that persist over time and are not progressive. In contrast, homozygous *DJ-1* knockout rats show an increase in call intensity compared to wild type control rats indicating differential effects of the PD mutations on rat USVs. It has been reported that both PINK1 and DJ-1 knockout rats develop nigrostriatal dopamine cell loss beginning at approximately 8 months of age. This suggests that the early vocalization deficits in the *PINK1* knockout rats may be associated with dopaminergic dysfunction preceding cell loss and/or extranigral dysfunction. In the present study analysis of nigrostriatal dopamine call counts is currently ongoing to confirm dopamine cell loss as well as to help in understanding the relationship between dopamine cell loss and alterations in USVs. In addition, the *PINK1* knockout rat will be a useful model to study the mechanisms associated with early vocalization deficits as it relates to PD.

Table 1: The effect of PINK1 knockout or DJ-1 knockout in rats on ultrasonic vocalizations: Call Intensity (dB)

Age (months)	**Pink1 Knockout**			**DJ-1 Knockout**		
	Max	**Mean**	**Top 10**	**Max**	**Mean**	**Top 10**
2	↓	↓	↓	NM	NM	NM
4	↓	↓	↓	NM	NM	NM
6	↓	↓	↓	NM	NM	NM
8-9	↓	↓	↓	NS	↑	NS

↓ represents significant decrease compared to wild type and heterozygous control rats. ↑ represents significant increase compared to wild type control rats ($p<0.05$). NS= non-significant, NM=not measured at that age.

Table 2: Ultrasonic vocalization deficits in aging and toxin and genetic models of Parkinson disease

Model	**Affect on Intensity (dB)**
Aging Rats (32 months)	↓ Max and Average
Unilateral 6-OHDA Rat	↓ Max
Alpha-Synuclein Overexpressing Mouse	↓Range
PINK1 Knockout Rat	↓ Max, Average
DJ-1 Knockout Rat	↑ Average

6-OHDA= 6-hydroxydopamine

Intensity differences were also observed in the *DJ-1* knockout rats however, the *DJ-1* knockout rats displayed an increase in intensity. In the DJ-1 experiment rats were tested only at one age and therefore it is unclear whether the alterations in USV intensity develop over time and eventually become deficits. Current studies are aimed to characterize the time course of USV alterations in this mutant line. Of interest is that sensorimotor function was also measured in both types of mutant rats and the alterations observed in sensorimotor function are consistent with the USV results in both types of knockout rats.

Our laboratory has also observed reduced intensity in rat USVs in other models of PD, including the classic

unilateral 6-OHDA rat and the alpha-synuclein overexpressing mouse [6,7]. Decreased vocalization intensity is frequently observed in PD patients and significantly compromises the ability of patients to communicate. The inability to communicate has a dramatic negative effect on the quality of life for patients.

It is well established that PD is a systemic disorder affecting multiple brain structures and neurotransmitters systems. The vast number of non-motor symptoms and cranial motor symptoms that develop in the disease and are not responsive to dopaminergic therapy highlights the broad effect the disease has on the brain and body. Therefore, genetic models of PD are potentially the optimal models to use for studying the pathological mechanisms underlying early pathological events in PD. Importantly, as has been suggested in the human literature, vocalization deficits may be an early behavioral biomarker for PD and defining the pathophysiology related to voice deficits is essential to developing new treatment targets for PD-related dysphonia.

V. CONCLUSION

Although the majority of cases of PD are sporadic, the discovery of specific mutations in genes that cause familial forms of PD has led to the development of a novel class of models for PD, genetic models. The strength of the genetic models over the acute toxin models is that they can provide valuable information on early dysfunction and progression of the disease. In the present study we show that the genetic models may be highly useful in studying vocalization deficits and potential therapeutics in PD. Future studies will focus on identifying the key brain structures pathological mechanisms involved in vocalization deficits in PD.

REFERENCES

[1] Rusz J, Cmejla R, Ruzickova H, Ruzicka E. Quantitative acoustic measurements for characterization of speech andvoice disorders in early untreated Parkinson's disease. The Journal of the Acoustical Society of America 2011;129:350.

[2] Stewart C, Winfield L, Hunt A, Bressman SB, Fahn S, Blitzer A, Brin MF. Speech dysfunction in early Parkinson's disease. Movement disorders 1995;10:562.

[3] Harel B, Cannizzaro M, Snyder PJ. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study. Brain and cognition 2004;56:24.

[4] Harel BT, Cannizzaro MS, Cohen H, Reilly N, Snyder PJ. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment. Journal of Neurolinguistics 2004;17:439-453.

[5] Rusz J, Cmejla R, Ruzickova H, Klempir J, Majerova V, Picmausova J, Roth J, Ruzicka E. Evaluation of speechimpairment in earlystages of Parkinson's disease: a prospective study with the role of pharmacotherapy. J Neural Transm 2013;120:319-3296.

[6]Ciucci MR, Ahrens AM, Ma ST, Kane JR, Windham EB, Woodlee MT, Schallert T. Reduction of dopamine Synaptic activity: degradation of 50-kHz ultrasonic vocalization in rats. Behav Neurosci, 2009;123:328-336 doi: 10.1037/a0014593.7.

[7]Ciucci MR, Ma ST, Fox C, Kane JR, Ramig LO, Schallert T. Qualitative changes in ultrasonic vocalization in rats after unilateral dopamine depletion or haloperidol: a preliminary study. Behav Brain Res, 2007;182:284-289 doi: 10.1016/j.bbr.2007.02.020. 8.

[8]Guo JF, Wang L, He D, Yang QH, Duan ZX, Zhang XW, Nie LL, Yan XX, Tang BS. Clinical features and [11C]-CFT PET analysis of PARK2, PARK6, PARK7-linked autosomal recessive early onset Parkinsonism. *Neurol Sci.* 2011;32:35-40.

[9] Bonifati V, Dekker MC, Vanacore N, Fabbrini G, Squitieri F, Marconi R, Antonini A, Brustenghi P, Dalla Libera A, De Mari M, Stocchi F, Montagna P, Gallai V, Rizzu P, van Swieten JC, Oostra B, van Duijn CM, Meco G, Heutink P. Autosomal recessive early onset parkinsonism is linked to three loci: PARK2, PARK6, and PARK7. *Neurol Sci. 2002;*23 Suppl 2:S59-S60.

[10] Levine MS, Cepeda C, Hickey MA, Fleming SM, Chesselet MF. Genetic mouse models of Huntington's and Parkinson's diseases: illuminating but imperfect. Trends Neurosci. 2004;27:691-697.

[11] Chesselet MF, Richter F. Modeling of Parkinson's disease in mice. *Lancet Neurol.* 2011;10:1108-1118.

[12]Fleming SM, Salcedo J, Fernagut PO, Rockenstein E, Masliah E, Levine MS, Chesselet MF. Early and progressive sensorimotor anomalies in mice overexpressing wild-type human alpha-synuclein. J Neurosci. 2004;24:9434-9440.

[13]Valente EM, Abou-Sleiman PM, Caputo V, Muqit MM, Harvey K, Gispert S, Ali Z, Del Turco D, Bentivoglio AR, Healy DG, Albanese A, Nussbaum R, González-Maldonado R, Deller T, Salvi S, Cortelli P, Gilks WP, Latchman DS, Harvey RJ, Dallapiccola B, Auburger G, Wood NW. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science. 2004; 304(5674):1158-1160.

[14] Bonifati V, Rizzu P, van Baren MJ, Schaap O, Breedveld GJ, Krieger E, Dekker MC, Squitieri F, Ibanez P, Joosse M, van Dongen JW, Vanacore N, van Swieten JC, Brice A, Meco G, van Duijn CM, Oostra BA, Heutink P. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science. 2003;* 299(5604):256-259.

ACOUSTICAL ANALYSIS OF VOCAL TREMOR IN PARKINSON SPEAKERS

C.Mertens[1], J.Schoentgen[1,2], F.Grenez[1], S.Skodda[3]

[1]Laboratory of Images, Signals and Acoustics, Universite Libre de Bruxelles, Brussels, Belgium
[2]National Fund for Scientific Research, Belgium
[3]Department of Neurology, Knappschaftskrankenhaus, Ruhr University of Bochum, Germany

***Abstract* : The objective is to present a method to analyze vocal tremor in sustained speech sounds and compare the results with a perceptual scoring of the degree of tremor. The vocal cycle lengths are tracked by salience analysis and dynamic programming. The cycle length time series is then split into four components by empirical mode decomposition: jitter, neurological tremor, physiological tremor and intonation. Neurological tremor cues are obtained on the basis of the instantaneous frequencies and amplitudes of empirical modes. The results report tremor size, tremor frequency and bandwidth of a corpus of vowels sustained by Parkinson speakers as well as their correlations with the perceived degree of tremor assessed via pairwise comparison.**

***Keywords* : speech analysis, laryngeal assessment, vocal tremor, empirical mode decomposition**

I. Introduction

The framework of the presentation is the assessment of disordered voices. The assessment of voice and laryngeal function is based on auditory ratings and acoustic analyses of speech sounds. Acoustic feature-based assessment methods are indeed popular because they are non-invasive and enable clinicians to monitor the voice of patients quantitatively.

Disorders of phonation are often a consequence of the inability of vocal folds to vibrate regularly. Larger than normal disturbances of the periodicity of the glottal source signal are therefore observed frequently as a consequence of organic, neurological or functional disorders of the larynx.

Fast, small and involuntary cycle-to-cycle perturbations of vocal cycle lengths are designated as vocal jitter and involuntary low-frequency modulations of the vocal cycle lengths are referred to as vocal tremor. The latter have physiological (breathing, cardiac beat and pulsatile blood flow) or neurological causes. Conventionally, vocal jitter and tremor are tracked in sustained speech sounds in which small cycle length perturbations are less likely to be masked by intonation or accentuation.

The objective of the study that is presented here is to measure vocal tremor frequency and size in normal speakers and patients suffering from neurological diseases. The analysis relies on the tracking of the vocal cycle lengths in sustained voiced speech sounds by means of a temporal method that is not based on strong assumptions with regard to the regularity of the speech cycles and their periodicity [1]. The obtained cycle length time series is then decomposed further into a sum of oscillating components by empirical mode decomposition [2] [3]. According to their frequency, these modes are then assigned to four categories: intonation, physiological tremor, neurological tremor and vocal jitter. The length time series components that are so obtained are then further analyzed with a view to obtaining the neurological tremor size and frequency as well as jitter size.

Vocal tremor and jitter cues are correlated with relative scores assigned by three judges to the perceived degree of tremor. The relevance of numerical cues of vocal cycle lengths perturbations may indeed be evaluated by their ability to predict subjective scores that are obtained via the auditory assessment of the vocal timbre. In this study, comparative judgments [4] of the perceived degree of tremor in all possible signal pairs are used to rank voiced speech samples sustained by 15 speakers according to perceived tremor level.

The speakers have been Parkinson patients. Parkinson's disease is a degenerative disorder of the central nervous system. During the initial stages of the disease, the symptoms are shaking, rigidity and slowness of limb motion. Possible vocal symptoms of the disease are vocal frequency tremor and hoarseness [5].

II. Method

A. Corpus

The corpus comprises sustained vowels [a] produced by 15 Parkinson speakers (3 female and 12 male speakers between 50 and 75 years of age). The signals have been recorded at a sampling frequency of 44.1 kHz in WAV format in the same recording environment and by means of the same equipment at the Department of Neurology of Bochum University Clinic. A stable speech signal

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

fragment of 3 seconds has been selected from each vowel for analysis.

B. Perceptual assessment by pairwise comparison of tokens

In the framework of a rating session, all possible pairs of the 15 stimuli are presented randomly to a listener who is asked to designate the token of the pair with the highest perceived level of tremor. The listener can also designate the two tokens of a pair as equally perturbed. The total number of pairs is equal to 105 ($15 \cdot 14/2$). The software that presents the pairs one after the other increments by one the score of each token that is designated by the listener as the most perturbed. The increment is equal to 0.5 when both tokens are declared having the same level of tremor. At the end of a listening session, the speech samples are ranked according to their total score. The advantage of scoring by pairwise comparison is that the overall ranking of the stimuli is obtained on the base of the ability of the listeners to compare two stimuli rather than on the base of their ability to categorize a single stimulus according to a subjective scale. Three male listeners familiar with the analysis and scoring of disordered voices, but without any training in neurology have participated in the rating sessions. The three participants report normal hearing.

C. Tracking of the vocal cycle lengths

The vocal cycle length tracking is based on a temporal method, which does not rest on the assumptions that the signal is locally periodic and the average cycle length known a priori. The vocal frequency is assumed to be between $60Hz$ and $400Hz$. The cycle length tracking rests on the detection in the speech signal of cycle patterns that characterize the same glottal event. The selection of the final length time series among several candidate series relies on dynamic programming that extracts a cycle sequence the length perturbations of which are minimal [1]. The cost function involves the second order differences of successive speech cycle durations as well as the cycle peak saliences. A peak is defined as a signal sample the amplitude of which is larger than its neighbors and the peak salience is defined as the length of the longest temporal interval over which a peak is a local maximum. The so obtained cycle length time series is then constant-step resampled for further processing.

D. Empirical mode decomposition

The vocal cycle length time series is analyzed via empirical mode decomposition (EMD). EMD breaks up iteratively a signal $x(t)$ into a sum of intrinsic mode functions $c_i(t)$ (IMF) and a residue $r(t)$.

$$x(t) = \sum_{i=1}^{M} c_i(t) + r(t) \tag{1}$$

An IMF is an alternating zero-mean function with respect to the horizontal axis and the residue is monotonic. A desirable property of empirical mode decomposition is that the speech cycle time series can be perfectly recovered by summing the empirical modes. Another property is that the method does not rely on basis functions that are fixed a priori. The orthogonality between consecutive IMFs is not guaranteed theoretically, however. Therefore, mode mixing may occur, which means that 2 successive IMFs may share the same frequency components.

E. Instantaneous frequency and amplitude of an IMF

1) Overview: Each mode function $c_i(t)$ is analyzed with a view to obtaining its instantaneous amplitude and frequency. Each mode function $c_i(t)$ is rewritten as the product of two components.

$$c_i(t) = a_i(t) cos(\theta_i(t)) \tag{2}$$

The first component, $a_i(t)$, is the time varying mode function envelope (AM component or instantaneous amplitude) and the second, $cos(\theta_i(t))$ is an FM function with unit amplitude and $\theta_i(t)$ the instantaneous phase. Published results show that a necessary condition of (2) is that $c_i(t)$ has to be mono-component and narrow band so that the spectra of the AM and FM components are non-overlapping.

2) Empirical AM-FM decomposition: Empirical AM-FM decomposition is iterative and involves the following steps.

a) Initialization : $y(t) = c_i(t)$ and $a_i(t) = 1, \forall t$
b) Detection of the local maxima of the absolute value signal $|y(t)|$
c) Determination of the envelope $e(t)$ of $|y(t)|$ by cubic spline interpolation of the positions and amplitudes of the local maxima
d) Updating : $a_i(t) \rightarrow a_i(t) \cdot e(t)$ and $y(t) \rightarrow \frac{y(t)}{e(t)}$
e) Testing whether all maxima of $|y(t)|$ have amplitude ≤ 1. If yes, stop. If no, loop to b)

Output $a_i(t)$ corresponds to the time-varying envelope of $c_i(t)$ and the oscillating component $cos(\theta_i(t))$ is obtained via a division, $\cos(\theta_i(t)) = \frac{c_i(t)}{a_i(t)}$.

3) Computation of the instantaneous frequency: The instantaneous frequency $f_i(t)$ can be obtained by a numerical differentiation of instantaneous phase $\theta_i(t)$ after phase unwrapping. The numerical phase differentiation relies on a 6-order polynomial.

$$f_i(t) = \frac{1}{2\pi} \frac{d\theta_i(t)}{dt} \tag{3}$$

Fig. 1, illustrates the time-frequency evolution of the instantaneous frequencies and amplitudes that have been extracted for each IMF. The abscissa is time, the ordinate is the frequency and the z-axis is the amplitude coded via the gray level. One observes that the slowest modulation frequencies are characterized by large amplitudes.

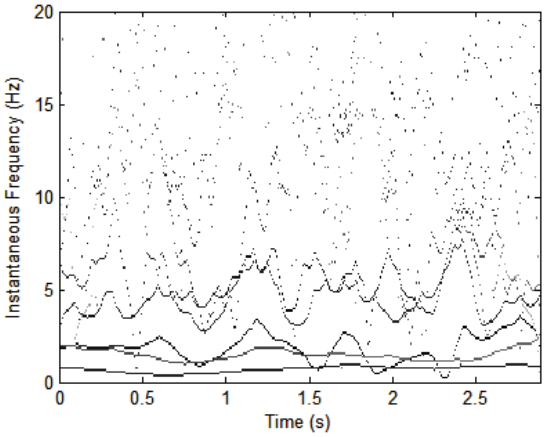


Fig. 1 : Time-frequency representation of IMF amplitudes and frequencies for a fragment of vowel [a] sustained by a Parkinson speaker.

F. Categorization of vocal cycle length perturbations

The next step consists in grouping and adding IMFs in four categories which are : intonation, physiological tremor, neurological tremor, and vocal jitter. Individual modes are assigned to one of the four categories on the base of their weighted average instantaneous frequency. The weights are the instantaneous amplitudes.

- The residue of the empirical mode decomposition is assigned to intonation.
- The IMFs with a mean instantaneous frequency below 3Hz are assigned to physiological tremor.
- The IMFs with a mean instantaneous frequency comprised between 3Hz and 15 Hz are assigned to neurological tremor.
- The other IMFs are assigned to jitter.

Fig. 2, illustrates slow and fast cycle length perturbations for a fragment of a vowel [a] sustained by a Parkinson speaker.

G. Vocal cues

1) Vocal frequency F_0: The vocal frequency is computed via the inverse of the average of the intonation time series.

2) Perturbation levels: Vocal jitter level σ_{jit}, neurological and physiological tremor modulation depths, σ_{neur} and σ_{physio}, total tremor level σ_{tre} and total perturbation level σ_{pert} are respectively related to the standard deviation of the jitter time series, the physiological and

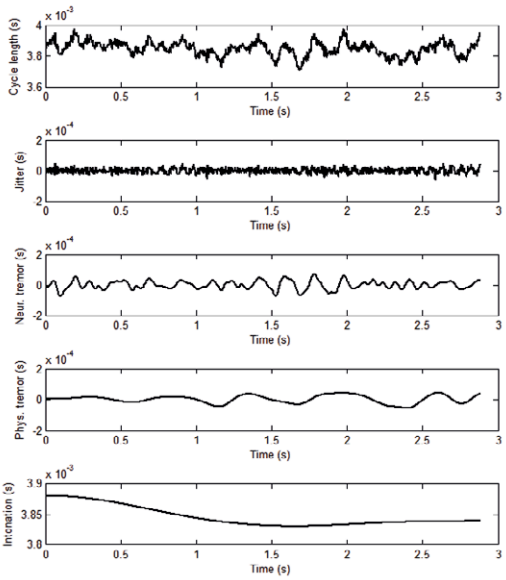


Fig. 2 : Categorization : Vocal jitter, physiological tremor, neurological tremor and intonation time series for a fragment of vowel [a] sustained by a Parkinson speaker

neurological tremor time series, the sum of physiological and neurological tremor time series or the sum of the jitter and tremor time series, divided by the average of the intonation time series.

3) Neurological tremor modulation frequency: The neurological tremor modulation frequency is obtained via the instantaneous frequencies and amplitudes of IMFs in the neurological tremor category and summarized via a weighted instantaneous frequency probability density. A sliding window of $0.1s$ is placed at the beginning of the vocal tremor time series and moves to the right with overlapping. For each position, several instantaneous frequency candidates are available. A weighted probability density estimate of the frequency content of that interval is estimated by means of a Gaussian kernel. That estimate is normalized by means of the average energy of neurological tremor in the interval.

The final time-frequency representation, illustrated in Fig. 3, is obtained via the concatenation of the density estimates. On observes that the typical neurological tremor frequency is centered around $5Hz$ but varies in time.

The trajectory of the typical neurological tremor frequency has been tracked via a scalar quantization technique [6]. The final typical neurological tremor frequency f_{neur} and the typical neurological tremor frequency bandwidth b_{neur} are obtained by means of the weighted average and standard deviation of this trajectory. The weights are the instantaneous modulation depths of the typical neurological tremor frequency (red trajectory curve). The instantaneous modulation depths equal the fraction, selected by the scalar quantization, of the average modulation

r / p	J_{aver}	f_{neur}	b_{neur}	σ_{physio}	σ_{neur}	σ_{jit}	σ_{tre}	σ_{pert}
J_{aver}		0.23	0.09	0.42	0.78	0.33	0.84	0.79
f_{neur}	0.39		0.66	0.23	0.25	0.26	0.33	0.33
b_{neur}	0.73	0.00		0.27	-0.02	0.42	0.07	0.16
σ_{physio}	0.11	0.39	0.32		0.15	-0.03	0.51	0.42
σ_{neur}	0.00	0.35	0.91	0.58		0.52	0.92	0.93
σ_{jit}	0.22	0.34	0.11	0.90	0.04		0.42	0.63
σ_{tre}	0.00	0.22	0.78	0.05	0.00	0.11		0.96
σ_{pert}	0.00	0.22	0.56	0.11	0.00	0.01	0.00	

Table I : Pearson's linear correlation coefficients (above the diagonal) and the corresponding probability values (below the diagonal) between the degree of perceived vocal tremor and the vocal perturbation cues. J_{aver} designates the average listener scores, σ_{physio}, σ_{neur}, σ_{jit}, σ_{tre} and σ_{pert} designate physiological tremor depth, neurological tremor depth, vocal jitter level, total tremor depth and the total perturbation depth. f_{neur} and b_{neur} refer to the typical modulation frequency and the bandwidth of the neurological tremor

depths in the window.

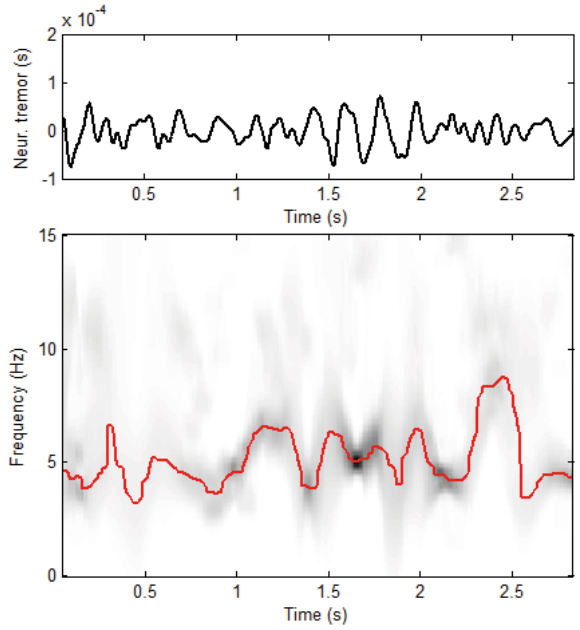


Fig. 3 : Neurological tremor time series (in temporal domain) for a fragment of vowel [a] sustained by a Parkinson speaker and its time-frequency representation

III. Results and conclusion

The neurological tremor depth in % has been correlated with perceptual scores that have been obtained via an pairwise auditory assessment of the vocal tremor level. One observes a good inter-judge agreement of listeners (Pearson's $r > 0.84$). Correlation coefficients of the perceived degree of vocal tremor level and the measured vocal cues are given in Table I. The cues are the vocal jitter and tremor modulation depths σ_{jit}, σ_{neur}, σ_{physio}, σ_{tre}, the total perturbation level σ_{pert} and the modulation frequency f_{neur} and frequency bandwidth b_{neur} of the neurological tremor. One observes that the neurological tremor level σ_{neur} and the total tremor level are correlated with the average listener scores ($r = 0.78$). One observes also a poor correlation ($r = 0.33$) between the perceptual scores and vocal jitter level σ_{jit}, which suggests that hoarseness and vocal tremor levels may be analysed and perceptually assessed separately.

References

[1] C.Mertens, F.Grenez, L.Crevier-Buchman, and J.Schoentgen, "Reliable tracking based on speech sample salience of vocal cycle length perturbations," in *Proceedings 11th Annual Conference of the International Speech Communication Association INTERSPEECH, Makuhari (Japan)*, 2010.

[2] N. E.Huang, Z. Shen, S. R.Long, M. C.Wu, H. H.Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H.Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceeding of the The Royal Society*, 1998.

[3] G.Rilling, P.Flandrin, and P.Goncalves, "On empirical mode decomposition and its algorithm," in *Proceedings of the 6th IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing, Grado (Italy)*, 2003.

[4] A.Kacha, F.Grenez, and J.Schoentgen, "Voice quality assessment by means of comparative judgments of speech tokens," in *Proceedings of the 5th Annual Conference of the International Speech Communication Association INTERSPEECH, Lisbon (Portugal)*, 2005.

[5] L.Cnockaert, J.Schoentgen, P.Auzou, C.Ozsancak, L.Lefebvre, and F.Grenez, "Low-frequency vocal modulations in vowels produced by parkinsonian subjects," *Speech Communication*, vol. 50, no. 4, pp. 288–300, 2008.

[6] K.Sayood, *Introduction to data compression*. The Morgan Kaufmann Series in Multimedia Information and Systems, 2012.

ANALYSIS AND EXPERIMENTS OF THE LOMBARD EFFECT IN PEOPLE WITH PARKINSON'S DISEASE

Panikos Heracleous[1], *Jani Even*[2], *Carlos Ishi*[2], *Masaki Kondo*[3],*Kyoko Takanohara*[3], *and Kazuya Takeda*[1]

[1]Department of Media Science, Graduate School of Information Science, Nagoya University, Japan
[2]ATR, Intelligent Robotics and Communications Laboratories, Japan
[3]Kyoto Prefectural University of Medicine, Kyoto, Japan

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disease with many motor symptoms or various motor and non-motor symptoms, including speech disorders. Although antiparkinson drugs and surgical method such as deep brain stimulation exist to treat some of PD's motor symptoms, therapies for speech impairment are not effective and satisfactory, leaving the area open for research. The current study aims at taking advantage of the Lombard reflex to improve the speech loudness of PD patients. As a first step, the experience of the Lombard reflex by Japanese PD people was confirmed, and the perception of PD patients' speech was evaluated by several listeners. In a following step, methods based on masking sound will be used for intensive training and for self-training of PD patients. Preliminary results indicate persistence of the Lombard reflex after training using masking noise. It might be possible that after intensive training, PD patients speak louder without using masking noise.

***Index Terms*— Parkinson's disease, speech loudness, Lombard reflex, rehabilitation.**

1. INTRODUCTION

To date, several studies reported that PD is the second most common neurodegenerative disease in developed countries [1]. It affects 1–2% of people over the age of 60 years, but the symptoms can develop before the age of 40 years [2]. Among PD patients, 70% develop speech impairments, such as dysarthria [3]. A main symptom of dysarthria is monotony of pitch and loudness. Dysarthria worsens in the late stage of the disease and can cause loss of communication and social isolation. Although other disease symptoms (i.e., akinesia,rigidity, tremor) can be treated using dopamine replacement therapy or antiparkinson drugs, dysarthria worsens for most patients.

This study investigates the effect of the Lombard reflex on PD patients. People without PD attempt to increase the intelligibility of their speech when talking in a noisy environment. During this process, several speech characteristics change, including increased loudness. This phenomenon is known as the Lombard reflex [4].

The current study aims at taking advantage of the Lombard reflex to improve speech intelligibility of PD people. During the first stage of the study, the main question was to determine whether Japanese PD patients could experience the Lombard reflex. The final goal is to develop a new rehabilitation technique applicable in speech therapy of PD patients.

Parkinsonian dysartria, or hypokenetic dysarthria, is a speech production disorder resulting from PD. The main characteristics of dysarthria are monotonous and reduced pitch and loudness, variable rate, short rushes of speech, imprecise consonants, and a breathy and harsh speech. Each of the speech production subsystems, respiration, phonation, articulation, resonance, and prosody, may be affected in dysarthria [5]. Some characteristics of PD people include impairment in the ability to sustain prolonged vowel phonation, perception that they are hypernasal, and affected prosody of speech.

Several methods have been introduced for the treatment of dysarthria. Treatment methods include speech therapy, pharmacologic treatments, and surgery. The three main speech therapy methods include the use of devices, treatment focused on speech prosody, and the Lee Silverman Voice Treatment (LSVT) [6, 7]. LSVT seems to be one of the most effective therapies [8]. LSVT focuses on increasing vocal loudness and has an intensive treatment of one month. Today, many PD patients receive speech therapy treatment from speech therapists. During these treatments, patients are asked to sing or speak louder. Intensive speech therapy treatments show improvements in speech loudness and in other aspects of speech, but the effect of most PD treatments on dysarthria remains unsatisfactory.

2. METHODS

2.1. Subjective evaluation of PD speech

Twelve listeners evaluated the speech produced by the PD patients. To generalize the experiment, six Japanese and six non-Japanese listeners were used. The authors were also in-

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

terested in investigating the possible differences concerning the Japanese and non-Japanese listeners. The age of the listeners was between 25–36 years and they all had normal hearing. The aim of this experiment was to investigate whether other listeners can perceive the speech changes that occurred in the PD patients because of listening to the masking sound. It is possible that some speech characteristics change under the Lombard conditions, but these characteristics cannot be observed by the listeners. In such a case, the Lombard effect might not be an effective way to deal with the speech disorders of the PD patients.

Each stimulus consisted of a vowel produced under a clean environment and of the same vowel produced while the patients were listening to the masking sound (i.e., a vowel-pair). The task of the listeners was to select the vowel instance that better matches a specific characteristic. In particular, the listeners had to select the louder vowel, the longer vowel, and the more natural vowel between the two vowels of the vowel-pair. In total, twenty vowel-pairs were played back through a headset.

2.2. Database and recording conditions

Three Japanese PD patients (i.e., a male and two female), who agreed and accepted the experimental conditions, participated in the experiments. Their ages were 76–82 years old, and all three participants used a wheelchair. However, they were able to read and speak, although their speech loudness was low. The participants were patients of the Kyoto Prefectural University of Medicine, Japan, and stayed in the hospital to receive pharmacological treatment. The severity stage of PD disease (Hohen and Yahr stage) for the 1st patient was 4, for the 2nd patient it was 4, and for the 3rd patient is was 3. During their stays in the hospital, they also received speech therapy treatment from a professional speech therapist. The patients were instructed by the speech therapist to speak with increased speech loudness or to sing their favourite song. The speech therapy treatment also included training of facial/lip muscles.

The recording took place in the office of the speech therapist. Some background noise of 35 dB(A) sound pressure level (SPL) also existed. Two close-talking microphones, a distant-talking microphone, and a microphone array were used to capture the patients' speech. A portable computer with special recording software installed and a 16-channel sound card was used. Another portable computer for displaying the words that had to be read was used.

Several kinds of noise were considered for masking. Considering the condition of the patients and their ages, using music for the masking sound was finally decided. The music was selected to be comfortable and enjoyable for the patients. After the experiments, all patients expressed their satisfaction regarding the selected music.

Before the experiments, the speech therapist explained the experimental procedure to the participants. Their positions on the wheelchair were also adjusted to ensure that they sat as comfortably as possible.

Each participant sat in front of the laptop at a distance of 50 cm and was instructed to read the utterances displayed on the screen. The experiment consisted of three sessions, as follows:

- The patient read the five Japanese vowels a, i, u, e, o without listening to masking sound.
- The patient read 120 short Japanese words while listening to masking sound through headphones at 70 dB(A) SPL. The words were selected from the Japanese Diagnostic Rhyme Test (JDRT) [9], and include words such sai, zai, etc.
- The patient read the five Japanese vowels a, i, u, e, o without listening to masking sound.

One of the female patients participated the experiments in two different days. In the second case, the patient read short utterances instead of short words. The rest of the experimental procedure was as described before.

During the experiments, the speech therapist was also present and continuously monitored the patients' condition by asking them questions or by watching their reaction. In one case, a family member of the participant was also present.

3. RESULTS

This section introduces the analysis results. Three characteristics that indicate experience of the Lombard reflex are investigated. In particular, the changes that occurred when there was masking sound in comparison with the case when masking sound was absent are examined. For analysis, the signal received by a close-talking microphone was used.

Figure 1, Figure 2, and Figure 3 show the average power spectrum of the five Japanese vowels when the three patients were listening to masking sound compared with the case when masking sound was absent. As is shown, when masking sound was present, in all cases the power was increased in the middle frequency range (around 1–6 kHz), indicating higher vocal effort and higher vocal fold tension, which reflects well the Lombard reflex.

Table 1 shows the p-values of two-paired t-tests [10]. statistical test. The aim of this analysis was to investigate whether the power differences without and with masking sound are statistically significant. Also, the authors were interested in investigating the significance of the power differences of the first session (i.e., without masking sound) and the third session (i.e., without masking sound, but after a 20-minute training session with masking sound). The latest result will give some information about the persistence of the Lombard reflex.

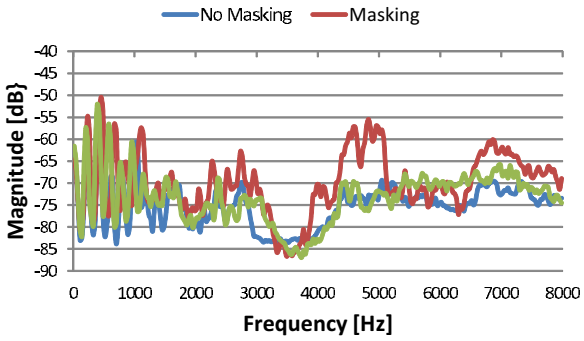


Fig. 1. Average power spectrum of the five Japanese vowels in the case of the 1st PD patient.

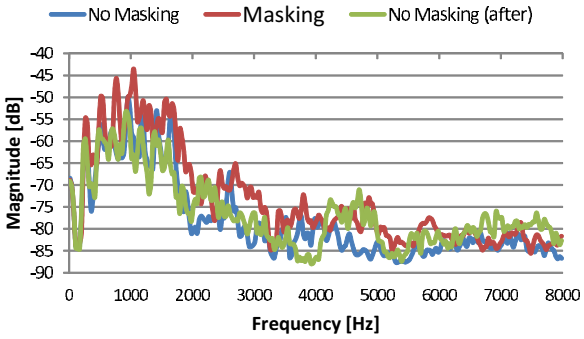


Fig. 2. Average power spectrum of the five Japanese vowels in the case of the 2nd PD patient.

As Table 1 shows, in all case the power differences without and with masking sound are statistically significant. Moreover, in the case of the 1st and 2nd patients, the differences without masking noise before and after training are also statistically significant. This is a very important result, because it might be an evidence that after intensive training using masking sound, the speech loudness of a PD patient remains louder. This, however, requires additional experiments with more patients and follow-up data.

The analysis of pitch showed that when the patients were listening to masking sound, F0 was increased. The mean F0 value for the five vowels was 213 Hz and the standard deviation was 12.6. When masking sound was absent, the mean F0 value was 196.5 Hz and the standard deviation was 5.8. Therefore, higher F0 and also variability were observed when masking sound was used. The difference was tested using a paired t-test. The two-tailed P value was 0.0132. By conventional criteria, this difference is considered to be statistically significant.

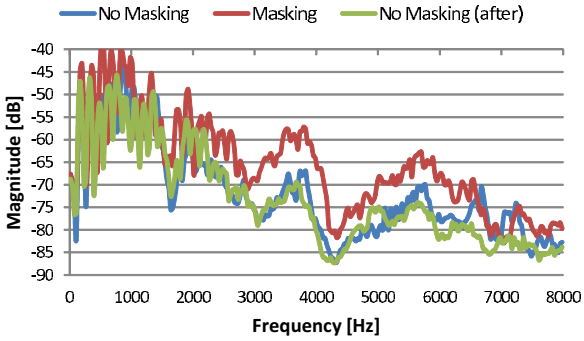


Fig. 3. Average power spectrum of the five Japanese vowels in the case of the 3rd PD patient.

Table 1. p-values of paired t-tests for power analysis

Patient	Recording condition	
	Without masking noise (before training) vs Masking noise	Without masking noise (before training) vs Without masking noise (after training)
P01	<0.0001	0.0002
P02	<0.0001	0.024
P03	<0.0001	0.23
P01 (sentences)	<0.0001	-

When Lombard reflex is experienced by a talker, vowel durations also increase. For this reason, the vowel durations when listening to masking sound were compared with the vowel durations when speech was produced without masking sound. When masking sound was used, the mean duration of the five vowels was 0.4956 seconds. In the case of no masking sound, the mean duration was 0.3169 seconds. The two-tailed P value was 0.0143. This difference is statistically significant.

Table 2 shows the results of the subjective tests in the case of Japanese listeners. As shown in 2, the six listeners perceived that the vowels produced under the Lombard conditions were louder by 95.8 %. In addition, the listeners perceived that the duration of the Lombard vowels were longer by 90.8 %. Regarding speech naturalness, only 31.7 % of the Lombard vowels were selected to be more natural compared to the non-Lombard vowels. Based on the discussion of the results with the four listeners, the possible reason for this might be the longer duration of the vowels, which make them less natural.

Table 3 shows the results of the subjective tests in the case of the non-Japanese listeners. As shown in 3, the listeners perceived the vowels produced under the Lombard conditions

Table 2. Results of the subjective evaluation of the PD patients' speech by Japanese listeners (rate of the selected Lombard vowel [%]).

Listener	Speech characteristic		
	Loudness	Duration	Naturalness
M01	95	75	40
M02	100	90	50
F01	100	90	20
F02	95	90	25
F03	100	100	25
F04	85	100	30
Average [%]	95.8	90.8	31.7

Table 3. Results of the subjective evaluation of the PD patients' speech by non-Japanese listeners (rate of the selected Lombard vowel [%]).

Listener	Speech characteristic		
	Loudness	Duration	Naturalness
M01	95	90	45
M02	95	95	25
F01	95	85	50
F02	95	90	30
F03	100	90	20
F04	100	100	35
Average [%]	96.7	91.7	34.2

as louder by 96.7 %. The vowels under the Lombard conditions were perceived as longer by 91.7 %. Finally, only 34.2 % of the vowels under the Lombard conditions were selected as being more natural. The results obtained in the case of the non-Japanese listeners are very similar to the case of the Japanese listeners who evaluated the data. After conducting a two-tailed paired t-test, the differences were considered statistically not significant.

In the case of the PD patients, the most important and critical speech characteristic for communication is speech loudness. A trade-off between naturalness and loudness in the PD patients should be allowed and accepted. The results reported in this section show that speech changes occurred while listening to the masking sound and that such changes can be clearly perceived by other listeners. This is a promising result that confirms the hypothesis that the Lombard effect can be used to improve the verbal communication of the PD patients.

4. CONCLUSIONS AND FUTURE WORK

This study focuses on the Lombard reflex in the case of PD patients. The results obtained show that Japanese PD patients experience the Lombard reflex when listening to masking sound while talking. As a result, the speech loudness, pitch, and vowel duration increased. Based on these observations, the next study will focus on developing new speech rehabilitation methods for PD people. Experiments using more participants and further analysis are currently in progress.

Acknowledgement

This work was partially supported by the collaborative study of Toyota Motor Corporation and Green Mobility Collaborative Research Center, Nagoya University.

5. REFERENCES

[1] C. M. Tanner, M. Brandabur, and E. R. Dorsey, "Parkinson disease: A global view," *Parkinson Report*, 2008.

[2] C.D. Marsden, "Parkinson's disease," *J Neurol Neurosurg Psychiatry*, vol. 57, pp. 672–682, 1994.

[3] L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with parkinson's disease and multiple sclerosis:a survey," *Folia Phoniatr Logop*, vol. 12, pp. 9–17, 1994.

[4] A. E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[5] G. M. Schulz and M.K. Grant, "Effects on speech therapy and pharmacologic and surgical treatments on voice and speech in parkinson's disease," *J. Communi. Disorder*, vol. 33, pp. 59–88, 2000.

[6] "Other key interventions. parkinson's disease. london: Royal college of physicians," *The National Collaborating Centre for Chronic Conditions*, pp. 135–146, 2006.

[7] C.M. Fox, L.O. Ramig, M.R. Ciucci, S. Sapir, D.H. McFarland, and B.G. Farley, "The science and practice of lsvt/loud: neural plasticity-principled approach to treating individuals with parkinson disease and other neurological disorders," *Seminars in Speech and Language*, vol. 27 (4), pp. 283?99, 2006.

[8] G. Ebersbach, A. Ebersbach, D. Edler, O. Kaufhold, M. Kusch, A. Kupsch, and J. Wissel, "Comparing exercise in parkinson's disease-the berlin big study," *Movement Disorders*, vol. 25 (12), pp. 1902–1908, 2010.

[9] M. Fujimori, K. Kondo, K. Takano, and K. Nakagawa, "On a revised word-pair list for the japanese intelligibility test," *In Proc. of International Workshop Frontiers in Speech and Hearing Research*, pp. 103–108, 2006.

[10] J. F. Box, "Guinness, gosset, fisher, and small samples," *Statistical Science*, vol. 2, pp. 61–66, 1987.

CHARACTERIZING VOCAL TRACT CENTRALIZATION AND ASYMMETRY IN AMYOTROPHIC LATERAL SCLEROSIS

Pedro Gómez-Vilda[1], Ana Rita M. Londral[2], Luz Rocha[2], Mamede de Carvalho[2], Victoria Rodellar-Biarge[1]
[1]NeuVox Laboratory, Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28223 Pozuelo de Alarcón, Madrid, Spain
[2]Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, Lisbon, Portugal
e-mails: pedro@pino.datsi.fi.upm.es, arml@campus.ul.pt

Abstract: **Amyotrohpic Lateral Sclerosis (ALS) is one of several neurodegenerative diseases which may benefit from speech processing for monitoring purposes. It is well known that ALS produces clear correlates in speech when the neuromotor units affecting oropharyngeal and mandibular muscles are suffering a specific deterioration. These correlates affect phonation, articulation, prosody and rhythm and can be precisely estimated using signal processing methods. This paper is oriented to give a measurement on how articulation is being affected by mapping the voiced fragments of running speech on the vowel triangle and proposing an asymmetry index which illustrates the degree of articulation deterioration. The index is based on the comparison of the actual vowel triangle gravity centre estimated from the pattient's utterance against a normative one, producing a deviation vector which illustrates the migration of the vowel space towards a neutral vowel near [æ]. An explanation for this fact may be found in the difficulties faced by the patient to activate precise articulation positions for the mandibular and hypoglossal muscles, and in the muscle atrophy in itself. Examples from a longitudinal study case are presented.**

Keywords: **neurological disease monitoring, speech processing, dysarthria, vowel triangle**

I. INTRODUCTION

The use of vowel representation spaces is of most importance in many fields, as in characterizing speech from patients affected by neurological disease (ND). The purpose of the present work is to study vowel space characterization in amyotrophic lateral sclerosis (ALS). This is a very severe and rapidly advancing neuromuscular disease of unclear origin, with no effective treatment to halt progression. The deterioration of the neuromotor system involved in respiration, phonation, swallowing and lingual and oro-facial muscle function degenerates in a rapidly progressive dysarthria. ALS patients usually have spastic-flaccid dysarthria, characterized, among other symptoms, by defective articulation, reduced speech rhythm, imprecise consonant production, marked hypernasality and vowel intelligibility decay [1]. The present paper is intended to explore the early detection of speech limitations, as well as to provide the speech therapist with objective measurements to evaluate disease progression and optimize rehabilitation efforts [3]. The paper concentrates in the description of vowel characteristics (vowel color) in the F1-F2 formant space, also known as vowel triangle [4], as a possible marker of ALS dysarthria [5], [6]. The Vowel Space Area (VSA) and the Formant Centralization Ratio (FCR) are coefficients defined to estimate the vowel span range and positioning produced by a given speaker [7], [8]. The relative deviation of these coefficients from normative statistics in a specific case may be used to evaluate the distortion of the FCR, in terms of distance and angle. The deviation can be defined as a vowel distribution asymmetry coefficient (VDAC). This measurement shows a semantic value in monitoring speech deterioration. The paper presents the results from a longitudinal study case in ALS showing the evolution of the vowel space asymmetry coefficient.

II. ARTICULATION AND NEUROMOTOR PATHWAYS

The relationship between facial and oral neuromotor systems and articulation gestures is of crucial relevance for the purposes of the present study. The main concepts involved are summarized and simplified in Fig. 1. This figure represents schematically how the neuromotor system (1) activates different muscular systems through neuromotor pathways related with speech (subthalamic sections of cranial nerves IX, X, V, corresponding to glosopharyngeal, vagus and trigeminous nerves, respectively), as the laryngeal, oro-pharyngeal and facial muscles [10]. Especially relevant to phonation and articulation are the laryngeal muscles (5) controlling larynx position and glottal closure, the hypoglossal muscles (3) controlling tongue forward, backward and upward movements, and the mandibular muscles (4) controlling upward and downward movements of jaw. The nasopharyngeal switch controls the closing or opening from pharynx to nasal airways affecting nasal vowels and stop consonants. Specifically, systems (3) and (4) are the most interesting ones for our study. The

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

relative positions of different vowels on the vowel triangle as a function of their two first formants (F1 and F2) are given in Fig. 1.b). It is known that F1 is controlled mainly by the degree of opening of the vocal tract, depending greatly on the mandibular system (4), whereas F1 is mainly controlled by the forward-backward position of tongue, depending greatly on the hypoglossal system (3).

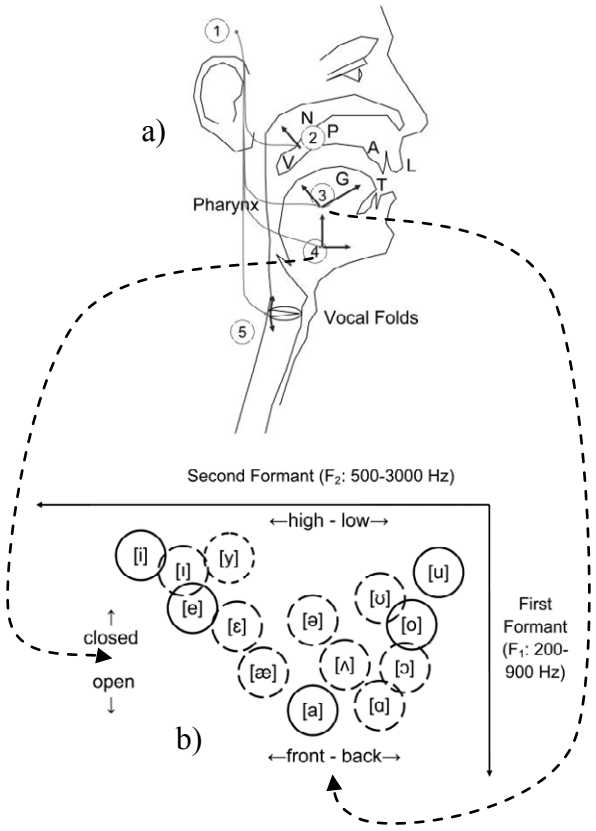


Fig. 1 Neuromotor structures (a) in relation to vowel representation spaces on the vowel triangle (b). 1) Subthalamic sections. 2) Velo-pharyngeal switch. 3) Hypoglossal system. 4) Mandibular system. 5) Laryngeal system.

Although there are also interdependencies between F1 and (3), and F2 and (4), these are thought to be less relevant, and in any case they can be formalized as an interdependence matrix. Consequently, the distortions observed in time and position of articulation correlates given by formants may be used to infer which neuromotor systems are affected and to which extent. As it will be discussed in section V, the relationship between articulation and speech neuromotor systems is of most importance for this kind of studies, as articulation correlates may be estimated from the speech signal through acoustic inversion, and therefore the neuromotor systems affected and their degree of dysfunction may be inferred.

III. Methods

Recordings from 20 male and 20 female subjects of the phonetically balanced sentence in Spanish /es hábil un solo día/ (IPA transcription [9]: ɛsɑ βɪlʊ n̪sɔ lɔ ðiæ) were taken at a sampling rate of 16 kHz. These recordings were used to estimate a prototype of the vowel triangle in the domain of the first two formants F_2 vs. F_1 corresponding to pairs $\{F_1(n), F_2(n)\}$, where n is the discrete time index. Formant distributions for each gender were used to estimate the distribution quantiles

$$q_i^{\theta} = \arg\left\{ \int_{\nu=q_i^{0}}^{\infty} \gamma_i(\nu)d\nu \Big/ \int_{-\infty}^{\infty} \gamma_i(\nu)d\nu < \theta \right\} \qquad (1)$$

where $\gamma_i(\nu)$ is the probability density funcion of the formant i in frequency ν, and θ is the specific quantile threshold (for instance $\theta=0.03$ would correspond to a 3% quantile). In the present study the following definitions apply: $\theta_1=0.03$, $\theta_2=0.5$ and $\theta_3=0.97$. Using these definitions the VDAC would be defined in module and argument as

$$M_A = \left[\left(\frac{2q_1^{\theta_2}}{q_1^{\theta_3} + q_1^{\theta_1}} - 1 \right)^2 + \left(\frac{2q_2^{\theta_2}}{q_2^{\theta_3} + q_2^{\theta_1}} - 1 \right)^2 \right]^{1/2}$$
$$\varphi_A = \arctan\left(\frac{2q_2^{\theta_2} - q_2^{\theta_1} - q_2^{\theta_3}}{2q_1^{\theta_2} - q_1^{\theta_1} - q_1^{\theta_3}} \right) \qquad (2)$$

The study consisted in analyzing recordings from a woman affected with ALS in contrast to a control healthy woman. This case study consisted of five recordings from a patient with ALS, during clinical assessment at the Neurology Department of Hospital de Santa Maria, in Portugal. Recordings were taken at approximately 3-month intervals, these being referred to as HA_T0 (November 2011), HA_T1 (January 2012), HA_T2 (March 2012), HA_T3 (July 2012) and HA_T4 (October 2012). In all cases the recordings contained utterances of the sentence in Portuguese */tudo vale a pena quando a alma não é pequena/* (IPA: [tʊ ðʊ ɥɑ ɭ ɐ pInæ˜ kwændʊ aaɭ mɐ nɐ ˜ɛ˜ pkenæ˜]). The first recording (HA_T0) was taken when clinical evaluation (ALSFRS [10]) indicated that a high score in bulbar related functions was already present. The third recording (HA_T2) was later eliminated from the analysis due to low acoustic quality. The results of the study conducted on these recordings are given in the next section.

IV. Results

The recordings for the longitudinal study, undersampled at 8kHz were processed to extract the formant positions F1-F2 each 2 ms using an 11-pole LPC lattice-ladder inverter to separate vocal tract and glottal source components [4]. The plots in Fig. 2 help in establishing a comparison among the vowel spaces

produced during the different patient evaluations, and to derive resolving conclusions.

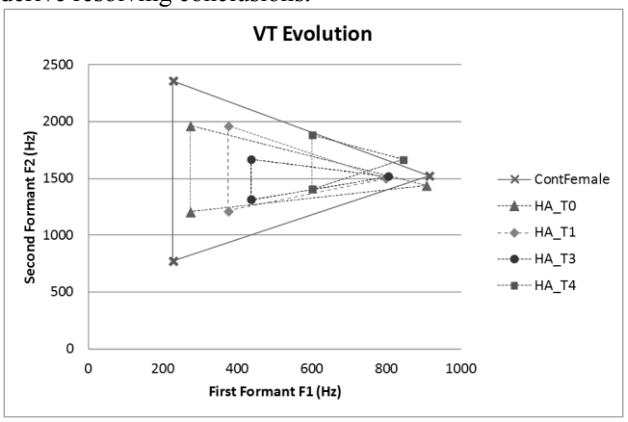


Fig. 2 Comparison between the vowel triangle from control subject (ContFemale), and the ALS patient in four different evaluations chronologically ordered from less severity to most severity (HA_T0, HA_T1, HA_T3 and HA_T4).

The VDAC for the database reference and the five cases studied is represented in modulus and argument in Fig. 3 below.

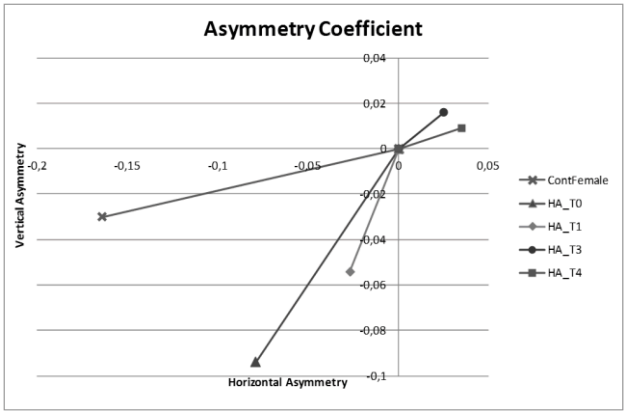


Fig. 3 VDAC for the control subject and the four different evaluations in Fig. 2.

Table 1. Comparison of vowel space coefficients for the control and four longitudinal study cases

Case	M_A	φ_A (deg.)	VSA	FCR
Male Ave.	0.099	-122.250	150390	1.11
Female Ave.	0.142	-131.299	238980	1.07
Control	0.167	-169.634	545170	0.84
HA_T0	0.123	-130.045	239780	1.11
HA_T1	0.060	-116.565	158200	1.25
HA_T3	0.015	32.619	64545	1.5
HA_T4	0.036	14.421	81909	1.49

The progressive degradation of the vowel triangle can be clearly perceived, with strong differences between the results for HA_T0 (still comparable with the control subject) and the three last utterances. This indicates that a strong decay in articulatory ability of the patient took place from November 2011 to January 2012. The VDAC estimated for the control subject and the four stages of the longitudinal study can be compared to the VSA and FCR in Table 1.

V. Discussion

The most interesting fact to be stressed is that, as disease progresses, the vowel triangle shrinks towards a lower VSA and less centralized FCR, whereas the module of VDAC indicates that the vowel space does not show separated vowel distributions anymore, as pairs $\{F_1(n), F_2(n)\}$ tend to accumulate in a global cluster. The angle of VDAC swings from the third to the first quadrant in a progressive succession. This means that the orientation of the vowel median centroids is evolving from a central situation to a tilted position towards vowel [æ]. Producing frontal or rear vowels requires the operation of the hypoglossal and mandibular neuromotor systems, although these are not necessarily active in the neutral mid position open vowel given by [æ] (see Fig. 1). It seems that when neuromuscular activity is already severely impaired, this articulation position ([æ]) would be the only one possible for the patient, and the formant patterns of the different vowels would be fused towards this final position [12].

It is to be investigated if this methodology can also be applicable to other neurological diseases, as Alzheimer and Parkinson, although important functional differences in the correlates are to be found. In the specific case of ALS, a general deterioration of neuromotor systems is present, leading at some stage of the disease to an irreversible significant muscular atrophy. This can be especially noticed in the case of the hypoglossal neuromotor system, where gradual and significant losses of muscular mass may be observed along the disease progression.

VI. Conclusions

The results support the aim of producing objective measurements of speech degradation which may be useful to the speech therapist in grading the deterioration of the patient's ability to articulate understandable speech. The most important findings established in this sense as disease progresses are the following:

- The vowel triangle shrinks, especially in F2. As a result of this fact, the number of distinguishable vowels is drastically reduced.

- The vowel triangle centroids evolve towards [æ]. This fact may serve as a clear indication of disease progression.
- Coefficients as VSA and FCR developed to detect the neurological speech deterioration in Parkinson's Disease are also good indicators in ALS.
- There is a good agreement between the estimations of the VSA and FCR in comparison to the VDCA in relation with disease progression.

This last finding is in good agreement with the discussion in the previous section. It is well known from literature [12] that F1 is very much related to the degree of opening of the vocal tract ([i] and [u] corresponding to the more closed extremes, whereas [a] gives the more open extreme), F2 being more related to the articulation position (where [u] is considered a back vowel whereas [i] is a frontal, and [a] would be a middle vowel). Producing frontal or rear vowels requires the operation of the hypoglossal and mandibular neuromotor systems which need not be active in the neutral mid position open vowel represented by [æ]. Under severely impaired neuromuscular activity this would be the only articulatory position and the relative colouring of the different vowels would be fused towards this final position. An important task to fulfil in the near future is the processing of a large database containing longitudinal studies as the one described to extend the statistical significance of the findings produced in this study.

There are other aspects of ALS dysarthric speech which have not been checked in the present study, as estimating the degree of hypernasality due to the failure of the levator veli palatini, palatoglosus and palatopharyngeous neuromuscular structures acting on the naso-pharyngeal switch. This would require a spectral detector to model the zeroes in the vocal and nasal tract anti-resonances. Consonantal dynamics could be traced using neuromorphic speech processing [4]. Dysprosody could also be characterized using well-known pitch tracking methods. These tasks are left for future research. Another important task to be accomplished is the estimation of the biomechanical parameters of phonation in ALS patients from glottal source correlates. The present paper covered only aspects related with non-nasal articulation.

Acknowledgments: This work is being funded by grants TEC2009-14123-C04-03 and TEC2012-38630-C04-04 from Plan Nacional de I+D+i, Ministry of Science and Technology of Spain.

REFERENCES

[1] Ball, L. J., Beukelman, D.R. and Pattee, G.L., "Timing of speech deterioration in people with amyotrophic lateral sclerosis", *Journal of Medical Speech-Language Pathology*, vol. 10 No. 4, 2002, pp. 231–235.

[2] Tomik, B. and Guiloff, R., "Dysarthria in amyotrophic lateral sclerosis: A review", *Amyotrophic Lateral Sclerosis*, Vol. 11, 2010, pp. 4-15.

[3] Weismer, G., Martin, R., Kent, R. D. and Kent, J. F., "Formant trajectory characteristics of males with amyotrophic lateral sclerosis", *J. Acoust. Soc. Am.* Vol. 91, 1992, pp. 1085-1098.

[4] Gómez-Vilda, P., Ferrández-Vicente, J. M., and Rodellar-Biarge, V., "Simulating the Phonological Auditory Cortex: From Vowel Representation Spaces to Categories", *Neurocomputing*, Vol.114, 2013, pp. 63-75.

[5] Yunusova, Y., "Articulatory Movements During Vowels in Speakers With Dysarthria and Healthy Controls", *J. Speech, Lang. and Hear. Res.*, vol. 51, 2008, pp. 596-611.

[6] Bongioanni, P., "Communication Impairment in ALS Patients: Assessment and Treatment", in *Amyotrophic Lateral Sclerosis*, M. Maurer (ed.), 2012.

[7] Sapir, S., Ramig, L. O., Spielman, J., Fox, C., "Acoustic Metrics of Vowel Articulation in Parkinson's Disease: Vowel Space Area (VSA) vs. Vowel Articulation Index (VAI)", *Proc. of MAVEBA11*, C. Manfredi, (ed.), Florence University Press, 2011, pp. 173-175.

[8] Sapir, S., Ramig, L. O. and Fox, C., "Formant Centralization Ratio: A proposal for a New Acoustic Measure of Dysarthric Speech", *J. Speech, Lang. and Hear. Res.* Vol. 53, 2010, pp. 114-125.

[9] International Phonetic Alphabet (IPA): Available from http://www.arts.gla.ac.uk/IPA/ipachart.html

[10] Jürgens, U., "Neural pathways underlying vocal control", *Neurosci. and Behav. Rev,* Vol. 26, 2002, pp. 235-258.

[11] Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B. and Nakanishi, A., "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function BDNF ALS Study Group (Phase III)", *J. Neurol Sci.* vol. 169, No. 1-2, 1999, pp. 13-21.

[12] Dromey, C., Jang, G-O. and Hollis, K., "Assessing correlations between lingual movements and formants", *Speech Communication*, vol. 55, 2013, pp. 315-328.

DETECTING REPEATED SPEECH: A POSSIBLE MARKER FOR ALZHEIMER'S DISEASE

A. Barney[1], D. Nikolic[1], V. Nemes[2], P. Garrard[2]
[1]Institute of Sound and Vibration Research, University of Southampton, Southampton, UK
[2]St George's, University of London, London, UK
ab3@soton.ac.uk, d.nikolic@soton.ac.uk, drvandanemes@gmail.com, p.garrard@sgul.ac.uk

Abstract – **A common abnormality noted by caregivers in patients with Alzheimer's disease is perseverative behavior – the tendency to repetitive behavior including in speech. This study investigated the potential to identify repeated speech segments from large sets of recorded data. Motif discovery techniques were applied to identify repeated phrases in test data recorded from a scripted text. The initial results indicate that the approach adopted has potential for detecting repetitive speech patterns. The use of repeated phrases as a cognitive performance marker could have the unique benefit of quantifying disease severity clinically, providing an invaluable adjunct to the range of biological indices currently available.**

***Key words*: Alzheimer's disease, repeated speech, motif discovery.**

I. INTRODUCTION

It is estimated that over the next 25 years the number of people in Europe with cognitive impairment will grow rapidly, mainly due to larger numbers of longer living elderly people developing Alzheimer's disease (AD) [1] with an associated rise in the cost of long term care. The magnitude of the predicted increase means that even a modest reduction in the burden of disease or its associated functional disability will bring significant socio-economic benefits. The development of disease modifying treatments is therefore a major goal of current biomedical research, and a growing number of candidate drugs are under development worldwide, many of them at the clinical trials stage.

Two factors will be critical to deriving maximum benefit from disease modifying therapies: first, the ability to make accurate diagnosis in the early stages of the condition; and secondly the availability of robust, objective biomarkers for determining efficacy in individual patients.

To be useful aids to diagnosis and measurement of disease progression, dementia biomarkers must correlate with measures of cognitive performance. At present these follow three approaches: i) global measures of deterioration, using estimates made either by clinicians or caregivers; ii) abbreviated cognitive assessments for bedside or office use, such as the mini mental state examination (MMSE), Addenbrooke's cognitive examination (ACE), and Montreal cognitive assessment (MoCA); and iii) comprehensive batteries of formal neuropsychological tests. Currently used measures of cognitive performance all have limitations: global measures of deterioration are based on subjective judgment and quantified using coarse-grained interval scales; the content of bedside cognitive batteries is fixed, and thus subject to learning or practice effects, which may mask disease progression; formal psychometric measures are time-consuming to complete, sensitive to fluctuations in a patient's level of motivation and effort, as well as to learning effects, and reflect performance in a laboratory, rather than "real world" context [2].

Among the many abnormalities noted by caregivers in patients with AD, one of the commonest is perseverative behavior – the tendency to make the same statement, ask the same question, or carry out the same action repeatedly over the course of the day [3]. This type of behavior is likely to be related to a decline in day-to-day memory (the commonest first symptom of AD): it is assumed that failure to store new information in episodic memory results in incomplete updating of conscious experience, and hence to repetition. This phenomenon is widespread among patients with established dementia, probably increases with progression of the condition, and is widely regarded as a sensitive indicator of disordered cognition [4]. To date, the frequency and severity of such repetitions has been quantified only in terms of caregivers' estimates [3]; more discriminating quantitative measures (such as the mean length of time between repetitions, the complexity of the repeated behavior, and the number of times the repetition is observed over the course of a day) have not been acquired, despite the potential value of such information.

II. METHODS

This study investigated the potential to identify repeated speech segments from large sets of recorded data. Motif discovery techniques were applied to identify speech segments in test data recorded from a scripted text with one or more repeated sections included.

The proposed methodology is illustrated in Fig. 1. In order to preserve anonymity and privacy for the speakers and their interlocutors, speech data was recorded using an accelerometer attached to the temporomandibular joint, amplified and sampled at a frequency of 16 kHz. This recording method offered high rejection of noise and the speech of others while preserving sufficient of the subjects speech to try to identify repeated sections. The voice segments extracted from recordings were bandpass filtered, denoised and divided into frames for feature extraction.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

A set of 90 features per frame were calculated including statistical measures, Mel-frequency cepstral coefficients, perceptual linear predictive cepstral coefficients and prosodic measures. Principal component analysis (PCA) was then applied on the feature vector and the first principal component only retained for processing. Motif detection was undertaken by using PAA/SAX conversion [5] to transform the data in a way more convenient for manipulation using symbolic words and organizing them into buckets containing all time-series subsequences with the same hash value. Searching for a motif was then performed within those smaller groups of subsequences rather than on the whole time series. This rearrangement of the data is beneficial as it narrows and accordingly speed-up the search. A search algorithm that calculated sections of a similarity matrix based on the Approximate Distance Map (ADM) algorithm [6] was then used to identify candidate motifs likely to be associated with repeated speech patterns.

III. Testing

A proof of concept experimental test was carried out on 5 healthy subjects. The subjects were instructed to read aloud scripts containing short, embedded, repeated questions and statements. The acceleration and speech signals were simultaneously recorded, pre-amplified and passed to a host computer. Audio recording via a conventional headset microphone served as a reference to validate the accelerometer data. Positions of the repeated motifs in the recordings were marked by listening to the recorded speech and saved for verification purposes. The approximately 4 minute long recordings were then processed as described and the percentage of correctly detected patterns was found to be between 57-71% depending on the subject and the choice of algorithm parameters.

IV. Discussion

The initial results indicate that the adopted approach has potential for detecting repetitive speech patterns. Further improvements are underway including fine-tuning of the algorithm parameters with aim of improving the detection rate. If successful, the technique could form the basis of a device for monitoring performance in, for instance, trials of disease modifying drugs for AD.

In the next stage of the study we will record bone-conducted speech in patients with possible or probable AD to establish the feasibility and tolerability of the proposed methodology in a natural environment. Future work will include larger subject populations, with the aims of: correlating rates of perseveration with stage and progression in dementia; calibrating treatment effects from established interventions; and defining distinct patterns of perseverative speech. We believe the cognitive performance marker outlined here could potentially have the unique benefit of quantifying severity clinically, providing an invaluable adjunct to the range of biological indices currently available.

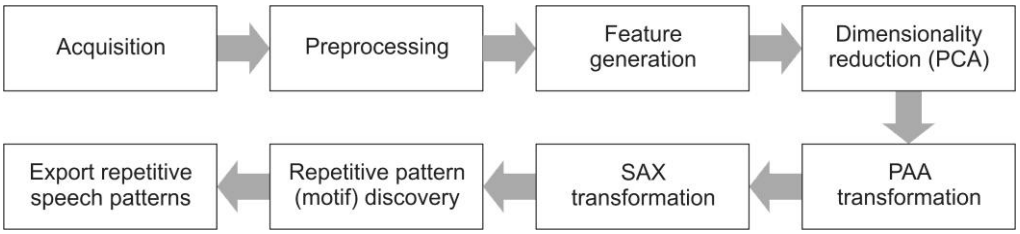


Fig. 1. Block scheme of the system for detection of the repetitive speech patterns from the bone-conducted speech signal

References

[1] Brayne, C., (2006) Incidence of dementia in England and Wales – The MRC cognitive function and ageing study. *Alzheimer Disease & Associated Disorders*, 20**,** S47-S51.

[2] Cohen, G. (1996) *Memory in the real world,* Hove, E. Sussex, Psychology Press.

[3] Pekkala, S., Albert, M.L., Spiro, 3rd A., Erkinjuntti, T., (2008) Perseveration in Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 25(2), 109-114.

[4] Bayles, K.A., Tomoeda, C.K., Mcknight, P.E., Helm-Estabrooks, N., Hawley, J.N., (2004) Verbal perseveration in individuals with Alzheimer's disease. *Seminars in Speech and Language*, 25**,** 335-347.

[5] Lin, J., Keogh E., Lonardi, S., Patel, P., (2002) Finding motifs in time series. *In Proc. of the 2nd Workshop on Temporal Data Mining*, 53-68.

[6] Shasha, D., Wang, T., (1990) New techniques for best-match retrieval. *ACM Trans. on Information Systems*, 8(2), 140-158.

ABNORMAL RHYTHMS OF SPEECH IN PATIENTS WITH IDIOPATHIC PARKINSON'S DISEASE

A. Bandini [1,2], F. Giovannelli[3], M. Cincotta[3], P. Vanni[3], R. Chiaramonti[3], A. Borgheresi[3], G. Zaccara[3], C.Manfredi[1]

[1] Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
andrea.bandini@unifi.it, claudia.manfredi@unifi.it
[2] Department of Electrical, Electronic and Information Engineering (DEI) "Guglielmo Marconi", Università di Bologna, Bologna, Italy
[3] Unit of Neurology, Florence Health Autority, Ospedale "Nuovo San Giovanni di Dio", Firenze, Italy

Abstract: **Parkinson's disease (PD) involves impairments of voice an speech (hypokinetic dysarthria). Dysprosody is one of the most common features of PD speech, that includes alterations of rhythm and velocity of articulation. The aim of this study is the evaluation of rhythm alterations of speech in Parkinsonian patients during a sentence repetition task. 13 PD patients (9 male and 4 female) and 9 healthy controls (4 male and 5 female) were tested. Results show significant differences between the two groups as far as parameters T_{inter} (time interval between two consecutive sentences) and D (percent of speech time with respect to sentence duration) are concerned. In particular, T_{inter} is larger in PD patients while D is higher in the control group. These preliminary results show that PD patients may exhibit longer pauses between adjacent sentences and a lower percentage of "speech time" during a whole repetition period. Thus, the decrease of D leads to an increase of NSR (Net Speech Rate, defined as the number of syllables per second). This study confirms that speech in PD patients is characterized by short rushes followed by inappropriate pauses.**

Keywords : **Parkinson's disease, dysprosody, hypokinetic dysarthria, speech rhythm, speech motor performance**

I. INTRODUCTION

Idiopathic Parkinson's disease (PD) is a neurodegenerative illness that involves neurons in the zona compacta of the substantia nigra of the midbrain and other pigmented nuclei [1-3]. This pathology is associated with a great variety of motor (tremor, stiffness, bradykinesia, postural instability) and non-motor symptoms (depression, cognitive impairments, sleep and mood disorders), that significantly reduce the quality of life of patients. [4,5]. The main signs of the disease include impairments of voice and speech (hypokinetic dysarthria), which affect about 70% of PD patients [6].

These subjects might present alterations related to all speech dimensions (i.e. respiration, phonation, articulation and prosody). Therefore, PD patients might present reduced variability of pitch and loudness, hoarseness, reduced stress, imprecise consonant articulation, inappropriate speech silence and speech rate alterations [7]. Such variations may debut in the early stage of the disease [8].

These reasons, together with the non-invasivity of the acoustic measurements of voice, have brought researchers to identify features of voice and speech that could discriminate PD patients from healthy controls, as an aid to early diagnosis and tracking of the disease progression.

One of the most common speech impairments in Parkinson's disease is dysprosody, that includes alterations of rhythm and velocity of speech, velocity of articulation and speech/pause ratio [7]. Many authors conducted studies on prosodic patterns of voice in PD patients, assessing parameters related to speech rate. Skodda et al. [9] found a speech rate variation (defined as the number of syllables per second) that follows the evolution of the disease. This modification is characterized by an articulatory acceleration in the early stages of the disease, followed by a slowing in advanced stages. Speech rate measures were also used to test the dopaminergic therapy effects on Parkinsonian voice. However, no significant differences were found during this pharmacological treatment [10]. A widespread task, used to evaluate speech rhythm disorders, is the syllable repetition task (oral diadochokinetic test). It was shown that PD patients tend to have less control on rhythm stability during this task, with a tendency to increase the pace of repetition. These results reflect a dysfunction at the level of basal ganglia that control the temporal regulation of a motor sequence [11]. Rhythm alterations derived from the oral diadochokinetic test were found also by Rusz et al. [12], where PD patients presented a fewer syllables per second with respect to healthy control subjects. Moreover, the evaluation of prosody from the reading of a text or a free monologue showed differences in the mean number of pauses between PD patients and controls.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

The aim of this study is to identify the presence of rhythmic alterations in the parameters related to speech rate in PD patients during a sentence repetition task.

II. METHODS

Subjects: 13 patients with idiopathic Parkinson's disease were recruited at the Department of Neurology of the Hospital "San Giovanni di Dio" of Firenze, Italy. Patients' age ranged from 53 to 83 years (mean: 72 years; standard deviation: 9.7 years), 9 patients were male, while the other 4 were female. At the moment of the experiment, disease duration ranged from 3 to 12 years (mean: 7.2 years, SD: 2.9 years). All PD patients were under levodopa medication and were tested during their "on" state.
A group of 9 healthy subjects was tested as control group (age: 34-73 years, mean: 59.6 years, standard deviation: 12.8 years), 4 male and 5 female.
All subjects were Italian native speakers. Signed informed consent was obtained from all the participants.

Experimental settings: Each subject was asked to repeat a standardized Italian vocalic sentence ("*Il bambino ama le aiuole della mamma*") at least 10 times, as spontaneously as possible, at comfortable loudness.
Speech signals were recorded on a standard personal computer using Audacity software (version: 2.0.3) with a Shure SM58 microphone and a Tascam US-144 board. The samples were digitized at a frequency of 44.1 kHz. The microphone, fixed on a boom, was positioned at a distance of 5 cm from the subject's mouth. The experiments were conducted in a quiet room of the "San Giovanni di Dio" hospital, and subjects were required to remain seated during the test.

Analysis: The acoustic analysis of the acquired signals was carried out at the Biomedical Engineering Laboratory, Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy.
As a first step an automatic voiced-unvoiced segmentation was performed in order to identify and "isolate" sentences in the whole signal. The algorithm implemented [13] splits the whole signal in short frames of the same length whose energy is evaluated and stored in an "energy vector". The Otsu's method applied to the energy histogram allows finding two thresholds that allow the separation between two classes (voiced and unvoiced frames). An example is reported in Fig. 1 where the test sentence is successfully segmented into two voiced parts.
Afterwards the following parameters were extracted:

- $T_{sentence}$: sentence duration (in seconds), calculated as the time interval between the beginning of a sentence and that of the next one;
- T_{inter}: inter-sentence duration (in seconds), calculated as the time interval between the end of a sentence and the beginning of the next one;
- T_{pause}: pause duration (in seconds), calculated as the sum of breaks inside a sentence;
- D: Duty Cycle, defined as the percent of voiced time with respect to the sentence duration;

$$D = \frac{T_{sentence} - T_{inter} - T_{pause}}{T_{sentence}} \times 100 \qquad (1)$$

- NSR: Net Speech Rate in syllables/s, defined as the number of syllables of the sentence, divided by the effective speech time ($T_{sentence} - T_{inter} - T_{pause}$)[13].

A customized algorithm was developed under Matlab R2012a tool for the extraction of these features, while the voiced-unvoiced segmentation was performed using the software BioVoice2 [14].
A two-tailed t-test was applied to assess the significance of differences between the two groups. Statistical analysis was performed with Microsoft Excel 2010.

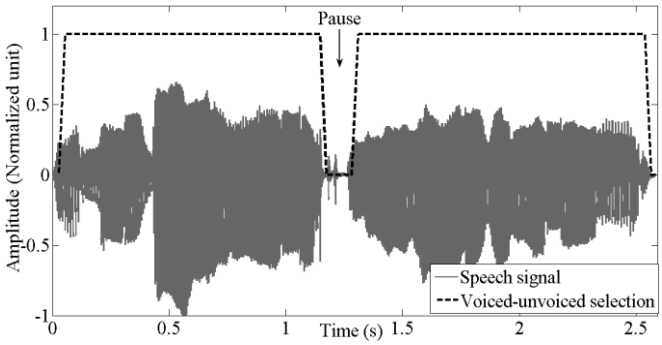


Fig. 1: Voiced-unvoiced segmentation performed on the standardized sentence used in this study.

III. RESULTS

Mean values and standard deviations of parameters are reported in Tab. I. As not all participants had breaks inside sentences, the parameter T_{pause} was considered only for D and NSR calculation.
From these results it can be noticed that significant differences exist between the two groups for parameters T_{inter} and D. In particular, the time interval between two adjacent sentences is larger in PD patients ($p = 0.01$), while duty cycle is larger in the control group ($p = 0.007$). Also NSR is larger in PD patients, although this difference is not significant ($p = 0.057$). No significant differences were found in the $T_{sentence}$ parameter.
A weak positive correlation exists between disease duration (number of years) and $T_{sentence}$ ($R = 0.50$), T_{inter} ($R = 0.41$) and NSR ($R = 0.20$), while the correlation with D is negative ($R = -0.35$).

At visual inspection no significant trend or pattern (increasing or decreasing) during the sentence repetition task was found in these four parameters.

Table I: Summary of the results with mean values, standard deviations and significance of differences between control and PD patients

	Controls		PD patients		
	Mean	S.D.	Mean	S.D	p
$T_{sentence}$ (s)	2.96	0.54	3.44	1.25	0.24
T_{inter} (s)	0.55	0.28	0.94	0.38	**0.01**
D (%)	79.59	7.79	65.66	13.94	**0.007**
NSR (syll/s)	6.14	0.66	7.00	1.32	0.057

IV. Discussion

The results of this study show that PD patients exhibit an alteration of prosodic patterns of speech during a sentence repetition task. With respect to control subjects, PD patients have longer pauses between two consecutive sentences and a lower percentage of "speech time" during an entire repetition period. Decrease of duty cycle leads to an increase of NSR, as shown in eq. (1). Therefore, PD patients tend to repeat the same sentence in a shorter time period than healthy controls, but at the expense of a longer recovery time, since no significant difference was found in the $T_{sentence}$ parameter.

This findings are in accordance with other studies. Skodda et al. [9] showed that PD patients have an articulatory acceleration in the early stages of the disease, followed by slowdown in advanced stages. Our findings may be due to the presence of an early stage group of Parkinsonians as disease duration is similar to that in Skodda et al. [9] (7.2 years vs 6.44). It will be highly interesting to carry out a new test on these patients to look for possible slowdown of speech rate. Furthermore, the increase of speech rate seems to be prevalent in male patients. The larger number of male patients with respect to females seems to substantiate this argument, although hereafter it will be necessary to take into account a gender-based analysis.

At first glance, our results on T_{inter} (time interval between two consecutive sentences) could be in contrast to findings in [9] where PD patients (both males and females) showed a less percentage of pauses than healthy controls. In fact, in our study, the pause between two sentences is on average higher in PD patients than in control group. However, it should be noted that in [9] the task is not the repetition of a sentence, but the reading of a passage composed by 4 sentences.

Other studies show that alterations in speech rhythm could be pointed out by the widespread syllable repetition task (repetition of syllables /pa/ or /pa/ /ta/ /ka/). It was shown that this test can reveal vocal pace variations, with articulatory acceleration [11][14]. Thus, it will be very interesting to correlate the variations of speech rate shown by our test with those obtained from the syllable repetition task.

The advantage of our method is that the chosen sentence (which consists mainly of vowel sounds) not only allows an automatic segmentation within the whole signal, but also the estimation of other acoustic parameters (fundamental frequency, jitter, shimmer, noise, etc.). In fact, as reported in [12], one of the most discriminative acoustic measures in the analysis of speech and voice of PD is the variation of fundamental frequency during a monologue. However a free monologue task would not be useful for an automatic (and possibly remote) identification of hypokinetic dysarthria since it does not allow for comparing participants. Moreover, Skodda et al. [16] showed that tVSA (total Vowel Space Area) and VAI (Voice Articulation Index), calculated from the first and the second formant frequencies of the vowels /a/, /i/ and /u/, are predictive of PD progression. Since our sentence ("*Il bambino ama le aiuole della mamma*") contains all these vowels, it would be possible to evaluate also these parameters.

Future work will concern the kinematic analysis of articulators (in particular jaw and lips) during speech. It was shown that, during a sentence repetition task, PD patients have lower amplitude and speed of lower lips in the articulation of syllables that contains plosive consonants [17] with respect to control subjects. The implementation of marker-less technology for face tracking would be beneficial, since it can be used together with acoustic measures of voice.

Thus future work will concern the automatic extraction of acoustic and kinematic measures from this sentence repetition task, in order to identify alterations in rhythmicity, acoustic and kinematic parameters related to voice articulation, to identify and quantify hypokinetic dysarthria in Parkinson disease.

V. Conclusion

This study presents some preliminary results on rhythmic alterations of speech in PD patients during a sentence repetition task. It was shown that patients with Parkinson's disease may present a higher speech rate but longer time between two repetitions with respect to control subjects. These differences could be useful for an automatic real-time classification of PD speech, in order to develop a classification algorithm for early diagnosis and for tracking the disease progression.

Further studies will concern the acoustic analysis of speech samples as related to kinematic marker-less analysis of articulators in order to find other differences related to dysprosody in Parkinson's disease.

REFERENCES

[1] P. Damier, E.C. Hirsch, Y. Agid and A.M. Graybiel, "The substantia nigra of the human brain II. Patterns of loss of dopamine-containing neurons in Parkinson's disease", *Brain*, vol. 199, pp. 1437-1448, 1999.
[2] H. Braak, K. Del Tredici, U. Rüb, R.A.I. de Vos, E.N.H. Jansen Steur, E. Braak, "Staging of brain pathology related to sporadic Parkinson's disease", *Neurobiology of Aging,* vol. 24, pp. 197-211, 2003.
[3] J. M. Dickson, R. A. Grünewald, "Somatic symptom progression in idiopathic Parkinson's disease" *Parkinsonism and Related Disorders*, vol. 10, pp. 487-492, 2004.
[4] D.W. Dickson, "Parkinson's disease and parkinsonism: neuropathology", *Cold Spring Harb Perspect Med*, vol. 2(8), pp. 1-15, 2012.
[5] K.R. Chaudhuri, D.G. Healy, A.H.V. Schapira "Non-motor symptoms of Parkinson's disease: diagnosis and management", *Lancet Neurol*, vol. 5, pp. 235-245, 2006.
[6] L. Hartelius, P. Svensson, "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey", *Folia Phoniatr Logop*, vol. 46, pp. 9-17, 1994.
[7] F.L. Darley, A.E. Aronson, Brown J.R, *Motor Speech Disorders,* Saunders, Philadelphia, PA, 1975, pp. 1-305.
[8] B.L. Harel, M.S. Cannizzaro, H. Cohen, N. Reilly, Snyder P.J., "Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment", *Journal of Neurolinguistics*, vol. 17, pp. 439-453, 2004.
[9] S. Skodda, H. Rinsche, U. Schlegel, "Progression of dysprosody in Parkinson's disease over time – a longitudinal study", *Movement Disorders*, vol. 24, pp. 716-722, 2008.
[10] S. Skodda, Visser W., U. Schlegel, "Short- and long-term dopaminergic effects on dysarthria in early Parkinson's disease", *J Neural Transm*, vol. 117, pp. 197-205, 2010.
[11] S. Skodda, A. Flasskamp, U. Schlegel, "Instability of syllable repetition as a model for impaired motor processing: is Parkinson's disease a "rhythm disorder"?", *J Neural Transm*, vol. 117, pp. 605-612, 2010.
[12] J. Rusz, R. Cmejla, H. Ruzickova, E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untereated Parkinson's disease", *J Acoust Soc Am*, vol. 129(1), pp. 350-367, 2011.
[13] S. Orlandi, P.H. Dejonckere, J. Schoentgen, J. Lebacq, N. Rruqja, C. Manfredi, "Effective pre-processing of long term noisy audio recordings: an aid to clinical monitoring", *Biomedical Signal Processing and Control*, vol. 8, pp. 799-810, 2013.
[14] A. Flasskamp, S.A. Kotz, U. Schlegel, S. Skodda, "Acceleration of syllable repetition in Parksinon's disease is more prominent in the left-side dominant patients", *Parkinsonism and Related Disorders*, vol. 18, pp. 343-347, 2012.
[15] C. Manfredi, L. Bocchi, G. Cantarella, "A multipurpose user-friendly tool for voice analysis: application to pathological adult voices", *Biomedical Signal Processing and Control*, vol. 4, pp. 212-220, 2009.
[16] S. Skodda, W.Grönheit, U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinon's disease", *PLoS ONE*, vol. 7(2), e32132, 2012.
[17] B. Walsh, A. Smith, "Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease", *Movement Disorders*, vol. 27(7), pp. 843-850, 2012.

Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms

Athanasios Tsanas[1,2]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

[2] Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, UK

Contact: tsanas@maths.ox.ac.uk, tsanasthanasis@gmail.com

Abstract: **Epidemiological studies suggest that lifetime prevalence of voice disorders is about 30% for the general adult population. Moreover, vocal performance degradation may be amongst the earliest indicators of a neurodegenerative disease onset, such as Parkinson's disease. Lacking alternative cost-effective biomarkers, biomedical speech signal processing has been gaining increasing impetus towards developing clinical decision support tools. Acoustic analysis of speech signals provides a convenient, automatic, accurate, robust, inexpensive, scalable approach assisting medical diagnosis and symptom severity monitoring. Nevertheless, the algorithmic tools developed for biomedical speech signal processing are spread across different software platforms, hindering direct algorithmic comparisons and the further development of this impending field. This study brings many biomedical speech signal processing algorithms together under the same software platform, and has led to the development of a practical free toolkit which can be accessed over the Internet using a simple application.**

Keywords: **Acoustic analysis, decision support tool, speech signal processing, sustained vowels**

I. INTRODUCTION

Neurodegenerative disorders and speech performance degradation have been closely associated at least since the 1970's. For example, the vast majority of Parkinson's disease (PD) subjects experience some problems associated with their voice [1]. In addition, the lifetime prevalence of general voice disorders is approximately 30% for the adult population, and have a substantial impact on a person's personal and professional life [2]. Regardless of pathological cause, voice disorders are characterized by the malfunction of one or more components in the vocal production mechanism, leading to poor vocal quality. Depending on the pathological cause, characteristic symptoms may include reduced or increased loudness, vocal tremor, and breathiness amongst others.

In speech clinical practice, human experts assess a subject's voice quality using *sustained vowel* phonations, and/or *conversational speech*. Sustained vowels, where the subject is asked to prolong his phonation for as long as possible and as steady as possible (in terms of pitch and amplitude), are particularly practical because they circumvent linguistic artifacts and are considered sufficient for many voice assessment applications [3].

The lack of a sufficient number of experts to perform vocal assessments has prompted the development of *biomedical speech signal processing algorithms*, to objectively and automatically characterize clinically useful properties of the speech signals. There is considerable research on the topic of developing clinical decision support tools using speech signals, in particular to study

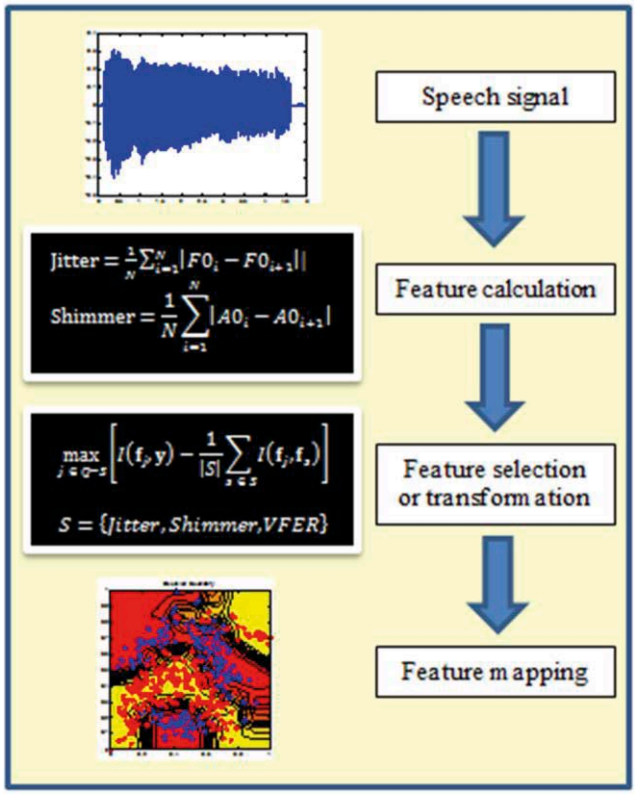


Fig. 1. Methodology for biomedical speech signal processing.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

neurodegenerative disorders [4-11]. The nature of this multi-disciplinary domain has attracted the attention of speech experts, phoneticians, clinicians, mathematicians, and engineers, and the algorithmic tools are dispersed in the literature across research areas. In many cases there is no readily freely available implementation of the algorithms; alternatively, the source code is provided by its developers in diverse software platforms, hindering their widespread use by practitioners.

This study describes the framework for the development of an automated clinical decision support tool, bringing together many of the biomedical speech signal processing algorithms which were previously scattered in the research literature. It provides the conceptual basis for the development of the current state of the art biomedical speech signal processing algorithms, highlighting the intricate characteristics of the speech signals which the algorithmic tools aim to characterize. Finally, there are some suggestions about directions future work could take in the quest of extracting clinically useful information from the speech signals which may not be adequately quantified using the currently available algorithmic tools.

II. METHODS

The methodology of a clinical decision support tool is summarized in Fig. 1. It consists of processing the original raw time series (speech signal) to extract distinctive, clinically useful properties (*feature calculation*), selecting or transforming the computed speech signal properties (*feature selection* or *feature transformation*), and mapping the final compact feature subset to the clinical outcome we want to associate the speech signal with (*feature mapping*). It is critical to emphasize that the developed algorithms have only been validated in the sustained vowel /a/ setting; it is not clear how meaningful and useful the output of the presented algorithmic tools is for other settings.

This study focuses on the computation of the speech signal properties and very briefly outlines the other two components.

A. Computation of features

This section summarizes briefly the most widely used biomedical speech signal processing algorithms, clustering them into algorithmically-related families. For a detailed overview, see Tsanas [11]. In principle, any signal processing tool and any time series analysis tool could be used to extract characteristics from the speech signal which might be useful for clinical applications.

Many speech signal processing algorithms rely on the accurate computation of the fundamental frequency (F0) which is critical to characterize speech signals [12]. Defining F0 is not straightforward in non-periodic signals (i.e. all pathological cases) and the development of robust F0 estimators is intensively pursued [13]. On the current evidence, the NDF [14] and SWIPE F0 estimators [15] appear very promising in processing sustained /a/ vowels, based on extensive comparisons against reference synthetic speech data created by a state of the art physiological model [11].

The first algorithmic group builds on physiological evidence that the vocal folds' oscillating pattern is nearly periodic in healthy voices and substantially departs from periodicity in pathological cases [3]. Two of the most well-known algorithms, *jitter* and *shimmer*, belong to this group [16], [3]. Jitter quantifies F0 deviations, and shimmer quantifies amplitude deviations. There is no unique formal definition of jitter and shimmer, and researchers have developed many *jitter variants* and *shimmer variants*. Similarly, the *Recurrence Period Density Entropy* (RPDE), the *Pitch Period Entropy* (PPE), the *Glottal Quotient* (GQ), and other F0-related measures [11] build on the concept of quantifying the extent of aperiodicity in the vocal folds' oscillating pattern.

The second general algorithmic group comprises signal to noise ratio (SNR) approaches. The rationale is that incomplete vocal fold closure leads to the creation of aerodynamic vortices, leading to increased acoustic noise. *Harmonic to Noise Ratio* (HNR), *Detrended Fluctuation Analysis* (DFA), *Glottal to Noise Excitation* (GNE), *Vocal Fold Excitation Ratio* (VFER), and *Empirical Mode Decomposition Excitation Ratio* (EMD-ER) express this concept algorithmically.

Linear Predicting Coding Coefficients (LPCC) is a generic signal processing tool: the underlying concept is that pathological voices are be more irregular and hence more difficult to predict based on past samples. Wavelets are another generic tool to analyzing non-stationary time series signals: we proposed using wavelet decomposition to analyze the F0 contour, and use the wavelet coefficients as features [17]. Lastly, *Mel Frequency Cepstral Coefficients* (MFCCs) are the gold standard in speaker recognition, and have recently shown great promise in biomedical applications [18], [9], [10].

The key information of the biomedical speech signal processing algorithms included in the toolkit is summarized in Table 1.

B. Feature selection or feature transformation

A common problem in applications with many features is the *curse of dimensionality*: a compact feature subset may lead to improved performance and promotes *interpretability* by means of inferring the most prominent properties of the speech signals for the investigated problem [19]. There are two main approaches: feature selection (selecting a subset of the original features), and feature transformation (transforming the original features to develop new, more predictive features). This falls outside the scope of this study; we refer to Guyon et al. [20] and Hastie et al. [19] for a comprehensive and authoritative overview.

Table 1: Summary and key information of the biomedical speech signal processing algorithms in the toolkit

Measure	Motivation	Features
Jitter & Jitter variants	The vocal folds are affected in voice disorders, and jitter aims to capture instabilities of the oscillating pattern of the vocal folds quantifying the cycle-to-cycle changes in *fundamental frequency*	One for each variant
Shimmer & shimmer variants	Shimmer aims to capture instabilities of the oscillating pattern of the vocal folds quantifying the cycle-to-cycle changes in *amplitude*	One for each variant
Recurrence Period Density Entropy (RPDE)	Quantifies the stochastic component of the deviation of vocal fold periodicity	1
Pitch Period Entropy (PPE)	In speech disorders it is very difficult to sustain stable pitch due to incomplete vocal fold closure. PPE quantifies the impaired control of stabilised pitch.	1
F_0-related measures	Summary statistics of F_0, differences from expected F_0 in age- and gender- matched controls, variations in F_0	Three for each F_0 estimation algorithm
Harmonics to Noise Ratio (HNR) & Noise to Harmonics Ratio (NHR)	In speech pathologies there is increased noise due to turbulent airflow, resulting from incomplete vocal fold closure. HNR and NHR quantify the ratio of actual signal information over noise.	4
Detrended Fluctuation Analysis (DFA)	Quantifies the stochastic self-similarity of the noise caused by turbulent airflow	1
Glottal to noise excitation (GNE)	Extent of noise in speech using energy and nonlinear energy concepts	6
Vocal fold excitation ratio (VFER)	Extent of noise in speech using energy, nonlinear energy, and entropy concepts	9
Empirical mode decomposition excitation ratio (EMD-ER)	Signal to noise ratios using EMD-based energy, nonlinear energy and entropy	6
Linear Predicting Coding Coefficients (LPCC)	Quantify deviations of the prediction of the current data sample as a function of the preceding samples. In pathological voices this deviation is expected to be larger.	10
Wavelet measures	Quantify deviations in F_0 (obtained using any F_0 estimation algorithm)	180
Mel Frequency Cepstral Coefficients (MFCC)	Voice pathologies lead to decreased control of the articulators (vocal tract) , and the MFCCs attempt to analyse the vocal tract independently of the vocal folds	12-42, depends on extracted components and the use of delta and delta-delta coefficients

C. *Feature mapping*

So far, we have described the methodology towards extracting speech signal properties (features), and determining a robust representation of the extracted information. In order to develop a complete decision support tool it is necessary to associate the computed features with the clinical outcome of interest. In order to achieve this, a database of *labeled* data is required: the user has to provide a database of speech signals where each speech signal (the *.wav file) corresponds to the known clinical outcome. Then, a supervised learning *mapping algorithm* associates the features with the clinical outcome (response). The aim is to subsequently use *unlabeled* data and interrogate the mapping algorithm to provide estimates of the labels. One suggestion is using *Random Forests*, a robust statistical machine learning tool which has shown impressive performance without requiring any hyper-parameter optimization. We refer to Hastie et al. [19] for details.

III. RESULTS

The toolkit requires a single input (*.wav file) and provides an output vector with the computed features along with the corresponding feature names. Optionally, it can also output the F0 contour for visualization and further processing.

IV. DISCUSSION

This study brings together many biomedical speech signal processing algorithms under a common software platform, resulting in the development of a new toolkit. This may facilitate further advances in extracting clinical-

ly useful information from speech signals, enabling practitioners easily apply known algorithms to their data, and researchers compare new algorithmic concepts to the existing literature. Hitherto, the existing speech signal processing algorithms were dispersed in the literature across research areas, a fact which reflects the multidisciplinary nature of this field. Although some researchers have open-sourced their algorithmic contributions, there was no standardized software platform to use these tools so far.

The toolkit has only been validated on sustained vowel /a/ phonations; its functionality may be extended to additional vowels and their interaction, for example to compute the vowel space area and related algorithms. The default settings have been optimized for biomedical applications in Parkinson's disease [9-11], but the toolkit is flexible to allow experienced users adjust additional parameters. Currently, there is no provision to extend the analysis to conversational speech, although this might be an interesting approach to pursue in the future.

A critical aspect of biomedical speech signal processing is *interpretability*. That is, determining the most distinctive properties of the speech signals for the investigated application, be means of observing the (jointly) most predictive subset of speech signal processing algorithms. To this end, it is often advisable to investigate many diverse feature selection algorithms to gain insight into the pathophysiological voice characteristics which are mostly predictive of the investigated clinical outcome [21]. Moreover, given a sufficiently large sample size, it is often useful to stratify the data by gender [9], [22].

Future studies could further investigate applying robust time-series analysis tools to this domain. However, probably the most promising approach towards developing new biomedical speech signal processing tools may be studying closely the physiology of the investigated application. Different voice disorders have different pathophysiological characteristics, and improved acoustic results might be obtained by developing new *physiologically-informed* algorithms targeting specific characteristics.

REFERENCES

[1] R. Pahwa, and E. Lyons. (Eds.) *Handbook of Parkinson's Disease*, 4th edition, Informa Healthcare, USA, 2007

[2] N. Roy, R.N. Merrill, S.D. Gray, E.M. Smith, "Voice disorders in the general population: Prevalence, risk factors, and occupational impact", *Laryngoscope,* Vol. 11, pp. 1988-1995, 2005

[3] I.R. Titze, *Principles of Voice Production*. National Center for Voice and Speech, Iowa City, US, 2nd printing, 2000

[4] Praat: doing phonetics by computer (Version 5.1.15) [Computer program], by P. Boersma and D. Weenink. Retrieved from http://www.praat.org/, 2009

[5] J.I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruíz, S. Aguilera-Navarro, P. Gómez-Vilda, "An integrated tool for the diagnosis of voice disorders", *Medical Engineering and Physics*, Vol. 28 (3), pp. 276-289, 2006

[6] S. Skodda, W. Gronheit, U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease", *Plos One*, 7(2): e32132. doi:10.1371/journal.pone.0032132, 2012

[7] J. Rusz, R. Èmejla, H. Ružièková, J. Klempíø, V. Majerová, J. Picmausová, J. Roth, E. Rùžièka, "Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy", *Journal of Neural Transmission*, Vol. 120, pp. 319-329, 2013

[8] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests", *IEEE Transactions on Biomedical Engineering*, Vol. 57, pp. 884-893, 2010

[9] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity", *Journal of the Royal Society Interface*, Vol. 8, pp. 842-855, 2011

[10] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease", *IEEE Transactions on Biomedical Engineering*, Vol. 59, pp. 1264-1271, 2012

[11] A. Tsanas: *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*, D.Phil. thesis, University of Oxford, Oxford, UK, (2012)

[12] M.G. Christensen and A. Jakobsson, *Multi-pitch estimation*, Synthesis lectures on speech and audio processing, Morgan & Claypool Publishers, 2009

[13] R.M. Roark, "Frequency and Voice: perspectives in the time domain," *Journal of Voice*, Vol. 20, No. 3, pp. 325-354, 2006

[14] H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi, T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Interspeech*, pp. 537-540, Lisbon, Portugal, September 2005

[15] A. Camacho, J.G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, Vol. 124, pp. 1638-1652, 2008

[16] R.J. Baken, R.F. Orlikoff, *Clinical measurement of speech and voice*, San Diego: Singular Thomson Learning, 2nd ed., 2000

[17] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity", *International Symposium on Nonlinear Theory and its Applications* (NOLTA), pp. 457-460, Krakow, September 2010

[18] J.I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters", *IEEE Transactions on Biomedical Engineering*, Vol. 53, pp. 1943-1953, 2006

[19] T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd ed., 2009

[20] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications*, Springer, 2006

[21] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Robust parsimonious selection of dysphonia measures for telemonitoring of Parkinson's disease symptom severity", *7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (MAVEBA), Florence, Italy, pp. 169-172, 2011

[22] R. Fraile, N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, C. Fredouille, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex", *Folia Phoniatrica et Logopaedica*, Vol. 61, pp. 146-152, 2009

Session I:
MODELS AND ANALYSIS (I)

PRESSURE AND VELOCITY IN A MODEL OF LARYNGEAL VENTRICLE

Fariborz Alipour

Department of Communication Sciences and Disorders
University of Iowa, Iowa City, Iowa, USA, alipour@iowa.uiowa.edu

Abstract: **A self-oscillatory computational phonatory model with ventricular folds was developed based upon prior validated vibration and airflow models. Using this model, pressure and velocity distributions were investigated. Vibration of the vocal folds was modeled with a finite element method and laryngeal flows and pressures were simulated with the solution of unsteady Navier-Stokes equations using a finite volume method. The results suggested that true vocal folds self-oscillated regularly with reasonable amplitude and mucosal waves. In each cycle of oscillation, a flow vortex was generated and moved downstream. Pressures were planar in most of the flow region. There were large gradients in the glottal region. The centerline velocity was highest during glottal closing and sharply dropped when near the center of the flow vortex.**

Keywords: **finite-element, vocal folds, glottal pressure, velocity pattern.**

I. INTRODUCTION

The laryngeal ventricle is the cavity just above the true vocal folds that separates them from the ventricular (false) folds. The understanding of aerodynamic and biomechanical conditions that cause ventricular folds to oscillate or intervene with the true vocal fold oscillations can reveal the possible detrimental effects of this supraglottic structure. During phonation, glottal flow passes by this ventricle as a pulsating jet and might influence the air pressure in the ventricle and in turn affect the aerodynamic forces on the true (TVF) and false vocal folds (FVF). The ventricular folds are usually further apart than true vocal folds and do not oscillate, but they have non-phonatory tasks of air flow control such as breath holding and articulation during glottal stops. However, when they grow abnormally or adducted due to neuromuscular disease, they might oscillate and cause voice disorders. Due to its aerodynamic and acoustic effects on phonation [1, 2] this structure has been the subject of many recent experimental and computational investigations. The purpose of this study was to develop an oscillatory computational model that predicts pressure and velocity distributions within a model of the laryngeal airway.

II. METHODOS

A biophysical phonatory model was built upon a previously developed and validated computational model. Vibration of the vocal folds was modeled with finite-element model [3] and the laryngeal flow was simulated with the solution of unsteady Navier-Stokes equations using finite volume method [4]. The details of the finite-element formulation and solution can be found in Alipour et al. [3] and details of the air flow solution during phonation can be found in Alipour et al. [4]. They are briefly described here. The vocal fold motion is defined by two-dimensional displacement vector which is continuously distributed through the vocal folds. It is assumed a planar oscillatory motion for the vocal folds (in coronal plane). The displacement vector can be defined as

$$\psi = \xi(x,y,z,t)\vec{i} + \eta(x,y,z,t)\vec{k} \qquad (1)$$

Where ζ and η are the lateral (x) and vertical (z) components of the displacement vector ψ. The vocal folds including the true and false folds are divided into 15 thin layers. The mesh design is shown in **Fig. 1**, where each layer discretized into 84 triangular elements in the true vocal fold and 20 triangular elements in the ventricular (false) folds. The vocal folds tissue include body (TA muscle), cover, ligament, and ventricular, each with separate mechanical properties. Within each element, the displacement field is described as a linear function of local coordinates. Once the potential and strain energy are described in terms of forces and deformations that were related to displacement field, the governing equations is a second order (in time) matrix differential equation for each element.

$$M\ddot{\psi} + D\dot{\psi} + K\psi = F \qquad (2)$$

With the matrix coefficients of M (mass), D (damping), and K (stiffness) that are 6x6 matrices. F is force vector. When these equations are combined for all elements within each layer, the resulting global system of equations includes all the nodes in the layer and in matrix form is similar to Eq. 2 except that displacement and force vectors have twice as many elements as the number

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

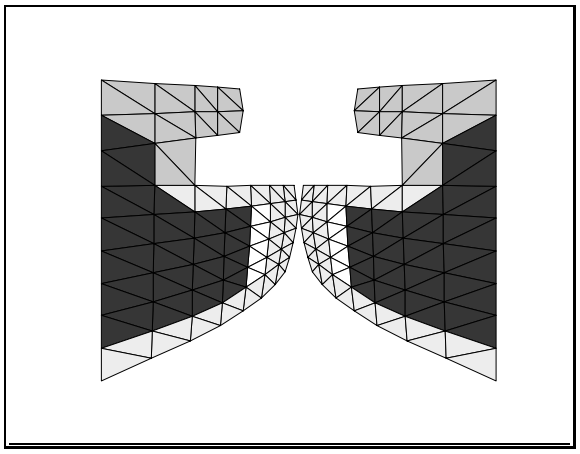


Fig. 1 – F**inite-element mesh design**

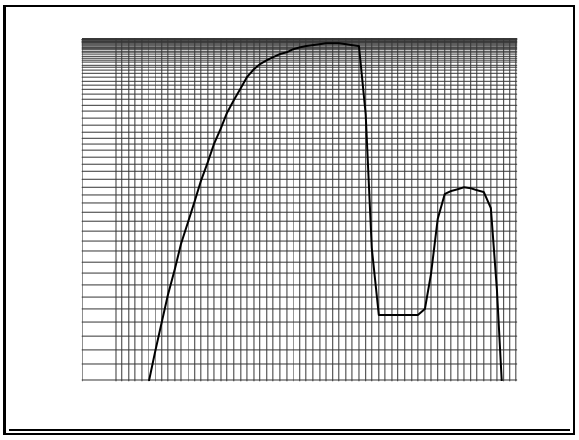


Fig. 2 – Portion of the flow mesh design

of nodes. At every time step (50 microseconds), the global matrices are first calculated; then the pressure distribution from the airflow is obtained and used to calculate the forcing vector. Once the solution is found for the displacement of the vocal folds, the boundary conditions are enforced by looking at the nodal degrees of freedom. The nodes that are stationary or fixed are excluded from the solution scheme after the assembly process. Whenever the nodes on the medial surface of the vocal fold approached the corresponding nodes on the other side, a distance test is performed to see if it is less than a small value (0.001 cm). When this happened a soft impact with a small amount of penetration is applied and nodal displacements are updated accordingly. When the vocal folds touch each other, the contact nodes lose one degree of freedom.

The governing equations for the air flow are the Navier-Stokes (N-S) equations

$$\nabla \cdot v = 0$$
$$\frac{\partial v}{\partial t} + v \cdot \nabla v = -\frac{1}{\rho}\nabla p + \frac{\mu}{\rho}\nabla^2 v \qquad (3)$$

Where v is the velocity vector, μ and ρ are viscosity and density of the air. The first step in the air flow solution is generating the flow grid. Since the flow domain was assumed to be 2D, a non-uniform rectangular grid is selected such that regions of higher velocities and larger pressure gradients contained more grid points. Since the glottal gap continually changes during each cycle of oscillation, a logarithmic distribution of grids is designed to ensure the presence of a number of grid points in the region near closure. The glottis that is used in the two-dimensional flow is calculated by averaging the glottal gap in the longitudinal direction. **Fig. 2** shows a portion of the 150 x 80 non-uniform staggered grids used for flow domain. The Equations 3 are solved with a cell-centered finite volume method. In this method the governing equations are integrated over an element such as a quadrilateral with sides denoted by "west-east" perpendicular to the x axis, and "south-north" perpendicular to the y axis. Applying the Green's Theorem and approximating the integration over the quadrilateral domain results in a set of algebraic equations to be solved. The grid is called staggered because the pressure and velocity components are not calculated at the same location. The pressure is calculated at the center of the cell (finite volume), velocity component *u* is calculated at the east and west faces, and velocity component v is calculated at the north and south faces. These calculations are required for the balance of momentum within each cell. The corner velocities and pressures can be calculated by interpolation. A 'shadow method' with a large source term simulates the vocal folds in the flow domain to avoid the complexity of grid movement. Please refer to Alipour *et al.* [4] for further details.

In this method the N-S equations are integrated numerically over every finite control volume of the flow domain, resulting in a set of algebraic equations to be solved. The N-S equations are elliptic and can be solved only with a specified velocity at the inlet. This necessitates the knowledge of flow rate *a priori*. However, in the computational biophysical model used here the lung pressure is known and flow rate is to be determined. In this model, at every time step the wall function is updated from the nodal coordinates and the Bernoulli equation is used to approximate the air flow rate and then N-S equations are solved with a known inlet velocity with an iteration method until convergence is achieved. The N-S solution provides the pressure distributions that are needed for the calculation of the aerodynamic forces on the vocal folds.

III. RESULTS

To examine more closely the velocity profiles and pressure distributions in the glottal region in time and space, five frames of the glottal cycle are selected as shown in **Fig. 3** with their coronal contours. Frames 3 and 5 are during opening of the glottis when the glottis is convergent, frame 7 is later in the cycle when the glottis is near maximally open with a shape that is slightly divergent (6 degree), and frames 9 and 11 are during closing of the glottis when the glottis is divergent. The number in the legend refers to the frame number.

Centerline velocity profiles from the previous frames are plotted in **Fig. 4.** Examination of the centerline velocities suggests that they are dependent upon three factors, (1) glottal width, (2) glottal angle, and (3) the location of the downstream vortex. For frame 3, the glottal width is narrow and the glottal angle highly convergent. The corresponding centerline velocity increases throughout the glottis due to the convergence, then decreases throughout the ventricle region, and then quickly decreases past the FVFs. The newly created vortex is still within the ventricle-FVF region, and the centerline velocity falls rapidly just past its center (beyond the FVF). For frame 5, the glottis is less convergent than in frame 3, since the glottis is opening. Here the centerline velocity again increases throughout the glottis proper (due to the convergence), but then decreases more slowly through the ventricle-FVF region, and continues to gradually reduce in velocity up to approximately the axial location of the center of the vortex, at which point it then quickly decreases. Because the vortex is near the centerline, it tends to hold the centerline velocity to higher values until just past its center, at which point the centerline velocity drops rapidly. This is also the case for frames 7, 9, and 11.

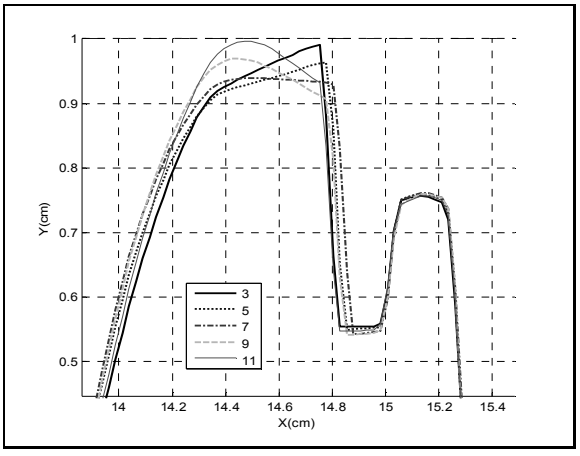


Fig. 3 – Five selected frames of a portion vocal wall

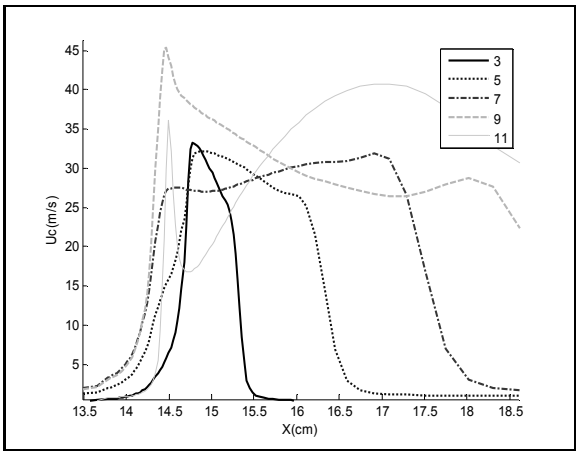


Fig. 4 – Center velocities during a cycle

The corresponding centerline pressure distributions are shown in **Fig. 5**. During initial opening (frame 3), pressure decreases gradually due to the convergence angle, and curiously stays slightly negative until past the FVFs, showing a slight dip at the exit of the FVF region. The centerline pressure distribution for frame 5 slightly later in the glottal opening is less steep within the glottis due to the less steep convergence angle and larger glottal width, and goes directly to near zero pressure at the exit of the glottis proper (not being influenced by the ventricle-FVF region). In frame 7 the glottis is widest with relatively high flow. The transglottal pressure at this time is relatively low and the centerline pressure drop is very gradual within the glottis, despite the 6 degree divergence.

During closure of the glottis (frames 9 and 11), the transglottal pressures are larger than for the other conditions of the cycle, and have steep drops to negative pressures essentially at the maximum glottal constrictions, followed by pressure recovery within the

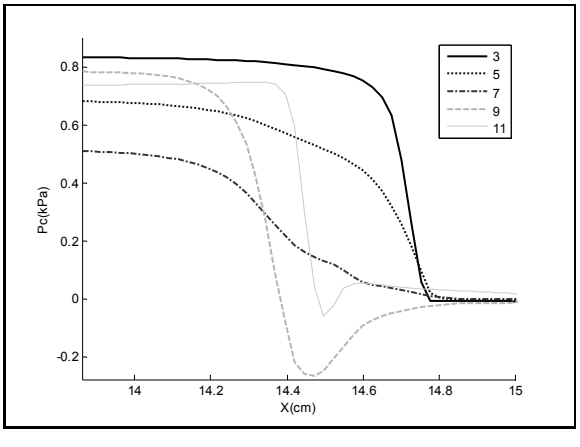


Fig. 5 – Center pressures during a cycle

glottis. Frame 11 with the narrower width and greater divergence has a faster centerline pressure recovery within the glottis.

IV. Discussion

Fig. 3 shows the 5 phases of vocal fold motion studied here. The glottis shows expected behavior – convergent glottal shaping during glottal opening and divergent shaping during glottal closing. The glottal angles ranged from 42 degrees convergent to 54 degrees divergent. Thus, the elements of a robust mucosal wave are present. The three dimensionality of the vocal fold motion suggests the inclusion of an anterior-posterior mode of oscillation as well as the typical medial-lateral and vertical modes, indicating that the model is capable of demonstrating the all primary modes of oscillation of typical interest.

The effective glottal duct length is longest during glottal opening (approximately 0.4 cm during the convergent glottal shaping) and shortest during glottal closing (approximately 0.3 cm during divergent glottal shaping), a point important for calculating total force on the glottal walls.

The centerline velocities are dependent upon three factors, the glottal width, glottal angle, and the location of the downstream vortex. The velocity increases as expected throughout a convergent glottis, and tends to decrease throughout a divergent glottis, and decreases past the true vocal folds when within the ventricle-FVF region. However, quite interestingly, the centerline velocity is highly dependent upon the location of the vortical structure that is produced at the exit of the glottis and is convected downstream. The vortex stays near the centerline in this model, and tends to keep the centerline velocity high [5,6] until it falls rapidly at the spatial location of the center of the vortex. Thus, the vortex appears to have dictated much of the behavior of the nearby centerline velocity. Thus, as seen in **Fig. 4**, the shape of the centerline axial velocity distribution appears short for convergent glottal shapes but longer for divergent shapes. The relative acoustic effect of these dynamic changes needs to be investigated. It is expected that more detailed vortex structure would be achieved using higher densities of flow mesh at the cost of more computational time.

The basic shape of the dynamic pressure distributions shown in **Fig. 5** are expected if they follow behavior consistent with static configurations. That is, the results for the convergent glottal shaped conditions have a decrease in pressure throughout the glottis until the end of the glottis where the pressure rapidly goes toward atmospheric pressure (with some variation as mentioned above). The divergent cases tend to show a pressure dip near the minimum constriction area, which is typical, although the dip is absent for the largest opening, 6 degree divergent case. The transglottal pressure (seen essentially on the y-axis in **Fig. 5**) also is consistent with expectation, with higher values during glottal closing, and lower values during glottal opening, and the lowest value for the most open condition. It is noted that there are no pressure dips near the exit of the glottis.

V. Conclusion

This report is a verification study of a dynamic computational model using nominal values for tissue and air. It is an extension of the "biophysical model" reported earlier [3] The extension is the inclusion of the ventricle, false vocal folds, a modification of the thyroarytenoid muscle distribution, and an update of the tissue properties. It reports the pressure distributions, velocity profiles, and vortex generation using this model. In addition, instead of using a Bernoulli expression to generate intraglottal pressures, the time-dependent Navier-Stokes equations are used. The motion of the vocal folds is three dimensional, but the calculation of the aerodynamics is two dimensional in that, at every time update, the glottal area is averaged along the glottal length to provide the equivalent glottal diameter over which the pressures are calculated and then applied back onto the medial surfaces of the vocal folds.

References

[1] F. Alipour and R.C. Scherer "Ventricular pressures in phonating excised larynges," *J. Acoust. Soc. Am.*, vol. 132, no. 2, pp. 1017-1026, 2012.

[2] F. Alipour and E.M. Finnegan "On the acoustic effects of the supraglottic structures in excised laryngesl," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2984-2992, 2013.

[3] F. Alipour, D.A. Berry, and I.R. Titze "A finite-element model of vocal-fold vibration," *J. Acoust. Soc. Am.*, vol. 108, no 6, pp. 3003-3012, 2000.

[4] F. Alipour, C. Fan, and R.C. Scherer "A numerical simulation of laryngeal flow in a forced-oscillation glottal model," *J. Computer Speech and Language*, vol. 10, pp. 75-93, 1996.

[5] C. Zhang, W. Zhao, S.H. Frankel, and L. Mongeau "Computational aeroacoustics of phonation, Part II: Effects of flow parameters and ventricular folds." *J. Acoust. Soc. Am.*, Vol. 112, pp. 2147-2154, 2002.

[6] X. Zheng, S. Bielamowicz, H. Luo, and R. Mittal "A computational study of the effects of false vocal folds on glottal flow and vocal fold vibration during phonation," *Ann. Biomed. Eng.,* Vol. 37, pp. 625-642, 2009.

This project was supported by grant R01DC009567 from the NIDCD, US National Institute of Health.

CONTRIBUTION OF PARANASAL SINUSES TO THE ACOUSTICAL PROPERTIES OF THE NASAL TRACT

M. Havel [1], J. Sundberg [2]

[1] Dept. of ORLHNS, Section Phoniatrics, University of Munich, Germany, miriam.havel@med.uni-muenchen,de
[2] Dept. of Speech, Music and Hearing, KTH "Royal Institute of Technology", Stockholm, Sweden, jsu@csc.kth.se

Abstract: **The contribution of the nasal and paranasal cavities to vocal tract resonator properties is unclear. Here we investigate resonance phenomena of paranasal sinuses with and without selective occlusion of the middle meatus, and the sphenoidal as well as the maxillary ostium in a cadaveric situs.**

Nasal and paranasal cavities of the thiel-embalmed cadaver were excited by sine-tone sweeps from a earphone in the epipharynx. A microphone at the nostrils picked up the response. Different conditions with blocked and unblocked middle meatus and sphenoidal ostium were tested. Additionally, infundibulotomy was performed allowing direct access to and selective occlusion of the maxillary ostium.

Response curves showed high reproducibility. A marked dip was observed after removing single sided occlusion of the middle meatus and the sphenoidal ostium. A marked low frequency dip was also detected after removal of occlusion of maxillary ostium following infundibulotomy.

Reproducible frequency responses of nasal tract can be derived from cadaver measurements. Marked acoustic effects of the maxillary sinus appeared only after direct exposure of the maxillary ostium following infundibulotomy.

Keywords: **voice, resonance, vocal tract, paranasal sinuses**

I. INTRODUCTION

The acoustical properties of the nasal tract are complicated by the pairwise arranged paranasal sinuses which are connected to it via narrow and complex orifices, the so-called ostia. The orifices of the maxillary and frontal sinuses open up to a common duct accessible via the middle meatus, whereas the ostia of the sphenoidal sinuses are directly connected to the posterior parts of the nasal cavity. The detailed anatomy of this system shows a great interindividual variability as well as a marked left/right asymmetry. The mucosa lining of the sinonasal tract with its considerable swelling potential provides additional uncontrollable confounder in acoustical measurements, especially when examined in vivo [1]. The result is an extremely complex frequency response of the nasal tract varying both within and between individuals. The complexity is even greater if also the oral cavity is included as in the consonants /m, n/ and in nasalized vowels [2, 3]. The aim of this work was to test a method to investigate the resonance properties of these sinuses by selectively occluding the middle meatus and sinus ostia in the stable anatomical condition of a cadaver.

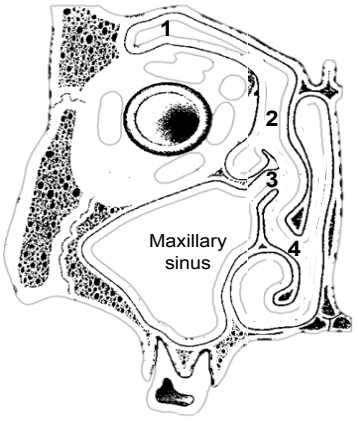


Fig. 1a: Schematic illustration of the osteomeatal complex (coronal plane): (1) frontal sinus, (2) frontal recess, (3) maxillary ostium covered by the uncinate process, (4) middle meatus flanked by middle and lower concha.

II. METHODS

Nasal and paranasal cavities of a male thiel-embalmed cadaver were acoustically excited by a sine sweep (range 200-4000 Hz, duration 18 s) generated by the Tone®-software (by Svante Granqvist). The earphone producing the sine sweep was hermetically sealed in the epipharynx by means of plasticine.

As a baseline condition, both middle meatus and sphenoidal ostia were occluded under endoscopic control using maltodextrin food thickener, a jelly-like viscous mass containing of cornstarch powder and water, routinely used in the assessment of swallowing disorders [4]. In the experimental conditions, the occlusion was removed by targeted suction.

The acoustical response to the sine sweep was picked up by a microphone placed close to the nostril and the effects of various blockage conditions were analyzed.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

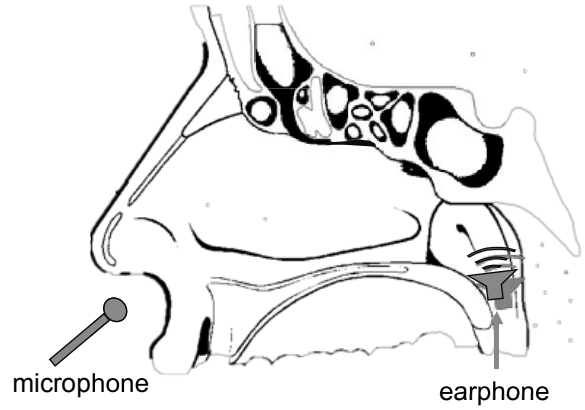


Fig. 1b: Schematic illustration of the placement of the earphone and the microphone during the sine-sweep measurements in the cadaveric situs (sagittal plane).

Additionally, to provide a direct access to the maxillary ostium and enable its selective occlusion, infundibulotomy (removal of uncinate process) was performed according to the criteria of functional endoscopic sinus surgery. Each condition was measured twice.

III. Results

The reproducibility of the sine sweep response curves was tested by repeated recordings. Fig. 2 shows two response curves obtained under the same condition (both middle meatus and sphenoidal ostia occluded); in the graph one curve has been shifted by two decibels for the sake of visibility. As illustrated in the figure, the discrepancy was generally negligible.

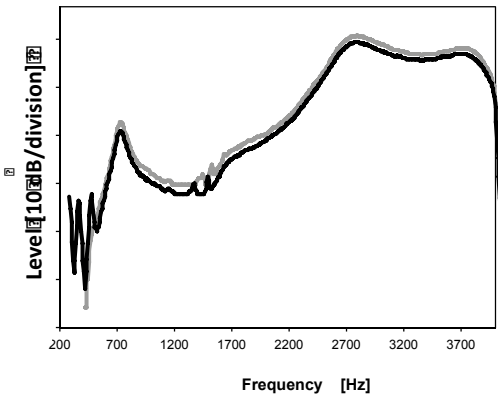


Fig. 2: Response curves obtained from repeated sine sweep excitation of the nasal tract while the middle meatus and sphenoidal ostia were occluded. The amplitude of one curve was displaced by two dB.

A clear resonance peak can be observed at 720 Hz and other, heavily damped peaks near 1400 Hz. Two other blunt peaks can be observed at 2800 Hz and 3800 Hz. These resonance frequencies show some similarity with those of an open-closed cylindrical resonator of 12 cm length [(2n-1)*F=34000/48=700 Hz].

Fig. 3 shows the response curves recorded for two cases: both middle meatus and sphenoidal ostia occluded (grey, also shown in fig. 2) as well as one obtained after removing the occlusion from the right middle meatus and the right sphenoidal ostium (black).

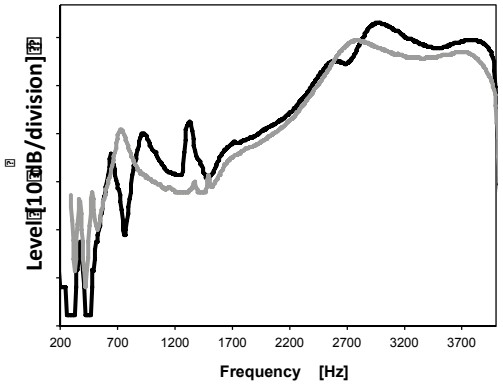


Fig. 3 Response curve obtained from sine sweep excitation of the nasal tract with the occlusion of one middle meatus and one sphenoidal sinus removed (black). The grey curve is the same as in Fig. 2.

In the latter case the resonance at 720 Hz seems to be divided into two peaks by a marked zero at 770 Hz. Another peak appeared at 1330 Hz and two blunt peaks at 3000 Hz and 3800 Hz.

The orifice of the maxillary sinus is hiding behind a structure called processus uncinatus. In functional endoscopic sinus surgery this structure is routinely removed in cases of suspected maxillary pathology. This surgical procedure allows inspection of the ostium and the sinus. After removal of processus uncinatus the two curves shown in fig. 4 were obtained, one with the ostium open, the other with the ostium occluded.

In the first case a marked dip appeared at 530 Hz presumably reflecting the absorbent effect of the now exposed maxillary sinus. The fact that this dip occurred only after removal of the processus uncinatus suggests that, under normal conditions, the maxillary sinuses may have a minor influence on radiated nasal sounds.

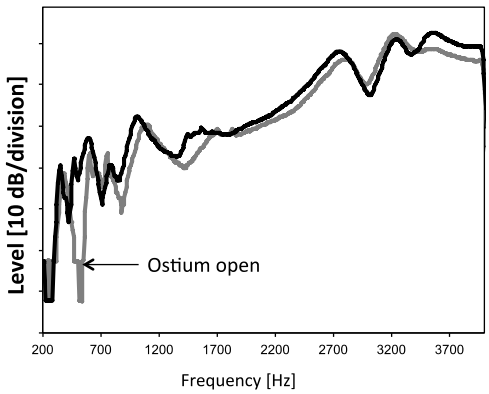


Fig. 4: Response curves obtained from sine sweep excitation of the nasal tract after infundibulotomy, with (black) and without (grey) occlusion of the maxillary ostium.

IV. Discussion

The nasal tract is an extremely complex and also unstable resonator as mentioned, see e.g. [2]. This is due to the intricate sinonasal anatomy as well as the in vivo cyclically changing shape of the mucosa. By using the stable anatomical condition of a cadaver we could obtain quite reproducible response curves to the sine sweeps (see Fig. 1).

We examined the nasal tract in isolation, without including the vocal tract. Thereby it was important to achieve a complete sealing of the velopharyngeal opening, as even a tiny leakage would have influenced the resonance frequencies and widened their bandwidths considerably. The results support the assumption that this condition was met; the bandwidth of the 720 Hz peak shown in Fig. 2 amounted to 90 Hz approx., a typical value for vocal tract formants in that frequency range. However, the bandwidths of some other peaks and dips observed were extremely wide, e.g. those appearing above 2000 Hz. This could be expected, given the very high circumference to area ratio of the nasal tract [5].

Contrary to a common assumption, the paranasal cavities function as acoustic absorbents, producing minima in the transfer function, rather than enhancing particular frequency regions in the radiated sound [6]. The frequencies of such minima, observed under various experimental conditions, represent important information.

After the removal of the occlusion of the right middle meatus and the right sphenoidal ostium we observed a marked dip appearing at 720 Hz. It does not seem likely that this dip was caused by the maxillary sinus as our findings suggest that the ostium of this cavity contributed to the sine sweep response only after direct exposure following infundibulotomy. Rather the 720 Hz dip would have emanated from the sphenoidal cavity. Dang & Honda have found examples of a similar resonance frequency for this cavity in their experiments [7, 8].

An effect of the maxillary sinus could be observed in terms of a marked zero at 530 Hz only after surgical exposure of the ostium following infundibulotomy. Assuming that the maxillary sinus can be considered as Helmholtz resonator with a neck area of 15 mm^2, neck length of 1 mm, and a volume of 12,000 mm^3, its lowest resonance frequency should be close to 600 Hz. This value is not very far from the frequency of the zero observed. Moreover it is similar to the antiresonance frequencies of the maxillary sinuses estimated from acoustic measurement by Dang and Honda [8]. Nevertheless, the importance of close agreement with previous findings of resonance frequencies of the paranasal sinuses should not be exaggerated, given the substantial intra- and interindividual variability of the sinonasal tract morphology.

V. Conclusion

Acoustical responses of sinonasal tract can be derived from cadaver experiments showing good reproducibility. Minor acoustical effects are assumed for maxillary sinus in natural condition since their orifices are shielded from nasal cavity. A direct exposure of the maxillary ostium following surgical intervention seems to introduce a marked effect in the response curve. However, the variable morphology of sinonasal tract impedes a direct attribution of response curve features to specific sinuses.

References

[1] J. Linqvist-Gauffin, J. Sundberg, “Acoustic properties of the nasal tract,” *Phonetica*, vol. 33(3), pp. 161-168, 1976.

[2] M. Båvegård, G. Fant, J. Gauffin, J. Liljencrants, “Vocal tract sweeptone data and model simulations of vowels, laterals and nasals,” *STL-QPSR,* vol. 34(4), pp. 043-076, 1993.

[3] T. Pruthi, C.Y. Espy-Wilson, B. H. Story, “Simulation and analysis of nasalized vowels based on magnetic resonance imaging data,” *J Acoust Soc Am*, 121(6), pp. 3858-73, 2007.

[4] R. J. Dewar, M. J. Joyce, “Time-dependent rheology of starch thickeners and the clinical implications for dysphagia therapy,” *Dysphagia*, 21(4), pp. 264-9, 2006.

[5] G. Bjuggren, G. Fant, “The nasal cavity structures,” *STL-QPSR*, vol. 5(4), pp. 005-007, 1964.

[6] T. Koyama, “Experimental study on the resonance of paranasal sinus,” *J Otolaryngol Jpn.,* 69(6), pp. 1177-1191, 1966.

[7] J. Dang, K. Honda, H. Suzuki, “Morphological and acoustical analysis of the nasal and the para-nasal cavities, “ *J Acoust Soc Am,* 96(4), pp. 2088-100, 1994.
[8] J. Dang, K. Honda, “Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation, “ *J Acoust Soc Am,* 100(5), pp. 3374-83, 1996.

COMPARISON OF COMPUTED AND MEASURED ACOUSTIC CHARACTERISTICS OF AN ARTIFICIALLY LENGTHENED VOCAL TRACT

V. Radolf[1], J. Horáček[1], A. M. Laukkanen[2]

[1] Institute of Thermomechanics, AS CR, Prague, Czech Republic, radolf@it.cas.cz
[2] Speech and Voice Research Laboratory, School of Education, University of Tampere, Tampere, Finland

Abstract: **The First formant frequency F1 of a human vocal tract (VT) prolonged with a glass tube was measured for a VT model with solid walls and a real human VT, and the results were compared to a mathematical model. The experiments with the VT model confirmed the legitimacy of an assumption of the solid walls in the mathematical simulations of acoustical characteristics of the artificial VT plexiglass model prolonged with the tube. The measured F1=73 Hz well corresponded to the computed value 78 Hz. However, the *in vivo* experiments in human, performed in a similar way, showed a much higher value of F1 at about 200 Hz, that is caused by soft tissues on the walls of real human VT cavities.**

Keywords: **biomechanics of voice, phonation into tubes, formant frequency**

I. INTRODUCTION

The objective of the present paper was to verify some results of mathematical modelling of the acoustic characteristics of the human vocal tract prolonged with a tube. Phonation into tubes or straws are used for voice training and therapy purposes, see e.g. [1,2,3].

For that purpose, similar experiments were performed *in vivo* and *in vitro* in order to obtain real physiological data for normal and some extreme ways of human phonation when the acoustic impedance of the vocal tract (VT) is artificially increased by prolonging the VT with different tubes or straws. The measurements were focused mainly on subglottal and oral air pressure values. Moreover the formant frequencies were evaluated from the spectra of the acoustic pressure measured outside the VT and inside the mouth cavity [4,5].

One of the main questions concerns the lowest formant frequency F1 of the acoustic cavity consisting of the VT prolonged at the lips with a tube. According to Story et al. [6], this formant frequency F1 for phonation into the so-called resonance glass tube should be in the range of 200-300 Hz. These results are strongly influenced by modeling of yielding walls of the VT that substantially increase F1. If the assumption of a solid VT wall is applied to the mathematical model [7], F1 decreases even below 100 Hz. It is hardly possible to find this low-frequency resonance in the spectrum of measured acoustic signal, when it is excited by vocal folds vibration with fundamental frequency over 150 Hz. For that reason special *in vitro* and *in vivo* experiments were designed, that enabled to measure the transfer function of the VT prolonged with a resonance tube without an influence of vibrating vocal folds.

II. METHODS

The *in vitro* measurement set up (see Fig. 1) consisted of a vocal tract model for vowel [u:] made of plexiglass, closed at the glottis end and connected to a glass resonance tube (264 mm in length, inner diameter 6.8 mm) at the lips. The acoustic pressure was measured by a special microphone probe B&K 4182 at the lips position and by a miniature microphone B&K 4138 at the open tube end. The model was excited by white noise signal from a woofer MTC 5 1/4" placed outside. The recording was made with a PC controlled measurement system B&K PULSE 10 using 32.8 kHz sampling frequency. The same measurement set-up was used for *in vivo* measurements, and the subject was a female voice trainer.

Such acoustic systems as those used *in vivo* and *in vitro* set-ups can be described by a matrix form equations in the frequency domain as, see eg [6,7]:

$$\begin{bmatrix} p_L \\ W_L \end{bmatrix} = T_1 \cdot \begin{bmatrix} p_G \\ W_G \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} . \begin{bmatrix} p_G \\ W_G \end{bmatrix}, \quad (1)$$

$$\begin{bmatrix} p_T \\ W_T \end{bmatrix} = T_2 \cdot \begin{bmatrix} p_L \\ W_L \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} . \begin{bmatrix} p_L \\ W_L \end{bmatrix}, \quad (2)$$

where p and W are acoustic pressure and volume velocity, respectively, T_1 is transfer matrix of the vocal tract, T_2 is transfer matrix of the tube and the indices G, L, T mean the position of glottis, lips and open tube end, respectively. The velocity is zero ($W_G = 0$) at the closed VT end at the glottis position, and thus from (1) and (2) we get

$$p_L(\omega) / p_T(\omega) = a_1 / (a_2 a_1 + b_2 c_1). \quad (3)$$

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

When the radiation impedance Z_{RAD} at the tube output is neglected considering the small diameter of the tube, the denominator in eq. (3) is the same as in the transfer function W_T / W_G and p_T / W_G of the whole system (VT+tube):

$$W_T(\omega)/W_G(\omega) = 1/(a_2 a_1 + b_2 c_1). \quad (4)$$

The resonance frequencies (peaks) of the measured transfer function p_L / p_T thus should be approximately equal to the peaks of transfer function W_T / W_G of the whole system. The numerator $a_1(\omega)$ is the same as a denominator in the transfer function of the vocal tract alone W_L / W_G and p_L / W_G when neglecting Z_{RAD} at the lips:

$$W_L(\omega)/W_G(\omega) = 1/a_1 . \quad (5)$$

The antiresonance frequencies (zeros) of the measured transfer function (3) then should lie very close to the peaks of the transfer function of the vocal tract without a tube.

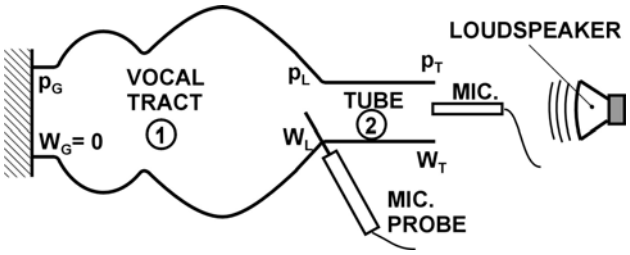


Fig.1 Schema of the experimental set-up and mathematical modeling.

III. Results

Frequency spectrum of the measured transfer function p_L / p_T for the artificial VT model prolonged with the tube and the transfer function of the same model computed according to eq. (3) are compared in Fig. 2. The first resonance peak occurs at the frequency of 73 Hz and 78 Hz for the measured and computed results, respectively (see also Tab. 1). The first two antiresonances, which according to eq. (4) are the formants of the vocal tract alone, can be well recognized in the measured spectra. The second of these first two computed and measured antiresonance frequencies differ by 8%.

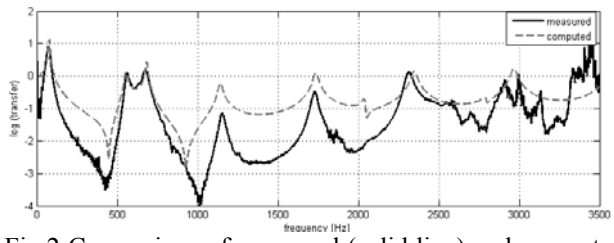


Fig.2 Comparison of measured (solid line) and computed (dashed line) transfer functions for the artificial vocal tract prolonged with the tube.

Frequency spectrum of the transfer function p_L / p_T measured for the real human vocal tract prolonged with the tube is shown in Fig. 3. First resonance frequency without any clearly visible peak can be found at about the frequency of F1=200Hz. For the other resonance frequencies see Tab. 1. The antiresonance frequencies cannot be recognized in the measured spectrum.

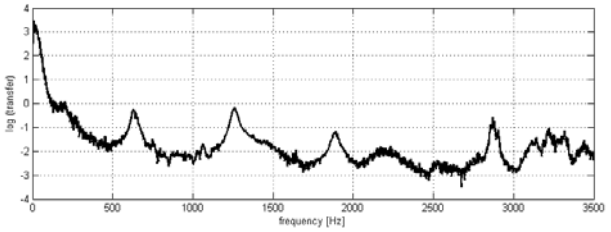


Fig.3 Measured transfer function of the real human vocal tract prolonged with the tube.

Tab.1 Measured and computed resonance frequencies in [Hz] for the artificial vocal tract model and for the real human vocal tract when both of them are prolonged with the resonance tube.

VT model - computed	VT model - measured	human VT - measured
78	73	200
555	560	630
684	672	
1142	1153	1262
1735	1729	1889
2347	2313	

IV. Discussion

First six resonance frequencies of the measured transfer function p_L / p_T for the artificial VT model differ only slightly from the computed results, see Tab. 1. This frequency range corresponds to a validity of 1D mathematical model of the VT for vowel [u:] with 62 mm maximum width in the mouth cavity, that could cause the existence of first transversal acoustic mode above 2.7 kHz.

The first formant frequency F1 at about 70-80 Hz of the artificial VT model with solid walls is much lower than F1 of about 200 Hz measured in real human VT. The difference in the formant frequencies is caused by the soft tissues inside the real human VT, and it is much smaller for the higher formant frequencies [8].

V. Conclusion

The experiment with artificial human vocal tract confirmed the legitimacy of using the solid wall mathematical model [7] to simulate acoustical resonance

properties of the artificial model of human VT prolonged by the tubes. Similar *in vivo* experiments carried out in real human VT showed a substantial difference between the first resonance (formant) frequency F1 of the solid wall VT model and F1 of real human VT acoustic cavities covered by soft tissues. The results confirmed the approach of Story et al. [6] showing that the VT model with yielding walls is really necessary to consider for mathematical modeling of the real human vocal tracts prolonged with the tubes.

ACKNOWLEDGEMENT

The study was supported under the grants of the Czech Science Foundation No P101/12/P579 and by the Academy of Finland (grants No. 1128095 and 134868).

REFERENCES

[1] I.R. Titze, E.M. Finnegan, A.M. Laukkanen, S. Jaiswal, "Raising lung pressure and pitch in vocal warm-ups: The use of flow-resistance straws," *Journal of Singing*, vol. 58, pp. 329-338, 2002.

[2] A.M. Laukkanen, "About the so called "resonance tubes" used in Finnish voice training practice. An electroglottographic and acoustic investigation on the effects of this method on the voice quality of subjects with normal voice," *Scandinavian Journal of Logopedics and Phoniatrics*, vol. 17 (34), pp. 151-161, 1992.

[3] S. Simberg, A. Laine, "The resonance tube method in voice therapy: description and practical implementations," *Logopedics Phoniatrics Vocology*, vol. 32, pp. 165-170, 2007.

[4] V. Radolf, A.M. Laukkanen, J. Horáček, J. Veselý, D. Liu, "Air-pressure, vocal fold vibration and acoustic characteristics of phonation during vocal exercising. part 1: measurement in vivo," *Engineering Mechanics*, 2013, in print.

[5] J. Horáček, V. Radolf, V. Bula, J. Veselý, A.M. Laukkanen, "Air-pressure, vocal folds vibration and acoustic characteristics of phonation during vocal exercising. Part 2: Measurement on a physical model," *Engineering Mechanics*, 2013, in print.

[6] B.H. Story, A.M. Laukkanen, I.R. Titze, "Acoustic impedance of an artificially lengthened and constricted vocal tract," *Journal of Voice*, vol. 14, pp. 455-469, 2000.

[7] T. Leino, A.M. Laukkanen, V. Radolf, "Formation of the actor's/speaker's formant: A study applying spectrum analysis and computer modeling," *Journal of Voice*, vol. 25, pp. 150-158, 2011.

[8] T. Vampola, J. Horáček, J.G. Švec, "FE modeling of human vocal tract acoustics. Part I: Production of Czech vowels," *Acta Acustica united with Acustica*, vol. 94, pp. 433-447, 2008.

[9] V. Radolf, "Direct and inverse task in acoustics of the human vocal tract.," *PhD . thesis*, Czech Technical University in Prague, 95 p. , 2010.

APPENDIX

The mathematical model is based on an analytical solution of 1D wave equation for acoustic wave propagation in the vocal tract cavity [9]:

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{1}{A}\cdot\frac{\partial A}{\partial x}\cdot\frac{\partial\varphi}{\partial x}-\frac{1}{c_0^2}\cdot\left(\frac{\partial^2\varphi}{\partial t^2}+c_0\cdot r_N\cdot\frac{\partial\varphi}{\partial t}\right)=0 \quad (6)$$

where φ is the flow velocity potential, x is longitudinal coordinate along the vocal tract measured from the vocal folds to the lips, t is time, r_N *is* specific acoustic resistance per an unite length, $A(x)$ is the cross-sectional area of the cavity and c_0 is speed of sound.

Relation between the acoustic pressure p and the volume velocity W at the input and output of each conical acoustic element can be described by the transfer matrix as

$$\begin{bmatrix} p_{OUT} \\ W_{OUT} \end{bmatrix}=\begin{bmatrix} a & b \\ c & d \end{bmatrix}\cdot\begin{bmatrix} p_{IN} \\ W_{IN} \end{bmatrix}, \quad (7)$$

where the elements of the transfer matrix are

$$a=\frac{\xi_0}{\xi_0+L}\cdot\left(\cosh(\gamma L)+\frac{1}{\gamma\,\xi_0}\cdot\sinh(\gamma L)\right),$$

$$b=-\frac{c_0\rho_0(r_N+j\,k)\cdot\xi_0}{A_{IN}\cdot\gamma(\xi_0+L)}\cdot\sinh(\gamma L),$$

$$c=A_{OUT}\cdot\frac{\left(1-\gamma^2\xi_0(\xi_0+L)\right)\cdot\sinh(\gamma L)-\gamma L\cdot\cosh(\gamma L)}{\gamma(\xi_0+L)^2\cdot c_0\rho_0(r_N+j\,k)},$$

$$d=\frac{A_{OUT}}{A_{IN}}\frac{\xi_0}{\xi_0+L}\cdot\left(\cosh(\gamma L)-\frac{1}{\gamma(\xi_0+L)}\cdot\sinh(\gamma L)\right),$$

L is length of the element, A_{IN} and A_{OUT} are the cross-sectional areas of the element input and output, respectively, ρ_0 is fluid density, γ is a complex exponent given by the formulas:

$$\gamma=\alpha+j\beta, \quad (8)$$

$$\alpha=\frac{r_N}{\sqrt{2+2\cdot\sqrt{1+(r_N/k)^2}}},\ \beta=\frac{k}{2}\cdot\sqrt{2+2\cdot\sqrt{1+(r_N/k)^2}},$$

$k=\omega/c_0$ is the wave number, ω is angular frequency of harmonic signal, j is imaginary unit: $j=\sqrt{-1}$. The coefficient ξ_0 is defined by input and output radius of the element R_{IN} and R_{OUT}, respectively,

$$\xi_0=\frac{R_{IN}}{R_{OUT}-R_{IN}}\cdot L\cdot \quad (9)$$

Frequency dependent viscous losses were considered as

$$r_N=\frac{1}{R}\cdot\sqrt{2k\mu/c_0\rho_0}, \quad (10)$$

where μ is dynamic air viscosity.

A GERMAN PARALLEL ELECTRO-LARYNX SPEECH – HEALTHY SPEECH CORPUS

Anna K. Fuchs, Martin Hagmüller

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

anna.fuchs@tugraz.at, hagmueller@tugraz.at

***Abstract*: In this paper, we describe the German parallel Electro-Larynx speech – Healthy speech (ELHE) Corpus which has been recorded in our recording studio. 3 female and 4 male healthy subjects recorded up to 500 sentences spoken one time with healthy speech (HE) and one time using the Electro-Larynx (EL) device. With these recordings, differences and similarities between EL and HE speech can be analyzed. Statistical analyses have shown that the length of EL sentences is longer than for HE sentences. Moreover, the fundamental frequency f_0 of EL speech depends on the EL device and the variance of f_0 is larger for HE speech due to the missing changing patterns of EL speech. Finally, analyses of signal-to-noise ratios (SNR) have shown that there are three different levels inherent in EL speech: silence, direct-radiated noise from the EL device (DREL) and speech (corrupted with DREL). For HE speech only two levels (silence and speech) can be distinguished. This corpus can be used to analyze healthy speech compared to disordered (EL) speech and based on this knowledge to improve the disordered speech.**

***K*eywords: Parallel speech corpus, Electro-Larynx (EL), Signal-to-Noise Ration (SNR), Speech Disorders**

I. Introduction

The electro-larynx (EL) is a device which is used by patients who had to undergo an operation called laryngectomy. Within this operation the larynx and the containing vocal folds, which are the excitation of the speech production, are removed. The EL substitutes the excitation signal from the vocal folds but the properties of the resulting speech suffer from the unnaturalness of this excitation signal. The main problems related with EL speech are 1.) the constant fundamental frequency f_0 and the lack of its variation, 2.) the low-frequency deficit and the improper source spectrum and 3.) the noise which is produced by the EL device itself (DREL). Although the main problems related with EL speech are known, we still do not know exactly which components of the EL speech acoustic signal are contributing most to its abnormal quality. According to [1], adding f_0 information would bring most benefit. In addition, removing EL self-noise and correction for a lack of low-frequency energy would also be effective.

To improve EL speech, the differences between natural speech and EL speech need to be analyzed. Especially, the problem with the constant f_0 could be improved if we can find a way to map natural f_0 to the constant EL speech. In order to fulfill this task, a parallel corpus is recorded which contains natural speech utterances as well as the same utterances produced using the EL device. With this corpus, properties of EL speech can be investigated and machine learning procedures can be applied in order to enhance EL speech.

II. Methods

II-A. Recording Details

The speech material of the German parallel ELHE corpus consists of up to 500 different sentences. The sentences were organized in 10 sessions with approximately 50 sentences each session. Each utterance was spoken one time with healthy speech (HE) and one time with the EL device, in order to compare sentences. The speech material consists of phonetically rich sentences from different German speech corpora. Prosodic differences can be investigated because sentences with main focus on intonational and contrastive stress are included (statement vs. question, emphasis on different parts of the word,...). All in all, the subjects had to read up to (two times) 503 sentences. In total this corpus consists of 6030 sentences which are recorded two times, one time with natural speech and one time with EL speech.

The Austrian German native speakers have been healthy subjects with an average age of 26 years (female) and 36 years (male). Although the anatomy of laryngectomized people is slightly different, we recorded healthy subjects in order to built up this parallel ELHE corpus. According to [2], who carried out listening tests, there are no significant perceptual differences between EL speech produced by a patient or by a healthy subject. The subjects used a Servox Inton digital, a widely used device in Europe and the US. Three female (F01, F03 and F07) and four male speaker (M02, M04, M05 and M06) have been recorded. All recordings were carried out on-

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

site at the recording studio of the Signal Processing and Speech Communication Laboratory at Graz University of Technology. During the recordings the test subject was overseen by a supervisor. The supervisor had the control over the speech recording software in order to control and modify the recording process immediately. The used software was *SpeechRecorder* [3] which had been designed to record speech corpora. The test subjects were recorded sitting in a sound proof recording room. The supervisor could observe the test subject through a glass window. The test subject had to speak sentences displayed on a screen. The speech material was recorded with a headset microphone AKG HC 577 L with omni-directional pickup pattern. The head-mounted high-quality condenser microphone was chosen to ensure a consistent recording quality, since it guarantees a constant distance of about 2 cm from the corner of the mouth.

A laryngograph (Lar) provides a ground truth signal for the fundamental frequency of the healthy speech signal. These signals extract the impedance corresponding to the healthy intonation. The speaker had to carry a neck band with the laryngograph electrodes. The laryngograph signal captures the vibration of the vocal folds. The electrodes are loaded with a high frequency current. The measured impedance is proportional to the contact area of the closed vocal folds (if the vocal folds are open, no relevant information can be extracted). The high frequency oscillations are disturbed by a low-frequency component, which are caused by the vertical movement of the larynx. Therefore, a band-pass filter with linear phase respond and with a lower cutoff frequency of 50 Hz and an upper cutoff frequency of 2000 Hz is suggested. Both, microphone signals and laryngograph signals, were sampled at 48 kHz with 16 bit resolution. Descriptive statistics about the parallel ELHE corpus can be seen in table I.

II-B. Signal-to-Noise Ratio Estimation

Many signal-to-noise ratio (SNR) estimation techniques are based on the knowledge of the clean speech signal and the noisy signal. In this paper, we want to estimate SNR when only the noisy signal is available.

In healthy speech there are (ideally) two levels: speech level (SL) and noise/silence level (NL). Within EL speech we have to deal with three levels: speech level (SL), noise level of EL background noise (DREL-L) and noise/silence level (NL). This is illustrated in Fig. 1. To find DREL-L a threshold thr needs to be implemented. We use an iteratively changing one, which is able to find DREL-L automatically.

One standard SNR definition for speech signals is the averaged segmental SNR ($SSNR_A$). We calculate the arithmetic average of linear $SSNR$ values:

$$SSNR_A = 10 \cdot \log_{10}\left(\frac{1}{L}\sum_{k=0}^{L-1}\frac{s^2(k)}{n^2(k)}\right). \qquad (1)$$

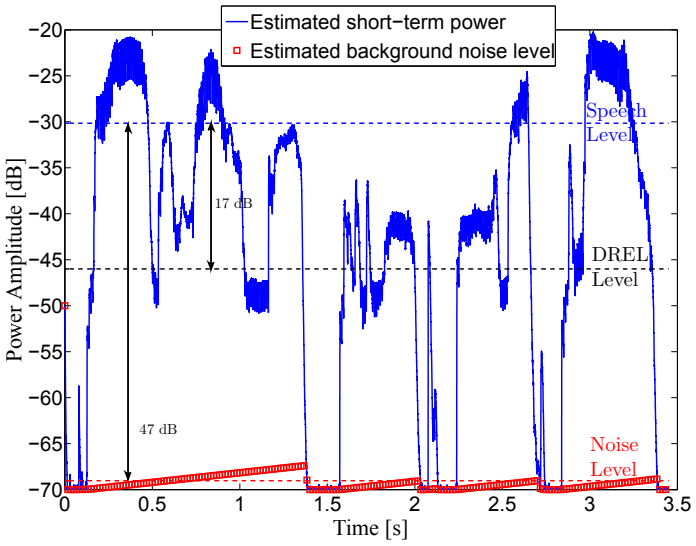


Fig. 1. *Short-term power $\overline{y^2(k)}$ and background noise level $\overline{n^2(k)}$ of one sentence spoken by a male speaker with the EL device; Three levels: Speech Level (SL), DREL Level (DREL-L) and Noise Level (NL); Speech-Noise Ratio: 47 dB – Speech-DREL Ratio: 17 dB.*

Within this notation $s(k)$ is the k-th value of the clean speech signal, and $n(k)$ is the k-th noise sample. We use a segment length of 1, which means that we do a sample-wise processing. Therefore, L is equal to the length of the signal $s(k)$. Furthermore, $SSNR_A$ is evaluated only on parts of the signal where speech is active, which is determined via a threshold. Values below this thresholds, which are associated to non-speech parts of the utterance, are neglected (see Fig. 2, VAD).

We can only observe the noisy speech signal $y(k)$, which is the sum of the speech signal $s(k)$ and the noise signal $n(k) \rightarrow y(k) = s(k) + n(k)$. To obtain $\overline{s^2(k)}$ in (1) we have to subtract the estimate of $\overline{n^2(k)}$ from the estimate of $\overline{y^2(k)}$. To estimate the short-term power of the signal $\overline{y^2(k)}$ and of the noise $\overline{n^2(k)}$, we use first-order IIR smoothing [4]. Whereas this algorithm works well for healthy speech, EL speech need to be handled with more care.

$$\overline{y^2(k)} = \begin{cases} (1-\gamma(k))y^2(k) + \gamma(k)\overline{y^2(k-1)}, & \text{if } \overline{y^2(k)} \geq thr \\ \overline{y^2(k-1)} & \text{otherwise} \end{cases} \qquad (2)$$

with γ being a smoothing constant, which differs for rising and falling signal edges in order to detect rising signal powers very rapidly. To estimate the background noise level, the short-term power of the signal computed in (2) is used:

$$\overline{n^2(k)} = \min\{\overline{y^2(k)}, \overline{n^2(k-1)}\}(1+\epsilon) \qquad (3)$$

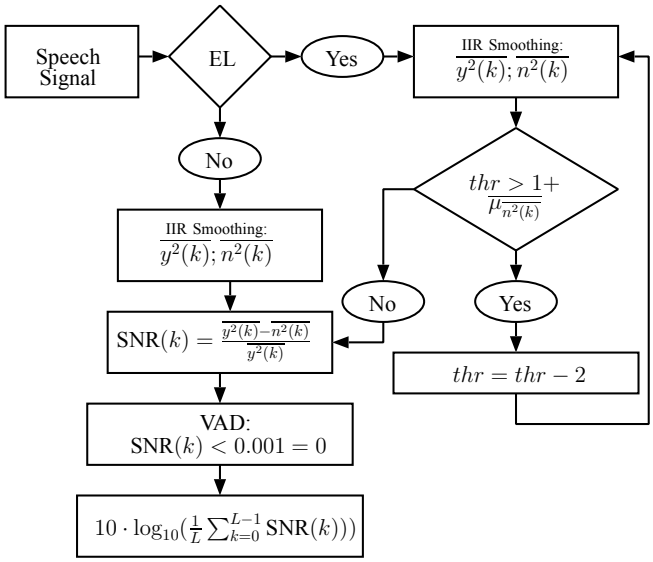


Fig. 2. *IIR smoothing for EL and HE speech with iteratively changing threshold thr (initialized with a fixed starting value); $\overline{\mu_{n^2(k)}}$ – estimated mean value of (3).*

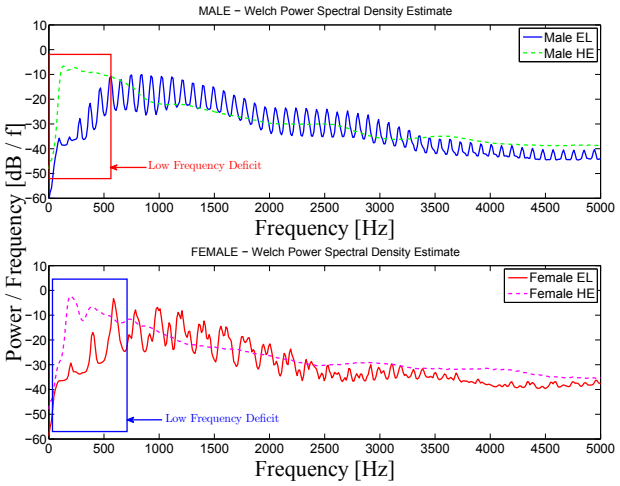


Fig. 3. *Spectral Power Density of EL speech and HE speech for averaged for all male (upper plot) and all female speakers (lowerplot).*

ϵ is a constant slightly larger than one to avoid that the result of the minimum operator is freezing at a global minimum.

For HE speech only the first part of (2) is used (see also Fig. 2). The threshold thr is only used for EL speech. The results of (2) and (3) depend on thr. Therefore, it is possible to find the optimal value for thr with an iterative loop. thr and other parameters are chosen empirically.

II-C. Power Spectral Density Comparison

Power spectral density (PSD) estimation was carried out on both speaking modes (EL and HE) and averaged per-utterance and per-gender. For PSD calculation we used Welch's method where the data is split into segments of length 70 ms, without overlap (Hamming window is applied), periodograms are computed and averaged. Fig. 3 shows the results for averaged female and male speakers PSD for both speaking modes.

III. Results

The whole corpus has been analyzed with respect to important speech related properties. In table I, one can see that speaking with the EL needs much longer than with healthy speech because adequate articulation improves intelligibility. Although the same sentences have been spoken, we have 4h30min of EL speech but only 2h44min of HE speech.

The mean fundamental frequency $\overline{\mu_{f_0}}$, as well as the standard deviation $\overline{\sigma_{f_0}}$, is estimated for each speaker and type of speech. Praat [5] has been used to extract f_0. $\overline{\mu_{f_0}}$ of healthy speech depends on the speaker. Due to the anatomy of the vocal folds female speaker have a higher $\overline{\mu_{f_0}}$ than male speaker and also $\overline{\sigma_{f_0}}$ is larger. For EL speech $\overline{\mu_{f_0}}$ depends on the adjustable EL device.

An important and necessary task during speech enhancement is to evaluate the enhanced speech. Subjective methods, e.g. listening tests, are the preferred method to evaluate results, but due to its inconvenience objective measure of speech quality are easier to apply. These measures are commonly based on the SNR or on some distance between the original speech and the "enhanced" speech. Preliminary SNR measurements based on IIR smoothing show that due to the DREL level, SNR values of EL speech are around 10 dB, whereas HE speech produces a SNR of around 40 dB.

In Fig. 3 PSD averaged over all sentences for each gender is illustrated. The spectral structure between EL and HE is completely different. The low-frequency deficit, which was also reported by [7], up to around 500 Hz can be seen for both genders. Moreover, the regular harmonics of the excitation signal which are responsible for the monotonic EL speech can be seen.

Additionally, the corpus has been analyzed according to its word statistic: The test sentences contain 3961 words, without counting multiple occurrences there are 1444 words. 1210 words only occur once. In order to implement a speech recognizer based on the recorded data a dictionary, which lists the phonetic transcriptions, needs to be set up manually. In Fig. 4 the symbol distribution according to the extended speech assessment methods phonetic alphabet (X-SAMPA) [6] is illustrated for the given text prompts. The distribution is typical for the German language where the most common phonemes are: the alveolar voiceless plosive [t], nasal [n] and fricative [s], as well as the open centered vowel [a], the open-mid front vowel [e], and the schwa [@], among others.

TABLE I

Statistical analyses of ELHE corpus: Number of sentences; Mean value and standard deviation of f_0 – $\overline{\mu_{f_0}}$,$\overline{\sigma_{f_0}}$; Signal-to-noise ration (SNR).

ID	Age		# Sentences	Length	$\overline{\mu_{f_0}}$	$\overline{\sigma_{f_0}}$	SNR
F01	28	EL	503	45min28s	192	7	17.95
		HE	503	29min57s	198	27	46.57
		Lar			196	27	
F03	31	EL	250	19min51s	199	6	9.02
		HE	250	13min48s	175	28	49.08
		Lar			174	28	
F07	18	EL	503	48min39s	199	2	8.28
		HE	503	26min53s	209	33	48.18
		Lar			209	33	
M02	38	EL	503	36min30s	99	4	16.97
		HE	503	24min55s	113	17	46.52
		Lar			112	17	
M04	50	EL	503	52min10s	93	1	18.83
			503	30min5s	140	30	52.62
		Lar			136	30	
M05	28	EL	503	45min56s	93	0	20.63
		HE	503	26min2s	138	28	52.31
		Lar			136	27	
M06	29	EL	250	19min32s	94	1	16.61
		HE	250	12min58s	119	20	53.32
		Lar			117	20	
Sum			6030	7h12min44s			

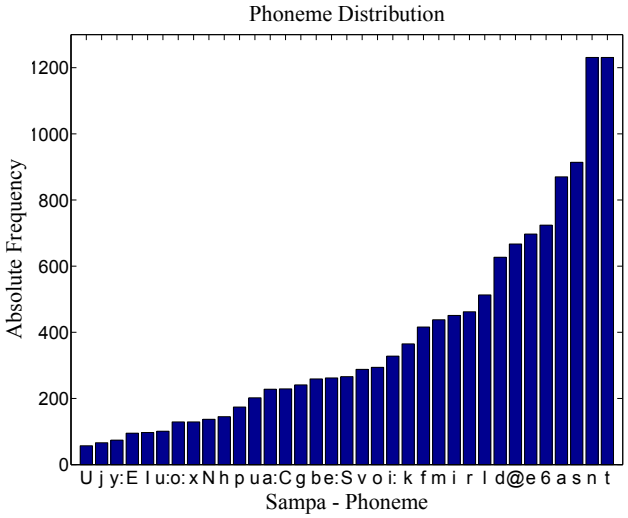


Fig. 4. *Phoneme distribution of the ELHE corpus.*

IV. Conclusion

In this paper, we introduced a new parallel ELHE corpus that consists of sentences spoken one time with an EL device and one time with healthy speech. Additionally to the healthy speech utterances, ground truth signals from the laryngograph are provided. The different kinds of sentences allow to analyze EL speech according to prosodic properties. With this corpus, the influence of different enhancement strategies can be evaluated. Speech recognition results, which represents a measure for speech intelligibility, can be compared directly between EL and HE speech. Mapping of prosodic properties from healthy speech to EL speech can be implemented using machine learning strategies.

We introduced the recording setup and material of this corpus and some statistic information on f_0 and word statistic. SNR estimation is carried out, in order to emphasize the spectral difference between healthy speech and EL speech.

V. Acknowledgments

The authors would like to thank HEIMOMED Heinze GmbH & Co.KG for their support.

VI. References

[1] G. S. Meltzner, and R. E. Hillman, “Impact of abnormal acoustic properties on the perceived quality of electrolaryngeal speech”, VOQUAL’03, pp. 73 – 78, 2003.

[2] M. Hagmüller, “Speech Enhancement for Disordered and Substitution Voices”, Ph.D. Thesis, Graz University of Technology, 2009.

[3] C. Draxler, and K. Jänisch, “SpeechRecorder - An universal platform independent multichannel audio recording software”, Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, pp. 559 – 562, 2004.

[4] E. Hänsler, and G. Schmidt, “Acoustic Echo and Noise Control: A Practical Approach”, John Wiley & Sons, New York, USA, 2004.

[5] P. Boersma, and D. Weenink, “Praat ver 4.06”, Software, downloaded from http://www.praat.org, 2007.

[6] J. Wells, “Computer-coding the IPA: A proposed extension of SAMPA”, Unpublished notes, http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm, Department of Phonetics and Linguistics, University College London, 1995.

[7] Y. Qi, and B. Weinberg, “Low-frequency energy deficit in electrolaryngeal speech”, J Speech Hear Res, Lisbon, Portugal, v. 34, pp. 1250 – 1256, 1991.

PHYSICAL SIMULATION OF VOICE TREMOR

R. Fraile[1], J. I. Godino-Llorente[1], M. Kob[2]

[1] Circuits & Systems Engineering Department, Universidad Politécnica de Madrid, Madrid, Spain, rfraile@ics.upm.es | igodino@ics.upm.es
[2] Erich-Thienhaus-Institut, Hochschule für Musik Detmold, Detmold, Germany, kob@hfm-detmold.de

Abstract: **Voice tremor is simulated using a high-dimensional discrete vocal fold model. Specifically, the effect of respiratory tremor on lung pressure and the effect of laryngeal tremor on vocal-fold stiffness and vocal-cord stress are covered. Reported results indicate that respiratory tremor causes amplitude modulation of the voice signal while laryngeal tremor causes both amplitude and frequency modulation. An analysis of the modulation present in the voice signal can thus be used to obtain cues related to the source of tremor.**

Keywords: **Voice tremor, Voice production modelling**

I. INTRODUCTION

Tremor can be defined as a rhythmical and involuntary oscillatory movement of a body part [1]. Voice tremor may occur in isolation from other types of tremor or together with additional manifestations of tremor. While the primary causes of tremor are still to be identified, the immediate reason for a tremulous movement is an abnormal behaviour of the muscles involved in that movement. Such abnormal behaviour may imply dystonia (e.g. vocal-fold dystonia in isolated voice tremor [1]) or irregular tension patterns [2]. According to the muscles involved, the sources of voice tremor may be roughly classified in three classes [3] [4]: (i) respiratory, (ii) phonatory and (iii) articulatory. Phonatory tremor is produced by irregular tension patterns in the laryngeal muscles. Among these, the thyroarytenoid (TA) muscles (Fig. 1) seem to play a more relevant role than other muscles [2].

From the acoustical point of view, voice tremor is better perceived from the phonation of sustained vowels than from running speech [5]. This is coherent with the nature of voice tremor: irregularity in muscular tension patterns is better noticed when a regular phonation is expected. In the speech waveform, the effect of tremor is two-fold: amplitude and frequency modulations happen [6] [7]. Both occur at a rate (in the range of 10 Hz) well below the fundamental frequency of voice [4] [7] and frequency modulation seems to be more relevant for perception [6] [7].

In this paper, we report on the results of computer simulation of voice tremor. While there presently is a large number of published works on voice production modelling and simulation, reported experiments on the simulation of voice tremor are scarce yet. Zhang and Jiang [8] simulated vocal tremor using a two-mass model. Lester et al. [9] have recently used a kinematic model of the vocal folds to analyse the relation between amplitude and frequency modulations. Barkmeier-Kraemer and Story [4] reported a preliminary experiment in which they observed the effect of an oscillatory subglottal pressure.

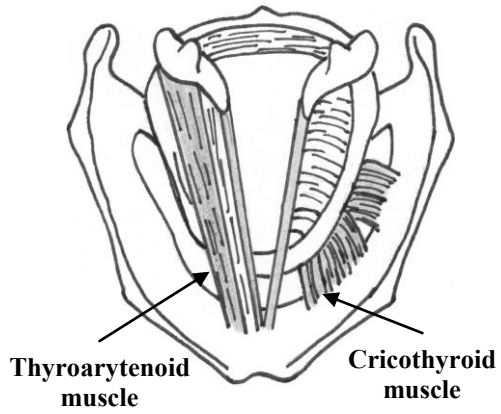


Figure 1. Thyroarytenoid and cricothyroid muscles

In our experiment, we have used a simulation model more complex than those in [4] and [8]. Namely, the multiple-mass model described in [10] has been utilized. The high dimensionality of the model allows simulating different sources of tremor. Particularly, the effect of irregularities in the tension of TA, cricothyroid (CT) (see Fig. 1) and respiratory muscles has been simulated. The output acoustic waveforms have been subsequently analysed to seek correlations between such irregularities and the changes in waveforms amplitudes and fundamental frequencies.

II. SIMULATION MODEL

A. Multiple-mass model

A detailed description of the full voice production simulation model can be found in [10]. We only include here a brief, qualitative description that serves as a framework for the tremor model.

The simulation model falls within the class of biomechanical models. It considers each vocal fold as a

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

set of 15 equal, parallel elements in the anterior-posterior direction. Thus, each element is a transverse section of one vocal fold. It is formed by two parts: the vocalis muscle and the mucosa. The masses modelling the vocalis muscle are assumed to include the vocal ligament, so the ligament stress is modelled too. Simulated mechanical properties to the vocal-fold tissues include: tissue elasticity, elasticity of the interfaces between adjacent tissues and compression forces tending to recover form after collision. Other simulated forces are: the tension of the vocal ligament, as caused mainly by the activity of the CT muscle, the pressure caused by the airflow crossing the glottis and the subglottal pressure.

The vocal tract has been simulated as a concatenation of 44 cylinders with diverse sections using the Kelly and Lochbaum model and adding a energy loss coefficient at each element. At the end of the last cylinder, whose radius corresponds to that of the open lips, the radiation on the acoustic wave is simulated assuming that no external acoustic wave arrives to the lips.

B. Tremor model

As mentioned in the introduction, the immediate causes of voice tremor may be in respiratory, phonatory or articulatory muscles. The modelling of articulatory tremor is out of the scope of this paper. That is, stable vocal tract and lip radiation of acoustic waves have been assumed. Respiratory tremor caused by tension irregularities in the respiratory muscles has been modelled by making the subglottal pressure (P_{sub} in [10]) variable instead of constant. Namely,

$$P_{\text{sub}}\ t\ = P_0 + p\ t \tag{1}$$

being P_0 = 700 Pa. Although the subglottal pressure has some relation with fundamental frequency, it does not appear to be a primary cause for changes in pitch [11] [12]. Consequently, a priori the only expected consequence of the instability of P_{sub} is amplitude modulation [4].

At glottal level, phonatory tremor is conjectured to be caused by irregular tensions in intrinsic laryngeal muscles (e.g. TA and CT). According to the results reported in [2], these are more related to vocal tremor than other extrinsic laryngeal muscles. The TA muscle is parallel to the vocalis muscle (Fig. 1) and, with respect to the CT muscle, it seems to play a secondary role in controlling the fundamental frequency [13]; the same can be said about the vocalis muscle [12]. In contrast, CT muscle activity is crucial for the determination of fundamental frequency [11] [12] [13]. The function of the CT muscle is twofold [15]: on the one hand, it affects elongation and tension of the vocal cords, on the other hand, it helps in modifying the stiffness of the vocal folds. This second function is shared with the TA muscle. In fact, activities of TA and CT muscles exhibit a moderately high degree of correlation [13].

Table I. Standard deviation of random variables modelling respiratory, TA and CT tremor.

Variable	Standard deviation
Lung pressure $p(t)$	40 Pa
Stiffness factor $k(t)$	0.15
Active stress $\sigma(t)$	4000 Pa

Since the activity of the TA muscle seems to be more closely related to tremor than that of the CT muscle [2] we may assume that the main mechanism that allows modulation of the fundamental frequency in tremor is the change in the stiffness of the vocal folds caused by irregular tension patterns in the TA muscle (and also CT, to a lesser extent). In fact, the activities of TA and CT muscles can be modelled in discrete voice production simulation models as changes in the mass and stiffness of the vocal folds [15] and such changes have been identified in [14] as a primary cause of changes in fundamental frequencies.

In the experiments reported here, the activities of the TA and CT muscles have been modelled as changes in the stiffness of the masses modelling the vocalis muscle (joint effect of TA and CT) and changes in tension of the vocal ligament (individual effect of CT). No changes in laryngeal shape or mass have been modelled so far. Change sin stiffness have been modelled as a multiplicative factor affecting the stiffness (k_s in [10]*):*

$$k_s\ t\ = k_{s0} \cdot\ 1 + k\ t \tag{2}$$

with k_{s0} having the same values specified in [10].

Last, the effect of TA and CT muscle irregularities have also been modelled as changes in the active stress experienced by the vocal cords (σ^{act}_{MAX} in [10]). Such tension has been made variable according to the following expression:

$$\sigma^{\text{act}}_{\text{MAX}}\ t\ = \sigma^{\text{act}}_{\text{MAX0}} + \sigma\ t \tag{3}$$

For all simulation runs, the constant terms in (1), (2) and (3) have kept the default values provided in [10]. The variable terms have all been modelled as zero-mean Gaussian random variables with the standard deviations specified in Tab. I.

Random values for $p(t)$, $k(t)$ and $\sigma(t)$ have been generated at a rate of 5 per second, hence modelling a tremor bandwidth of 5 Hz. At intermediate time instants a quadratic interpolation has been applied that ensures continuity of the first derivative. As a result of each simulation run, one voice sample has been generated having a duration of 1 second and a sampling rate equal to 8000 Hz.

III. RESULTS

In the first experiment, 10 synthetic voices were obtained simulating only respiratory tremor. Fig. 2 shows

the normalised standard deviations (standard deviations divided by means) of the peak amplitude of each voice period and the fundamental frequency as a function of the sample standard deviation of lung pressure[1].

As expected, according to simulation results, the variability caused by respiratory tremor in lung pressure has little effect on fundamental frequency [4]. In fact, for all 10 simulations fundamental frequency remains stable around 141 Hz. In contrast, the normalised standard deviation of the peak period amplitude of simulated voices (which is a measure of shimmer) is almost linearly related to the standard deviation of lung pressure, according to the results obtained using this simulation model. As for correlation between peak amplitude and fundamental frequency, a value of ρ=-0.22 was measured for the Pearson correlation coefficient.

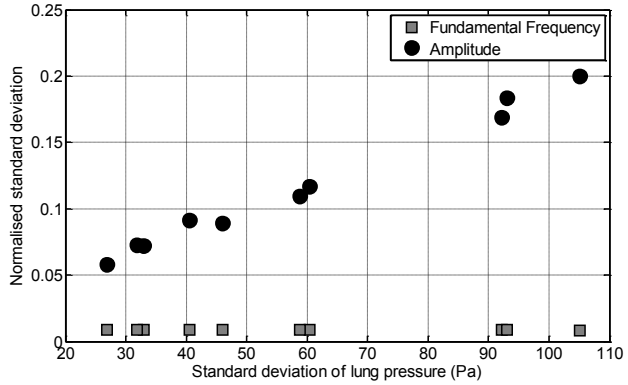


Figure 2. Normalised standard deviations of peak amplitude and fundamental frequency as a function of the standard deviation of lung pressure.

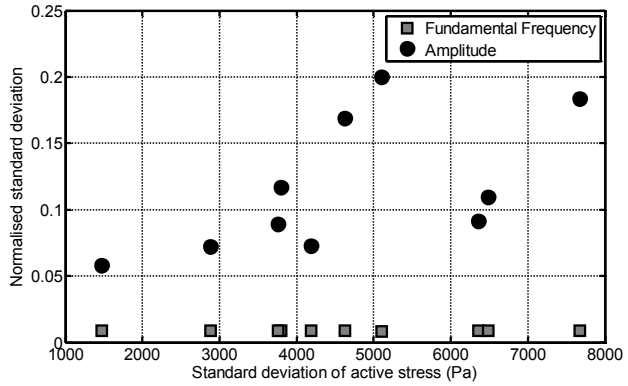


Figure 3. Normalised standard deviations of peak amplitude and fundamental frequency as a function of the standard deviation of the active stress.

[1] Note that while the standard deviation of the distribution function used for generating $p(t)$ is 40 Pa, the sample standard deviation for each simulation run may greatly differ, since only five random values are generated for each run (1 second at a rate of 5 Hz).

In the second experiment, 10 synthetic voices were obtained simulating only tremor in the active stress of the vocal ligament (mainly related to CT tremor). Fig. 3 shows the corresponding results. Again, the tremor induced in ligament stress does not affect the fundamental frequency of voice but, in contrast to Fig. 2, the linear relation between ligament stress variability and voice amplitude variability is lost. As in the first experiment, a low correlation was found between peak amplitude and fundamental frequency (ρ=-0.22).

Last, 10 additional voices were obtained simulating only tremor in the stiffness of the vocalis masses as a model for the joint effect of TA and CT tremor. Results are plotted in Fig. 4. They indicate that changes in stiffness have an effect both on voice amplitude and fundamental frequency. Also, a high correlation between peak amplitude and fundamental frequency was found in this case (ρ=0.87).

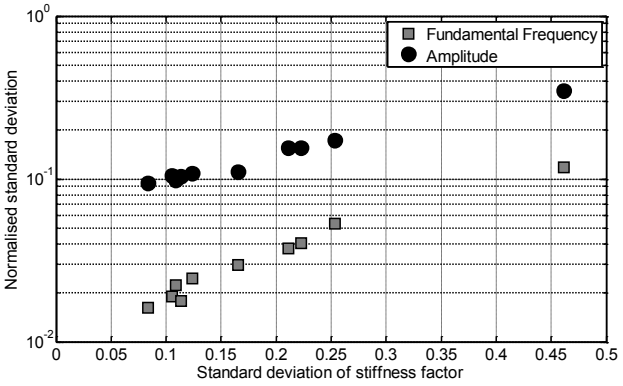


Figure 4. Normalised standard deviations of peak amplitude and fundamental frequency as a function of the standard deviation of the stiffness factor.

IV. Discussion

The effect of vocal tremor on the acoustic signal is a combination of amplitude and frequency modulation [6] [7]. In contrast, other perturbation measures such as noise seem not to be related to tremor [7]. As for modulation parameters, modulation amplitude has been reported to be much more significant for the perception of tremor than modulation frequency [7] [16]. This is coherent with the fact that physiologic and pathologic tremors happen within the same frequency range [1]. Thus, amplitude serves better the purpose of discriminating between them. Accordingly, we have chosen the standard deviations of both peak amplitude and fundamental frequency as relevant measures of tremor for our analysis.

We have independently simulated the effect of muscular tremor on three physical magnitudes affecting voice production at respiratory and laryngeal levels. As reported in previous studies [3] [4], we have found that respiratory tremor significantly affects voice amplitude and has little effect, if any, on fundamental frequency.

Fig. 2 also indicates that according to our simulation model there is a quasi-linear relation between the standard deviation of lung pressure and the standard deviation of voice amplitude. At laryngeal level, a tremulous behaviour of TA and CT muscles has a direct impact on the stress experienced by vocal cords and the stiffness of vocal folds. As indicated by our results, this affects both voice amplitude and fundamental frequency (Figs. 3 and 4).

While laryngeal tremor is likely to occur at the same time as other kinds of tremor [17] and laryngeal tremor itself affects several biomechanical parameters relevant for voice production (e.g. stress and stiffness), an independent simulation of such effects can help to identify different sources of tremor from the acoustical analysis of voice. For instance, if frequency modulation with modulation frequency up to 10 Hz is present in a voice signal, this indicates the presence of laryngeal tremor, as is the case for cerebellar voice tremor [3]. On the opposite, if no frequency modulation is present but amplitude modulation is detected, this is a cue of probable respiratory tremor and normal behaviour of the TA and CT muscles. When both amplitude and frequency modulations are present, a joint analysis of their modulation amplitudes can help to identify the source of tremor. Recently, Lester et al. [9] used this approach to analyse a case in which amplitude modulation was larger than frequency modulation. They concluded that for the analysed case laryngeal tremor was not the only source of vocal tremor. According to our results, the cross correlation between peak period amplitude and fundamental frequency may be an additional cue for identifying sources of tremor: for isolated laryngeal tremor such correlation should be higher than in the case when respiratory tremor is also present.

In the case of patients suffering from Parkison's disease, it has been reported that the primary source of vocal tremor is likely not to be at laryngeal level but at either subglottal or supraglottal levels or both [16]. While the analysis of tremor at supraglottal level has not be included in our study at this stage, our results for respiratory tremor indicate that amplitude modulation should be more relevant in tremulous voices produced by patients suffering from Parkinson's disease than frequency modulation.

REFERENCES

[1] G. Deuschl, P. Bain, and M. Brin, "Consensus statement of the Movement Disorder Society on tremor", *Movement Disorders*, vol. 13, n. S3, pp. 2–23, 1998.

[2] E. M. Finnegan, E. S. Luschei, J. M. Barkmeier, and H. T. Hoffman, "Synchrony of laryngeal muscle activity in persons with vocal tremor," *Arch. of Otolaryngology-Head & Neck Surgery*, vol. 129, n. 3, p. 313, 2003.

[3] H. Ackermann and W. Ziegler, "Cerebellar voice tremor: an acoustic analysis", *J. Neurology, Neurosurgery & Psychiatry*, vol. 54, n. 1, pp. 74–76, 1991.

[4] J. Barkmeier-Kraemer and B. Story, "Conceptual and clinical updates on vocal tremor", *ASHA Leader*, 2010.

[5] A. Lederle, J. Barkmeier-Kraemer, and E. Finnegan, "Perception of vocal tremor during sustained phonation compared with sentence context", *J. Voice*, vol. 26, n. 5, pp. 668.e1–e9, 2012.

[6] J. Kreiman, B. Gabelman, and B. R. Gerratt, "Perception of vocal tremor", *J. Speech, Lang. & Hearing Res.*, vol. 46, n. 1, pp. 203–214, 2003.

[7] S. Anand, R. Shrivastav, J. M. Wingate, and N. N. Chheda, "An acoustic perceptual study of vocal tremor", *J. Voice*, vol. 26, n. 6, pp. 811.e1–e7, 2012.

[8] Y. Zhang and J. J. Jiang, "Nonlinear dynamic mechanism of vocal tremor from voice analysis and model simulations", *J. Sound & Vibration*, vol. 316, n. 1, pp. 248–262, 2008.

[9] R.A. Lester, J. Barkmeier-Kraemer, and B.H. Story, "Physiologic and Acoustic Patterns of Essential Vocal Tremor", *J. Voice*, vol. 27, n. 4, pp. 422-432, 2013.

[10] R. Fraile, M. Kob, J. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and J. Gutiérrez-Arriola, "Physical simulation of laryngeal disorders using a multiple-mass vocal fold model", *Biomed. Signal Process. & Control*, vol. 7, n. 1, pp. 65–78, 2012.

[11] R. Collier, "Laryngeal muscle activity, subglottal air pressure, and the control of pitch in speech", *Haskins Lab. Status Rep. on Speech Res.* SR-39/40, 1974. Available: http://www.haskins.yale.edu/sr/SR039/SR03909.pdf (visited May, 2013)

[12] J. E. Atkinson, "Correlation analysis of the physiological factors controlling fundamental voice frequency", *J. Acoust. Soc. of Amer.*, vol. 63, n. 1, pp. 211–222, 1978.

[13] T. Shipp, E. T. Doherty, and P. Morrissey, "Predicting vocal frequency from selected physiologic measures", *J. Acoust. Soc. of Amer.*, vol. 66, n. 3, pp 678–684, 1979.

[14] M. Sawashima and H. Hirose, "Laryngeal gestures in speech production", *Annu. Bull. Res. Instit. of Logoped. & Phoniatr.*, n. 14, pp. 29–51, 1980.

[15] S. Y. Lowell and B. H. Story, "Simulated effects of cricothyroid and thyroarytenoid muscle activation on adult-male vocal fold vibration", *J. Acoust. Soc. of Amer.*, vol. 120, n. 1, pp. 386–397, 2006.

[16] J. Jiang, E. Lin and D.G. Hanson, "Acoustic and airflow spectral analysis of voice tremor", *J. Speech, Lang. & Hearing Res.*, vol. 43, n. 1, pp. 191-204, 2000.

[17] D. Wolraich, N. Redding, S.L. Khella, and N. Mirza, "Laryngeal tremor: co-occurrence with other movement disorders", J. Oto-Rhino-Laryng., Head and Neck Surgery, vol. 72, n. 5, pp. 291-294, 2010.

DIFFERENT IMPLEMANTATION TECHNIQUES TO INCLUDE TEETH IN MRI DATA FOR VOCAL TRACT MEASUREMENTS

L. Traser [1,2], T. Flügge[,3], M. Burdumy[1,4], R. Kammberger [5], B. Richter [1] M. Echternach[1]

[1] Institute of Musicians` Medicine, Freiburg University Medical Center, Germany
[2]Department of Otolaryngology, Freiburg University Medical Center, Germany
[3]Department of Craniomaxillofacial Surgery, Freiburg University Medical Center, Germany
[4]Department of Radiology, Medical Physics, Freiburg University Medical Center, Germany
[5]Institute of Microsystem Technology, Freiburg University, Germany

Abstract: **Magnetic resonance imaging (MRI) has been used to acquire three dimensional models of the vocal tract to analyse their acoustic properties. However, in MRI techniques, teeth do not generate a strong signal and cannot be discerned in images. Nevertheless, they might have an acoustic effect on formants. In this investigation we analyzed four different methods to introduce teeth into 3D models of the vocal tract that were segmented out of MRI scans. Acoustic measurements of the vocal tract where performed.**
The highest agreement of model and reality was reached for a laser scan and a dental impression while MRI with blueberry juice as a contrast agent and a special teeth MRI showed worse conformity. Our data confirm that the four techniques differ in how precise teeth are represented while the acoustic properties of vocal tract formants differ only by a small amount
Keywords : **MRI, teeth, vocal tract**

I. INTRODUCTION

The vocal tract and its associated resonance properties are very important for the modification of the voice source. However, there are still open research questions concerning modifications of vocal tract shape in special singing and speaking functions. In the last years, strenuous efforts have been made to investigate the vocal tract with the help of magnetic resonance imaging (MRI). This technology has the great advantage of no known risks for the subject, since this imaging technique does not use ionizing radiation. Thus, MRI has been used to acquire images of the three dimensional (3D) vocal tract or analyse fast vocal tract shape modifications using dynamic real time MRI in a two dimensional plane [1]. However, clinical magnetic resonance scanners are based on imaging the proton spin of hydrogen which is rarely present in teeth. As a consequence, teeth do not generate a strong signal in MRI and cannot be discerned in images. However, teeth might have an acoustic effect on formants, as they contribute to mouth closure. Hence, for a complete acoustic analysis of the vocal tract, there is a need of including teeth in vocal tract models.

II. METHODOS

In this investigation we analyzed four different methods to introduce teeth into 3D models of the vocal tract that were segmented out of MRI scans. The vocal tract structure was extracted using ITK-SNAP 2.4.0. (see figure 1).

Four methods for superimposing teeth into MRI material were chosen to be tested in three different subjects (Authors L.T., M.B., and M.E.).

The first method comprised an MRI scan where the subject had the mouth filled with blueberry juice, as introduced by Takemoto et al [2]. A T1-weighted 3-D VIBE-sequence was used with a voxel size of 1x1x1.3 mm [3], the subject was lying in a prone position with his head inside a 12-channel head coil. The blueberry juice generates a strong signal in the oral cavity around the teeth, outlining their surface contours. The dark areas void of signal inside these contours could then be segmented. Due to the uncomfortable position the scan was done in a recording time of 13 seconds. Second, a T2-weighted Turbo Spin-Echo sequence for dental imaging [4] with scan duration of about 5 minutes and a voxel size of 0.75x0.75x.8 mm3 was used. Due to the longer signal acquisition, the tissue and fluid surrounding the teeth could be used to segment the surface of the teeth. Third, a dental impression was performed and the teeth structure scanned. Last, a dental laser scan (iTero; Align Technology, USA) was acquired that calculates a digital teeth surface model. After segmentation of the 3D vocal tract model excluding teeth, all 4 tooth models were fused with the help of anatomical landmarks like the hard palate, jaw or dental roots using the Voxim software (IVS Solutions, Chemnitz, German) (figure 1 and 2). Then, a three-dimensional rapid prototype was printed out (figure 4). For each subject one 3D model was double printed.

The models were then rated by dentists and compared to the real mouth and teeth structure of the subjects (figure 3 and 4). The rating was done using a visual analogue scale of 10cm (with 0 is very bad and 10 is very good) in 4 categories (see table 1 and 2). Also an acoustic measurement of the 3D models was performed to analyze the formant frequencies.

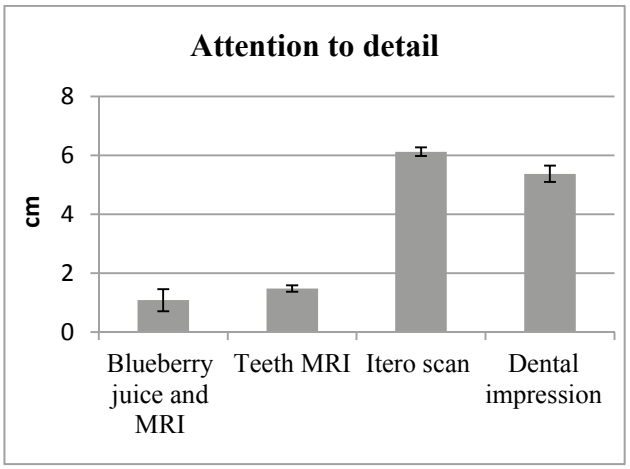
Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

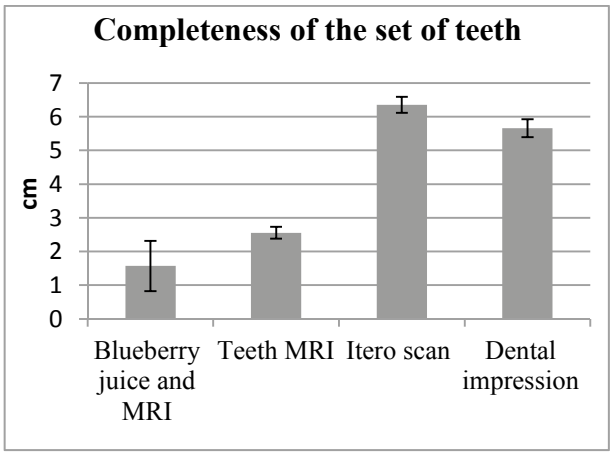
Figure 1: How a 3D Model is formed (from left to right): Preprocessed MRI image with active contour evolution driven by intensity regions; Segmented vocal tract in 2D; 3D Vocal tract model and 3D teeth model from scan are fused using anatomical landmarks from the original MRI material.

III. Results

The highest agreement of model and reality was reached for the laser scan and the dental impression for all 4 categories (see table 1 and 2). In comparison there was no significant difference between these both methods but the attention to details was better in the scan (p=0,044).
The models constructed out of the MRI material with blueberry juice showed the worst agreement. The model constructed out of the teeth MRI material was significant better than the MRI model that used blueberry juice as a contrast agent for all categories but the teeth position.
No statistical significant interaction could be shown between the subject and the model. There was no significant difference in the rating of the double models.

Preliminary results show that the acoustic properties of vocal tract formants differ only by a very small amount between the four reconstructions.





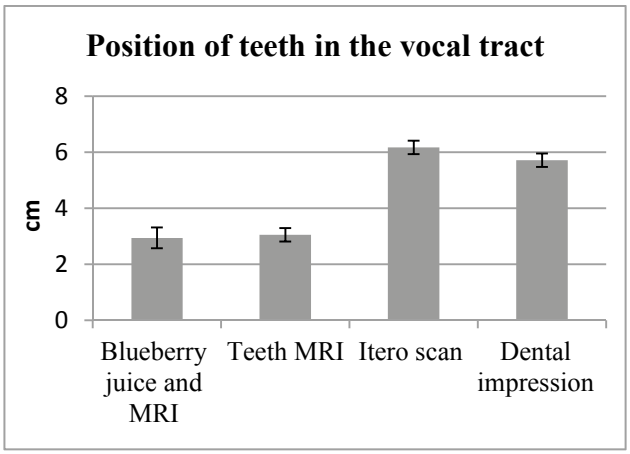


Table 1: Results of the rating using a visual analogue scale of 10cm (with 0 is very bad and 10 is very good) in the categories: I Completeness of the set of teeth, II Attention to detail and III Position of teeth in the vocal tract.

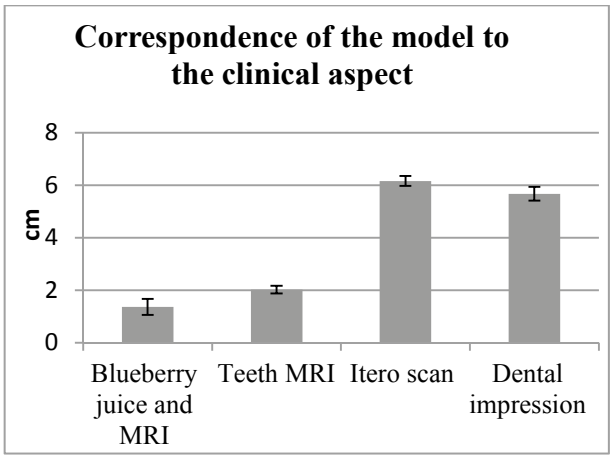


Table 2: Results of the rating using a visual analogue scale of 10cm (with 0 is very bad and 10 is very good) in the categorie: IV Correspondence of the model to the clinical aspect, presented with the standard error.

IV. Discussion

In order to measure vocal tract formant frequencies a detailed reconstruction of the vocal tract seems crucial. As a part of this, it seems necessary to implement teeth into the vocal tract structures since these could have an effect on formants. Our data shows that the four techniques to implement teeth into the vocal tract differ in how precise teeth are represented. The reliability of our data seems to be high when the double models show no significant differences in their rating results. Differences in acoustic properties between the 4 methods seem to be minor.

References

[1] Echternach, M., Markl, M., and Richter, B. (2012). "Dynamic real-time magnetic resonance imaging for the analysis of voice physiology," *Current opinion in otolaryngology & head and neck surgery*, 20, 450–457.

[2] Takemoto, H., Kitamura, T., Nishimoto, H., and Honda, K. (2004). "A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions," *Acoustical Science and Technology*, 25, 468–474.

[3] N. Rofsky, V. Lee, G. Laub, M. Pollack, G. Krinsky, D. Thomasson, M. Ambrosino, and J. W. (1999). "Abdominal MR Imaging with a Volumetric Interpolated Breath-hold Examination," *Radiology*, 212, 876–884.

[4] Hövener, J.-B., Zwick, S., Leupold, J., Eisenbeiβ, A.-K., Scheifele, C., Schellenberger, F., Hennig, J., et al. (2012). "Dental MRI: imaging of soft and solid components without ionizing radiation," Journal of magnetic resonance imaging. *JMRI*. 36. 841–846.

Figure 3: Real mouth and teeth structure of the subject LT.

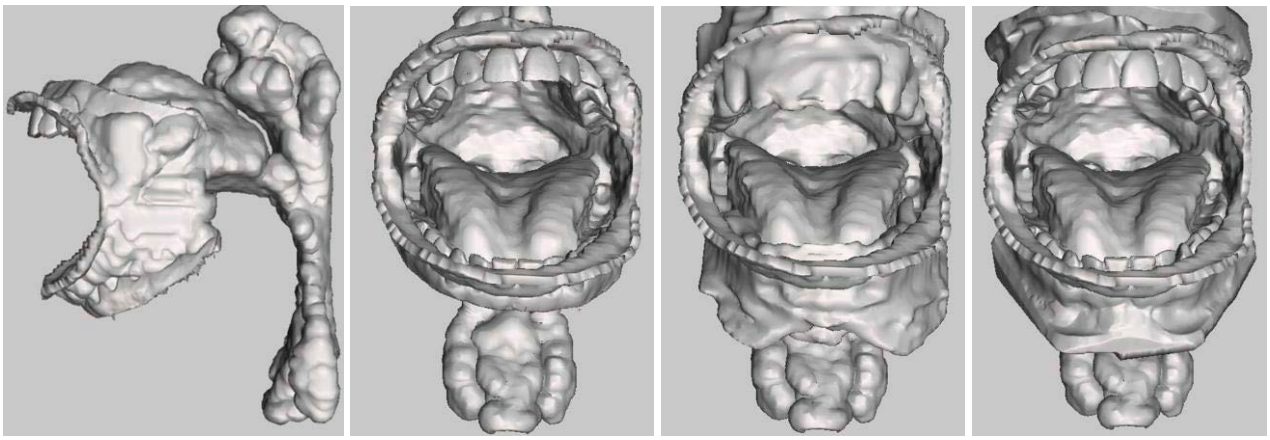


Figure 2: From left to right: Three-dimensional MRI vocal tract models with teeth generated by intraoral scan (lateral and frontal), MRI with blueberry juice and dental impression (frontal view).

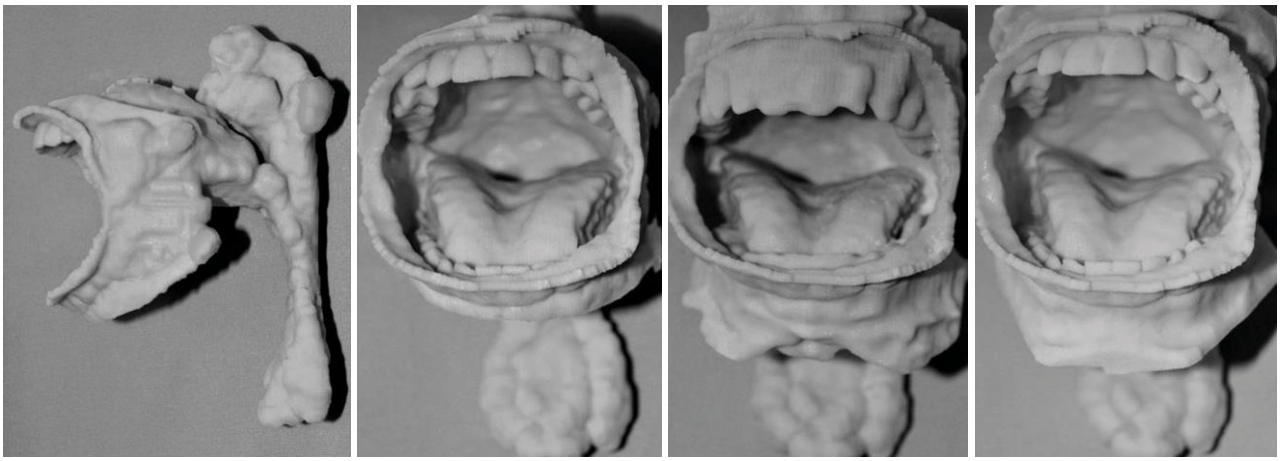


Figure 4: From left to right: Three-dimensional rapid prototype printed models of subject LT with teeth generated by intraoral scan, MRI with blueberry juice and dental impression.

RAPID MAXILLARY EXPANSION: A PRELIMINARY CONSONANT PHONETIC ANALYSIS

Bandini A[1], Biondi E[2], Lombardo L[2], Siciliani G[2], Manfredi C[3]

[1] Department of Information Engineering, University of Firenze, Firenze, Italy and Department of Electrical, Electronic and Information Engineering (DEI) "Guglielmo Marconi" - University of Bologna, Bologna, Italy
[2] Postgraduate School of Orthodontics, University of Ferrara, Ferrara, Italy
[3] Department of Information Engineering, University of Firenze, Firenze, Italy

Abstract: **Rapid Maxillary Expansion (RME) is an effective orthodontic treatment for increasing the transversal dimension of maxillary bone and therefore palatal volume. However, expansion of the palate can affect speech by altering the articulation points of the tongue and changing resonance by enlarging the oral cavity. We set out to compare the effect of two such devices on consonant and vowel sounds using a perceptive questionnaire and speech analysis software.**

Keywords : **RME, consonant, phonetic.**

I. INTRODUCTION

Rapid Maxillary Expansion (RME) is an effective treatment used widely in orthodontics, not only to counter transversal contraction of the maxillary bone, but also to make space in the upper arch for subsequent correction of dental crowding. Nowadays, the most popular of such devices is the Hyrax-type, which features two metal bands fitted over the upper first molars, and a median screw that activates 2 or 4 arms that extend towards the molars and premolars (Figure 1). RME increases the skeletal transversal dimensions of the palate exerting considerable force on the median suture, keeping it under tension and thereby bringing about new ossification. In this way up to 4.8 mm expansion can be gained at the molars, and up to 2.5 mm at the premolars, with no significant changes in the height of the palatal vault [1], and an increase in the maxillary arch perimeter of up to 6 mm [2]. This treatment does increase palatal volume by 21% [3]. These significant changes in palatal morphology can, however, affect speech. Although several orthodontics-related phonetics studies have been published, only two of these have focused on the alterations brought about by RME. The first of these studies , De Felippe 2010 [4], relied on patients' perceptions and a self-assessment questionnaire to investigate the impact of RME on speech, while the second, Stevens 2011 [5], involved phonetics analysis using speech analysis software. This was set up to compare Hyrax and Bite-Block RMEs at 6 acquisition times, analyzing fricative spectra for /s/ and /ʃ/, and vowel formants for /i/. This showed that application of either RME lowered the pitch of fricatives, and increased the frequency of the first formant while reducing the frequency of the second formant of the vowel sound /i/, alterations that returned to baseline levels once the appliance was removed. No differences between the two RMEs investigated were found.

In this study we set out to confirm that RME-induced oral cavity enlargement alters phonetics, and to investigate any differences between 2-arm and 4-arm Hyrax RMEs: it is plausible that the bulkier a device, the more it may interfere with speech.

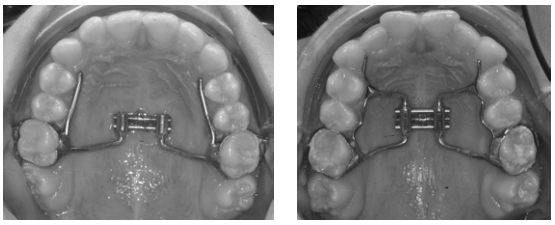


Fig.1 Rapid Maxillary Expanders: two-arms RME on the left ; four-arms RME on the right.

II. METHODS

Study inclusion criteria: 35 (21 females and 14 males, aged 7 to 14 years) RME patients in general good health were included in this study. Each underwent pre-treatment radiography, panoramic radiograph (OPT), lateral cephalogram, and cephalometric analysis. Dental impressions were taken before and after expansion.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Patients were assigned to two groups: Group A (Banded 2-arm Hyrax RME Philosophy®, Lancer Italia S.r.l.) and Group B (Banded 4-arm Hyrax RME Philosophy®, Lancer Italia S.r.l.). RMEs were activated using 12.5 mm or 14.5 mm screws, depending on the best fit in the individual's palate. Patients were treated, using the same protocol, at the University of Ferrara Dental Clinic and a private practice in Florence (Italy). Activation of the device was carried out by a single turn of screw per day, until the palatal cusp of the upper first molar came into contact with the buccal cusp of the lower first molar.

Dental casts were scanned (RevengOrthodontic professional 3D scanner, Nemotec, Italy) to obtain digital models, on which linear maxillary distances (cuspid to cuspid distance, molar to molar distance, cuspid height and molar height) and palatal volume were measured according to Gracco [3], using Rhinoceros® software. Patients were further subdivided on the basis of their palate size: small, medium or large.

Speech samples were collected using a high-quality microphone connected to a laptop computer (MacBook Pro, Mac Os X, version 10.6.7) equipped with PRAAT (version 5.3.02) recording software and 20 Mb recording buffer, at a 44.1 kHz sampling rate. All samples were recorded in a quiet room (background noise < 10 dB) with the microphone placed 5 cm below the patient's chin, directed 45° forwards and downwards. Two groups of 29 and 14 sentences, containing all the sounds of the Italian language, each repeated three times, were used as speech samples. We also recorded the pronunciation of prolonged vowel sound /i/ in 10 repetitions. Like Stevens [5], we conducted the recording 6 times, T0 before RME fitting; T1 15 minutes after RME fitting; T2 at one month; T3 at three months; T4 after RME removal (6 months after fitting); and T5 two months after RME removal.

One chosen sentence from every subject at each acquisition time was submitted to a group of 10 listeners, without any knowledge of phonetics and unaware of the aim of this study. Listeners filled out a perceptive questionnaire about speech acceptability, giving a mark from 1 to 5, according to the Likert scale. Pre-treatment scores were used to classify the test subjects as either "normal speakers" (score 1 to 1.9) or "people with pre-existing speech difficulties" (score 2 or above).

We subjected the recordings to two types of acoustic analysis:

1) Evaluation of phonetic changes during and after RME therapy. We analyzed 4 sentences containing fricatives /s/ and /ʃ/, and palatal consonants /ɲ/ and /ʎ/ spoken by 10 patients fitted with a 4-arm appliance using the above software. T0–T5 recordings were analyzed in triplicate, giving a total of 90 samples of each consonant, and 360 samples overall.

2) Comparison of the two kinds of RME (2 and 4 arms). Recordings of 4 sentences containing fricatives /s/ and /ʃ/ and palatal consonants /ɲ/ and /ʎ/ spoken by 12 patients, 6 from Group A (2 arms) and 6 from Group B (4 arms), were analyzed. We chose these consonant sounds because they involve articulation of the tongue on the palate, where the RME screw and arms are located. These sounds were analyzed at T1, just after bonding, when the speech impairment is reportedly higher [5].

Analysis protocol: Acoustic analysis of the recordings was carried out at the University of Florence Biomedical Acoustics Laboratory (Department of Information Engineering). For fricatives the power spectral density (PSD) was calculated using the Welch method (128-point window), i.e.,

- Power percentage in bands (2.5–8) kHz and (5–15) kHz.
- Spectral Moments (Mean, Standard deviation, Skewness, and Kurtosis) of the energy distribution.
- Pitch frequency in PSD.

We chose two frequency ranges for fricatives, 2.5-8 kHz and 5-15 kHz. These values are different from those used by Stevens [5], as young patients were expected to have a higher pitch than adults.

For palatal consonant analysis, the first three formant frequencies (F1, F2 and F3) were measured.

Software used: "Audacity" was used to extract consonant sounds from sentences; "Matlab R2012a" for fricative analysis; "BioVoice2" [6] to study palatal consonants; "PRAAT" for audio recording and to study palatal consonants; and "Rhinoceros" for digital measurement.

Statistical analysis: K-means cluster analysis was used to divide patients into three groups based on their linear and volumetric palate dimensions. Statistical analysis of phonetic results was performed using Matlab R2012a and MS Excel 2010.

III. Results

In our sample, 3D measurement of digital models showed that maxillary expansion had brought about a mean increase of 3.10 mm in inter-canine distance, 4.97 mm in inter-molar distance, and a mean volume increase of 744.18 mm³ (14,81%).

Experiment 1: The questionnaire showed a general perceived worsening of speech after RME placement, followed by a gradual improvement at T2 and T3. At RME removal, respondents noted a new speech impairment, which returned to pre-treatment at T5. Three of these children were judged to have "pre-existing speech difficulties, 4 a small palate before treatment and the remaining six a medium-sized palate. No correlation was found between palate size and speech sound trends.

In the acoustic analysis, we analyzed the trend of the mean values.

A. Fricatives

Both fricatives, /s/ and /ʃ/, showed similar tendencies, i.e., an initial reduction at T1 followed by an increase during treatment. Nevertheless, /s/ displayed an increase in

power in band (5–15 Hz), whereas /ʃ/ displayed stable values in both bands (2.5–8 Hz) and (5–15 Hz).
Pitch frequency, which approaches the Mean, the first Spectral Moment, decreased from T0 to T1, then increased until it reached a value greater than that registered at T0 (for /s/ from 6984.03 Hz at T0 to 8032.2 Hz at T5; and for /ʃ/ from 4453.73 Hz at T0 to 5927.6 Hz at T5), as shown in Figure 2.

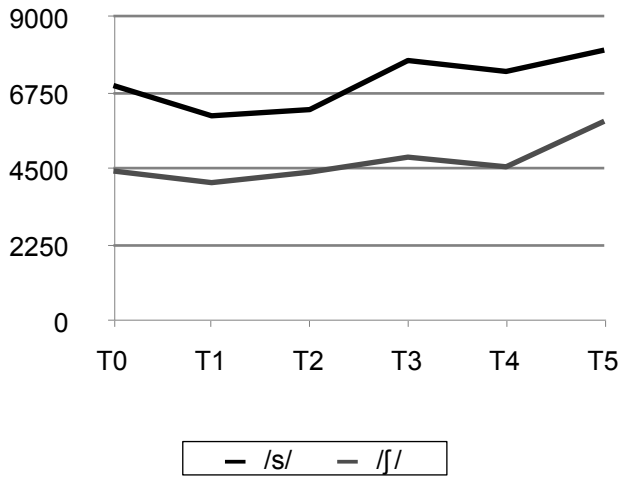


Fig.2 Fricatives /s/ and /ʃ/: trend of pitch frequency.

The standard deviation of /s/ decreased at the end of treatment, whereas the standard deviation of /ʃ/ increased. The Skewness and Kurtosis of both fricatives decreased to a value close to zero at T5.

B. Palatal consonants

Palatal consonant showed different behavior. In the lateral palatal consonant /ʎ/ (Figure 3), F1 remained stable (about 507 Hz) from T0 to T5. In contrast, F2 increased progressively from T0 to T2, then fell to a value less than that originally measured (from 1969.2 Hz at T0 to 1912.56 Hz at T5). F3 increased at T1, then decreased progressively to T5, whose value was lower than that of T0 (from 2936.72 Hz at T0 to 2838.3 Hz at T5).
Regarding the nasal palatal /ɲ/ (Figure 4), F1 remained almost stable, but the T5 value was higher than at T0 (from 428.63 Hz at T0 to 504.08 Hz at T5). F2 increased progressively from T0 to T2, then decreased to T3, subsequently remaining stable at a slightly higher value than at T0 (from 1727.37 Hz at T0 to 1834.53 Hz at T5). F3 remained nearly stable, albeit with fluctuating values; at T5 (3046.06 Hz) it was very slightly higher than at T0 (2999.16 Hz).
Experiment 2: The perceptive questionnaire indicated that Group B (4-arm expander) displayed greater speech impairment at T1 than Group A (2-arm expander). Indeed, the acoustic analysis showed lower pitch frequency, and lower standard deviation, for fricatives in the former. Likewise, regarding palatal consonants, Group B showed a higher formant frequency than group A.

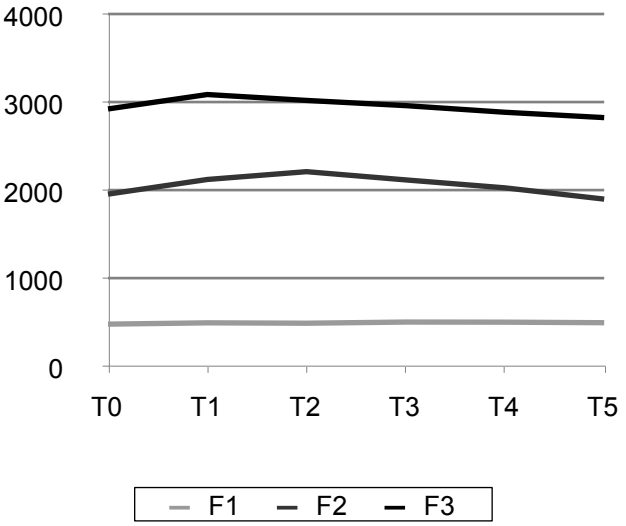


Fig.3 Palatal /ʎ/: trend of Formant F1, F2, F3 during RME therapy.

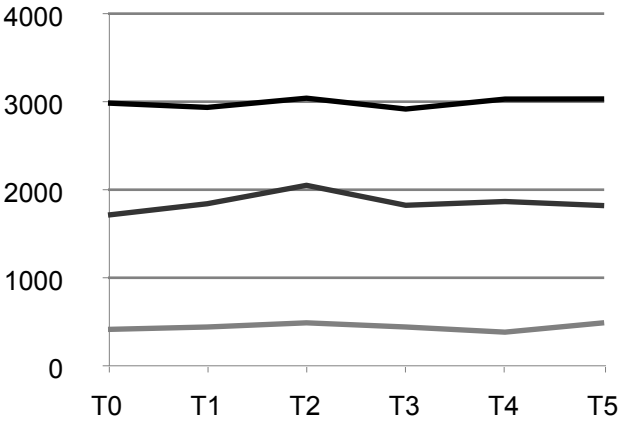


Fig.4 Palatal /ɲ/: trend of Formant F1, F2, F3 during RME therapy.

IV. Discussion

Our findings confirmed those by Hohoff [7], Stevens [5] and Khattab [8] showing that placement of an orthodontic device, such as lingual or vestibular brackets, or an RME, causes an immediate reduction in fricative frequency. In our sample both fricatives displayed the

same behavior: their frequency decreased at RME placement, then increased gradually during therapy to values higher than those measured originally. After device removal the frequency dropped, presumably due to temporary tongue disorientation immediately after RME removal, and then increased once again to a value higher than that measured pre-treatment. At the end of maxillary expansion, the spectral curve of both fricatives are more homogeneous and skewed (Skewness and Kurtosis are 0); after treatment the /s/ spectra is narrower than at the beginning, whereas the /ʃ/ spectra is wider. This means that the pitch is more stable than that recorded pre-treatment, but that the frequency is higher.

Berhman [9] and Bertino [10] have demonstrated that after surgical augmentation of the upper airways, both nasal vowels and consonants are modified in terms of a reduction in the frequency of formants. Likewise, Ungor [11] reported that subsequent to surgical reduction of the paranasal sinus in sinus lift surgery, in nasal vowels F1 decreased, whereas F2 and F3 increased. Harrington [12] explained that if a constriction in the palatal region occurs, F2, and subsequently F3, will have a higher value, whereas a high value of F1 requires a large oral cavity. Therefore, after palatal expansion, we would expect to see an increase or no change in F1, accompanying reduction in F2 and F3.

Indeed, we found that, for the lateral consonant /ʎ/, F1 remained constant, while F2 and F3 were lowered. However, the nasal /ɲ/ showed an increase in F2 frequency, but a very slight, almost imperceptible, increase of F1 and F3 frequencies. Nevertheless, this unexpected behavior is presumably ascribable to the fact that this nasal consonant is predominantly influenced by nasal, rather than oral, resonance.

The second experiment revealed a larger reduction in fricative frequency in the 4-arm expander group, with respect to the 2-arm group, a clear sign of greater speech impediment. Greater impairment, i.e., a higher formant frequency value, of even the palatal consonants was revealed in the 4-arm group. Like in the first experiment, after placement of both types of RME, formant frequency increased at F1 as a result of speech difficulty, but with the 4-arm group recordings showing greater consonant distortion, as predicted by the questionnaire findings and confirming our initial hypothesis.

V. Conclusion

This study reveals some preliminary results regarding consonant modification during and after maxillary skeletal expansion in growing children. In the short term we show that RME therapy causes modification of both fricatives and palatal consonants, which even 2 months after device removal are different from those measured pre-treatment. This appears to indicate that palatal expansion affects phonetics.

We noted that a 4-arm RME causes greater speech impairment than one with 2 arms, although this difference was only meaningful in the first month of application, since by the second or third month sounds had returned to baseline levels.

References

[1] Lagravere MO, Major PW, Flores-Mir C, “Long-term skeletal changes with rapid maxillay expansion: a systematic review”. Angle Orthod 2005; 75: 1046-1052.

[2] Lagravere MO, Major PW, Flores-Mir C, “Long-Term Dental Arch Changes After Rapid Maxillary Expansion Treatment: A Systematic Review”. Angle Orthod 2005; 75: 155-161 .

[3] Gracco A, Malaguti A, Lombardo L, Mazzoli A, Raffaeli R, “Palatal Volume Following Rapid Maxillary Expansion in Mixed Dentition”. Angle Orthod 2010; 80: 153-159.

[4] De Felippe NL, Da Silveira AC, Viana G, Smith B. “Influence of palatal expanders on oral comfort, speech, and mastication”. Am J Orthod Dentofacial Orthop 2010; 137:48-53.

[5] Stevens K, Bressman T, Gong S, Tompson BD, “Impact of a rapid palatal expander on speech articulation”. Am J Orthod Dentofacial Orthop 2011; 140:67-75.

[6] Manfredi C, Bocchi L, Cantarella G, “A multipurpose user-friendly tool for voice analysis: application to pathological adult voices”. Biomedical Signal Processing and control 2009; 4:212-220.

[7] Hohoff A, Seifert E, Fillion D, Stamm T, Heinecke A, Ehmer U. “Speech performance in lingual orthodontic patients measured by sonagraphy and auditive analysis”. Am J Orthod Dentofacial Orthop 2003; 123:146-52 .

[8] Khattab TZ, Farah H, Al-Sabbagh R, Hajeer MY, Haj-Hamede Y, “Speech performance and oral impairments with lingual and labial orthodontic appliances in the first stage of fixed treatment. A randomized controlled trial”. Angle Orthodontist 2013; 83(3): 519-526.

[9] Berhman A, Shikowitz MJ, Dailey S, “The effects of upper airway surgery on voice”. Otolaryngology-Head and Neck Surgery 2002; 127(1):36-42.

[10] Bertino G, Matti E, Migliazzi S, Pagella F, Tinelli C, Benazzo M, “Acoustic changes in voice after surgery for snoring: preliminary results”. Acta Otorinol Ital 2006; 26:110-114.

[11] Ungor C, Saridogan C, Yilmaz M, Tosun E, Senel FC, Icten O, “An acoustical analysis of the effects of maxillary sinus augmentation on voice quality”.Oral Surg Oral Med Oral Pathol Oral Radiol 2013;115:175-184.

[12] Harrington J, “Acoustic Phonetics”. The handbook of Phonetics Sciences, Wiley-Blackwell 2010.

Session II:
HIGH-SPEED IMAGING

THE EFFECT OF FRAME RATE OF HIGH-SPEED VIDEOENDOSCOPY ON THE ACCURACY OF CLINICAL VOICE ASSESSMENT

Dimitar D Deliyski[1,2,3,4], Stephanie RC Zacharias[2,4,5],
Alessandro de Alarcon[1,3], Maria E Golla Powell[2,4], Terri Treman Gerlach[6]

[1] Division of Pediatric Otolaryngology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[2] Communication Sciences Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[3] Department of Otolaryngology, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
[4] Department of Communication Sciences and Disorders, University Cincinnati, Cincinnati, OH, USA
[5] Division of Speech-Language Pathology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[6] Voice and Swallowing Center, Charlotte Eye Ear Nose and Throat Associates, Charlotte, NC, USA
dimitar.deliyski@cchmc.org, stephanie.zacharias@cchmc.org,
alessandro.dealarcon@cchmc.org, maria.e.golla@gmail.com, tgerlach@ceenta.com

Abstract: **This study investigated the impact of high-speed videoendoscopy frame rates on the assessment of nine vocal-fold clinically-relevant vibratory features. Results indicated mucosal wave magnitude and extent, aperiodicity, glottal edge, contact and loss of contact of vocal folds were the features most sensitive to frame rate. Frame rates of 8000 fps and higher are free from visually-perceivable feature degradation, and for rates of 5333 fps and higher, degradation is minimal. For rates of 4000 fps and higher, clinical assessments via visual ratings are not affected. Rates of 2000 fps and lower are inadequate for interpreting several clinical features and can lead to inaccurate functional assessment.**

Keywords: **High-Speed Videoendoscopy, Frame Rate, Clinical Voice Assessment, Laryngeal Imaging**

I. INTRODUCTION

High-speed videoendoscopy (HSV) is emerging as a potentially valuable assessment tool for visualizing the vocal folds in motion. However, lack of practical guidelines for the use of HSV is one barrier to wide-spread implementation of HSV in clinical settings. Voice researchers have access to ultra-high-speed cameras that can capture vocal-fold vibrations at rates over 20000 frames per second (fps) [1]. However, there is a large disparity between high-speed technologies available in a research setting and commercial HSV systems available in the voice clinic, with frame rates between 2000 and 5000 fps. Previous research suggests that an insufficient frame rate may bias the visual assessment of clinically relevant features [1-3]. Vibratory characteristics, such as mucosal wave, may become imperceptible at higher velocities due to temporal averaging inherent in HSV technology [1,3]. Additionally, cross-terms due to temporal aliasing may be substantial at lower frame rates [4,5]. Ultimately, if insufficient frame rate can alter a clinical feature the same way as a pathophysiological mechanism, then the accuracy and reliability of both visual and objective evaluation of HSV may be jeopardized.

The purpose of the current study is to investigate the threshold at which the HSV frame rate degrades clinically-relevant vocal-fold vibratory characteristics. The questions addressed in the study are: a) Which clinically-relevant vibratory features are most speed-demanding?; b) Does higher fundamental frequency require higher frame rates?; c) Which phonatory behaviors require higher frame rates?; d) Are the frame rate requirements different based on gender and pathology?; and e) What are the *recommended* and *minimum* frame-rate requirements for clinical assessment?

II. METHODS

Human Data: Fourteen adult vocally-normal speakers (7 male, 7 female) and 14 adults with voice disorders (7 male, 7 female) were recruited from the University of South Carolina and the Charlotte Eye, Ear, Nose, and Throat Associates. Data were collected using HSV at 16000 fps. Participants were recorded producing the vowel /i/ at six different phonatory behaviors (habitual pitch and loudness, high and low pitch, breathy and pressed phonation, and falsetto). A total of 168 recordings were collected. Recordings were reviewed for visual quality and a total of 159 tokens were used.

Data Pre-Processing: Each token was downsampled to form 17 frame-rate denominations (16000, 8000, 5333, 4000, 3200, 2667, 2286, 2000, 1778, 1600, 1333, 1143, 1000, 800, 615, 400, and 200 fps).

Vibratory Characteristics: Nine different vocal-fold vibratory characteristics were identified based on our current HSV clinical protocol, including: mucosal wave magnitude and extent, amplitude and phase asymmetry, aperiodicity, glottal edge, contact and loss of contact, and mucus bridges breaking (Table 1) [3].

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Table 1: Definitions of the nine vibratory characteristics used to compare the downsampled tokens to the original 16000-fps recordings.

Clinical Feature	Definition
Mucosal Wave Magnitude	Vertical propagation of the mucosa
Mucosal Wave Extent	Lateral propagation of the mucosa
Amplitude Asymmetry	Amplitude difference between left-right vocal-fold vibration
Phase Asymmetry	Phase difference between left-right vocal-fold vibration
Aperiodicity	Variation of the period of the glottal vibratory cycle
Glottal Edge	Smoothness and shape of the vibrating vocal-fold edges
Contact	Realization of contact of the vocal folds during vibration
Loss of Contact	Loss of contact of the vocal folds during vibration
Mucus Bridges Breaking	Release of mucus strand bridging the glottis

Visual-Perceptual Assessment: Two speech-language pathologists (SLP) and one otolaryngologist (ENT) were asked to compare two tokens of the same recording. Using custom-designed software with a specialized graphic user interface (Fig. 1), raters were asked to compare the downsampled video on the right, to the reference video on the left, which remained at a constant 16000 fps. Raters were blinded to gender, status (normal vs. pathology), phonatory behavior, and frame rate of the downsampled video; however, they were aware that each subsequent token presented on the right was at a lower frame-rate than the previous token. The raters were instructed to mark the first token at which a detectable difference between the two videos was noticed, and subsequently mark the first token at which differences between the two videos would result in a change of clinical rating (or the token was too degraded to rate). Once a change in clinical rating was indicated, a different randomized series of video tokens began.

Table 2: Stage 1 results: the features that were most sensitive to frame rate are bolded. Glottal edge also had the highest intra- and inter-rater reliability.

Clinical Feature	First Difference Noticed (FPS)
Mucosal Wave Magnitude	**5333**
Mucosal Wave Extent	**5333**
Amplitude Asymmetry	4000
Phase Asymmetry	4000
Aperiodicity	**5333**
Glottal Edge	**5333**
Contact	**5333**
Loss of Contact	**5333**
Mucus Bridges Breaking	4000

The study was conducted in two stages. The purpose of the first stage was to identify which of the nine features was most sensitive to frame rate. The purpose of the second stage was to use the feature that was most sensitive to frame rate to determine the thresholds at which visual differences are first noticed and clinical ratings change.

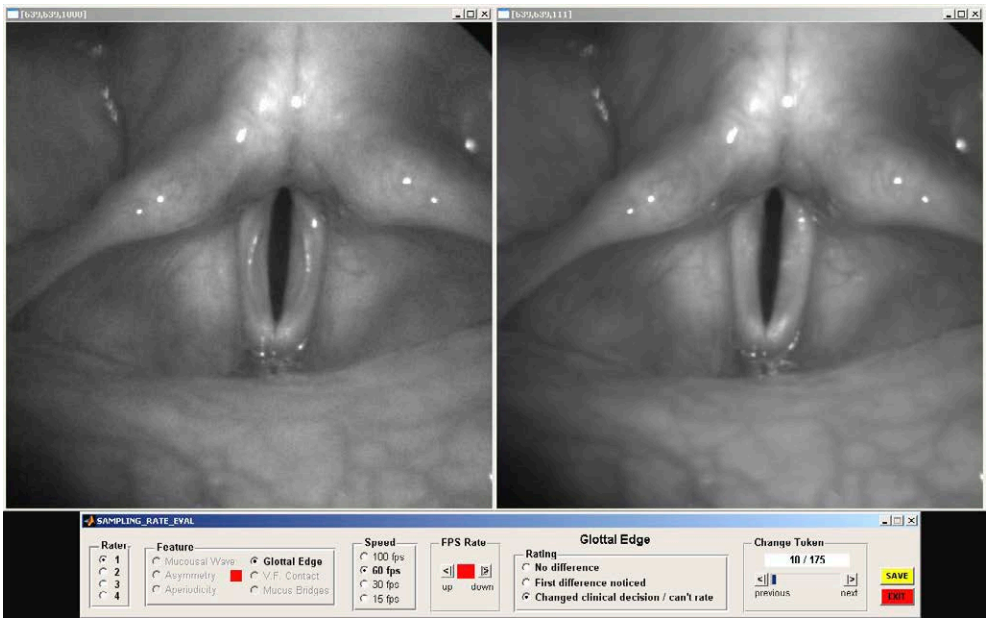


Figure 1: Customized graphic interface which allowed users to compare the reference video on the left (16000 fps) to the downsampled video on the right. Note the mucosal wave is visible on the left, but is not visible on the right.

Stage 1: For each of the 9 vibratory features previously identified (Table 1), 2 male and 2 female tokens (1 habitual; 1 high-pitch) that best represented each vibratory feature were selected for rating. Raters completed the thresholding sequence for each vibratory feature, and for each series of tokens. This stage was fully executed twice to assess for intra-rater reliability. The first trial sequences were presented at a *playback rate* of 60 fps, and the second was presented at 30 fps.

Stage 2: Based on the results of the Stage 1 experiment, the most sensitive of the nine previously-defined features was determined. All 159 of the original HSV tokens (including all 6 phonatory behaviors) were used to rate the most sensitive feature. The thresholding sequence described above was implemented for this stage as well. Ten percent randomized redundancy was built-in for intra- and inter-rater reliability assessment.

III. Results

Stage 1: A total of 7344 tokens were rated. Playback rate was not determined to affect judgments of any of the clinical features. Therefore playback rate for Stage 2 was defaulted to 60 fps. No differences were detected at 8000 fps for any vibratory feature by any of the 3 raters. Noticeable differences were first noted at 5333 fps for 6 features (Table 2). From these 6 features, glottal edge was chosen as the feature to be assessed for Stage 2 for the following reasons: a) The definition of glottal edge is most consistent across clinicians, b) Both intra- and inter-rater agreements were the highest for this feature, and c) This feature is always present and is visible in most tokens, allowing for use of the full dataset and without introducing a bias by prescreening for the presence of the feature.

Stage 2: A total of 8925 tokens were rated. No differences were detected at 8000 fps for any of the recordings by any of the raters. At 5333 fps differences were noted in 5% of the recordings. Frame rates of 4000 fps and above did not report any clinical rating changes. A 2.5% error rate was noted at 2667 fps, which increased to 6.1% at 2286 fps (Fig. 2).

An Analysis of Variance (ANOVA) was conducted across gender, behavior, and status. For first noticeable difference, statistically significant main effects were reported for gender and status ($p=0.001$), with a statistically significant interaction noted between gender and status ($p<0.001$). Statistically significant main effects for change in clinical rating were noted across all independent variables ($p<0.001$), with interaction noted between gender and status ($p=0.001$).

Intra-rater agreement was high, ranging from 94% to 100% for first noticeable difference, and from 69% to 81% for change in clinical rating. Inter-rater agreement for first noticeable difference was consistently high, ranging from 85% to 88%. For change in clinical rating inter-rater agreement was higher between the two SLPs and lower between the SLPs and the ENT (Table 3).

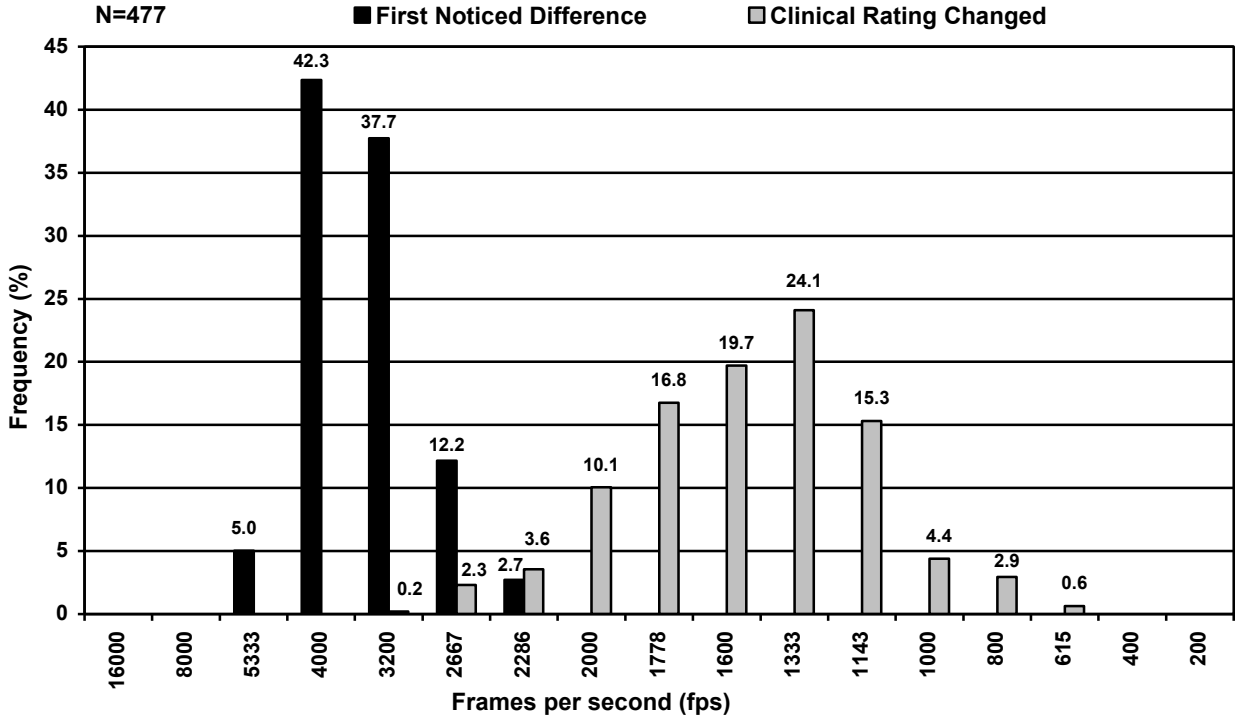


Figure 2: Visual rating thresholds for first noticeable difference and change in clinical rating in the glottal edge vibratory feature. No difference was noted at 8000 fps by any of the raters for any of the 477 recordings rated.

Table 3: Inter-rater agreement (%)

N=159	First Difference Noticed		Clinical Rating Changed	
	Rater 2 (SLP)	Rater 3 (SLP)	Rater 2 (SLP)	Rater 3 (SLP)
Rater 1 (ENT)	85%	88%	64%	69%
Rater 2 (SLP)		87%		76%

A mild correlation was found between fundamental frequency and the frame rate at which image differences were first noted. This frame rate – fundamental frequency relationship increased to a moderate level of correlation for changes in clinical rating (Table 4).

Table 4: Relationships (Spearman's Rho) between frame rate and fundamental frequency.

N=159	First Difference Noticed	Clinical Rating Changed
Spearman's Rho	0.31	0.64
p-value	0.0000	0.0000

IV. Discussion

Stage 1: All nine vibratory features (Table 1) were assessed across two phonatory behaviors (habitual and high pitch), and no visual differences were noted at 8000 fps for any of these features. This finding suggests that 8000 fps is sufficient to accurately assess all clinically relevant vibratory features without concerns of image degradation.

Stage 2: As with Stage 1 findings, no visual differences were noted at 8000 fps across gender, status and phonatory behavior, with fundamental frequencies varying in a wide range (from 72 to 1000 Hz). To date, very few clinical studies have used frame rates of 8000 fps, or above. Interestingly, the original videokymography system developed by Švec and Schutte used a scan rate of 7812.5 lines per second [6].

Our study also found that frame rates as low as 4000 fps did not affect the clinical assessment of vibratory features. Therefore, it is recommended that future clinical HSV systems allow for rates of 8000 fps with a minimum requirement of 4000 fps.

Despite several statistically significant ANOVA main effects and interactions, group-mean differences for first difference noticed are not substantial to affect the recommended rate of 8000 fps. The only substantial group-mean difference for change in clinical rating was gender. Interestingly, gender effects indicate minimum required frame rates can be reduced by up to 20% when assessing adult males (i.e. 3200 fps instead of 4000 fps). Although, it is not practical to design separate protocols for men and women, the ability to decrease frame rates without jeopardizing clinical rating can be used to improve image quality for males that present with overly dark HSV images. This also means that pre-existing HSV data from male subjects at rates of 3200 fps and above do not pose reliability issues.

Intra- and inter-rater agreement was high due to the paired-comparisons design of the study. Findings from Table 3 indicate there may be clinical professional differences in ratings of vibratory features between SLPs and ENTs.

V. Conclusion

It is recommended that future clinical HSV systems allow for rates of 8000 fps with a minimum requirement of 4000 fps. HSV recordings at rates of 2667 fps or lower should be interpreted with great caution. Rates of 2000 fps and lower are inadequate for interpreting several clinical features and may lead to inaccurate functional vibratory assessment.

VI. Acknowledgements

Funding for this study was provided by the National Institutes of Health — NIDCD: R01-DC007640 "Efficacy of Laryngeal High-Speed Videoendoscopy". Special thanks to Habib Moukalled, Kimberly Hufnagel, Shannon Batson, and Heather Bonilha for contributions in early stages of this study.

References

[1] D. Deliyski, "Laryngeal high-speed videoendoscopy." in *Laryngeal Evaluation: Indirect Laryngoscopy to High-Speed Digital Imaging,* K.A. Kendall and R. J. Leonard Eds. New York: Thieme Medical Publishers, 2010, pp. 243-270.

[2] H.S. Shaw and D. D. Deliyski, "Mucosal wave: A normophonic study across visualization techniques," *J Voice,* vol. 22(1), pp. 23-33, 2008.

[3] D.D. Deliyski, P.P. Petrushev, H.S. Bonilha, T.T. Gerlach, B. Martin-Harris, R.E. Hillman, "Clinical Implementation of Laryngeal High-Speed Videoendoscopy: Challenges and Evolution," *Folia Phoniatr Logop,* vol. 60, pp. 33-44, 2008.

[4] T. Ikuma, M. Kunduk, A. McWhorter,"Mitigation of temporal aliasing via harmonic modeling of laryngeal waveforms in high-speed videoendoscopy," *J Acoust Soc America,* vol. 132(3), pp. 1636-1645, 2012.

[5] T. Ikuma, A. McWhorter, M. Kunduk, "Effects of frame rates and window size in objective analysis of high-speed videoendoscopy data using harmonic models." in *Proceedings of the 10th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research*. D.D. Deliyski Ed. Cincinnati, Ohio: AQL Press, 2013, pp. 49-50.

[6]. J.G. Švec, and H.K. Schutte, "Videokymography: High-speed line scanning of vocal fold vibration," *J Voice*, vol. 10(2), pp. 201-205, 1996.

GLOTTAL GAP TRACKING USING TEMPORAL INTENSITY VARIATION AND ACTIVE CONTOURS

Gustavo Andrade-Miranda[1], Juan Ignacio Godino-Llorente [1]

[1] Circuits and Systems Engineering Department, Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid, Spain
gxandrade@ics.upm.es, igodino@ics.upm.es

Abstract: The present work describes a new methodology for the automatic tracking of the glottal space from high-speed digital images of the larynx. This approach involves four steps: Firstly, the region with maximal glottal opening is found, by detecting the total intensity variation of the video in the *x* and *y* axis. Secondly, we create a standard template using the information obtained from manual segmentations. Thirdly, using normalized cross correlation the surface that represents the best matching between the template and the next frame is obtained. The matching area will be the initialization for the segmentation algorithm in each frame. Finally, using an algorithm based on active contours we obtain the glottal gap. This procedure is done iteratively until the last frame has been reached. The performance, effectiveness and validation of the approach is demonstrated even in high-speed recordings in which the images present an inappropriate closure of the vocal folds.

Keywords : **Glottis, Glottal gap, HSDI, ROI**

I. INTRODUCTION

Biomedical images play an important role for a precise, fast and reliable diagnosis of the vocal folds vibration. In this sense, laryngeal images provide visual cues about the vibratory patterns that commonly the acoustic measurement cannot provide. These images are recorded using videoendoscopic techniques. There are two basic videoendoscopic procedures that are used to capture the vibratory movement of the vocal folds: slow motion stroboscopy (SMS) and high speed digital imaging (HSDI). The HSDI systems record images of the larynx at a typical rate of 2000 frames/second, while the rate obtained in slow motion is only around 25 or 50 frames per second. HDSI illuminates using a continuous light whereas SMS uses a stroboscopic lamp to show the movement of the vocal folds. A clear advantage of the HDSI with respect to SMS is that the stills are not fuzzy and incorrectly illuminated. However, both methods present camera rotations, side movements of the laryngoscope, and movements of the patient, causing a delocalization of the vocal folds and the glottal gap that complicates the application of automatic image processing techniques. On the other hand, an accurate detection of the glottal gap and its tracking along time is required to objectively characterize the vibratory patterns of the vocal folds. This is usually carried out synthesizing different representations such as Glottal Vibration Profiles (GVP), Glottal Area Waveforms (GAW), Kymograms, and extracting some important measurements such as: the ratio of vibratory amplitude, ratio of periods of vibration, etc. It is known that these parameters are correlated with voice quality and health condition, and help the specialist to evaluate the phonation in an objective way. Currently, the previous task of identifying the glottal gap is usually carried out using semi-automatic methods. In this context, and with the exponential growth of computer power and the constant improvement of the algorithms used for image processing, the hard task of automatically segmenting the glottal space has achieved a dramatic advancement. However, many of the techniques found in the literature still have weaknesses that make them impractical in a clinical environment, in which the automatization and reliability are fundamental. The most common techniques reported in the literature to detect the glottal space are based on histogram [1], region growing methods [2], watershed [3] and active contour delineation methods [4].

The main issues with the aforementioned algorithms is that they do not take into account the temporal dimension of the problem and they do not consider that the glottis correspond less than 25% of the total image, so each frame is treated individually leaving aside the information obtained from the previous frames. The method proposed is based on analysis the intensity variations throughout the video sequence or part of it, finding the region of the greatest rate of change that will be our region of interest (ROI). The ROI help us to reduce the number of false detections, and could be iteratively updated to be tolerant for camera displacements. Meanwhile, the standard template combined with the cross correlation operation will provide the initialization for each frame. Among the advantages of the method are: high degree of adaptability to existing techniques and easy implementation.

The rest of the work is organized as follows: Section 2 develops the methodology implemented for the glottis tracking. Section 3 evaluates the results obtained using the new approach and finally in section 4 presents some conclusions.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

II. METHODS

A. ROI Detection

The displacements of the glottis are small between consecutives frames, so images taken at consecutive time instants are usually strongly related among them. The motion that we are going to evaluate is the deformation of the glottis along the time. The videos recorded for this purpose focus their attention in the glottal gap during all the sequence. For this reason, it is possible to consider the translation movements in a short period of time equal to zero. However, due the involuntary movements of the camera or the patient and the skills of the clinicians during the acquisition, the videos still present small translations that are significant when the number of frames that will be evaluated increases. Considering the aforementioned, establishing a criterion based on the spatial intensity profile changes to detect the ROI is reasonable. Finally, the size of the window to be tracked can be selected adaptively based on the variations of image intensity and inter-frame disparity in a set of frames. The algorithm takes advantage of the continuous light features presented in the HSDI to analyze the intensities variation of each frame in x and y coordinates, in order to find the area with the largest variability within the image. Such area will correspond to the glottis, since it is the region with more intensity variation over time. Firstly, the total columns variation intensity (TIV_c) is computed through the analysis of the intensity variation of each frame with respect to the average intensity variation of the N frames. TIV_c represents the region with the largest variability in the N frames evaluated, and its behavior resembles to a Gaussian-like function whose center coincides with the glottal gap. The last step is finding the stationary points or the points with the largest positive intensity variation, which will be used to establish the boundaries of the ROI in the x axis. After the analysis of 18 HSDI sequences with their respective TIV_c, it was observed that optimal cut-off values were below 40% of the maximum peak of the Gaussian, so we proceeded to use the method of gradient descent and we determined the stability values or abrupt intensity changes that meet the condition;

$$X_m - X_{m-1} \geq 0.009 \quad (1)$$

Where m is equal to the number of columns in the sequence and 0.009 is the selected threshold. Therefore, when the above condition is satisfied we say that X_m is the cutoff point on the x axis. To ensure that the ROI does not loss information, we add 5 pixels in the cut-off borders X_{cl} and X_{cr}.

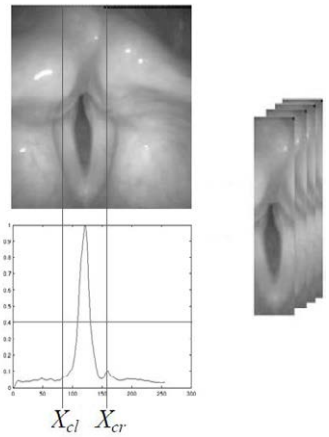


Fig. 1: Selection of the X_{cl} and X_{cr} cut-off points with N=300 frames

The procedure described above is repeated for the N frames of the video, and thus we obtain a new sequence which is a bounded version of the original one. The total intensity variation in rows (TIV_r) is computed following the same line used previously with a slightly different, since TIV_r use the result obtained of TIV_c as a starting point. The equations and thresholds are maintained for all the videos.

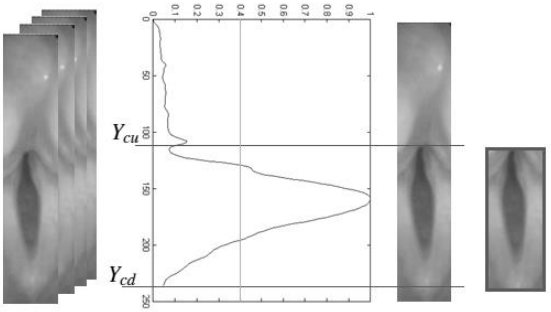


Fig. 2: Selection of the Y_{cu} and Y_{cd} cut-off points with N=300 frames

B. Standard template selection

The choice of the standard template was analyzed empirically from different types of glottis that were extracted by manual segmentations. The experimentation led us to decide choosing a template like the one showed in the Fig 3. The template is composed by a white background and a black foreground, and it has a glottis-like shape. The white background acts like an edge enhancer in order to highlight the glottis contour. The size of the template is 12x42 pixels.

Fig 3. Standard template

C. Cross Correlation Coefficient

The use of the correlation for object identification is an idea that goes back to the beginning of image processing and computer vision. It computes the similarity among an image $I(x,y)$ and a given template $T(x,y)$. The cross correlation coefficient [5] is the most common correlation estimator used for object localization. It quite stable to differences in the illumination, and it can be expressed as:

$$r(m,n) = \frac{\sum_i (I_i - I_m)(T_i - T_m)}{\sqrt{\sum_i (I_i - I_m)^2}\sqrt{\sum_i (T - T_m)^2}} \quad (2)$$

Where I_i is the intensity of the ith pixel in the image I, T_i is the intensity of the ith pixel in the template T, I_m is the mean intensity of image I, and T_m is the mean intensity of the template T. The correlation coefficients are normalized within the range [-1, 1], and if the two images are absolutely identical the value is 1; if they are completely uncorrelated is 0; and if they are completely anti-correlated, for example if one image is the negative of the other, is -1. The correlation coefficient has been selected due to its easy implementation and because it provides valuable information about the glottis and vocal folds contour, and the results obtained can be used for a better initialization of different algorithms found in the state of the art. The template obtained in section B is correlated with every frame of the sequence, and the regions over a threshold empirically established in 0.45 will be chosen like initialization for the segmentation algorithm. When the resulting correlation is below the threshold, the glottis is fully closed. The Fig 4 shows the initialization of three different frames belonging to the same video sequence.

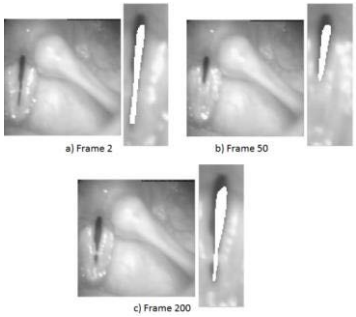


Fig. 4: Initialization of three different frames. a) Frame 2; b) Frame 50; and, c) Frame 200

C. Localizing region-based active contours

There are two main categories for the active contour models: edge-based and region-based. The edge-based utilize image gradients in order to identify the object's boundary. This type of highly localized image information is adequate in some situations. The main limitation of this model is that it usually incorporates the edge information ignoring other image characteristics. The second disadvantage is that it must be initialized close to the local minima of interest; in this way we avoid the snake to be trapped in other local minima. Meanwhile for the region-based models, the foreground and background are described statistically and this model tries to find the energy that best fits the image. The advantages of this technique include robustness against initial curve placement and insensitivity to image noise. However, techniques that use global statistics are usually not ideal for segmenting heterogeneous objects. In cases where the object to be segmented cannot be easily distinguished in terms of global statistics, region-based active contours may lead to erroneous segmentations. Glottis detection in laryngeal images has a certain degree of complexity because these images are heterogeneous and noisy at the same time. Its heterogeneity and noise can be solved using the local statistics approach proposed by Lankton and Tannenbaum [6]. The idea is to allow the foreground and background to be modeled in terms of smaller local regions, since foreground and background regions cannot be always represented with global statistics. This framework allows for correct conversion in cases of inhomogeneity, common in medical images. The analyses of local regions lead to the construction of a family of local energies at each point along the initial curve. In order to optimize the local energies, each point of the curve is considered separately and moves to minimize the energy computed in its own local region. The energy functional is given by:

$$E(\emptyset) = \int_{\Omega_x} \delta\emptyset(x) \left(\int_{\Omega_y} B(x,y)\, F\big(I(y), \emptyset(y)\big) dy \right) dx + \gamma \int_{\Omega_x} \delta\emptyset(x)\, \|\nabla\emptyset(x)\| dx \quad (3)$$

Where $\int_{\Omega_x} \delta\emptyset(x)$ allows the computation of the energy just around the curve using only contribution of the neighborhood's statistics. $B(x,y)$ ensures that the energy $F\big(I(y), \emptyset(y)\big)$ will operate only on local image information about x. The last term is only for regularization, and ensures the curve smoothness. The energy F can be modeled in three different ways: the uniform modeling energy, the means separation energy and the histogram separation energy. We choose to use the Chan Vese-model, which models the interior and exterior of region as constant intensities represented by their means. The number of iteration is setting in 300 and the radio of each local region is 7 pixels.

III. RESULTS

The methodology described in the previous section has been tested with five HSDI sequences, taken from the

database provided by Dr. Erkki Bianco y Gilles-Degottex. The resolution of the videos is 256x256 pixels and the sampling rate is 4000 frames/seconds. All the videos chosen have recorded under different conditions, such as: different illumination levels, contrast problem, presence of nodules, partial occlusion of the glottis, and lateral displacements of the camera. The ROI detection let us to eliminate more than 25% of non-relevant information found in the laryngeal images. However, variations in the number *N* will produce ROIs of different sizes. For that reason our method was set with *N*=100, which means that we recalculate the region of interest for each 100 frames. The template was chosen empirically after testing with different ones obtained by manual segmentation. In order to obtain the region with the best similarity among the template and the ROI, the next step is computing the cross correlation coefficient. This region was used like an initialization for the active contour algorithm. In some cases the initialization produce small errors, for instance, the presence of shadows that keep some similarity with the glottis features or fuzzy images in which the glottis cannot be distinguish easily. However in both cases, if the foreground and background of the regions have similar statistics the active contour based on local region evolves up to its disappearance. The algorithm presented was compared with a manual segmentation using the Pratt index [7]. This algorithm calculates a figure of merit that measures the similarity between boundaries, where 1 indicates that the two edges are equal and 0 that there is not similarity. The Fig 5 summarizes the results obtained from 5 HSDI sequences in which each of the shaded region represents a glottal cycle of the videos. The quality of the segmentation is analyzed using a 5-point scale directly linked with the Pratt index.

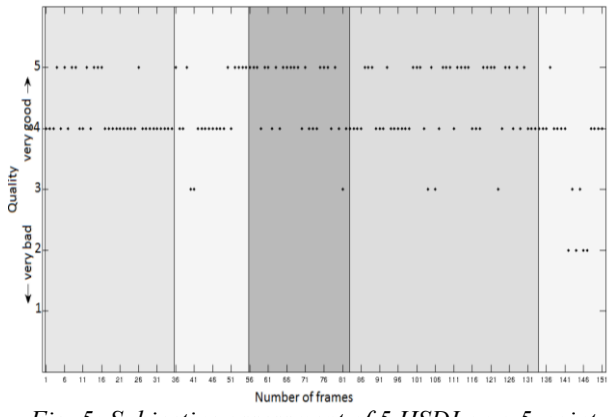


Fig. 5: Subjective assessment of 5 HSDI on a 5-point scale

IV. Conclusion

The present work proposes an alternative algorithm to the existent methods for the glottis tracking. One of the most important achievements is the fact that reduces the searching area, based only on temporal intensity variations. The weak point of the method is the empirical selection of the template; however the initialization obtained is satisfactory in most of the frames evaluated. The segmentation algorithms based on local region provide a good delimitation of the glottis. In order to illustrate the errors committed by the algorithm, the Fig 5 shows 4 frames belonging to the last sequence that reported a Pratt index under 0.5. The reason for obtaining such a low value is due to problems in the initialization, since these frames are divided in two or more regions and only one of them was located by the correlation. The motivation of this paper was to explore techniques that even being traditional for video tracking have not been considered previously in the state of art for the detection and tracking the glottal gap. The experimentation has shown that its use provides valuable information to detect and track the glottal space. The results obtained are very promising; however this algorithm needs to be tested in a wider set of conditions to ensure its generalization capabilities.

References

[1] D. D. Mehta, D. D. Deliyski, T. F. Quatieri, and R. E. Hillman, "Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings," Speech, Language and Hearing Research, vol. 54, no. 1, pp. 47 – 54, 2011.

[2] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Dollinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," Medical Image Analysis, vol. 11, no. 4, pp. 400 – 413, 2007.

[3] V. Osma-Ruíz, J. I. Godino-Llorente, N. Sáenz-Lechón, and R. Fraile, "Segmentation of the glottal space from laryngeal images using the watershed transform," Computerized Medical Imaging and Graphics, vol. 32, no. 3, pp. 193 – 201, 2008.

[4] B. Marendic, N. Galatsanos, and D. Bless, "New active contour algorithm for tracking vibrating vocal folds," in Image Processing, 2001. Proceedings. 2001 International Conference on, vol. 1, 2001, pp. 397–400.

[5] R. O. Duda and P. E. Hart, Pattern Classification and scene Analysis. John Wiley & Sons Inc, 1973.

[6] S. Lankton and A. Tannenbaum, "A localizing regionbased active contours," IEEE Trans. on Image Processing, pp. 2029–2039, 2008.

[7] I. E. Abdou and W. K. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detectors," Proceedings of the IEEE, vol. 67, pp. 753–763, 1979.

SPECTRAL ANALYSIS OF LARYNGEAL HIGH-SPEED VIDEOS: CASE STUDIES ON DIPLOPHONIC AND EUPHONIC PHONATION

P. Aichinger[1,2], I. Roesner[1], B. Schneider-Stickler[1], W. Bigenzahn[1], F. Feichter[1], A. K. Fuchs[2], M. Hagmüller[2], G. Kubin[2]
[1]Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Austria
[2]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
philipp.aichinger@meduniwien.ac.at

Abstract: **Laryngeal high-speed videos (LHSVs) are analysed in order to provide an objective diagnostic criterion for the detection of diplophonia. Diplophonia is a significant subset of irregular phonation and characterized by the presence of two pitches in the voice sound. In current clinical practice diplophonia is detected by auditive perceptual rating, which should be assisted by objective analysis whenever the presence of diplophonia is doubtful. Ten cases of diplophonic and ten cases of euphonic phonation are analysed by means of spectral video analysis (SVA). The non-unimodality measure NUM is introduced and serves as a classification feature. Estimates for the sensitivity (SE) and specificity (SP) of the presented classification paradigm are SE: 90% (95%$-CI$: [55.5%, 99.7%]) and SP: 80% (95%$-CI$: [44.4%, 97.4%]). The estimators reflect promising results. Taking into account the confidence intervals (CIs), increasing the sample size must be considered. As a consequence from these results, LHSVs should be investigated in clinical studies more intensively in order to develop and establish solid interpretation guidelines for clinical issues.**

Keywords: Laryngeal High-Speed Videos, Diplophonia, Diagnostic Study, Video Signal Processing, Voice Disorders

I. INTRODUCTION

Communication disorders may cost 154 to 186 billion dollars per year alone in the US, which equals 2.5 % to 3% of the US Gross National Product [1]. To decrease costs related to communication disorders, accurate and reliable voice assessment methods are needed [2]. Despite great efforts in conducting research on voice production, there is still a lack of such methods [3].

Diplophonia is a phenomenon in disordered voice, which is not well understood. It is characterized by the presence of two simultaneous pitches in the voice sound. Most commonly, diplophonia is understood as irregular phonation [4] or type 2 phonation [5], which does not provide an instrument for differential diagnosis (i.e., distinction between diplophonia and other kinds of voice disorders). The origins of the two pitches often remain hidden for clinicians, because standard stroboscopy does not allow for the investigation of double fundamental frequencies. As a consequence, the treatment decisions and treatment effect measurements are often difficult.

In contrast to stroboscopy, laryngeal high-speed videos (LHSVs) provide interpretable data, even for irregular phonation. However, clinical interpretation of LHSVs is still difficult, due to the lack of clinical research. It is unclear, how auditive perceptual ratings (e.g., the presence of diplophonia) relate to objective LHSV measures, which will be investigated in this study.

Several cases of diplophonic and euphonic phonation are analysed by means of spectral video analysis (SVA). It will be shown, that the presence of diplophonia relates to the presence of spatially distributed secondary oscillation frequencies in the LHSV. We encourage the use of LHSVs, whenever the presence of diplophonia is unclear.

The paper is structured as follows: The methods section describes the experimental setup, the extraction of NUM via SVA and the classification paradigm. The results section gives the classification results and individual SVA results for four representative subjects. The discussion and conclusion section concludes the paper.

II. METHODS

Ten diplophonic and ten euphonic phonations are examined by means of LHSVs at a frame rate $fr = 4\,kHz$. The videos are recorded by a phoniatrician with a rigid endoscopic camera (Richard Wolf GmbH., HRES ENDOCAM 5562). The camera is inserted into the mouth of the subject way back to the pharynx. Simultaneously to the video, audio recordings are taken with an AKG HC 577 L microphone and a TASCAM DR-100 recorder. The HC 577 L is an omnidirectional head worn condenser microphone. The DR-100 is a handheld recording device that supplies the microphone with the required phantom power and records the microphone signal as an uncompressed wav-file at a sampling rate of 48 kHz and a resolution of 24 bits. The

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

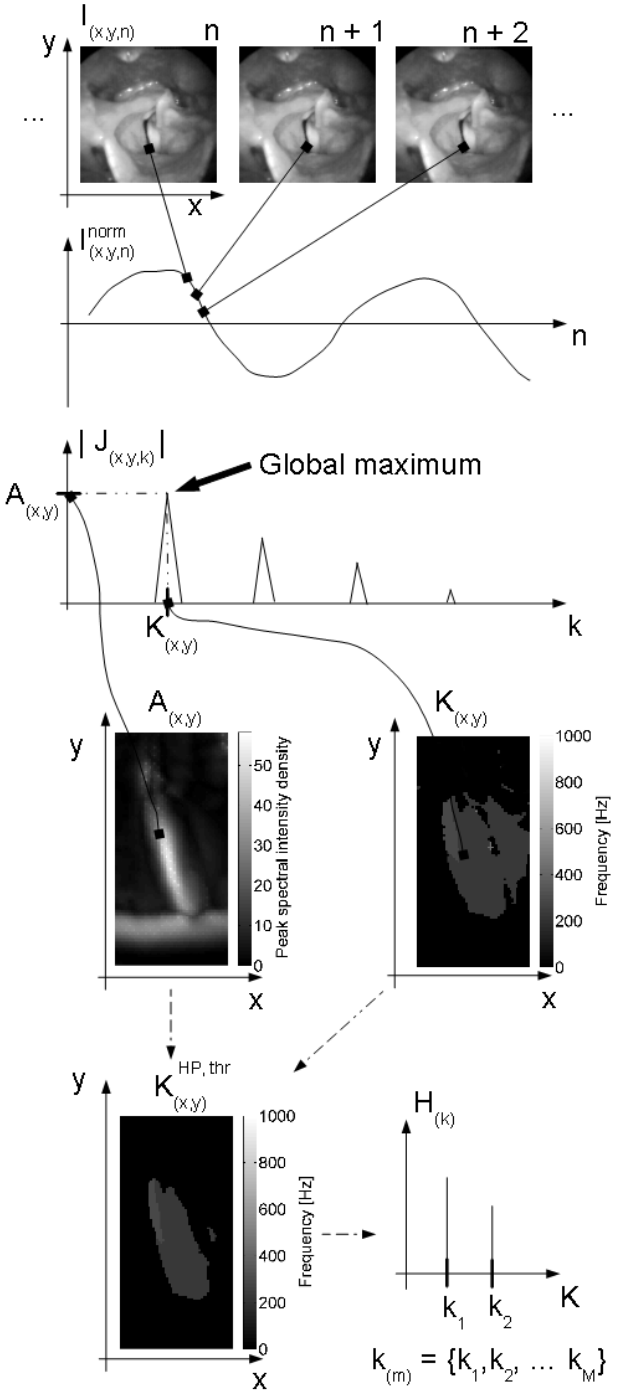


Fig. 1: Spectral video analysis (SVA): The oscillation frequencies $k_{(m)}$ of the vocal folds are measured from the LHSV.

audio records are manually synchronized to the videos by visual waveform matching.

For each phonation a stationary segment ($N = 500\,frames, 125\,ms\,at\,4\,kHz$) is recorded and analysed be means of SVA [6]. Fig. 1 visualizes the methodology of the SVA.

The video is a 3-dimensional array of intensity values $I_{(x,y,n)}$, where x is the lateral position, y is the sagittal position and n is the discrete time. The intensity time series are normalized pixel wise with respect to n.

$$I^{norm}_{(x,y,n)} = \frac{I_{(x,y,n)}}{\frac{1}{N} * \sum_{n=1}^{N} I_{(x,y,n)}} - 1 \tag{1}$$

The normalized time series $I^{norm}_{(x,y,n)}$ are windowed with a Kaiser window ($\beta = 0.5$). The windowed time series are transformed to the frequency domain via discrete Fourier transformation, giving $J_{(x,y,k)}$. For each pixel, the frequency with the maximal intensity spectral density is chosen, resulting in the peak frequency image $K_{(x,y)}$.

$$K_{(x,y)} = \underset{k}{\operatorname{argmax}}\{|J_{(x,y,k)}|\} \tag{2}$$

The peak intensity spectral density image $A_{(x,y)}$ relates the peak intensity spectral density to the x/y coordinates.

$$A_{(x,y)} = \max_{k}\{|J_{(x,y,k)}|\} \tag{3}$$

In order to suppress oscillations with irrelevant small amplitudes, the relevance threshold thr is introduced. Additionally, low frequency components are removed.

$$K^{HP,thr}_{(x,y)} = \begin{cases} K_{(x,y)} \cdots K_{(x,y)} \geq 70\,Hz \cap A_{(x,y)} \geq thr \\ NaN \cdots K_{(x,y)} < 70\,Hz \cup A_{(x,y)} < thr \end{cases} \tag{4}$$

The frequencies in $K^{HP,thr}_{(x,y)}$ are counted into the peak frequency histogram $H_{(K)}$. The distinct peaks in the histogram represent spatially distributed oscillation frequencies of the laryngeal tissue, which are referred to as "modes" in a SVA sense. M is the number of modes, which is a function of thr.

To provide a scale measure as a diagnostic criterion, the non-unimodality measure NUM is introduced. NUM is the minimal thr for which $M = 1$. We hypothesize that NUM can predict the presence of diplophonia, which will be tested.

$$NUM = min\{thr \in \mathbb{R}^{+} | M_{(thr)} = 1\} \tag{5}$$

In order to generate a baseline classification for the presence of diplophonia, a listening test on the audio records has been conducted. An expert classified the audio segments played back on an AKG K 271 MK II headphone. The experiment was blinded and randomized.

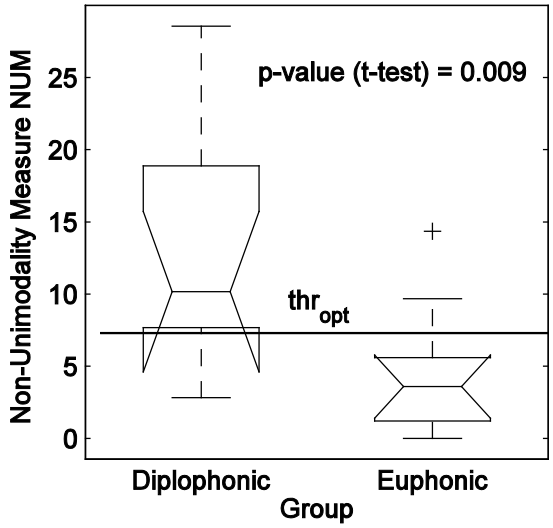


Fig. 2: Boxplot: Non-unimodality measure NUM, healthy versus diplophonic phonation.

Fig. 2 shows the boxplot of NUM for groups diplophonic and euphonic. NUM is higher in the diplophonic group, which indicates that it is a potential diagnostic criterion. The t-test (p = 0.009) reveals a significant difference of the means. However, the distributions do overlap, so it is impossible to realize perfect classification without false decisions. The optimal cut-off threshold thr_{opt} is found via a receiver operating characteristic (ROC) curve. At thr_{opt}, the sensitivity (SE) and the specificity (SP) of the test are optimal.

Fig. 3 shows the ROC curve of predicting diplophonia by means of a simple cut-off threshold classifier. The ROC curve depicts the SE (y-axis) and 1 - SP (x-axis) with respect to the cut-off threshold. Choosing a low cut-off threshold (e.g., 1) makes all of the presented cases classified as diplophonic, whereas a high threshold (e.g., 19) makes all of the cases classified as euphonic. The optimal threshold thr_{opt} is chosen as a trade-off between these two extremes of high SE and high SP. It is found by minimizing the distance from the curve to the virtual optimum point (i.e., SE = 1, SP = 1, at the upper left corner), and settles at 7.

III. Results

Table 1 illustrates the LHSV based classification of diplophonia, with the optimized cut-off threshold thr_{opt}. Cases with $NUM > 7$ are counted in the upper row (9 + 2 cases). Cases with $NUM < 7$ are counted in the lower row (1 + 8). The columns of the table represent the perceptual decision of the expert rater. Diplophonia is present in 9 + 1 cases, and absent in 2 + 8 cases. The presented table demonstrates high interrelation between the perceptual classification and NUM.

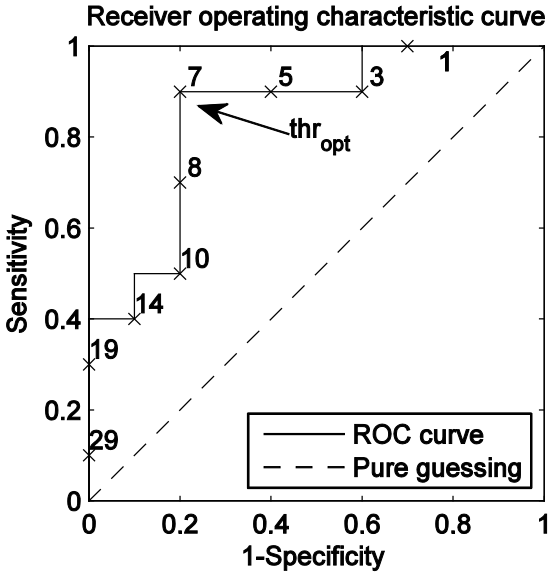


Fig. 3: Non-unimodality measure NUM as a predictor for diplophonia. The optimal cut-off threshold thr_{opt} is found by minimizing the distance from the curve to the upper left corner.

The SE and SP of the classifier are estimated as 90 % and 80 %. The CIs are found by using an iterative algorithm, based on the binomial distribution [7].

$$SE = 9 / (9 + 1) = 90\ \% \qquad (6)$$
$$CI(95\%) = [55.5\%, 99.7\%]$$

$$SP = 8 / (8 + 2) = 80\ \% \qquad (7)$$
$$CI(95\%) = [44.4\%, 97.4\%]$$

Table 1: Classification table: Non-unimodality measure NUM versus presence of diplophonia.

		Group:	
		Diplophonic	Euphonic
Test result:	$NUM > 7$	9	2
	$NUM < 7$	1	8

Figs. 4 and 5 show the SVA of four representative cases. Fig. 4 shows the peak frequency images $K_{(x,y)}^{HP,thr_{opt}}$. In the true positive case, the major part oscillates at 224 Hz. The upper left part of the image shows a secondary oscillation at 288 Hz, which is non-unimodal in a SVA sense.

The true negative case shows one relevant oscillation frequency (120 Hz) along the entire glottal region, which is unimodal in a SVA sense. The false positive case with a primary frequency of 320 Hz shows secondary oscillations (648 Hz, 968 Hz) at the upper right area of

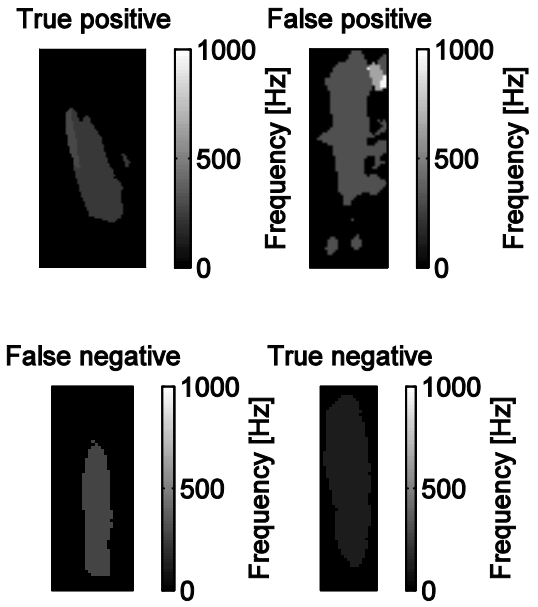


Fig. 4: Peak frequency images $K_{(x,y)}^{HP,thr_{opt}}$ of four representative cases.

the image. After visual inspection of the video, these secondary oscillation frequencies were revealed as artifacts from mucus reflections. The false negative case shows a very small area with above threshold oscillation (6 pixels at 368 Hz). Visual inspection revealed a slow glottis drift (i.e., movement of the glottis along x and y), resulting in spectral intensity density adversely spread. The false decisions of the presented cases are likely to be compensable in future versions of the analysis prototype by incorporating additional image processing techniques. Fig. 5 summarizes the observations from the peak frequency images into the peak frequency histograms $H_{(K)}$.

IV. Discussion and Conclusion

In this study, an objective diagnostic criterion for the detection of diplophonia was tested on a case group of ten diplophonic and ten euphonic subjects. It was shown that the NUM highly relates to the presence of diplophonia, as determined perceptually. The estimated values for SE and SP are promising, but hold large CIs because of the limited sample size. Thus, the analysis method must be further validated on a larger sample size before assisting the clinician in determining the presence of diplophonia.

The investigated method is more invasive than auditive perceptual expert ratings or objective acoustic methods (i.e., computer analysis of microphone signals). Nevertheless it is worthwhile to examine vocal fold oscillations by means of LHSVs. Compared to objective acoustic methods, LHSVs do not suffer from vocal tract

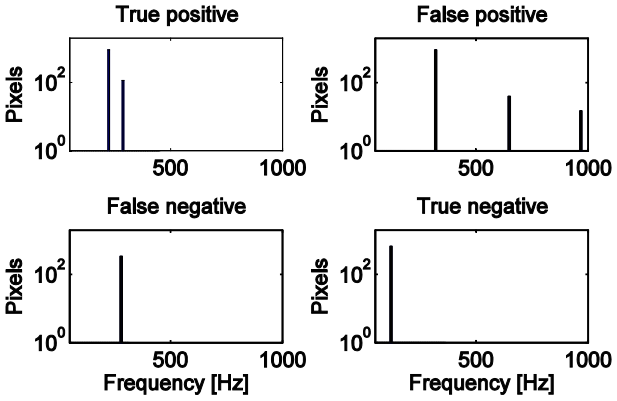


Fig. 5: Peak frequency histograms $H_{(K)}$ of four representative cases.

resonances, which makes LHSV analyses in general more accurate and less error prone. Besides, LHSVs provide spatial information on laryngeal vibration, which is not contained in the audio signal. Compared to auditive perceptual ratings, it is likely that the analysis of LHSVs will achieve more objective results in terms of reliability and validity, which must be tested in more extensive studies.

References

[1] R.J. Ruben, "Redefining the Survival of the Fittest: Communication Disorders in the 21st Century," *The Larnygoscope,* vol. 110, pp. 241–245, 2000.

[2] M. Kob and P. Dejonckere, "Advanced Voice Function Assessment - Goals and Activities of COST Action 2103," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 173–175, 2009.

[3] P. Aichinger, F. Feichter, B. Aichstill, W. Bigenzahn, and B. Schneider-Stickler, „Inter-device reliability of DSI measurement," *Logopedics Phoniatrics Vocology*, vol. 37, no. 4, pp. 167–173, 2012.

[4] D. Michaelis, M. Fröhlich, and H. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[5] I. R. Titze, "Workshop on acoustic voice analysis: Summary statement," National Center for Voice and Speech, Free books, pp. 1–36, 1995.

[6] S. Granqvist and P. Lindestad, "A method of applying Fourier analysis to high-speed laryngoscopy," *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3193–3197, 2001.

[7] M. Bland, "An Introduction to Medical Statistics," Oxford University Press, 2000.

CORRELATION BETWEEN VIDEO LARYNGOSTROBOSCOPY AND ACOUSTIC VOICE PARAMETERS

V.Uloza[1], A.Vegiene[1], R.Pribuisiene[1], I.Uloziene[1], V.Saferis [2]

[1]Department of Otolaryngology, Lithuanian University of Health Sciences (LUHS), Kaunas, Lithuania
[2]Department of Physics, Mathematics and Biophysics, LUHS, Kaunas, Lithuania
Virgilijus.ulozas@kmuk.lt

Abstract: **The purpose of this study was to evaluate quantitatively the basic parameters of the video laryngostroboscopy (VLS) and determine correlations between the basic VLS variables and acoustic vocal function parameters. Digital VLS recordings, acoustic voice assessment and calculation of dysphonia severity ondex (DSI) were performed for 108 individuals: 26 healthy and 82 patients. The VLS variables (glottal closure, regularity, mucosal wave on the affected/healthy side, symmetry of vibration, and symmetry of image) were rated and quantified on a visual analog scale. Correlations between the VLS parameters and results of acoustic voice analysis parameters and DSI measurements were tested using Pearson's correlation coefficient. The correlations of VLS variables and acoustic voice measurements were moderate and statistically significant. The DSI showed a tendency to statistically significant upper-limit moderate correlations to VLS parameters. As the result of discriminant analysis an optimum system of parameters discriminating normal and pathological voice groups was established: maximum phonation time, mucosal wave on the affected side, and glottal closure.**

Analysis of correlations between VLS and acoustic voice parameters provides more versatile approach into the pathophysiology of phonation and suggests the documented and measurable evidence of complex mechanisms of vocal function.

Keywords: **laryngostrobosopy, acoustic voice parameters**

I. INTRODUCTION

Video laryngostroboscopy (VLS) currently represents the most important and the most commonly used well-recognized method to visualize larynx and vocal fold vibrations. However, the main limitation of the VLS remains the subjective nature of the interpretation of examination results. Several attempts have been made to establish various methods of measurement and quantification of VLS findings suggesting different VLS rating forms and involving numerous diversity of VLS variables [1-2]. The most recent studies are directed to optimize the VLS evaluation for clinical settings, and therefore reduce the number of VLS parameters and reveal the most reliable judgments [3-4]. The other side of the shield is the lack of information about the correlations between data of VLS examination and other measurements of laryngeal phonatory function.

The purpose of this study was to evaluate quantitatively the basic parameters of the VLS and determine correlations between the basic VLS variables and acoustic vocal function parameters.

II. METHODS

Digital VLS recordings were performed for 108 individuals: 26 healthy and 82 patients with mass lesions of vocal folds and paralysis using XION EndoSTROB DX device (XION GmbH, Berlin, Germany) and a 90 degree rigid endoscope. The required sample size for achieving 80% power was fulfilled.

Six following VLS variables: (1) glottal closure, (2) regularity, (3, 4) mucosal wave on the affected/healthy side, (5) symmetry of vibration, and (6) symmetry of image were rated and quantified on a Visual Analogue Scale (VAS) two times with the time interval of one year by three laryngologists. To evaluate inter-rater and test-retest reliability of the VLS parameters, intra-class correlation coefficients (ICCs) were calculated.

Segments of at least 2-s duration of the sustained vowel /a:/ were analyzed using Voice Diagnostic Center lingWaves software, Version 2.5 (WEVOSYS, Forchheim, Germany) software. Acoustic voice signal data were obtained for: (1) fundamental frequency (Mean Fo, Hz), (2) standard deviation of Fo (SDF0, Hz), (3) maximum Fo (MaxFo, Hz), (4) minimum Fo (MinFo, Hz), (5) percent of jitter, (6) percent of shimmer, and (7) glottal noise energy (GNE).

Dysphonia severity index (DSI) was calculated using lingWaves VDC Vospector analysis. According to Wuyts et al., DSI is based on the weighted combination of the following selected set of voice measurements: highest frequency in Hz, lowest intensity in dB(A), maximum phonation time in sec, and jitter (%). The DSI for perceptually normal voices equals +5 and for severely dysphonic voices -5. The more negative the patient's DSI, the worse is his or her vocal quality [5].

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Correlations among the VLS parameters and results of acoustic voice assessment were tested using Pearson's correlation coefficient.

Discriminant analysis was used for classification in normal voice and pathological voice subgroups. Stepwise method, minimizing Wilks' lambda was used creating an optimal variables system Parameters that were used for analysis included acoustic voice parameters, DSI and VLS parameters. The level of statistical significance by testing statistical hypothesis was 0.05.

III. RESULTS

Generally, the ICC obtained from the total study group represented statistically significant ($p < 0.01$) ''moderate'' to ''almost perfect'' (ICC 0.46–0.90) inter-rater reliability for five VLS parameters evaluated. Some exception has been revealed only for the mucosal wave on the healthy side. The ICC of the inter-rater reliability was highest for symmetry of glottal image; the most problematic VLS parameter for rating was mucosal wave on the healthy side The ICC obtained during the first session and during the second session 1 year after remained on the same level and did not differ significantly, thus confirming substantial to almost perfect inter-rater reliability. The ICC calculated for individual raters demonstrated substantial to almost perfect intra-rater (test-retest) reliability for all raters (0.71–0.95, $p < 0.01$).

Table 1. Correlations between VLS parameters, acoustic voice parameters and DSI

Acoustic parameters	Deviance of VLS parameters					
	Glottal closure	Regularity	Mucosal wave on affected side	Mucosal wave on healthy side	Symmetry of vibration	Symmetry of glottal image
	R					
MeanFo	-0.09	-0.23*	-0.21*	-0.14	-0.28*	-0.24*
SD Fo	0.35*	0.31*	0.36*	0.27*	0.34*	0.34*
MaxFo	0.05	-0.13	-0.09	-0.02	-0.15	-0.11
MinFo	-0.24*	-0.36*	-0.35*	-0.25*	-0.39*	-0.33*
Jitter	0.50*	0.44*	0.49*	0.43*	0.42*	0.39*
Shimmer	0.53*	0.48*	0.53*	0.46*	0.48*	0.44*
GNE	-0.23*	-0.38*	-0.39*	-0.37*	-0.31*	-0.26*
MPT	-0.63*	-0.52*	-0.53*	-0.51*	-0.50*	-0.52*
MinInt	0.52*	0.56*	0.50*	0.47*	0.48*	0.47*
MaxTon	-0.62*	-0.69*	-0.67*	-0.54*	-0.69*	-0.64*
DSI	-0.69*	-0.67*	-0.68*	-0.62*	-0.64*	-0.61*

*p< 0.05

Abbreviations: R- Pearson's correlation coefficient, RGA- relative glottal area, MeanFo- fundamental frequency, SDFo- standard deviation of Fo, MaxFo- maximum Fo, MinFo- minimum Fo, GNE- glottal noise energy MPT- maximum phonation time, MinInt – minimum voice intensity, MaxTon – maximum tone, DSI – dysphonia intensity index.

Table 1 presents correlations between VLS and acoustic voice parameters. In general, the main objective acoustic voice parameters correlated significantly with VLS parameters assessed on VAS. As shown in Table 1, statistically significant moderate correlations between acoustic voice parameters, reflecting perturbations of voice signal, i.e. jitter; shimmer and SDFo, and VLS were revealed. GNE reflecting turbulent glottal noise energy and including pitch and amplitude perturbation showed statistically significant slight-to-moderate negative correlations with VAS assessed VLS parameters. However, correlations among GNE and VLS Glottal closure were week. Slight statistically significant correlations between MeanFo and VLS parameters were detected, except for Glottal closure and Mucosal wave on healthy side. Acoustic MinFo correlated slight-to-moderate with VLS parameters, however, MaxFo did not. Results in Table 1 also show correlations between VLS and DSI parameters. From the data presented in Table 5, all VLS parameters correlate significantly and moderate with DSI as an aggregate measure and with the separate components of the DS individually. However, the DSI as a compound measure that includes voice perturbation measurements and voice capabilities measurements demonstrate a tendency to statistically significant upper-limit moderate correlations to VLS parameters.

As the result of discriminant analysis using stepwise method, an optimum system of all parameters discriminating normal and pathological voice groups was established. This system included three variables: (1) maximum phonation time, (2) mucosal wave on the

affected side, and (3) glottal closure. Fisher's linear discrimination functions were used for classification. In the Table 2 the coefficients of the discrimination function are presented. The individual case is classified to the group which has the highest discriminant function value.

The mean classification efficiency reached 95.4 %, sensitivity 94.0 %, specificity 100 %, respectively. Cross validation results show correctly classified 94.5 % of cases confirming that classification rule works reliable.

Table 2. Fisher's linear discriminant function coefficients.

	Group	
	Normal voice (F1)	Pathological voice (F2)
MPT, sec	0.917	0.695
Glottal closure	0.147	0.219
Mucosal wave on affected side	0.027	0.202
(Constant)	-12.541	-17.828

Abbreviations: MPT-maximum phonation time

IV. DISCUSSION.

During decades, VLS remains the ordinary and inalterable method in clinical practice and research as the only clinically feasible test that allows visualization of the vocal folds and their vibratory function. The demand for objective and quantitative image analysis in clinical practice and research is obvious. Quantitative measures of the VLS variables could provide more reliable information for diagnostic requirements and be useful in monitoring a patient's treatment progress over time [1-4]. In this study, we have demonstrated that (1) quantification of basic VLS parameters is possible, rather simple, reliable, and clinically feasible; (2) both inter-rater and intra-rater reliability of quantitative VLS assessment reached moderate to almost perfect levels. The VLS rating form presented in this study is relatively simple, fast, does not require very special training and therefore is feasible for clinical use providing documentation and reproducibility of the VLS examination.

Results of the present study determined statistically significant moderate correlations between acoustic voice parameters reflecting perturbations of voice signal, i.e. jitter, shimmer and SDFo, and VLS parameters reflecting asymmetrical vibration pattern. Thus, our data confirm sensitivity of VLS measurements to vocal fold asymmetry assuming that severity and not just presence of asymmetry is the key aspect of the variable [6]. Acoustic GNE parameter related to turbulent glottal noise showed statistically significant slight-to-moderate negative correlations with VAS assessed VLS parameters. The general tendency was determined: the more deteriorated vibratory pattern was detected by VLS the more prominent deviances in acoustic voice signal were revealed.

These results support the current consensus that voice is a multidimensional phenomenon and cannot be described by one parameter, but should be investigated by means of voice quality and voice function analyses. Moreover, the different types of information about phonation obtained and accumulated from different sources should be considered as complementing each other. On the other hand, multiple statistically significant correlations between VLS parameters and acoustic voice measurements revealed in this study may show some overlapping and excess of information.

Discriminant analysis using stepwise method employed in this study and elaborated optimum system of VLS and acoustic voice parameters discriminating normal and pathological voice groups could serve as one of possible ways to make quantification of VLS assessment clinically feasible. This system included three variables, i.e. one acoustic and two VLS features: (1) maximum phonation time, (2) mucosal wave on the affected side, and (3) glottal closure and demonstrated high mean classification efficiency, sensitivity, and specificity.

However, further research using evolutionary computing (genetic algorithms) for selection of different type parameters and integrated search of clinically most relevant features is advocated. Future elaboration of automated methods of aggregation and analysis of different types of the most relevant information including VLS, acoustic voice measurements and clinical questionnaires data would create the necessary background for development of automated decision support systems for diagnostics of voice and laryngeal disorders [7].

V. CONCLUSION

Analysis of correlations between the vibratory pattern of the vocal folds obtained via VLS and acoustic voice parameters provides more versatile approach into the pathophysiology of phonation and suggests the documented and measurable evidence of complex mechanisms of vocal function.

REFERENCES

1. Dejonckere PH, Crevier L, Elbaz E, Marraco M, Millet B, Remacle M, Woisard V. Quantitative rating of video-laryngostroboscopy: a reliability study.*Rev Laryngol Otol Rhinol (Bord)*. 1998;119:259-260.
2. Rosen CA. Stroboscopy as a Research Instrument: Development of a Perceptual Evaluation Tool. *Laryngoscope* 2005; 115:423–428.
3. Kelley RT, Colton RH, Casper J, Paseman A, Brewer D. Evaluation of stroboscopic signs. *J Voice*. 2011;25(4):490-495.
4. Nawka T, Konerding U. The interrater reliability of stroboscopy evaluations. *J Voice.* 2012; 26(6):812.e1-10.
5.Wuyts FL, De Bodt MS, Molenberghs G, Remacle M, Heylen L, Millet B, Van Lierde K, Raes J, Van de Heyning PH. The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. *J Speech Lang Hear Res*. 2000; 43(3):796-809
6. Bonilha HS, Deliyski DD, Whiteside JP, Gerlach TT.Vocal fold phase asymmetries in patients with voice disorders: a study across visualization techniques. *Am J Speech Lang Pathol*. 2012; 21(1):3-15.
7.Verikas A, Gelzinis A, Bacauskiene M, Hallanderb M, Uloza V, Kaseta M. Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders. *Artif Intell Med.* 2010; 49:43–50.

CORRELATION ANALYSIS BETWEEN ACOUSTIC SOURCE, ELECTROGLOTTOGRAM AND NECK VIBRATIONS SIGNALS

Wolfgang Wokurek*, Manfred Pützer†
*Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland
†Klinische Phonetik, Institut für Phonetik,
Universität des Saarlandes, Saarbrücken, Deutschland
wokurek@ims.uni-stuttgart.de, puetzer@coli.uni-saarland.de

Abstract: **This study focuses on the transfer functions of the cartilage and tissue system conducting vocal fold vibrations to an acceleration sensor located in front of the larynx. Two reference signals of the voiced source are considered: Firstly, the volume velocity obtained by inverse filtering of the oral sound pressure is an indirect, but an acoustic source reference. Secondly, the electroglottogram (EGG) relatively displays the vocal fold tissue contact.**

Keywords: **acceleration sensor, electroglottogram, inversely filtered speech**

I. Introduction

Neck vibrations are of interest for various reasons. In voice dosimetry they allow quantitative collection of the voice activity (amplitude and duration) over days. The study of the subglottal cavity is a second area of research. There the resonance parameters (frequencies and bandwidths) were studied in individual, articulatory and phonatory contexts. Most studies use a scalar recording device. Efficient piezo transducers are available that result in large electric amplitudes. Obviously the direction of the vibration is lost and the scalar averages the spatial vibration in a known or unknown way.

The spatial capabilities of the sensor of this study were primary exploited in [7], [4]. The main notion of this study is to make spatial recordings of the neck vibrations and try to identify the aerodynamic path by its resonances. The unaccounted resonances will be assigned to the cartilage and tissue path. The result will be resonance parameters of system models connecting the vibration sources, volume velocity and electroglottogram (EGG), to each spatial vibration direction. This parameterization may be relevant for phonetic and medical purposes and will be evaluated in subsequent studies.

The neck vibration is recorded by a three dimensional acceleration sensor device touching the skin of the speakers neck in front of the cricothyroid ligament [7]. Movements of this tissue that originate in the vocal folds movements are considered. They are transduced to the cricothyroid ligament simultaneously by two separate ways: mechanically by cartilage contact (thyroid cartilage, arytenoid cartilages, cricoid cartilage), and aerodynamically by sound waves in the subglottal cavity (larynx below the vocal folds, trachea and bronchial tree).

Earlier studies of neck vibrations focussed mainly on the subglottal cavity, particularly on its resonance parameters [2], [1], [3], [5]. Conversely, in the context of the present study the subglottal cavity is the main component of the aerodynamic path from vocal fold vibration to the vibrations of the neck.

Inverse filtering is a challenging signal processing technique that may be successfully approached by sophisticated adaptive and iterative methods [6]. In this study a relatively simple framewise linear prediction is used. Particularly the lack of adapting and synchronizing the analysis windows to the fundamental cycle leads to artefacts in the subsequent system identification. This requires improvement in a further study.

II. Material and Methods

A. Material

One male speaker with no known voice and/or speaking problems was asked to produce the vowels [iː, aː] and [uː] at normal, low, and high pitch with modal, breathy and hoarse phonation quality. These phonation qualities are considered to cover a practical range of phonation varieties. The different vowels are checked to see any influence of vowel quality on

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

the laryngeal system. Finally, the sounds were uttered with a high and a low pitch to include variation along this dimension. For brevity, only [i:] and [a:] at normal pitch and modal phonation quality and an [a:] at normal pitch and breathy phonation quality are presented in this study.

B. Method: Sensor

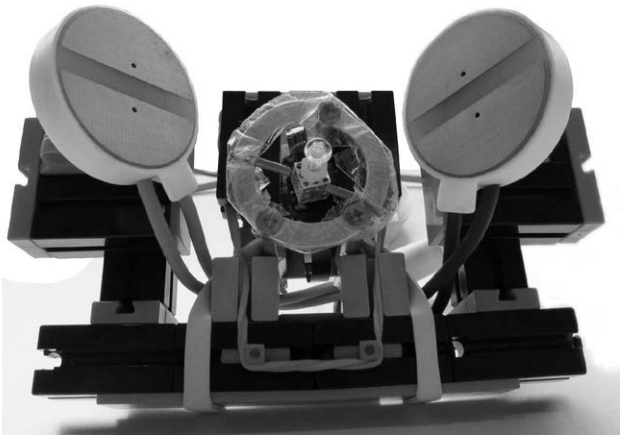


Figure 1: Acceleration sensor and electrodes for the electroglottograph

C. Method: Analysis

The oral sound pressure is analyzed by linear prediction of order 48 (the sampling rate is 48 kHz) and inverse filtered and integrated to achieve the volume velocity at the vocal folds as an acoustic source reference signal.

The sensor measures each spatial coordinate direction twice to improve the signal to noise ratio. The principle component analysis is applied to find the directions and the amplitudes of the vibration modes like in [7]. A segment stable for about a second is located manually in the sound of one of the sustained vowels. The temporal sample indices $n = 1, \ldots, N$ correspond to that segment. The samples $a_i(n)$ of the six sensor channels are arranged in columns

$$\mathbf{a}_i = (a_i(1), \ldots, a_i(N))^*, \quad i = 1, \ldots, 6 \qquad (1)$$

where * denotes transposition. The columns are put together to form the $N \times 6$ matrix of acceleration data

$$\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_6) \quad . \qquad (2)$$

Each row of $\mathbf{A}$ may be viewed as a sample of the six dimensional vector valued sequence of acceleration measurements.

To find independent modes of vibration in the acceleration vector sequence $\mathbf{A}$, the 6×6 correlation matrix is computed

$$\mathbf{R_A} = \mathbf{A}^*\mathbf{A} \quad . \qquad (3)$$

The eigen decomposition of this correlation matrix

$$\mathbf{R_A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^* \qquad (4)$$

results in the diagonal matrix

$$\mathbf{\Lambda} = \mathbf{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_6) \qquad (5)$$

of non-negative eigenvalues

$$\lambda_1 >= \lambda_2 >= \cdots >= \lambda_6 >= 0 \qquad (6)$$

and in the orthogonal matrix

$$\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_6) \qquad (7)$$

containing the eigenvectors $\mathbf{v}_i$ as columns.

According to Ineq.(6), the eigenvalues are arranged in the order of descending magnitude starting with the largest eigenvalue λ_1. The eigenvalue λ_1 represents the energy of the major vibration mode. The direction of the major vibration mode is given by the corresponding eigenvector $\mathbf{v}_1$. Similarly, the second vibration mode is given in energy and direction by λ_2 and $\mathbf{v}_2$.

The signal of vibration mode i, the principle component $\mathbf{pc}_i$, is computed by projecting the acceleration data $\mathbf{A}$ to the eigenvector $\mathbf{v}_i$

$$\mathbf{pc}_i = \mathbf{A}\mathbf{v}_i \quad . \qquad (8)$$

The principle component $\mathbf{pc}_i$ is a column vector and its elements are the samples of the acceleration along the eigenvector $\mathbf{v}_i$

$$\mathbf{pc}_i = (pc_i(1), \ldots, pc_i(N))^* \quad . \qquad (9)$$

In the following the first two principle components $pc_1(n)$ and $pc_2(n)$ are used.

The analysis method to obtain the transfer functions is calculating the cross correlations between the two source reference signals and the first two principle components of the acceleration sensor outputs $r_{Ref_k,Sensor_l}(\tau)$, as well as the autocorrelations of the source reference signals $r_{Ref_k}(\tau)$. The correlation sequences are transformed to freuqency domain

$$S_{Ref_k,Sensor_l}(f) = \mathcal{F}\, r_{Ref_k,Sensor_l}(\tau) \qquad (10)$$

and

$$S_{Ref_k}(f) = \mathcal{F}\, r_{Ref_k}(\tau) \qquad (11)$$

yielding the transfer functions

$$H_{Ref_k,Sensor_l}(f) = \frac{S_{Ref_k,Sensor_l}(f)}{S_{Ref_k}(f)} \quad . \qquad (12)$$

III. Results

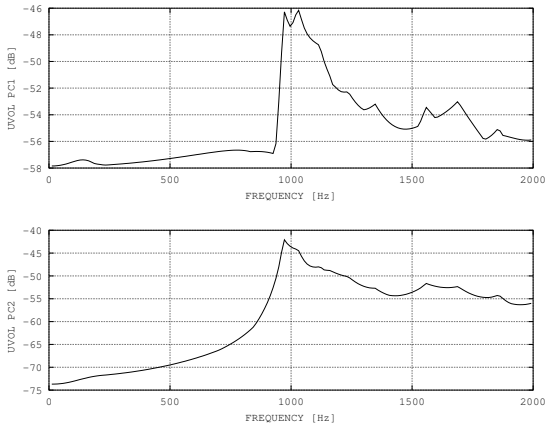


Figure 2: Transfer functions from volume velocity to the first and second principle components of acceleration. Vowel [a:], modal phonation quality

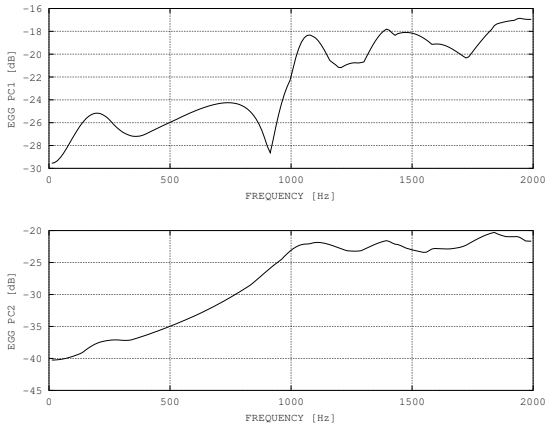


Figure 3: Transfer functions from EGG to the first and second principle components of acceleration. Vowel [a:], modal phonation quality

The amplitude spectra of transfer functions estimates are presented. Each figure shows the transfer function from the *source signal* (volume velocity at the glottis or EGG) to the first principle component of the acceleration signal at the top, and to the second principle component at the bottom. The amplitude axis is scaled in decibels. The levels contain many unknown but constant contributions from all converters (microphone, EGG, acceleration sensors, amplifiers and ADCs). Hence the amplitudes of different figures may be compared. The frequency range is from 20Hz - 2kHz. The lower limit is introduced by the sound card and the upper is in the range of the third resonance of the subglottal cavity.

Examples for modal phonation quality are shown in Figs. 2-5 and for breathy phonation quality in Figs. 6-7. The vowels [i:] and [a:] are shown for modal phonation and vowel [a:] for breathy phonation. All other varieties of the recorded material are omited for brevity.

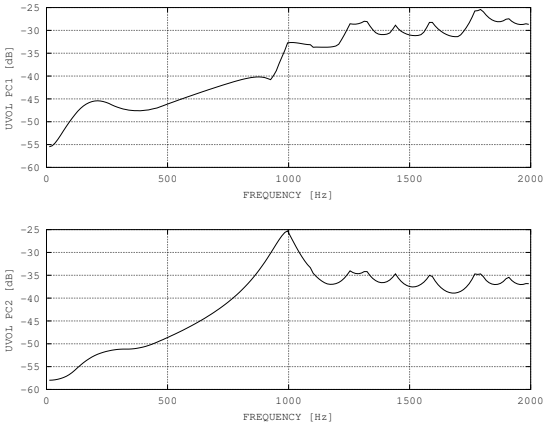


Figure 4: Transfer functions from volume velocity to the first and second principle components of acceleration. Vowel [i:], modal phonation quality

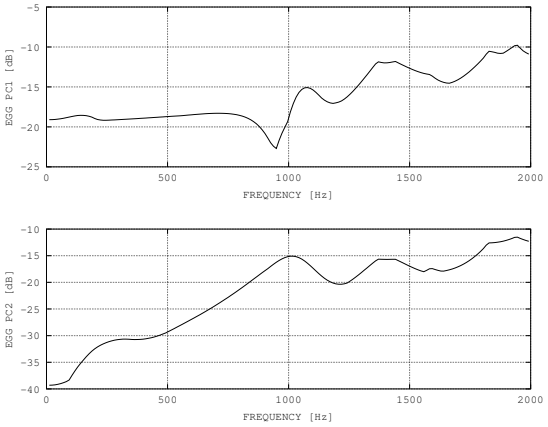


Figure 5: Transfer functions from EGG to the first and second principle components of acceleration. Vowel [i:], modal phonation quality

IV. Discussion and Conclusion

All transfer functions shown share the following landmarks (to various degrees): A resonance near 200Hz and one near 750Hz. A closely neighbored valey-peak pair near 1kHz. A resonance near 1.4kHz and one near 1.8kHz.

The lowest resonance near 200Hz was observed earlier and is attributed to the coupled sensor neck system. The sensor and its suspension alone have a resonance near 100Hz. When this system touches the neck, mass as well as elasticity and damping are added. The total effect might be the observed wide

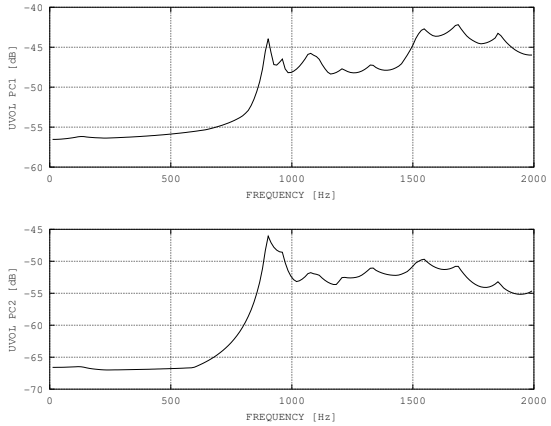


Figure 6: Transfer functions from volume velocity to the first and second principle components of acceleration. Vowel [a:], breathy phonation quality

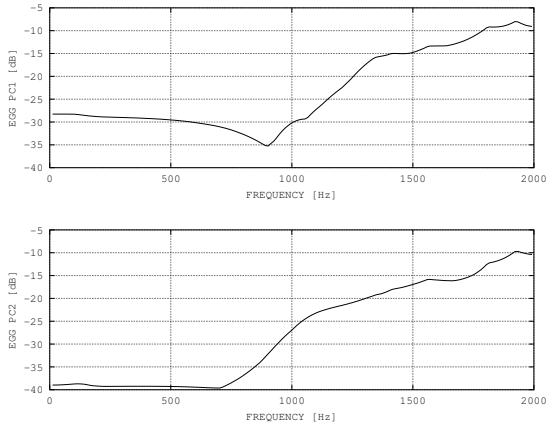


Figure 7: Transfer functions from EGG to the first and second principle components of acceleration. Vowel [a:], breathy phonation quality

peak. The peak looks more prominent in the first principle component in Figs. 3 and 4, in the second principle component in Fig. 5 and equally distributed in all other figures. For breathy phonation quality it occurs particularly weak. This lowest resonance is assumed to be translatory and might therefore be present in the directions of both principle components.

The resonance near 750Hz is attributed to the lowest resonance of the subglottal cavity. It stems from the scalar subglottal pressure and drives the sensor in and out. Hence, it is expected in the first principle compontent of the acceleration signal, where it is.

It should be noted that the first formant of the [i:] might contribute to the resonance near 200Hz and the first formant of the [a:] might contribute to that near 750Hz. This might occur, because no sophisticated procedure was used for inverse filtering.

A resonance near 1kHz was observed in our earlier work and no satisfying assignment was made. It occurs in both principle components and might stem from the rotational resonances of the sensor.

All figures show many resonance peaks above 1kHz. Two, near 1.4kHz and near 1.8kHz, might be attributed to resonances of the subglottal cavity. Whereas they are expected in the direction perpendiculat to the neck and hence in the first principle component, they might leak to the second principle component by the nature of our acceleration sensor device.

Currently the remaining resonances can not be assigned to the sensor device or the cartilage structure. The plan for future work is to measure the resonances of the sensor device and estimate that of the sensor tissue system. Then the unaccounted peaks will be attributed to the contact vibration paths in the larynx.

References

[1] X. Chi and M. Sonderegger. Subglottal coupling and its influence on vowel formants. *The Journal of the Acoustical Society of America*, 122(3):1735–1745, 2007.

[2] S. M. Lulich. *The Role of Lower Airway Resonances in Defining Vowel Feature Contrasts*. PhD thesis, MIT, 2006.

[3] A. Madsack, S. Lulich, W. Wokurek, and G. Dogil. Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs. LabPhon11, Wellington, Juli 2008.

[4] M. Pützer and W. Wokurek. Considering subglottal acceleration, electroglottographic and sound pressure signals of different phonation qualities. In *Proceedings of the 10th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research [AQL 2013*, Cincinnati, June 2013.

[5] K. N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1998.

[6] J. Walker and P. Murphy. A review of glottal waveform analysis. In *Progress in nonlinear speech processing*, pages 1–21. Springer, 2007.

[7] W. Wokurek and M. Pützer. Acceleration sensor measurements of vibrations of the larynx in patients with vocal fold adduction deficiencies. In *MAVEBA11*.

Special session: Acoustic analysis of newborn infant cry: an aid to early autism diagnosis?

Organizer: Dr. Maria Luisa Scattoni, Department of Cell Biology & Neuroscience,Istituto Superiore di Sanità, Roma, Italy and Dr. Silvia Orlandi, Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy

Introduction: Philip Sanford Zeskind, Director, Neurodevelopmental Research Levine Children's Hospital, Carolinas Medical Center, Charlotte, North Carolina, U.S.A.

DETECTION OF SUBCLINICAL NEUROBEHAVIORAL INSULT USING SPECTRUM ANALYSIS OF NEWBORN INFANT CRYING

Philip Sanford Zeskind

Department of Pediatrics, Levine Children's Hospital at Carolinas Healthcare System, Charlotte, North Carolina, USA, Philip.Zeskind@CarolinasHealthCare.org

Abstract: **Spectral analysis of the cry of the newborn and young infant has long been used as a sensitive assessment of neurobehavioral integrity. Whereas early studies described the spectral cry characteristics of infants with known nervous system damage, later work found the same measures of a higher fundamental frequency, shorter temporal components and higher threshold for elicitation to differentiate relatively healthy infants who are at risk for poor developmental outcomes. Measures of the cry sound may also detect which infants within an identified risk condition may have actually suffered nervous system insult and thus predict subsequent individual development. Perhaps most significant is the ability of infant cry analysis to detect insults to neurobehavioral integrity in the absence of abnormal signs on physical and neurological examinations. This has great implications for both the early identification of developmental disorders as well as research into the threshold for nervous system damage resulting from potential prenatal teratogens. Future translational comparisons of the vocalizations of human and rodent species may provide the avenue for studying the early identification of developmental disorders, including autism spectrum disorder.**

Keywords: **Infant crying, spectral analysis, diagnosis, autism, neurobehavior**

CRY ANALYSIS AND THE INFANT WITH KNOWN NERVOUS SYSTEM DISORDERS

The cry sound of the newborn and young infant has been likened to a biological siren, a high-pitched, wavering acoustic signal that reflects the organic state of the infant and then broadcasts that state to the social environment [1]. As an indicator of the nature of that state, variations in the cry sound may signal changes in infant homeostasis resulting from, for example, hunger or pain. Similarly, variations in the cry sound may also reflect the integrity of the young infant's neurobehavioral organization. In this capacity, listening to the infant's cry sound has long been used in clinical settings to support the differential diagnosis of brain damage. Primary among the relevant features of the cry has been the fundamental frequency, or basic pitch, of the sound. Whereas a lower-pitched, hoarse guttural cry sound has been used to support the diagnosis of trisomy genetic disorders, such as Down's syndrome, an unusually high-pitched sound characterizes the cries of infants with a wide range of other genetic disorders, such as maladie du cri du chat, and brain-damaging conditions, such as kernicterus or asphyxia [2]. A most salient hallmark of this higher-pitched cry is the presence of hyperphonation, an acoustic structure characterized by a sudden shift in frequency from the typical 400-600 Hz range to a sound over 1000 Hz. Other measures of the cry that have differentiated infants with known cases of nervous system damage include shorter expiratory sounds, shorter overall bouts of crying and higher thresholds for cry elicitation. In essence, this pioneering research described how infants who were already known, or highly suspected, to have nervous system damage also had distinctive characteristics in their cry sounds.

CRY ANALYSIS AND THE INFANT AT RISK

Subsequent research examining the diagnostic utility of spectral analysis of infant crying has focused on infants who show no routine signs of nervous system disorder yet are at risk for poor cognitive, intellectual and social development due to known nonoptimal prenatal conditions. Using the same measures of crying as those that were previously used to differentiate infants with known cases of brain damage, early studies showed that the analysis of infant crying could also detect the deleterious effects of a wide range of prenatal conditions on the newborn infant's neurobehavioral development in the absence of specific nervous system disorder [3]. Infants suffering from conditions ranging from preterm birth, atypical fetal growth and low birth weight to prenatal licit and illicit drug exposures, including maternal use during pregnancy of methamphetamine, methadone, cocaine, cigarettes and alcohol, have all been differentiated from comparison infants by various combinations of a more elaborate set of spectral and temporal characteristics of their cry sounds [4]. While different studies have used different cry characteristics and different ways of measuring the same cry characteristics, common findings among studies typically include infants producing cry sounds with a higher fundamental frequency, more frequent presence of hyperphonation, shorter expiratory sounds and more dysphonation (sonic turbulence). This continuing line of research has demonstrated the sensitivity and value of the analysis of infant crying as a concomitant measure of the infant's neurobehavioral integrity and known risk status.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

NEUROBEHAVIORAL BASES OF VARIATIONS IN INFANT CRY SOUNDS

In the absence of frank brain damage, an important question is what these variations in the cry sounds of the infant at risk signify about the integrity of the infant's nervous system. Physioacoustic models have delineated functional changes in several specific neural mechanisms producing variations in the above infant cry characteristics. These include vagal input to the laryngeal muscles controlling the pitch of the cry sound and the coordinated activity among brainstem, midbrain, and limbic systems, as well as autonomic and other neural systems controlling variations in the temporal morphology and threshold of crying [5,6]. In particular, variations in the pitch of crying originate in the lower brainstem, which controls the tension of laryngeal muscles through the vagal complex (cranial nerves IX–XII) and phrenic and thoracic nerves. Hyperphonation reflects instability in these neural control mechanisms and is often found in infants who suffer from poor autonomic and neurobehavioral regulation. Autonomic and central nervous system regulation of the respiratory cycle underlie the rhythmic temporal morphology of crying. Similarly, the threshold for the initiation of crying is directly related to integrity of the autonomic nervous system and its effects on the rhythmic organization of arousal [7]. Due to their bases in central innervation and autonomic modulation, individual differences in these measures of infant crying have been related to several other measures of neurobehavioral function in the infant, such as those regulating the infant's behavioral state and ability to orient and attend to both the social and nonsocial environments [8].

CRY ANALYSIS AND THE DETECTION OF SUBCLINICAL NEUROBEHAVIORAL CONDITIONS

A most significant utility of the analysis of infant crying is its sensitivity to subclinical neurobehavioral insult. The term "subclinical" in this case is used to refer to neurobehavioral insult in the absence of routine abnormal signs or known risk conditions. That is, the analysis of infant crying may detect neurobehavioral insult in infants who appear healthy and would otherwise go unnoticed as being at risk for poor development. For example, Prechtl's early research indicated that infants who have experienced high numbers of nonoptimal obstetric conditions are at increased risk for developing a variety of neurological disorders at 2 to 3 years of age, even if they do not show abnormal signs at birth [9]. This finding raises the important question of whether the subsequent development of neurological disorders did not occur until the child was 2-3 years of age or if there was a neurological insult present at birth but not detectable by routine measures of neurobehavioral integrity. A later study of two-day old infants with high numbers of the same nonoptimal obstetric conditions showed that their cry sounds had the same higher fundamental frequency, shorter initial expiratory sounds, shorter bouts of crying and higher threshold for cry elicitation as the cries of infants with known neurological disorders [9]. This finding was obtained even though all infants in the study were full term and full birth weight and showed no abnormal signs on routine physical and neurological examinations. In this sense, the analysis of infant crying identified subclinical neurobehavioral problems that would perhaps otherwise go undetected until two to three years of age.

Similarly, the analysis of infant crying has been used to detect subclinical neurobehavioral insults due to potential prenatal teratogens. For example, infants whose mothers consumed alcohol during pregnancy were differentiated from comparison infants by measures of the threshold, duration and fundamental frequency of crying [10]. Again, all infants were full term, full birthweight and showed no abnormal signs on routine physical and neurological examinations, including signs of fetal alcohol exposure. Importantly, the amount of maternal alcohol consumed during pregnancy was insufficient to result in any abnormal physical or neurological signs, yet was detected by analysis of the cry sound. While, for the purposes of the study, infants were selected based on a known potential teratogen, in the absence of abnormal signs, these infants would gone undetected as being at risk for poor development in a typical newborn nursery or clinic. These findings also have significance for research designed to examine the effects of potential teratogens. In this case, in a study of the threshold for "damage" due to prenatal alcohol exposure, the amount of maternal alcohol consumed in this study would have been incorrectly presumed to be "safe".

Fig. 1 shows a spectrographic presentation of a 2-sec cry expiration of a two-day old infant with no prenatal alcohol exposure. The spectrogram shows the typical pattern of phonation, a clear harmonic structure and a fundamental frequency at approximately 500 Hz. Fig 2 shows a spectrographic presentation of a 2-sec cry expiration of a two-day old infant with prenatal alcohol exposure. In addition to an average fundamental frequency of about 750 Hz and the sonic turbulence of dysphonation, hyperphonation in the middle of the segment is evident. The sudden shift in pitch was to a fundamental frequency in excess of 1900 Hz. The dominant frequency, evident in the darkest harmonic at the top of the spectrum, exceeded 3800 Hz. This would be the frequency with the highest amplitude and most notable sound produced by the infant. As a component of the biological siren, these high-pitched sounds have been shown to affect adult caregivers' physiological and behavioral responses to the infant, thus resulting in the nature of the cry sound also contributing to the infant's future developmental pathway [1].

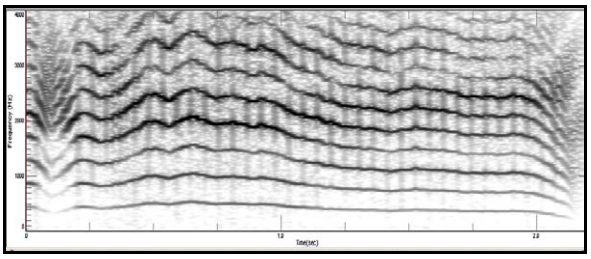


Fig. 1 Spectrogram of cry of non-exposed infant

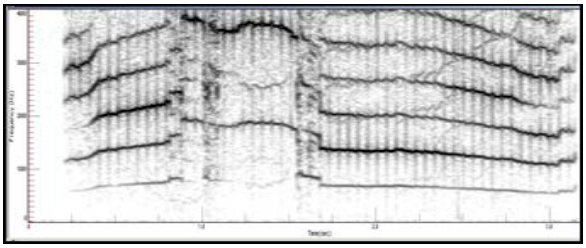


Fig. 2 Spectrogram of cry of alcohol-exposed infant

An intriguing potential of infant cry analysis is its ability to provide a window into understanding the different developmental outcomes of individual infants within a "risk" category. The concept of "risk" is a probabilistic one; that is, while some infants within a risk category eventually show poor outcomes, others do not. Similarly, while some infants within a risk category show atypical cry sounds, others do not. As indicated briefly above, individual differences in measures of infant crying are associated with individual differences in other measures of neurobehavioral function and predictive of individual differences in infant developmental outcome [11,12]. In this way, the analysis of infant crying may be able to identify those infants within a risk category who have actually suffered some form of neurobehavioral insult. Sensitivity to individual differences in neurobehavioral integrity and developmental outcome is evidenced in a recent study of infants at risk for autism spectrum disorder (ASD). Infants were classified as being at risk for ASD based on the presence of the disorder in an older sibling [13]. First, as part of the typical paradigm in which the cries of a group of infants at risk were analyzed, infants at risk for ASD had pain-related cries at six months of age with a higher and more variable pitch than low-risk infants. Second, within the risk group, infants who at 36 months were classified with ASD had among the highest frequency cry sounds that were also more poorly phonated than those of infants not diagnosed with ASD. These results suggest that the analysis of infant crying may not only differentiate a group of infants at risk for developing ASD, but also differentiate which infants within the risk group are at greatest risk.

TRANSLATIONAL ANALYSES OF INFANT CRYING

Perhaps the next stage in infant cry analysis will be its benefit in translational analyses of the vocalizations of different species. The relatively long history of research into the analysis of infant crying has resulted in the development of a rich set of measures and methods. Of course, a limitation of human infant research is the lack of control over correlated variables that may also contribute to variations in infant cry sounds and, thus, confound results. Determining the effects of prenatal cocaine exposure on human infants, for example, is limited by the associated high-risk social environment and maternal polydrug-use. In contrast, studies of the vocalizations of other mammalian species can be conducted under conditions of better experimental control over dosage and environmental and genetic variables. The comparative literature, however, has historically developed fewer measures of vocalizations that are sensitive to nervous system function. The development of translational measures by which the cry sounds of human infants and distress vocalizations of other mammalian species can be directly compared may provide the opportunity for findings from the human and comparative literatures to inform one another. As such, the beginnings of an enhanced taxonomy and novel set of measures has recently been developed for comparison between human infant cry vocalizations and rat pup ultrasonic vocalizations (USVs) [14]. Preliminary research using this taxonomy has found corresponding measures of the acoustic structure of human infant cries and rat pup USVs to be sensitive to the adverse effects of prenatal cocaine exposure. With cutting edge work on the development of ASD in both human [13] and mouse vocalizations [15], future work may benefit from further development of translational measures of mouse and other mammalian vocalizations, as well.

CONCLUSION

Spectrum analysis of human infant crying provides a sensitive measure of the integrity of neurobehavioral organization. Measures of the fundamental frequency and temporal morphology of the cry sound not only differentiate infant groups, but may also help identify which infants within a risk sample have suffered neurobehavioral insult. The ability of these analyses to detect neurobehavioral insult in the absence of other abnormal signs has strong implications for early diagnosis and detection of the adverse effects of potential prenatal teratogens. Future translational comparisons of spectral features of cry sounds may benefit the development of methods for early detection of developmental disorders.

REFERENCES

[1] P.S. Zeskind, “Infant crying and the synchrony of arousal,” in *Evolution of Emotional Communication: From Sounds in Nonhuman Mammals to Speech and Music in Man*, in E. Altenmuller, S. Schmidt and E. Zimmerman, Eds, New York: Oxford University Press, 2013, pp. 155-174.

[2] O. Wasz-Hockert, J. Lind, V. Vuorenkoski, T. Partanen, and E. Valanne. T*he Infant Cry: A Spectrographic and Auditory Analysis*. London: Heinemann, 1968.

[3] P.S. Zeskind and B.M. Lester, “Analysis of infant crying”, in *Biobehavioral Assessment of the Infant,* L. Singer and P.S. Zeskind, Ed, New York: Guilford, 2001, pp.149-166.

[4] L. LaGasse, A. Rebecca Neal and B.M. Lester, “Assessment of infant cry: Acoustic analysis and parental perception”, *Mental Retard and Devel Dis Res Reviews*, vol 11, pp. 83-93, 2005.

[5] H.L. Golub, “A physioacoustic model of the infant cry”, in *Infant crying: Theoretical and research perspectives,* B.M. Lester and C.F.Z. Boukydis, Eds. New York: Plenum, 1989, pp. 59-82

[6] B.M. Lester, “A biosocial model of infant crying”, in *Advances in infant research*, L. Lipsitt, and C. Rovee-Collier C, Eds. Norwood, NY: Ablex, 1984, pp. 167–212.

[7] P.S. Zeskind, T.R. Marshall and D. M. Goff, “Cry threshold predicts regulatory disorder in newborn infants”, *J. Ped. Psychology*, 21, 1996, pp. 803-819.

[8] P. Zeskind, “Production and spectral analysis of neonatal crying and its relations to other biobehavioral systems in the infant at risk,” in *Infants born at risk: Physiological, perceptual, and cognitive processes,* T. Field and A. Sostek, Eds. New York: Grune & Stratton, 1983, pp. 23-44.

[9] P.S. Zeskind and B.M. Lester, “Acoustisc features and auditory perceptions of the cries of newborns with prenatal and perinatal complications”, *Child Dev*el, 49, pp. 580 – 589, 1978.

[10] P.S. Zeskind, C.D. Coles, K.A. Platzman, and P. Schuetze, “Cry analysis detects subclinical effects of prenatal alcohol exposure in newborn infants”, *Inf Beh and Devel*, 19, pp. 497-500, 1996.

[11] L. Huntington, S.L. Hans and P.S. Zeskind, “The relations among cry characteristics, demographic variables and developmental test scores in infants prenatally exposed to methadone”, *Inf Beh and Devel*, 13, 533-538, 1990.

[12] B.M. Lester, “Prediction of developmental outcome from acoustic cry analysis in term and preterm infants”, *Pediatrics*, 80, pp. 529 –534. 1987.

[13] S.J. Sheinkopf, J.M. Iverson, M.L. Rinaldi and B.M. Lester, “Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder”, *Autism Research*, 5, 331-339, 2012.

[14] P.S. Zeskind, M.S. McMurray, K.A. Garber, J.M. Neuspiel, E.T. Cox, K.M. Grewen, L.C. Mayes and J.J. Johns, “Development of translational methods in spectral analysis of human infant crying and rat pup ultrasonic vocalizations for early neurobehavioral assessment”, *Frontiers in Psychiatry/ Child and Neurodevelopmental Psychiatry*, 2, pp. 1-16, 2011.

[15] M.L. Scattoni, L. Ricerri and J.N. Crawley, “Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters”, *Genes, Brain and Behavior*, 10, pp. 44-56, 2011.

ON THE APPLICATION OF GENETIC SELECTION OF A CUSTOMIZED FUZZY MODEL FOR THE CLASSIFICATION OF INFANT CRY PATTERNS

Alejandro Rosales-Perez[1], Carlos A. Reyes-Garcia[1], Jesus A. Gonzalez[1] and Orion F. Reyes Galaviz[2]

[1]Biosignals Processing and Medical Computing Laboratory, Computer Science Department
Instituto Nacional de Astrofisica Optica y Electronica (INAOE), Tonantzintla, Puebla, Mexico,
{arosales,kargaxxi,jagonzalez}@inaoep.mx
[2] Electrical and Computer Engineering, University of Alberta, Edmonton, Canada, orionfrg@yahoo.com;

Abstract: **Infant crying is an innate communication mode, driven by the Central Nervous System. Although more limited, it is similar to adult's speech. The crying wave carries useful information, as to determine the physical and psychological state of the baby, as well as to detect possible physical pathologies from very early stages of life. In this work, the study of the infant crying is approached through an automatic infant cry recognition process. For this purpose, we first process the infant cries in order to extract the acoustic features, with which vectors of each wave are formed, which, in place, are treated as the sample patterns. Next, we propose to use Genetic Selection of a Fuzzy Model (GSFM) for the classification of infant cry patterns. GSFM selects a combination of feature selection methods, type of fuzzy processing, learning algorithm, and its associated parameters that best fit to the data. The implementation stage as well as some experiments and some results are shown, which obtained very high recognition accuracy.**

***Keywords*: Feature extraction, Pattern recognition, Infant Cry Classification, Model Selection, Genetic Algorithms**

I. INTRODUCTION

Infant crying is at birth the only communication means of babies, which, while very restricted, takes the roll of adult's speech. Through crying, the baby shows his/her physical and psychological state. Several studies have been performed in order to distinguish between different kinds of cries. The crying wave carries useful information, as to detect possible physical pathologies from very early stages of life. Those studies bring the opportunity of helping babies by understanding their needs or by detecting specific diseases, in which case the appropriate treatment can be provided and future complications prevented. For example, world-widestatistics indicate that from one thousand new born, 1 or 2 present deep hearing loss or severe hearing loss. Nevertheless not all of them receive diagnosis and opportune treatment. This fact generates a very serious problem, because the beginning of an opportune treatment is delayed. The earlier the deafness diagnosis the better guaranteed the possibilities of rehabilitation and acquisition of the language. In cases like those the application of non-invasive tools, like infant cry analysis, to produce early diagnosis could help to provide the needed treatment.

The first works with infant cry were initiated by Wasz-Hockert since the beginnings of the 60s [5]. In 1964 the research group of Wasz-Hockert showed that the four basic types of cry can be identified by listening: pain, hunger, pleasure and birth [5]. Since then, many other studies related to this line of research and to automatize the recognition of cries have been reported. Sergio D. Cano carried out and directed several works devoted to the extraction and automatic classification of acoustic characteristics of infant cry. In one of those studies, in 1999 Cano presented a work in which he demonstrates the utility of the Kohonen´s Self-Organizing Maps in the classification of Infant Cry Units [6]. In [7] a radial basis function (RBF) network is implemented for infant cry classification in order to find out relevant aspects concerned with the presence of Central Nervous System (CNS) diseases. First, an intelligent searching algorithm combined with a fast non-linear classification procedure is implemented, establishing the cry parameters which better match the physiological status previously defined for the six control groups used as input data. Finally the optimal acoustic parameter set is chosen in order to implement a new non-linear classifier built on a radial basis function network, an ANN-based procedure which classifies the cry units into two categories, normal or abnormal class, as the ones shown in (Fig. 1) and (Fig. 2).

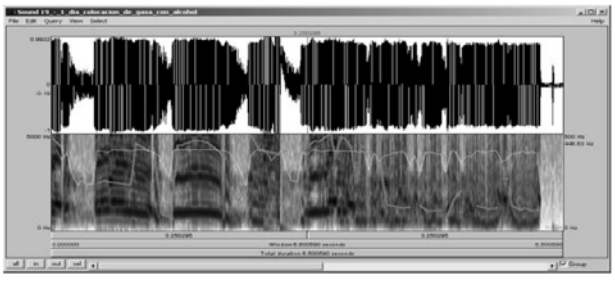


Figure 1: Wave form and Spectrogram of a Normal Cry

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

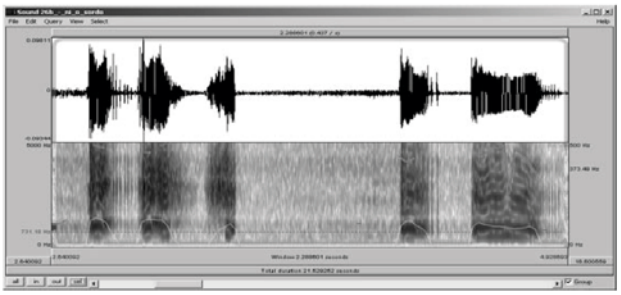


Figure 2: Wave form and Spectrogram of an Abnormal Cry

The infant cry automatic classification process is, in general, a pattern recognition problem, similar to Automatic Speech Recognition (ASR). The goal is to take the wave from the infant's cry as the input pattern, and at the end obtain the kind of cry or pathology detected on the baby [8], [9]. Generally, the process of Automatic Cry Recognition is performed in two steps (Fig. 3). The first step is known as signal processing, or feature extraction, whereas the second is known as pattern classification. In the acoustical analysis phase, the cry signal is first normalized and cleaned, and then it is analyzed to extract the most important characteristics in function of time. The cleaning of the signal is applied in order to eliminate irrelevant and undesirable information, like background noise, channel distortion, and particular characteristics of the signal. Some of the more used techniques for the processing of the signals are those to extract: linear prediction coefficients, cepstral coefficients, pitch, intensity, spectral analysis, and Mel´s filter bank. The set of obtained characteristics is represented by a vector, which, for the process purposes, represents a pattern. The set of all vectors is then used to train the classifier. Later on, a set of unknown feature vectors is compared with the knowledge that the computer has to measure the classification output efficiency.

For our purposes, the study of the infant crying is performed through an automatic infant cry recognition process. In this work, besides applying some of the above mentioned acoustic extraction techniques, and forming the set of pattern vectors, we propose the use of a Genetic Selection method to generate a customized Fuzzy Model (GSFM) for the classification of infant cry. GSFM selects; the best combination of feature selection methods, the more adequate type of fuzzy processing, an appropriate learning algorithm, and its associated parameters that best fit the data. The viability of the proposed technique in the infant cry classification task is supported by the results obtained through experiments. In this paper we show the implementation stage as well as experiments and some results, in which up to 99.42% in recognition accuracy was obtained. Although the results were obtained for identifying cause of crying and differentiate between normal and pathological cry, it is our hypothesis that having available the right sample data base for training, our proposed method will perform well in the early identification of other infant disorders as those belonging to the Autism Spectrum Disorders.

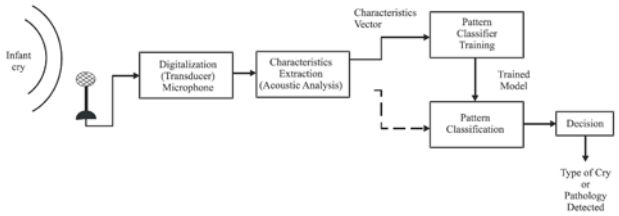


Figure 3. Infant´s Cry Automatic Recognition Process

II. METHODOS

Acoustic Processin:g: The acoustic analysis implies the application and selection of filter techniques, feature extraction, signal segmentation, and normalization. With the application of these techniques we try to describe the signal in terms of its fundamental components. One cry signal is complex and codifies more information than the one needed to be analyzed and processed in real time applications. For this reason, in our cry recognition system we use a feature extraction function as a first plane processor (Fig. 4). Its input is a cry signal, and its output is a vector of features that characterizes key elements of the cry's sound wave. We have been experimenting with diverse types of acoustic features, emphasizing by their utility Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC). [1].

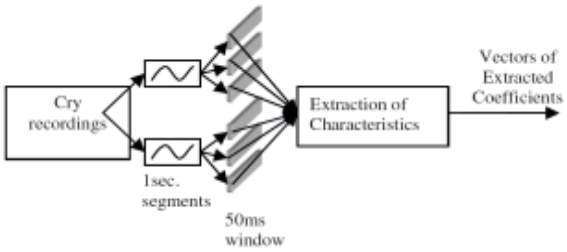


Figure 4. Feature Extraction Process

Linear Predictive Coefficients: Linear Predictive Coding (LPC) is one of the most powerful techniques used for speech analysis. It provides extremely accurate estimates of speech parameters, and is relatively efficient for computation. Based on these reasons, for some experiments, LPC are used to represent the crying signals. Linear Prediction is a mathematical operation where future values of a digital signal are estimated as a linear function of previous samples. In digital signal processing, linear prediction is often called linear predictive coding (LPC) and can thus be viewed as a subset of filter theory [1].

Mel Frequency Cepstral Coefficient: Are perceptual characteristics that can be obtained like filtered signals through different frequency scales. The Mel spectrum operates on the basis of selective weighting of the frequencies in the power spectrum [8]. High order frequencies are weighted on a logarithmic scale whereas lower order frequencies are weighted on a linear scale. This technique pretends to simulate the properties the ear has as a filter, which is more sensitive to some frequencies than to others [9]. The set of values for n features may be represented by a vector in an *n*-dimensional space. Each vector represents a pattern which is expected to contain distinguishing MFCC features as shown in Fig. 5 and Fig. 6 [9].

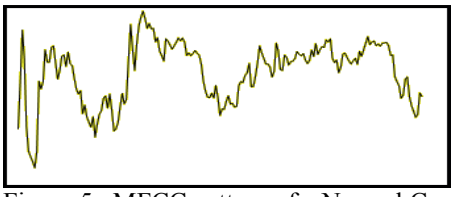


Figure 5. MFCC pattern of a Normal Cry

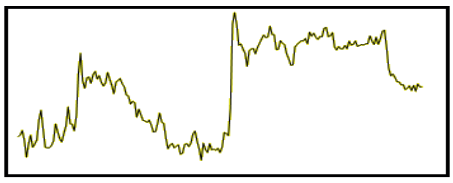


Figure 6: MFCC pattern of a Hipo-acoustical Cry

Pattern Recognition Techniques: Several pattern recognition techniques have been used for the classification task. Among these techniques, artificial neural networks have been one of the most widely used [7,8,9,10].With the same purpose, several hybrid approaches that combine fuzzy logic with neural networks [2,4,9], fuzzy logic with support vector machines [11] or evolutionary strategies with neural networks [9] have also been explored. Most of these works have reported favorable results in infant cry recognition. Alternative approaches that combine genetic algorithms with fuzzy logic and neural networks have been proposed [2,4]. Nonetheless, most of the works only determine the parameters for a specific learning algorithm. In this work we propose to explore the use of the Genetic Selection of a Fuzzy Model (GSFM) approach for infant cry classification. GSFM was recently proposed and it was applied to acute leukemia subtypes classification. A genetic algorithm is used in GSFM for selecting the right combination of a feature selection method, the type of fuzzy processing, a learning algorithm, and their associated parameters that better fit to a data set.

Data Set Description: For the present experiments we worked with samples of infant cries. The infant cries were collected by recordings done directly by medical doctors, the cry samples were carefully labeled at the time of the recording with references like infant age and the reason for the cry. Then each signal wave was divided in segments of 1 second, each segment represents a sample. Next, acoustic features were obtained by means of Frequencies in the Mel scale (MFCC), by the use of the freeware program Praat v4.0.8. Every 1 second sample is divided in frames of 50-milliseconds and from each frame we extract 16 coefficients, this procedure generates vectors whit 304 coefficients by sample. The infant cry corpus has 340 samples of cries of asphyxia, 192 for pain, 350 for hunger, 879 cries of babies who are deaf and 157 of normal cries. Pain and hunger cries come from normal babies, so they are also part of the normal cries collection. Table 1 shows the di_erent data sets and the number of samples of each case. These data sets were used in our experiments.

Tabla 1. Description of infant cry data sets

Data set	No. Samples	Samples by class
Asphyxia vs Normal and Hungry	847	Asphyxia: 340 Normal and Hungry: 507
Deaf vs Normal and Hungry	1386	Deaf: 879 Normal and Hungry: 507
Hungry vs Pain	542	Hungry:350 Pain:192

Genetic Selection of a Fuzzy Model: The process of the construction of a fuzzy model is shown in Figure 7. A labeled data set is the input. Given that each sample is described by a set of N features, and that N is usually large, the first step is to reduce the dimensionality of the data set. This task is done by applying a feature selection method. Then, the subset of selected features is converted into fuzzy values, which is the fuzzifying step. Next, the parameters of fuzzy membership are fitted to reduce the overlapping degree. Finally, with the fuzzy features a fuzzy classifier is built. Given a pool of feature selection methods, fuzzy processing and learning algorithms, GSFM [2] selects the combination of them that minimizes the error.

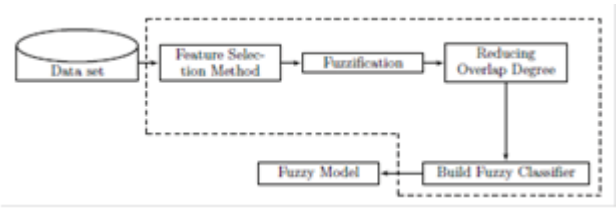


Figure 7: Process for building a model

III. RESULTS

We performed several experiments considering binary classification. First, we considered the binary problems

identifying between asphyxia and normal3 cries, deaf and normal, and finally hungry and pain cries. For each experiment, we used GSFM to determine the best model. The evaluation was done using 10 fold cross validation. This technique divides the data set into 10 disjoint subsets and in each fold a subset is left apart for testing and the remaining subsets for training. As evaluation metrics we used accuracy (ACC), true positive rate (TPR), true negative rate (TNR) and area under the ROC4 curve (AUC) [3]. Table 2 shows the best obtained results in our experiments and the reported by Rosales-Perez et al. [4].

Table 2: Percentual classification results for each experiment. Results are the average of using 10 fold cross validation. Reported results by Rosales-Perez et al. [4] are also shown. The best result is shown in bold font for each case:

ID	Data Set	Accuracy		TPR		TNR		AUC	
		GSFM	[17]	GSFM	[17]	GSFM	[17]	GSFM	[17]
1	Asphyxia vs Normal	**90.68**	88.67	85.29	**90.00**	**94.29**	87.78	**95.79**	92.85
2	Deaf vs Normal	**99.42**	97.55	**100.00**	98.75	**98.42**	95.47	**100.00**	99.75
3	Hungry vs Pain	**97.96**	96.03	**99.43**	95.59	95.26	**96.67**	**98.89**	98.35

Table 3 describes the obtained models for each case. For each model, the feature selection method (FSM), type of membership function (TMF), number of linguistic properties (NLP), the learning algorithm (LA), and its associated parameters are shown.

Table 3: Selected models for infant cry data sets using GSFM

ID	FSM	TMF	NLP	LA	Parameters
1	Correlation	Trapezoid	3	FKNN	nn = 7 sm = correlation
2	InfoGain	Bell	7	FDT	cv = 0.87
3	InfoGain	Gauss	3	FKNN	nn = 5 sm = chord

IV CONCLUSION

In this work we show a method that can be used to support opportune medical diagnosis on babies at early stages of life, as a noninvasive technique. We described the design and implementation of GSFM which allows selecting an adequate model to differentiate between different types of cries with precise results. Among the main advantages of the adopted approach is the fact that our system releases the user from having to determine the right combination of model components. Our experimental results show that our approach outperforms results reported in the literature from methods that only consider one learning algorithm.

REFERENCES

[1]. Lederman, D.: Automatic classification of infants cry, Ben Gorion University, M.Sc. Thesis P pp. 1-11 (2002).

[2]. Rosales-Perez, A., Reyes-Garcia, C., Gomez-Gil, P., Gonzalez, J., Altamirano, L.: Genetic selection of fuzzy model for acute leukemia classification. In: Batyrshin, I. Sidorov, G. (eds.) Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 7094, pp. 537{548. Springer Berlin / Heidelberg (2011)

[3]. Beck, J., Shultz, E., et al.: The use of relative operating characteristic (roc) curves in test performance evaluation. Archives of pathology & laboratory medicine 110(1), 13 (1986)

[4]. Rosales-Perez, A., Reyes-Garcia, C., Gomez-Gil, P.: Genetic fuzzy relational neural network for infant cry classification. In: Martinez-Trinidad, J., Carrasco-Ochoa, J., Ben-Youssef Brants, C., Hancock, E. (eds.) Pattern Recognition, Lecture Notes in Computer Science, vol. 6718, pp. 288-296. Springer Berlin / Heidelberg (2011)

[5].Wasz-Höckert, O., Lind, J., Vuorenkoski, V., Partanen, T., Valanne, E. (1968) "The infant cry a spectrographic and auditory analysis". Clinics in Devel. Medicine. 29.

[6]. Sergio D. Cano, Daniel I. Escobedo y Eddy Coello, El Uso de los Mapas Auto-Organizados de Kohonen en la Clasificación de Unidades de Llanto Infantil, Grupo de Procesamiento de Voz, 1er Taller AIRENE, Universidad Catolica del Norte, Chile, 1999, pp 24-29.

[7]. Sergio D. Cano Ortiz, Daniel I. Escobedo Beceiro, Taco Ekkel2, "A Radial Basis Function Network Oriented for Infant Cry Classification", Proc. Of 9th Iberoamerican Congress on Pattern Recognition, Puebla, Mexico, 2004.

[8]. Orozco, GJ, Reyes CA, "Mel-Frequency Cepstrum Coefficients Extraction from Infant Cry for Classification of Normal and Pathological Cry with Feed-Forward Neural Networks," Proc. International Joint Conference on Neural Networks. Portland, Oregon, USA, pp. 3140-3145, 2003.

[9]. Orion Fausto Reyes-Galaviz, Sergio Daniel Cano-Ortiz, Carlos Alberto Reyes-García. "Evolutionary-Neural System to Classify Infant Cry Units for Pathologies Identification in Recently Born Babies". Proceedings of the Special Session MICAI 2008, Pg. 330-335. Eds. Alexander Gelbukh & Eduardo Morales. IEEE Computer Society. ISBN: 978-0-7695-3441-1.

[10] Ekkel, T. "Neural Network-Based Classification of Cries from Infants Suffering from Hypoxia-Related CNS Damage", Master Thesis. University of Twente, 2002. The Netherlands.

[11]. Barajas-Montiel, S. and Reyes-Garcia, C., "Fuzzy Support Vector Machines for Automatic Infant Cry Recognition", in Intelligent Computing in Signal Processing and Pattern Recognition, LNCIS, ed. Huang, De-Shuang et.al., Springer Berlin / Heidelberg, isbn: 978-3-540-37257-8, pp. 876-881, vol: 345, 2006.

EARLY DIAGNOSIS IN AUTISM SPECTRUM DISORDERS: SUGGESTIONS FROM ANIMAL MODELS

S. Orlandi[1,2], C. Manfredi[1], A. Guzzetta[3] and M.L. Scattoni[4]

[1] Dept. of Information Engineering, University of Firenze, Firenze, Italy, silvia.orlandi@unifi.it
[2] Dept. Electronic and Information Engineering "Guglielmo Marconi", University of Bologna, Bologna, Italy
[3] Dept. of Developmental Neuroscience, Stella Maris Scientific Institute, Pisa, Italy, a.guzzetta@inpe.unipi.it
[4] Dept. of Cell Biology and Neurosciences, Istituto Superiore di Sanità, Roma, Italy, marialuisa.scattoni@iss.it

Abstract: **Autism Spectrum Disorders (ASDs) have an estimated prevalence rate of 1/88 suggesting that ASD represents a significant public health problem. Although causes of ASD are still unknown, the strongest evidence appears to be genetic. In this regard, studies on mouse models bearing mutations identified in ASD candidate genes could be extremely helpful to identify early behavioral traits that cannot be studied in human infants since at present ASD cannot be diagnosed reliably before two years of age.**
Our team recently characterized early phases of neurobehavioral development in several animal models of autism detecting quantitative and qualitative abnormalities in their vocal and motor repertoires. These results suggested us to assess early vocal and motor repertoires in high-risk infants (siblings of ASD children) and in control infants. An approach linking experimental findings obtained in animals to human infants might be successful in order to identify early markers of ASD. Aim of our project is to develop an automatic system to record newborn cry and movements during the first six months of life with a specific protocol. Our acoustic analysis is focused on fundamental frequency (F0), number of cry-episodes and F0 shapes. Preliminary results showed some differences concerning fundamental frequency between normal and high-risk cases. Moreover high-risk subjects emitted a lower number of cry-episodes than control subjects. Following the mouse methodological approach, we are also investigating melody in high risk infants searching for early indicators of ASD.

Keywords : **Infant's cry, Autism Spectrum Disorders, Autism Animal model, acoustic analysis**

I. INTRODUCTION

Autism Spectrum Disorders (ASDs) are a group of complex disorders of brain development characterized at different levels. by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors. They include autistic disorder, childhood disintegrative disorder, pervasive developmental disorders not otherwise specified and Asperger syndrome [1].

In most cases, ASDs signs appear in the first year of life, when it is not yet possible to carry out a reliable diagnosis. This is because the diagnostic tests currently used and considered gold-standard for ASD are based on the identification of behavioral symptoms which are more evident after the 24th month of life. Etiology of ASDs is still unknown, although several studies identified genes mutations, copy number variants and abnormal neuro-transmission in specific brain areas such as limbic system, amygdala, cerebellum and sub cortical areas that mediate motor control [2].

Preclinical research on ASD currently represents an emerging field in translational neuroscience. In fact, animal models of ASD can provide translational tools to identify neurochemical markers and behavioral patterns that cannot be studied in human infants since at present ASD cannot be diagnosed reliably before two years of age. In this regard, mouse models bearing mutations identified in ASD candidate genes and that exhibit a clear autistic-like phenotype would be most promising. An approach linking experimental findings obtained in animals to human infants might be successful in order to identify as early as possible vulnerable behavioral patterns associated with alteration in selected genetic and biochemical markers.

Our team recently characterized early phases of neurobehavioral development in several animal models of autism [3-5]. In particular, to address communication deficits, we investigated ultrasonic vocalizations detecting qualitative abnormalities in their vocal repertoire that may resemble the atypical vocalizations and the monotonic tone found in some autistic infants [6,7]. These data collected on animal models suggested us to assess early vocal and motor repertoire in infants siblings of ASD children and in a population of about 200 healthy infants. Previous studies showed that some age-specific motor and vocal traits are altered in ASD children [8-10]. In most of these studies general movements (GMs) and infant crying analyses were assessed by home-videos of children's first birthday party, but this method presents some limits. First of all, existing data refer to the assessment of a restricted number of infants (e.g. 10-12) [10]. Moreover, the method may be not reliably standardized because of methodological differences in the quality of recordings and in the setup of the observations.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Finding links between GMs and cry analysis is desirable and of great relevance since they both reflect the development and the integrity of the central nervous system and can be exploited for early clinical diagnosis enabling a more effective intervention of several pathologies since they are easy to perform, cheap and marker-less. Although perceptual analysis of movements and crying is carried out in specialized clinics, the lack of automatic tools requires to clinicians a great deal of time with a large margin of error. It is therefore important to develop semi-automatic qualitative methods to support the clinical diagnosis allowing for feature extraction and perceptual analysis with marker-less techniques.

This paper presents a new tool for the management of patient data, tests and reports, acquisition of audio and video data, their editing and analysis to support clinicians with perceptual diagnosis.

II. METHODS

This work is linked to the Italian grant project "Young Researcher 2008: Non-invasive tools for early detection of autism spectrum disorders", aiming to detect early markers of Autism Spectrum Disorders (ASDs) through the study of infant crying and GMs during the first six months of infant's life. This project aims to identify normative ranges for acoustical and motor parameters in a population of about 200 healthy newborn/infants, both male and female (control group). The control group will be compared with 15-20 "high-risk" newborn/infants, i.e siblings of children already diagnosed with ASD [11, 12].

Infants were audio and video recorded at home five times during the first six months of life according to a specific protocol [12]: at 10 days, 6, 12, 18 and 24 weeks of life. The protocol also includes clinical assessment performed using a set of questionnaires (Italian Questionnaire of Temperament; Bayley Scales of Infant Development; the first child vocabulary, MacArthur - Bates Communicative Development Inventory and the Modified Checklist for Autism in Toddlers (*M-CHAT*)). Informed consent was obtained from parents. The protocol was approved by the local ethical committee (Istituto Superiore di Sanità, Roma, Stella Maris Hospital, Pisa and Bambin Gesù Hospital, Roma, Italy).

Cry analysis was carried out by the estimation of acoustic parameters such as fundamental frequency (F0), intensity, resonance frequencies of the vocal tract and length of each cry episode [13].

GMs analysis was performed with a perceptual technique that is often used by clinicians to diagnose motors problems associated with impairment of the central nervous system such as the early diagnosis of cerebral palsy [14].

A. ARAD System

According to the project [11, 12] and in co-operation with ISER tech srl (Prato, Italy), we developed a new tool for audio/video acquisition and analysis for contact-less diagnosis, in particular in neonatology area, named ARAD (Acquisition, Reporting and Analysis for Diagnosis), shown in Fig. 1. ARAD is designed for home use to minimize the discomfort for the involved subjects and the impact of the external environment on children habits. Hence, the basic requirement is the ease of transport and assembly of the system. It includes a laptop connected to a high-speed USB webcam (Logitech HD pro webcam C910) able to provide a 1280x1024 pixel video stream, a sound board (Tascam US-144-MK2) and a professional microphone (Shure SM58).

ARAD allows the management of personal data and medical history of the patient, data acquisition, personalized test editing, audio/video editing and reporting in a single software tool. The storage of data patient is managed through a centralized database structured to guarantee privacy and personal data protection. Personal data management is integrated with the centralized database that allows the user viewing reports of patients.

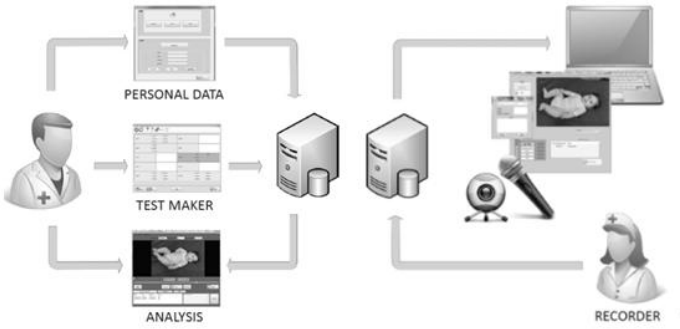


Fig. 1-ARAD System

The system consists of four parts:

a) Personal Data Management System

an integrated management system to share data and medical history with dedicated interface and for patient's clinical storyboard. It allows importing and exporting signals, images and tests from external sources.

b) Test maker

an interactive environment for the creation and run of specific clinical tests with modular customizable structure. It is possible to insert multimedia contents, images, audio / video screenshots and monitor the results.

c) Recorder

an integrated environment for capturing multichannel audio/video with the possibility to enter notes and contextualized information through the wizard and flexible setup of the system.

d) Analyzer

tool devoted to the perceptual analysis and audio/video editing. Provides the ability to easily cut /copy/evaluate sequences of interest and enter clinical assessments without the need to resort to the use of other software.

B. Analysis tool

ARAD is equipped with a devoted tool for the analysis of audio and video recordings. It provides objective cry parameters and simplifies GMs perceptual analysis. Specifically:

a) Cry analysis

On the selected crying frames time-frequency analysis is carried out according to the methods used in [11, 13, 15]. Extracted parameters are: fundamental frequency of cry excerpts (F0), vocal tract resonance frequencies, number of cry-episodes, vocal percentage, number and length of voice breaks.The recorded sound is band-pass filtered by a Butterworth filter of order 5 and a cut-off frequency of 50–1000 Hz.

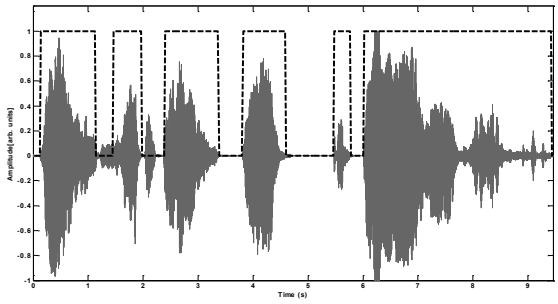


Fig. 2-An example of V/UV detection

Detection of crying episodes (voiced frames lasting at least 150ms) is performed using a robust newly developed Voiced/Unvoiced (V/UV) detection procedure [15]. An accurate detection of starting and ending points of the events allows avoiding incorrect splitting of a single event into several intervals that may occur in the case of irregular and quasi-stationary signals as newborn infant cries are, disregarding noise and silence while retaining those irregularities that may be a diagnostic index of possible pathologies or malfunctioning. An example of V/UV selection is shown in Fig. 2. On selected voiced frames F0 is estimated with a two-step procedure already found successful when analyzing newborn cries [14].

In each cry episode, the fundamental frequency presents a well-defined trend (melody). Based on mouse data [3, 6]) and previous human studies [16], four typical patterns are detectable in the infant crying: symmetrical (frequency rising and falling around a central peak), rising (frequency peak appears near the end of the episode), falling (frequency peak appears at the beginning), and plateau (with an almost constant frequency). The automatic extraction of these patterns is under study.

b) General movements analysis

On selected video frames that contains the GMs the clinician can built, compile and export specific tests. It is also possible to enter number of motion sequences and of abnormal movements, as well as the duration and the time instant at which they occurred.

. III. RESULTS

At present we have collected data from 75 children, namely 66 control (CC) and 9 high-risk cases (HRC). Acoustic analysis was performed only on 3 high-risk cases that were found positive to the GMs analysis qualitatively carried out by clinicians. We analyzed 30 seconds of cry for each infant at 10 days, 6, 12, 18and 24 weeks of life (10000 CC and 474 HRC hunger cry-episodes). Preliminary results show that, in high-risk subjects:

1. the number of cry-episodes is 37% lower than in control subjects;
2. the number of vocalic zones is 20% lower than in control subjects;
3. F0 is about 50 Hz lower than in control subjects and shows less variability.

The ontogenetic profile of F0 in the first 6 months of life is shown in fig. 3.

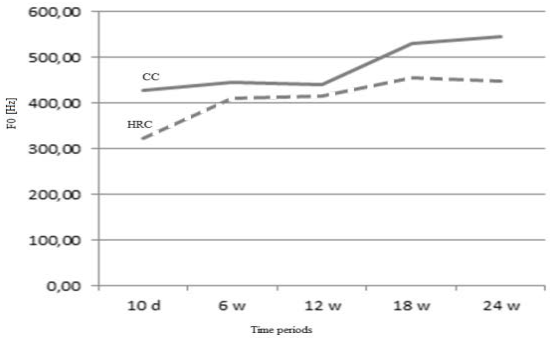
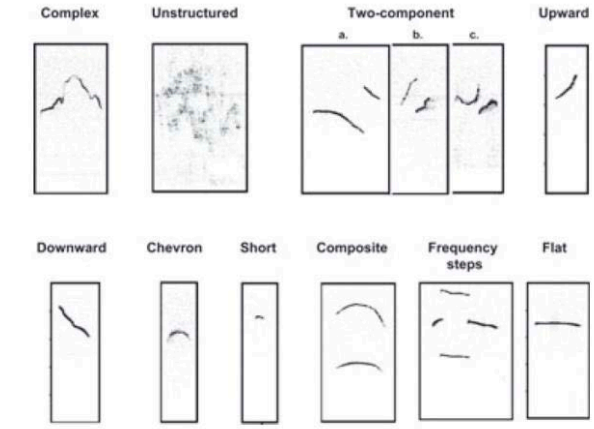


Fig. 3-F0 in first 6 months of infants'life

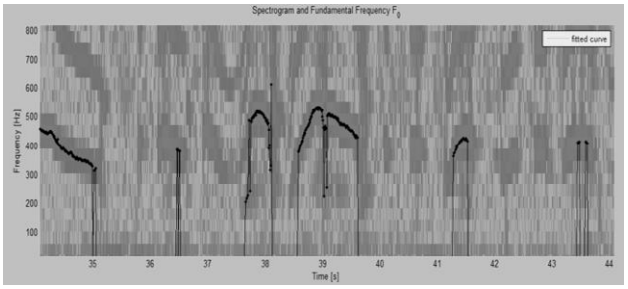
IV. DISCUSSION

Mice and humans both vocalize through the vocal folds and the ultrasounds and crying emitted by their pups/infants are structurally similar and have the same communication value. Thus future studies will be devoted to correlate infant crying of high-risk infants with the ultrasonic vocalizations emitted by animal models of ASD in terms of sound patterns. Our studies on animal models of ASD revealed that pups have a restricted vocal repertoire when tested at postnatal day 8 (10 waveform patterns and an example of typical newborn cry are

illustrated in Fig. 4). We are also investigating melody in high risk infants searching for early indicators of ASD.



(a)



(b)

Fig. 4, (a) typical sonograms of ultrasonic vocalizations [6] and (b) an example of baby's cry melody at 10 days

ACKNOWLEDGMENT

Supported by the Italian Ministry of Health Grant (GR3), Young Researcher 2008, "Non-invasive tools for early detection of Autism Spectrum Disorders". The authors thank ISER Tech srl for its contribution to this work.

REFERENCES

[1] Diagnostic and statistical manual of mental disorders (4th ed., text rev.). Washington, DC: APA, 2000

[2] Bauman ML, KemperTL, Neuroanatomic observations of the brain in autism: a review and future directions. Int J Dev Neurosci, 2005, 23(2,3) pp 183-187

[3] Romano E, Michetti C, Caruso A, Laviola G, Scattoni ML. Characterization of neonatal vocal and motor repertoire of reelin mutant mice. PLoS One. 2013 May 21;8(5):e64407

[4] Michetti C, Ricceri L, Scattoni ML. Modeling social communication deficits in mouse models of autism. Special Interest Section .Animal Models in Autism, Autism-Open Access, 2012

[5] Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. Genes Brain Behav. 2011 Feb;10(1) pp 44-56

[6] Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in the BTBR T+tf/J mouse model of autism. PLoS One. 2008 Aug 27;3(8):e3067

[7] Scattoni ML, McFarlane HG, Zhodzishsky V, Caldwell HK, Young WS, Ricceri L, Crawley JN. Reduced ultrasonic vocalizations in vasopressin 1b knockout mice. Behav Brain Res. 2008 Mar 5;187(2), pp 371-8

[8] Phagava H, Muratori F, Einspieler C, Maestro S, Apicella F, Guzzetta A, Prechtl HF, Cioni G General movements in infants with autism spectrum disorders. Georgian medical news, 2008, 156, pp. 100-105

[9] Esposito, G., Venuti, P., Developmental changes in the fundamental frequency (f0) of infants' cries: A study of children with Autism Spectrum Disorder, Early Child Development and Care, 2010, 180 (8), pp. 1093-1102

[10] Baranek GT, Autism During Infancy: Retrospective Video Analysis of Sensory-Motor and Social Behaviors at 9–12 Months of Age, J Autism Dev Disord, 1999, 29(3), pp. 213-224

[11] Orlandi S, Manfredi C, Bocchi L, Scattoni ML, Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis, ConfProc IEEE Eng Med BiolSoc, 2012, pp. 2953-2956

[12] Bocchi L, Orlandi S, Manfredi C, Puopolo M, Guzzetta A, Vicari S, Scattoni ML, Early Diagnosis of Autism Spectrum Disorders – Design of the Data Acquisition and Management System, 5th European Conf. of the Int.Fed. Med. Biol.Eng, Budapest, Hungary, IFMBE 2011, Proc.37, pp. 187-190

[13] Manfredi C, Bocchi L, Orlandi S, Donzelli GP, High-resolution cry analysis in preterm newborn infants, Med. Eng. &Phys, 2009, 31(5), pp. 528-532

[14] Einspieler C, Prechtl HFR, Bos AF, Ferrari F, Cioni G, Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. Cambridge University Press, 2004

[15] Orlandi S, Dejonckere PH, J. Schoentgen J, Lebacq J, Rruqja N, Manfredi C, Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring, Biomedical Signal Processing and Control, 2013, 8(6), pp. 799-810

[16] Wermke K, Mende W, "Musical elements in human infants' cries: in the beginning is the melody", Musicae Scientiae, 2009, 13, pp. 151-173

PREMATURE INFANTS' CRY MAINTAINS ABNORMALITIES AT TERM: A SONOSPECTROGRAPHIC STUDY

D. Lenti Boero[1], C. Lenti[2]
[1]Department of Social and Human Sciences University of Valle d'Aosta, Aosta, Italy, d.lentiboero@unida.it
[2] Operative Unit Child Neuropsychiatry AO San Paolo Hospital, University of Milan, Italy, carlo.lenti@unimi.it
of Milan, Milan, Italy

Abstract: **Preterm birth is a risk event for neurological integrity and psychomotor development of children. Spectrographic analysis of cries has been proposed as a diagnostic aid in many pathologies. Presumably, prematurity differentially affects nervous development in relation to the amount of time of extrauterine life before correct conceptional age. The aim of present study is to compare the cries of severe premature (birth age <= 33 weeks) and moderate premature (birth age => 34 weeks) infants recorded at corrected gestational age with normal controls.**
Keywords: **Preterm birth, infant cry, spectrographic analysis.**

I. INTRODUCTION

Premature survival has been incremented with the development of infant health care in the last 20 years, but preterm birth is still a risk event for neurological integrity and psychomotor development of children [1,2,3,4]. Many factors make a preterm baby prone to develop a neurological handicap but the predictive value of clinical and instrumental examination is still matter of debate [5,6]. The spectrographic analysis of the cry has been proposed as a diagnostic aid in many pathologies, and reviews of infant cry studies confirm the hypothesis that pathologies affecting the cry-production system do indeed result in changes in the acoustic characteristics of the cry [7,8,9,10,11,12]. As regards as premature infants, previous studies compared them at earlier conceptional age with at term controls [13], and authors believed that the differences found could have been ascribed to an effect of early neurological maturation [8], but see [14]. Presumably, prematurity differentially affects nervous development in relation to the amount of time of extrauterine life before correct conceptional age. The aim of present study is to investigate, at term, the cry outcome of cries from severe premature (birth age <= 33 weeks) and moderate premature (birth age => 34 weeks) infants.

II. METHODOS

Subjects and data collection: Thirty-three premature infants (18 males and 15 females) were recruited in the nursery of a local hospital in Milan (Italy), fourteen of them were twins, all the infants had normal blood parameters and normal brain ultrasound scan. Inclusion criteria were absence of ventilatory aid and absence of any pathology beyond prematurity. The sample was subdivided into two groups: 24 severe premature (G.A. <= 33 weeks, mean +- S.D = 30.66, +- 1.88; mean weight at birth +- S.D = 1548.12 +- 447.7), and 9 moderate premature (G.A. => 34 weeks, mean +- S.D = 34.33 +- 0.5; mean weight +- S.D = 2051.66 +- 362.83). A control group of 32 infants (19 M; 13 F) was recruited for comparison. All the infants were singularly taken in a quiet room adjacent the nursery, and induced into the awake state [15] by gentle touch on skin surface, cries were induced by manipulation stimuli during neurological examination performed following the criteria of [16] at the same hour of the day, and recorded at the corrected gestational age (between 38 and 40 weeks) on Sony digital audio tape DT-90 by means of DAT sound recorders (Sony TCD D7 and TASCAM DAP1) and of a Sennheiser ME66 unidirectional microphone positioned between two to five cm from the mouth of the crying babies. The entire cries uttered were recorded; in order to sample all the possible variability of each infant cry we examined the first and the last six cry units, and sampled six cry units along the entire cry at regular time intervals.

Sound analysis: High resolution sonograms were produced and measured by means of Raven 1.3 and Praat 5.1.21 on MacBookPro computer.

Sound parameters: 1) qualitative. a) voicing: voiced, voiceless, and partially voiced [11,12,17], b) melodic contours: rising, rising-falling, double rising or double rising-falling, flat, falling falling-rising and eventually vibrato contour, as defined in [18] as fundamental frequency showing a continuous saw-like line. 2) quantitative (measured only on the fundamental frequency of voiced cries). a) time (in MSc): length of wail, silent interval to the next wail, and time for reaching the maximum frequency; b) frequency (in Hz): starting, end, maximum and minimum frequency on the fundamental, dynamic gamma (difference from maximum to minimum frequency in each wail), and peak frequency, defined as the frequency with the maximum energy in the spectrogram.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Statistics: Data were analysed by means of GLIM3 and SPSS.

III. Results

Weight means +-SD were 3231.91+-367.87 gr., 3107.11+-359.39 gr., and 3276.09+-228.02 gr., respectively for severe preterms, moderate preterm and normal controls; the severe preterms did not differ from moderate nor from control as regards as weight at term conceptional age (F =,0.306, d.f. = 1,54, P = 0.582, F = 0.762, df = 1,31, P = 0.389, respectively for controls and moderate preterms). Moderate preterms did not differ from controls (F = 2.957, df = 1,39, P = 0.093).

Voicing. All groups had significantly less voiceless and partially voiced cries than voiced ones (GLIM chi-square = 55.63, df = 1, P < 0.0001 and chi-square = 1043.6, df = 1, P < 0.0001, respectively for both groups of prematures together and normal controls). The severe premature infants at term had a significantly higher percentage of voiceless and partially voiced cries than the normal controls (GLIM, chi-square = 26.81, df = 1, P < 0.0001). The moderate premature infants had a percentage of voiceless and partially voiced cries similar to the normal ones (GLIM, chi-square = 3.57, df = 1, P > 0.05), and significantly lower than the severe prematures (GLIM, chi-square = 3.92, df = 1, P < 0.05).

Melodic contours. Severe prematures had a higher amount of vibrato contours than normal controls (GLIM, chi-square = 29.87, df = 1, P < 0.0001). Moderate prematures had an amount of vibrato contours similar to the normal subjects (GLIM, chi-square = 2.08, df = 1, P > 0.1), and significantly lower than severe prematures (GLIM, chi-square = 6.11, df = 1, P < 0.025).

Time and frequency parameters. Quantitative parameters are shown in tab. 1.

Both mean length of cries and starting mean frequency at term were related with birth weight (F= 10,806, df = 1,63, P = 0.0017 and F = 7.677, df = 1,63, F = 0.0073 respectively for the length of cries and stating frequency).

Time parameters. Severe preterm infants had significantly shorter cries and cry intervals then controls (SPSS univariate F-tests: $F_{1,1217}$ = 87.40, P < 0.000; and $F_{1,1217}$ = 33.08, P < 0.000, respectively for length of cries and intervals between cries). Analogously as the severe subjects, the moderate preterm had significantly shorter cries and cry intervals then controls (SPSS univariate F-tests: $F_{1,1067}$ = 17.97, P < 0.000; and $F_{1,1067}$ = 5.11, P =0.024, respectively for length of cries and intervals between cries). The two preterm groups were not significantly different between them as regards as the latter parameters ($F_{1,330}$ = 1.86, P = 0.173: and $F_{1,330}$ = 1.04, P = 0.30, respectively for length of phonation and intervals between cries), (fig. 1).

Tab. 1. Quantitative parameters of preterm infants.

	severe preterm		moderate preterm		controls	
	mean	+- S.D.	mean	+- S.D.	mean	+- S.D.
lenght of wails	710	554	805	660	1077	611
wail interval	282	314	330	517	406	270
t. for max frequency	252	323	302	512	308	372
starting F0	413	101	390	75	381	80
end F0	385	104	362	69	350	73
maximum F0	551	106	536	70	493	81
minimum F0	342	79	334	73	324	71
dynamics of F0	209	105	202	73	167	120
peak frequency	1412	854	1565	964	1310	656

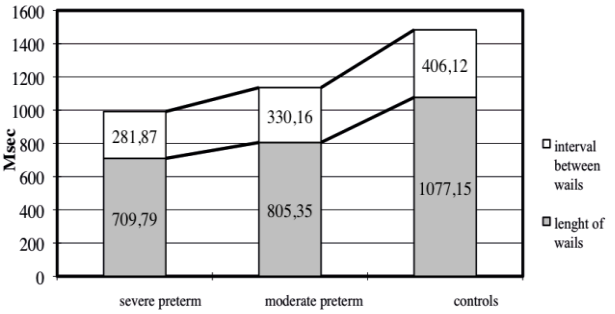


Fig. 1. Time parameters and respiratory cycle in preterm infants and controls

Frequency parameters. Both preterm groups had significantly higher fundamental frequencies than normal controls (SPSS multivariate ANOVA, Pillais trace, $F_{4,1495}$ = 26.22, P< 0.000, and $F_{4,1337}$ = 7.09, P< 0.000, respectively for severe preterm and controls, and moderate preterm and controls). The severe premature group had significantly higher frequencies than the moderate group (SPSS multivariate ANOVA, $F_{4,329}$ = 0.048, P < 0.048). The peak frequency was significantly higher in the two preterm groups than in controls (SPSS univariate ANOVA, $F_{1,1498}$ = 4.46, P = 0.035, and $F_{1,1498}$ = 10.76, P = 0.001, respectively for severe preterm and controls, and moderate preterm and controls), but it was similar in the two preterm groups ($F_{1,326}$ = 1.84, P = 0.176), (mean +- SD = 1412.14 +- 853.60; 1565.42 +- 963.72; 1310.42 +- 655.54, respectively for the severe, the moderate preterm and the normal controls), Fig. 2.

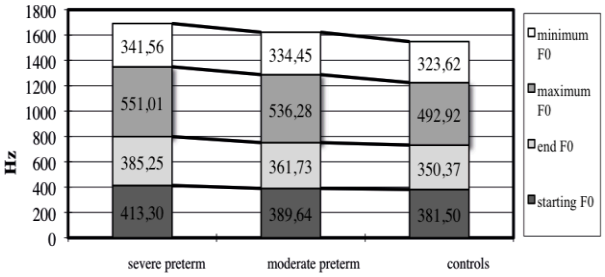


Fig. 2. Mean frequency parameters of preterm infants and controls

IV. Discussion

The aim of present study was to investigate the hypothesis that prematurity might differentially affects nervous development in relation to the amount of time of extrauterine life before correct conceptional age. We believe that this hypothesis was confirmed. Previous cry research performed on pathological infants showed that the cry abnormalities are not pathognomonic of a specific illness, but are univocally dependent from the insult on the nervous motor controls caused by the pathology itself [8,11]. However, as regards as preterm infants, cry comparison with normal subjects was performed at birth age , thus at different gestational ages, and utilising only the first cry units uttered after the stimulus [13,19], and some authors believe that the differences found could have been ascribed to an effect of early neurological maturation [8]. This approach and related conclusions was criticised by [14], who found that the LTAS (long term average spectra) of preterm infants analysed at term conceptional age had an higher fundamental frequency than control infants born at term. In present study we chose a different approach and examined other parameters than the only quantitative ones on the fundamental.

Because the infant cry is a nonlinear phenomenon resulting from the interaction between nervous motor control, vocal system (larynx and vocal tract), and acoustic energy provided by the air flow from the lungs [20,21,22], nonlinearities produced can lead to highly complex and individually variable acoustic output, making any single cry unit a unique acoustic event as regards as shape, intensity variation, frequency dynamics, and harmonics. However, a most common feature of the infant cry of normal infants is the steadiness of the melodic contour showing a continuity of frequency transition from a time window to another according to a clear direction in the spectrogram [23].

The qualitative parameters we took into account mostly examined the above aspects. Though all infants recorded in the first days of life show anomalies in voicing [23], the fact that sever preterm had significantly more voiceless cries than moderate preterm and controls, suggests that in the severe group discrepancies between the amount of acoustic energy from the lungs, and the ability of the phonatory apparatus to cope with it are present [11,12,21,24]. Also, the severe group had significantly more vibrato contours then the other two groups, because vibrato contours are defined in cry literature as fluctuations in the fundamental frequency that do not show any stabilisation toward some constant value [8], this datum suggests that in some way the integrity of the phonatory chain is altered in severe preterm. In other pathological groups, non-steady patterns are found in greater amount even when no other neuroradiological and/or clinical signs are present [8,11]

Quantitative parameters. Time: in human infants the length of each cry unit is determined by both respiratory peripheral and motor control central factors [8]. Both severe and moderate preterm in our study had significantly shorter cries than normal ones, and did not show any difference between them. Because a criterion for inclusion in our study was that absence of intubation and respiratory assistance, results different from other studies [25] might be ascribed to this factor, however, central causes cannot be excluded.

Frequency. Frequency parameters are related to both the body mass, that imply the size of the larynx and the vocal folds, and to the tension of the vocal folds themselves, that modify, within a certain range, the length, and consequently the frequency of the acoustic output. The body-masses of our three samples of infants were not significantly different at corrected G.A., however, the premature infants had significantly higher fundamental frequency parameters, and the regression analysis demonstrated that the lower the body-mass at birth, the higher the fundamental frequency at term. Asphyxiated infants with midbrain MRI diagnosed lesions have significantly lower, and not higher, fundamental frequency [12], thus the higher fundamental frequency of preterm infants can be considered of a functional type. [26] found an impaired cry response to pain stress in cries of preterm infants, our finding could also be related to a higher difficulty for preterm to cope with stress, progressively increasing with earlier premature birth. The higher frequency parameters and shorter times for maximum frequency in severe preterm suggest a lesser ability in controlling the shape of the melodic contours.

V. Conclusion

Our study suggests that the earlier the birth the more alteration might be found in the nervous motor control of early phonation. Because premature infants are prone to delay motor and language development [3] the control of phonatory development might be an important step for the investigation of language impairment in the premature: a blueprint for future research.

Financial support: This study was supported in 1995 and from 2001 to 2003 by a grant of the MURST (Ministry for University and Scientific Research)", by founds from University of Milan given to Carlo Lenti, and by the "Pierfranco and Luisa Mariani Foundation".

REFERENCES

[1] J.J. Volpe, *Neurology of the newborn*, 3rd ed. Philadelphia: WB Saunders, 1995.

[2] A.E. Anderson, S.R. Wildin, and M. Woodside, "*Severity of medical and neurologica complications as a determinant of neurodevelopmental outcome at 6 and 12 months in very low birth weight infants*", J. of Child Neurology, vol. 11, pp. 215-219, 1996.

[3] A. Sansavini, M. Rizzardi, R. Alessandroni and G. Giovanelli, "*The development of Italian low- and very-low- birthweight infants from birth to 5 years: the role of biological and social risks*" Int. J. of Behav. Devel., vol. 19 (3), pp. 533-547, 1996.

[4] C.M. McCarton, I.F. Wallace, M. Divon and H.G. Vaughan, "*Cognitive and neurologic development of the premature, small for gestational age infant through age 6: comparison by birth weight and gestational age*", Pediatrics vol. 98, pp. 1167-1178, 1996.

[5] J.M. Perlman, "*White matter injury in the preterm infant: an importat determination of abnormal neurodevelopmental outcome*", Early Hum. Devel., vol. 53, pp. 99-120, 1998.

[6] A.H. Whitaker, J.F. Feldman and R. Van Rossem, "*Neonatal cranial ultrasound abnormalities in low birth weight infants: relation to cognitive outcomes at six years of age*", Pediatrics vol. 98, pp. 719-729, 1996.

[7] O. Wasz-Hockert, J. Lind, V. Vuorenkoski, T.J. Partanen, E. Valanne, *The infant cry: a spectrographic and auditory analysis* (Spastics Int. Med. Pub.), London: Heinemann, 1982.

[8] M. Koivisto, "Cry analysis in infants with Rh haemolytic disease," *Acta Paed. Scan. Supp.,* vol. 335, pp. 1-73, 1987.

[9] M.J. Corwin, B.M. Lester and H.L. Golub, "*The infant cry: What can it tell to us?*" Current problems in Pediatrics, vol. 26, pp. 325-334, 1996.

[11] D. Lenti Boero, G. Weber, M.C. Vigone, and C. Lenti, "*Crying abnormalities in congenital hypothyroidism: a preliminary spectrographic study*", J. of Child Neurol., vol. 15 (9), pp. 603-608, 2000.

[12] D. Lenti Boero, "*Neurofunctional spectrographic analysis of the cry of brain injured asphyxiated infants: a physioacoustic and clinical study*", in: 6th International Workshop. Models of vocal emissions for biomedical applications. Proceedings. C. Manfredi Ed., Firenze: Firenze University Press, p. 3-6, 2009.

[13] K. Michelsson, "*Cry analysis of symptomless low birth weight neonates and of asphyxiated newborn infants*" Acta Paed. Scan. Supp., vol. 216, pp. 10-45, 1971.

[14] A.M. Goberman and M.P. Robb, "*Acoustic examination of preterm and full-term infant cries: the long-term average spectrum"*, J. of Speech, Lang., and Hearing Res., vol. 42, pp. 850-861, 1999.

[15] H.F.R. Prechtl, "*Assessment methods for the newborn infant, a critical evaluation*", in P. Stratton, Ed., Psychobiology of the human newborn. Chichester: John Wiley & Sons Ltd., 1982, pp. 21-52.

[16] L. Dubowitz, V. Dubowitz and E. Mercuri, 1999. *The neurobiological assessmentof the preterm and full term newborn infant*, Cambridge: McKeith Press.

[17] Liebermann P. "*The physiology of cry and speech in relation to linguistic behavior*", in B.M. Lester and C.F.Z. Boukydis, Eds. Infant crying: Theoretical and research perspectives. New York: Plenum Press, pp 29-58, 1985.

[18] P. Sirvio, & K. Michelsson, Sound spectrographic cry analysis of normal and abnormal newborn infants. *Folia Phon.*, vol. 28, pp. 151-173, 1976.

[19] J.L. Tenold, D.H. Crowell, R.H. Jones, T.H. Daniel, D.F. McPherson, and A.N. Popper, "*Cepstral and stationarity analyses of full-term and premature infants' cries*" J. of Acous. Soc. of Am., vol. 56, 975-980, 1974.

[20] J.D. Newman, "*Investigating the physiological control of mammalian vocalizations*", in J.D. Newman, Ed. The physiological control of mammalian vocalization. New York and London: Plenum Press, pp.1-7, 1988.

[21] I.R. Titze, *Principles of voice production*. New Jersey: Prentice Hall, 1994.

[22] W.T. Fitch, J. Neubauer and H. Herzel, "*Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production*" Anim. Behav., vol. 63, pp. 407-418, 2002.

[23] D. Lenti Boero, C. Volpe, A. Marcello, C. Bianchi and C. Lenti, "*Newborns crying in different situational contexts: discrete or graded signals?*" Perc. Motor Skills vol. 86, 1123-1140, 1998.

[24] W. Mende, K. Wermke, S. Schindler, K. Wilzopolski and S. Hock, "*Variability of the cry melody and the melody spectrum as indicators for certain CNS disorders*", Early Child Devel. & Care, vol. 65, pp. 95-109, 1990.

[25] A.T. Cacace, M.P. Robb, J.H. Saxman, H. Risemberg and P. Koltai, "*Acoustic features of normal-hearing pre-term infant cry*, Int. J. of Ped. Otorhinolaryng., vol. 33, pp. 213-224, 1995.

[26] C. Johnston, B. Stevens, K. Craig and R. Grunau, "*Developmental changes in pain expression in premature, full-term, two- and four-month-old infants*", Pain, vol. 52, pp. 201-208, 1992.

A NEW TOOL FOR AUDIO AND VIDEO ANALYSIS: AN AID TO CONTACT-LESS CLINICAL DIAGNOSIS IN NEWBORNS

S.D. Barbagallo[1], S.Orlandi[1,2], C. Manfredi[1]

[1] Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[2] Department of Electrical, Electronic and Information Engineering (DEI) "Guglielmo Marconi", Università di Bologna, Bologna, Italy

Abstract: **Recently, and especially in newborn infants, traditional techniques for diagnosis and assessment of neurological disorders are complemented by contactless methods based on the assessment of parameters obtained both from automatic and perceptual analysis of audio and/or video recordings that result in semi-quantitative evaluation of the patient's status. This paper presents a software tool conceived for helping clinicians in the perceptual analysis of neurological impairments in newborn based on audio and video recordings. The system is also provided with devoted software for automatic objective analysis of newborn cry signals.**

Keywords: **Perceptual analysis, newborn cry, general movements, computer aided diagnosis, objective voice analysis.**

I. INTRODUCTION

Recently, and especially in newborn infants, traditional techniques for diagnosis and assessment of neurological disorders are complemented by contact-less methods based on the assessment of parameters obtained both from automatic and perceptual analysis of audio and/or video recordings that result in semi-quantitative evaluation of the patient's status.

Contact-less techniques provide advantages in terms of comfort and safety of the patient with respect to sensor-based/invasive methods, but the amount of recorded data is often prohibitive and highly time consuming for perceptual analysis even for trained and qualified clinicians, as it must be performed in accordance with strict protocols. Moreover, devices and software tools commonly used by clinicians are heterogeneous and not specifically designed for clinical use that is clinicians use different hardware and software tools to manage patient data, record and process audio and video signals to obtain parameters of interest.

To the authors' knowledge, to date there is no software tool that integrates in one system all the components required to successfully perform this kind of analysis.

This paper presents a new tool that addresses this need, particularly critical in neonatal neurology. It allows managing patient data and recordings, processing and analyzing audio and video signals, thus providing an aid to perceptual analysis for contact-less early diagnosis of neurological impairments such as cerebral palsy or autism spectrum disorders in newborns.

In particular, autism spectrum disorders (ASDs) are complex disorders of brain development. In most cases ASDs symptoms appear in the first year of life, when it is not yet possible to carry out a reliable diagnosis. This is because diagnostic tests for ASD are mainly based on the identification of behavioural symptoms that are more evident and recognizable after the 24th month of life. Therefore new approaches are needed and searched for early diagnosis to provide more effective support from caregivers.

Recent studies support the hypothesis of a strict relationship between autism and cry [1] [2], but at present no reliable method for a fully automated analysis is available. Crying is the first and primary method of communication among humans: parents are often able to distinguish a painful cry from a hunger or sleepy one. Thus, acoustic analysis of newborn cry is of relevance to identify parameters that can be indicators of neurological pathologies, such as fundamental frequency (F0), intensity, vocal tract resonance frequencies, length and shape (melody) of cry episodes, and has gained great scientific interest in the last years. [3], [4]. Moreover, motor disorders such as hypotonia, motor apraxia and dyspraxia are main symptoms of neurological disorders and of autism [5], therefore great interest is paid to newborn spontaneous (i.e. not induced or stimulated) movements, named general movements (GMs).

GMs are a set of foetal and newborn spontaneous movements that occur from the 10th week of postmenstrual age to about the 6th month of gestational age. GMs are assessed by expert clinicians that visually analyze and score their variety and complexity. In high risk subjects (i.e. newborns with possible brain dysfunctions or siblings of an already diagnosed autistic child), deficiency in this qualities or almost complete lack of movements is related to a possible neurological

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Fig. 1 First window of the tool for audio signal pre-processing. The voiced part of the signal are detected by the implemented algorithm and then merged for further elaboration.

impairment such as cerebral palsy [6], [7] and autism spectrum disorder [8]. As both cry and the GMs are influenced by neurological impairments, they have become the subject of growing clinical interest as early, non-invasive and low cost diagnostic methods.

II. METHODS

Besides becoming a computer-aided diagnosis tool, the software here proposed was developed with the aim of guiding the clinician throughout the process of contact-less assessment and diagnosis with the aid of a user-friendly environment.

The software is organized into two main interfaces: the first one, where the signal is pre-processed for the extraction of frames of clinical interest and the second one where the signal is processed to obtain the parameters. This module allows handling both audio and video signals.

The first interface is designed for managing both kinds of signal such as playing the signal, inserting text notes in relation to a particular event and adding reference markers (e.g. point out a relevant event). The selection and extraction of relevant signal segments is simply made by selecting their starting and ending points: the software automatically merges into a single file all the selected segments one after the other. This option is particularly useful in the assessment of GMs from video recordings as prescribed in the protocol [7], and it is essential in the elaboration of audio signals. Fig.1 shows the first interface for an audio signal.

As concerns audio recordings, a unique feature of the new tool is the possibility to automatically select the voiced parts only. This is performed by implementing a newly developed robust algorithm [5].

The second window allows for parameter extraction from the signal's frames selected in the first window. For audio signal, the software performs newborn cry analysis by the estimation of acoustic parameters such as fundamental frequency (F0), the first three resonance frequencies of the vocal tract (F1, F2 and F3), the power spectral density (PSD) and the spectrogram [9].

F0 is estimated by means of a two-step procedure that was shown to outperform other methods thanks to the adaptive procedure implemented for the local definition of the length of each signal frame on which the acoustic parameters are estimated: the higher the F0 the shorter the length of the window. In this way the high variability of this kind of signal, that is typically non-stationary, is taken into account. Estimation of vocal tract resonances is carried out by finding maxima in the PSD obtained by means of a parametric approach [9] [10] (Fig. 2).

For video signal the system offers a movements' analysis tool that allows clinicians to perform a semi-automatic analysis of patient's movements from recorded video clips.

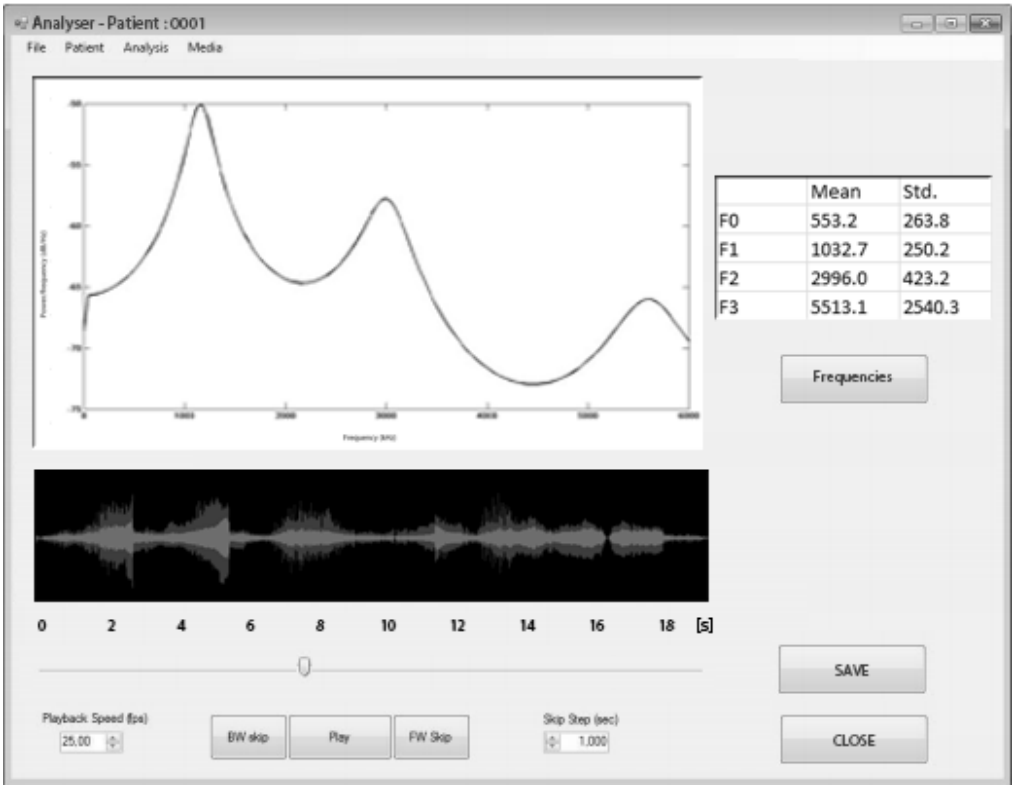


Fig. 2 Second window of the tool for audio signals analysis. The window shows the PSD, F0 and resonance frequencies F1 and F2 (mean and standard deviation) of the cry episode displayed in the lower plot.

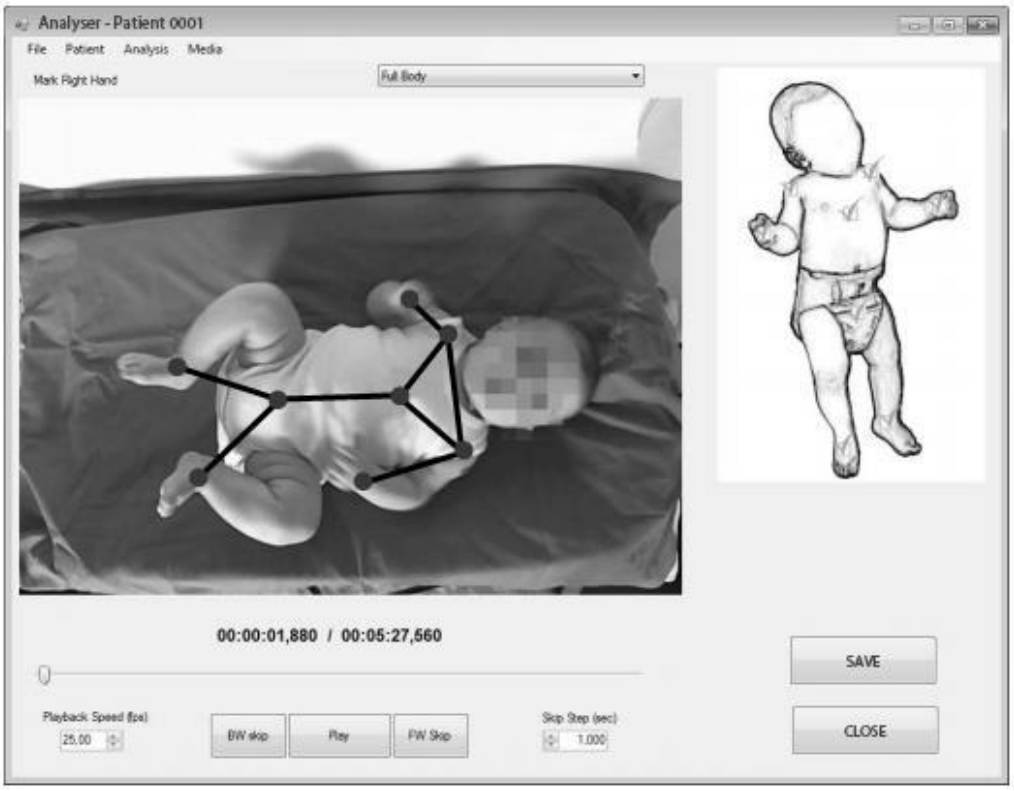


Fig. 3 Second window of the tool for video signals analysis. The body model on the right hand of the window guides the user to the choice of the points of interest to be tracked.

Tracking of movements of a specific section of the body is achieved by manually selecting the points of interest on a video frame with the aid of a body model. The system collects x and y coordinates values of each point on the image plane and saves it on a .csv file that can be handled with any software such as Matlab or Excel to extract and track movements' features, such as speed and acceleration (Fig.3).

III. Results and Conclusions

To our knowledge the software here described is the first tool that allows for analyzing both cry and general movements in newborn infants within a single framework.

At present few systems are available for cry recording and analysis, in particular for the study of F0 [11] [12] [13], but no tool exists for the analysis of GMs. As recent studies [14] highlight the possibility of identifying some typical signs of autism in the first year of life from joint audio and video analysis, this tool could give a significant support to early diagnosis of this disorder.

In addition to providing an appreciable decrease in investigation time, costs and manual errors, the tool is a first step towards the growing clinical interest for marker-less monitoring and diagnosis.

The whole system is under further development to include additional options and features. It will undergo testing in the clinical centers as well as a specific assessment according to ISO standards for the design of tools for medical use with human applications.

With few modifications the proposed tool could be used for the screening of a wide range of peri- and post-natal pathologies and could also be adapted for home care monitoring.

REFERENCES

[1] P. Venuti, G. Esposito and Z. Giustu, "A qualitative analysis of crying and vocal distress in children with autism," *Journal of Intellectual Disability Research,* vol. 48, pp. 4-5, 2004.

[2] S. Orlandi, L. Bocchi, C. Manfredi, M. Pupolo, A. Guzzetta, S. Vicari and M. L. Scattoni, "Study of cry patterns in infants at high risk for autism," in *7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Florence, Italy, August 25-27, 2011.

[3] S. D. Cano-Ortiz, C. A. Reyes-Garcia, O. F. Reyes-Galaviz, D. I. Escobedo-Beceiro and J. D. Cano-Otero, "Emergence of a New Alternative on Cry Analysis: The Fuzzy Approach," in *V Latin American Congress on Biomedical Engineering CLAIB 2011 May 16-21* , Habana, Cuba, 2011.

[4] B. Mampe, A. D. Friederici, A. Christohe and K. Wermke, "Newborns' Cry Melody Is Shaped by Their Native Language," *Curr. Biol.,* vol. 19, no. 23, pp. 1994-1997, 2009.

[5] X. Ming, M. Brimacombe and G. C. Wagner, "Prevalence of motor impairment in autism spectrum disorders," *Brain and Development,* vol. 29, no. 9, pp. 565-570, October 2007.

[6] C. Einspieler, P. Marschik, A. Bos, F. Ferrari, G. Cioni and H. Prechtl, "Early markers for cerebral palsy: insights from the assessment of general movements.," *Future Neurol,* vol. 7, no. 6, pp. 706-717, 2012.

[7] C. Einspieler, H. Prechtl, A. Bos, F. Ferrari and G. Cioni, Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants, Cambridge University Press, 2004.

[8] C. Einspieler, A. Kerr and H. Prechtl, "Abnormal general movements in girls with Rett disorder: The first four months of life," *Brain & Development,* vol. 27, no. S8-S13, 2005.

[9] A. Fort and C. Manfredi, "Acoustic analysis of newborn infant cry signals," *Medical Engineering & Physics,* vol. 20, no. 6, pp. 432-442, 1998.

[10] A. Fort, A. Ismaelli, C. Manfredi and P. Bruscaglioni, "Parametric and non-parametric estimation of speech formants: application to infant cry," *Medical Engineering & Physics ,* vol. 18, no. 8, pp. 677-691, 1996.

[11] C. Manfredi, L. Bocchi, S. Orlandi, L. Spaccaterra and G. P. Donzelli, "High-resolution cry analysis in pretermi newborn infant," *Med. Eng. & Phys.,* vol. 31, no. 5, pp. 528-532, 2009.

[12] S. Orlandi, L. Bocchi, G. P. Donzelli and C. Manfredi, "Central blood oxygen saturation vs crying in preterm newborns," *Biomed. Signal Proc. And Control,* vol. 7, no. 1, pp. 88-92, 2012.

[13] B. Reggiannini, X. Li, H. F. Silverman, S. J. Sheinkopf and B. M. Lester, "A flexible analysis tool for the quantitative acoustic assessment of infant cry," *J Speech Lang Hear Res.,* 2013.

[14] L. Bocchi, S. Orlandi, C. Manfredi, M. Puopolo, A. Guzzetta, S. Vicari and M. L. Scattoni, "Early Diagnosis of Autism Spectrum Disorder - Design of the Data Acquisition and Management System," in *5th European Conf. of the Int. Fed. Med. Biol. Eng*, Budapest, Hungary, 2011.

Session III:
SINGING VOICE

VOICES FROM THE HISTORY

Luigi Dei

Department of Chemistry "Ugo Schiff" – University of Florence
luigi.dei@unifi.it

I would like to start my lecture with a poem by Emily Dickinson, some verses that are very wonderful and evocative. I believe it's the best way to introduce my very strange and peculiar excursion along the voice's paths: « Silence is all we dread. There's Ransom in a Voice - But Silence is Infinity. Himself have not a face. » The not conventional use of capitals, typical of Dickinson's poetry, allows associating silence to infinity and giving to the voice the emphasis of the ransom. Therefore, silence as a lifeless thing and lack of communication, horrifies us; nevertheless, a single voice, even ours, is sufficient to redeem from infinity, indescribable world without any sounds and dimension. The silence has not any face, contrariwise the voice is a face or thousands different faces. The voice is a kind of *mare magnum* in which we navigate all the life. Our existence is this sailing, seemingly eternal, where the voice, generated or listened to by us, rocks us, it scans our sad or happy moments, and builds, with discretion, – almost as we do not realise – our history, but perhaps even the other History, that with H capital. Even event with great relief occurred in the past century is a station where we stop pondering to understand – maybe only to know –. This station has got a loudspeaker constantly switched on that irradiates, with the voice, some words, but more often the notes of a song. Of the private stations, as our birth, infancy, adolescence, youth, maturity and old age, and from those public of the historical events that unrolled themselves with us and around us, we recall many things. Indeed, I am thinking that, perhaps, the voice of that very loudspeaker throws us again to those tracks. The first wailing and the first crying when our voice begins to take shape, in the sound spectrum of which somebody is hypothesising the waves typical of the mother's voice are present. The calm sunsets of our first days of life gladdened by more or less famous lullabies. The songs of our young loves, then those of the more stable and long-lasting ones, and probably of those prohibited of the adulterine affairs. The melodies that befriended the birth of our progeny and afterwards, their youth, our maturity; their maturity, our old age; our end in the silence, their crying that accompanies us in the nothing. From the birth to the death it's a continuous yachting gently pushed by the wind of the voice, by the various and different voices that crowd our hearing, voices that could appear as noise and on the contrary they are amazing signals of life. It's so intense and overwhelming the experience of listening to a voice that we terribly labour to remember. We memorize much better images and characters, than timbre and sound of voices. In the dreams we have mainly visions, moving images, but voices and sounds difficultly populate the dreamlike world. And this is a clear proof that the hearing of a voice is an unique experience.

The very history, at least until the tape recorder arrived in the twentieth century, is made of written sources, even of drawings, paintings, statues, never of vocal sources. We have an idea about the human features of Socrates, Cesar, Dante, Charles the Great, Columbus, Napoleon, Garibaldi but what about their voices? Therefore, if the voice, that quality that makes unique the *homo sapiens sapiens* also for its very extended selection of possible articulations, remained not reproducible for millennia, drowned as the American poet wrote "in the infinity of the silence", this ceased to occur since 1934 when the AEG and Telefunken patented the first tape recorder – the Magnetophohon K1 – and in this way the technological possibility to memorize and reproduce the voice, and more in general sounds and music, was born. This marvellous invention, fruit of human brilliance and creativity, as well as the poems by Dickinson, allows us telling the history from 1840 until today thanks to re-hearing voices, more exactly songs that accompanied the big transformations and the turbulent stirrings of more than seventy years of world history. The twentieth century is the century of the ransom too, of the voices coming from the entire social classes excluded from all that finally demand rights, dignity and sovereignty and that try to make concrete act, form in evolution, the power of the matter of the three cornerstones of the French Revolution: *liberté*, *egalité*, *fraternité*. Struggles for the emancipation, decolonisation, demolition of the racial hate, exaltation of human rights, negation of every integralism, social justice: how many historical and philosophical essays have they been written on this subject? But what's better than a splendid voice can tell us our history, the history of a process still in evolution, but that, we have known for long time, shall make us to triumph, shall allow overcoming every obstacle and at the end we will succeed, we will be right, we shall overcome …

The twentieth century that leaves us great hopes and at the same time warns us about the drama of the war, but it gives us also a new idea, that takes force in the people, about the non-unavoidability of conflicts: hence the twentieth century reclaims at great voice that the pacifism and not the wars can be the engine of the history. Several wars certainly remain in various regions of the world, but maybe some antibodies are developing so that we can wish they succeed in growing to defeat war viruses. Perhaps two world wars have created a break between past and future, perhaps the bloody wars that crossed centuries

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

and millennia of civilisation are going to give way. This is surely a good thing, even if we must always pay attention to wars that are very far from us and against those of the present and future, wars without any bombs and trenches, the new imperialisms and oppressions that germinate in the palaces of finance and capitals that rule the world. To remember this message that the twentieth century left for us in a bottle, we open another bottle brought to us by the waves of a deep darkness: it does not contain messages, but rather a resounding voice, moving, warm and sensual and all we must think still today that the lamp-post under which we could never go back to embrace our loved is always in ambush, that everybody could become again Lili Marlen …

Under that lamp-post the soldier and Lilì have never met again, but the History went ahead and after the tragedy of the war some new seasons opened for our Europe, contradictory seasons, made of economical growth and wealth, but even of strong movements of workers emancipation. An Europe that was divided by the Cold War. The Europe of the capitalism and of the Trade Unions from one side, and on the other side the Europe of the socialism and communism that blanks out freedom and kills the democracy. If I wanted remember with some voices the Fifty years of the past century, if I wanted individuate the song stations of reconstruction after the war, of the seeds for united Europe of nowadays, I would like to fly over France at the beginning of '50 and over Italy at the end of these years, to discover a little and bashful signal of cosmopolitan integration, of abandon of imperialistic colonialism: in front of the not unified Europe, people of the third world begin to look out of the window. A woman of berber origin, poor, humble, born in the middle of a street conquers the new and old world with simple songs of love, with her incomparable voice of a *Piaf*, a little sparrow able to charm the hearts by means of the simple and mysterious force of her voice …

The voice that thanks to the radio, to the amplification and recording succeeds in mocking the oceans, the voice that succeeds in moving in real time *hic et nunc* all the mankind, wanderers of every genre who walk along the paths of all our globe. It's the very voice, indeed, that enchants, coming out from cardboard cones that vibrate thanks to the electromagnetic induction. The radio at the beginning, and then the television, glorify the voice, that spoken by the speakers of the new newspapers without printing press, as well as that sung which cheers up, amuses and excites the citizens of the world. That world that starts to be within rang of loudspeaker or fluorescent screen. The information, magically transported by the human voice, cross the oceans at an extraordinary velocity, widen the knowledge horizons of peoples, independently on the grade of education. The great literature goes into the houses of everybody thanks to the voices and the movements of the actors. The voices that burst into our little world ransom the silence, open incredible horizons, drag all of us to the dreams heaven, the sonorous wind kidnaps the little minds just barely made literary and makes them to fly in the infinite sky, in the blue painted of blue …

It's the year 1958 and the Italian song *Volare* crosses in short order the ocean and remains at the first place in the rank in the United States for five weeks! – the only Italian song in the whole history – just in the country where Elvis Presley e Frank Sinatra, two *voices* par excellence and antonomasia, are a big hit! The voices that run themselves after from country to country, from continent to continent, but they do not succeed in passing the iron curtain. Despite the beginning of the destalinisation, a portion of Europe is still off limits: the year before the Italian song *Volare* becomes famous, one of the masterpieces of the world literature of the 2oth century is printed for the first time here in Italy: *Doctor Zivago*. It's the only great and true voice coming from the East of Europe, it isn't a sung voice, it's mute, but wonderfully stentorian. And this very voice remembers us that the history and the nature are a whole. The history that very quickly, with a rapid movement almost similar to the "toccata and fugue", we are caressing by means of these vocal stations, is not the history of the great personalities, but not even the history of the "without a name", it's something that transcends the mankind, with a sacredness which is in some way tragic but projected toward the future. A history that the Italian writer Italo Calvino reminds us "moves like the vegetable kingdom, like the wood that transforms itself when spring comes". A history of the 20th century for which I do not succeed in finding a voice in music, but rather a poetic voice, that of the Polish Nobel Prize for the literature and poetry Wisława Szimborska: "we arrived just to this point: I sit down under a tree, | on a bank of a river | in a sunny morning. | It's a meaningless event | and it shall not pass into the annals of history. | We don't deal with battles and treaties | of which we study the causes, | no killing of dictators worthy of memory …". The 20th century is this too and even though no songs are present to conserve the memory, nevertheless we must have remembrance. A century where the peace come after the war heralds a memorable costumes revolution that changed the way to face civilian questions, the rights of contemporary developed societies with high scientific and technological content. The 1968 that makes the new generations' attention-seeking to come to the fore, the will of progress, of social justice, of egalitarianism, of the laity. A period rich of unrests where the desire of individual and collective freedoms and aspiration for a better world chase themselves in an universe in which an amazing scientific and technological progress is being born quickly leading to the third revolution: after the first agricultural revolution and the second one industrial, here is the informatics revolution, that of bit and baud, the revolution that will make the world reachable by a "clic". A period during which transforming and making better the world becomes an ethical imperative, an aim so strong to permeate the song by the Liverpool's *beetles*, that invite Jude to take a sad song and making better, after having placed it into the heart …

Just one year after a far voice shall represent the reaching of a mirage for the man of every time. “That’s a small step for a man, one giant leap for mankind”. These word were pronounced by Neil Armstrong on July 20th, 1969. Then some seconds of silence and Neil will exclaim: “Magnificent desolation!” I look high at the sky and see a luminescent sickle. Man conquered the moon! And this time the achievement goes into the annals of the history thanks to these few words articulated by a voice and recorder for the eternity. But the ’68 will be even the beginning of the iron curtain crumbling away: Prague’s spring and then the invasion by the Red Army tanks will open the door for a slow but inescapable process that twenty years later shall lead to the downfall of the Berlin Wall and to the end of the Cold War. The world divided in two blocks is going to be fragmented in something much more complicated, a globalisation never seen in the history of mankind and difficultly framing in pre.-constituted schemes. The twentieth century opened with the spiritual *We shall overcome* and the third millennium makes its debut with a new drama, the international terrorism and the eternal contrast poverty-richness. New terrorism and finance wars jeopardize to make us again helpless. The new post-lamp of Lilì Marlen where the soldier never went back transfigure itself into the little daily things orphans of the new wars victims, “shirts in the closet, shoes in the hall, coffee cups on the counter, jackets on the chair, papers on the doorstep, but you’re missing” …

And now we are at present, at the paradox that in front of the victory of the capitalism, went out triumphing from the twentieth century, we assist to its strongest crisis, a crisis which is completely internal the range and the effects of which we are still struggling to decipher. All is occurring when the stars of East seem to be proceeded to shine with greater brightness and those of the old and white haired West appear to show a constant and, maybe, inescapable fainting. But here we must stop, we have to turn down the voice and be silent: the voices of the future shall narrate the history that, watered by the present, is going slowly beginning to germinate.

To conclude and still reiterate once again how magic be the power of the voice with respect to whatever other sound generator, I would like to make you listening to a very short musical piece from the *Tuba Mirum* by Giuseppe Verdi’s Requiem. It will be a *crescendo* of emotions from the first rings of the trumpet, to those of the successive from the various trumpets situated in different parts of the theatre almost to simulate echoes and bounces, until the resounding bursting of the whole orchestra. But absolutely the most intense emotion will arrive when more than one hundred women and men will make you feel their very voice, the mankind’s voice which is more powerful than every trumpet ring or orchestra *tutti*. The voice that will sing the *Tuba, mirum spargens sonum,* of the crack of doom, making us discovering that the admirable sound is not that spread by the crack of doom trumpets, but rather that of the human voice that we are, at this Congress, someway celebrating. In this short singing promenade along the paths of the twentieth century, we have listened to tuning verses in English, French, German, Italian and now, at the end, in a died language that is so living and much evocative …

SPECTRALLY ESTIMATED VOCAL TRACT LENGTHS OF SINGING VOICES AND THEIR CONTRIBUTING FACTORS

M. Sakaguchi, M. Kobayashi, R. Nisimura, T. Irino, H. Kawahara
Department of Design Information Sciences, Wakayama University, Wakayama, Japan
{s135021,s130043}@center.wakayama-u.ac.jp, {nisimura,irino,kawahara}@sys.wakayama-u.ac.jp

Abstract: **A new computationally efficient and reproducible method for relative vocal tract length estimation is applied to singing voices stored in the RWC music database for music research and mimicry singing database. The analyses indicated stylistic differences between classical singers and POP-song singers in terms of vocal tract length dependencies on voice fundamental frequency. The similar differences are also observed in different vocal expressions. For example, classical bass, baritone and alto singers seem to keep their vocal tract length relatively constant over their usable voice ranges. These findings suggest that the proposed vocal tract length estimation method is useful as an assisting tool for voice training.**

Keywords:— **Singing voice, vocal tract length, voice training, fundamental frequency**

I. Introduction

Huge physical variability of individual voices is a challenging obstacle for objective assessment of singing voices. The primary source of the variability is singers' physical dimension, which result in differences of vocal tract lengths. Effects caused by vocal tract length differences have to be normalized before applying further sophisticated analyses. The vocal tract length also varies due to intrinsic differences between vowels and voice fundamental frequency (F0) [1, 2]. Dependencies on these factors are also important measure for voice assessment.

The target of this article is to establish a procedure to measure vocal tract lengths objectively and non-invasively, from singing voices. We introduce a new computationally efficient vocal tract length estimation method based on an interference-free representation of periodic signals [3, 4]. The method was applied to speaking voices and illustrated highly reproducible results [5]. In this article, preliminary analysis results using the proposed method to singing voices are introduced.

II. Method

This section introduces the proposed vocal tract length estimation method. It is a two staged procedure. The first stage is minimization of spectral distance and the second stage is least square estimation of relative vocal tract lengths. More detailed descriptions can be found in our previous article [5].

A. Minimization of spectral distance

Proportional vocal tract shape change yields proportional dilation/compression of the vocal tract transfer function. The proposed vocal tract length estimation method searches the best stretching factor to minimize spectral difference between two voice samples. It is necessary to reduce disturbing factors for this simple idea to function properly.

The first disturbance is caused by harmonic structure of voiced sounds. Smoothing power spectra, using an F0-adaptive smoothing function in the frequency domain, eliminates this effect. A triangular smoother having the base width that is equal to two times of F0 [3] is used.

The next disturbance is global spectral slope difference. The difference is caused mainly by glottal closure details. This effect is equalized by removing smoothed spectral shape by using wider spectral smoother, for example a raised cosine smoother having 2000 Hz base width. The third disturbance is fine spectral details caused by sharp zeros caused by pyriform fossa and glottal source waveform. Smoothing spectral details reduces this factor, by using relatively narrow spectrum smoother, for example, a raised cosine smoother having 300 Hz base width.

In addition to this set of procedures, relevant frequency region for difference evaluation has to be selected. It is because in the lower frequency region, namely lower than two times F0, the spectral shape consists of strong effects of glottal waveform. It is also because in the higher frequency region, namely 4 kHz or higher region, individual three dimensional shape differences of vocal tracts introduces significant transfer function differences [6]. The best linear stretching factor of the frequency axes of the preprocessed target and the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

reference spectra, which minimizes preprocessed spectral distance, yields the vocal tract length ratio between these spectra. This provides the basis for estimating relative vocal tract lengths of individual singers.

B. least square estimation

All possible combinations of voice samples (let N represent the number of voices) provides $N(N-1)$ vocal tract length ratios $r_{n,m}, (1 \leq n, m \leq N)$ for determining N unknowns. They are relative vocal tract lengths $l_n, (1 \leq n \leq N)$. Logarithmic conversion of $r_{n,m} = l_n/l_m$ makes interrelations between unknowns yield the following set of linear equations.

$$\boldsymbol{r}_{\log} = H\boldsymbol{l}_{\log}, \tag{1}$$

where $\boldsymbol{r}_{\log}$ represents a vertical vector consisting of $\log(r_{n,m})$. The weight matrix H has 1 at n-th element and -1 at m-th element of corresponding row to $\log(r_{n,m})$. The logarithmically converted vocal tract lengths are stored as the elements of the vector $\boldsymbol{l}_{\log}$. All elements of the last row of H are set to 1 to represent normalization condition. The last element of $\boldsymbol{r}_{\log}$ is set to 0, accordingly.

Exponential conversion of the least square estimate of $\hat{\boldsymbol{l}}_{\log}$ provides relative vocal tract lengths. This procedure was tested using a Japanese vowel database consisting of 384 speakers ranging 6 years old to 56 years old. In this test, the averaged spectral distance of five vowels is minimized. The spectrally estimated vocal tract length ratios were distributed around the vocal tract length ratios calculated based on the solution of the set of linear equation, with the standard error value 0.009. This illustrates that the proposed method yields highly reproducible estimates for speech sounds. The same procedure was applied to analyze singing voices by different singers and in different singing styles.

III. Analyzed materials

Two sets of databases were used. The first one is RWC music database for computer music research [7]. It is a portion of a collection of 100 audio CDs with full of copyright cleared contents for academic research purpose. Singing voices are stored as one category of musical instruments. It consists of three sopranos, three altos, three tenors, three baritones, three basses and three popular singers' voices in various singing styles. The singers were instructed to sing the full range (for the singer) of chromatic scale in Japanese five vowels in designated strength, expression and styles. The other database was prepared for singing style analysis and synthesis for Japanese POP-songs [8]. Professional singers (two male and two female singers) sang several materials in two different singing styles. One style is their own singing style and the others are mimicry of well known singers'. In this analysis, recordings of isolated Japanese five vowels are used.

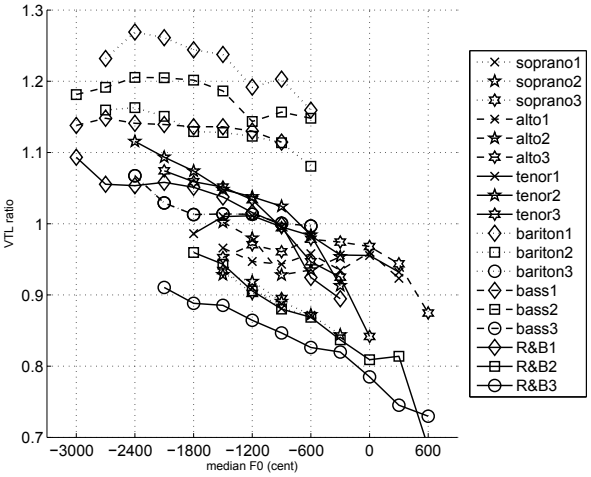


Fig. 1. Estimated relative vocal tract lengths for *forte* without *vibrato*. Each line represents each singer. Horizontal axis represents averaged F0 of grouped segments and vertical axis represents the relative vocal tract length.

IV. Analysis results

A waveform symmetry-based F0 extractor [9] was used to analyze F0s of the singing voices. Semi-automatically extracted each voiced segment consists of one isolated vowel. Segments in each singing style for each singer are grouped using frequency bins with 300 cent width. Averaged spectral distance between groups are minimized to calculate spectrum-based vocal tract length ratios. Relative vocal tract lengths are estimated from the ratios by using least square estimation mentioned above. Unlike usual speech analysis, sometimes (less than 2% of all possible combinations) the original vocal tract length ratio and that calculated from the estimated relative vocal tract lengths were found to be different significantly. These outliers were eliminated from the dataset and estimates were recalculated.

A. RWC database

Figure 1 shows estimated vocal tract lengths for *forte* without *vibrato*. Averaged F0 in this plot is represented in terms of cent, setting 440 Hz to 0 cent. Estimated vocal tract lengths of female singers are generally shorter than that of male singers. This is consistent with body size differences between male and female singers. Two bass

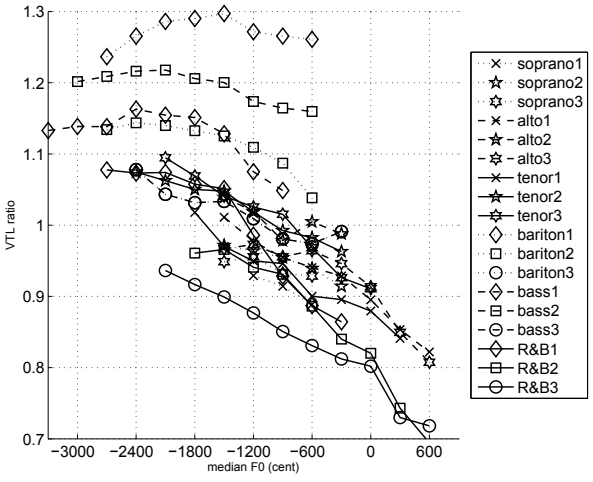


Fig. 2. Estimated relative vocal tract lengths for *piano* without *vibrato*.

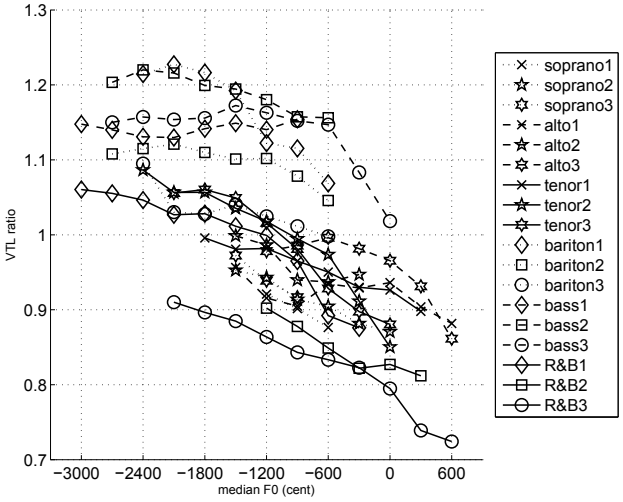


Fig. 3. Estimated relative vocal tract lengths for *forte* with *vibrato*.

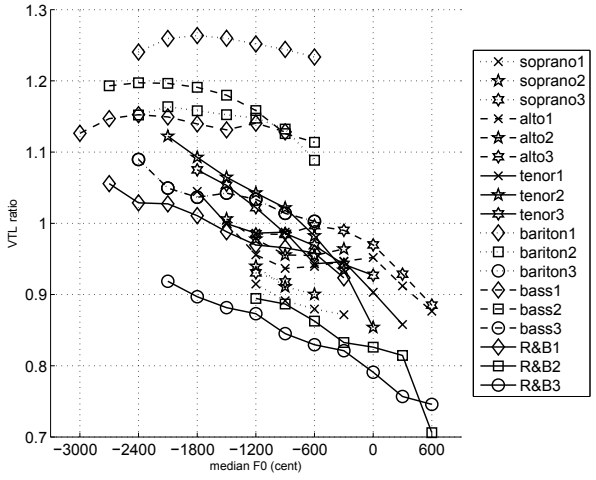


Fig. 4. Estimated relative vocal tract lengths for *forte* with *staccato*.

and two baritone singers show relatively F0 independent vocal tract lengths. Also three alto singers show similar trend. Other singers show systematic vocal tract length shortening when F0 increases. This shortening is consistent with raising of larynx when singing high-pitched notes [1]. In other words, relatively constant vocal tract lengths found in bass, baritone and alto may suggest that there exists intentional control to keep the lengths.

Figure 2 shows estimated vocal tract lengths for *piano* without *vibrato*. Similar trends found for Fig. 1 are observed. It is also true for Fig. 3 and Fig. 4, where *vibrato* and *staccart* are introduced respectively. This may indicate that vocal tract length variations are dependent on singer, F0 and relatively independent of vocal effort (intensity), lengths (such as *staccato* or *tenuto*) and *vibrato*.

B. Mimicry database

Figure 5 shows results using the mimicry database. G and I are male singers and M and U are female singers. Their dependencies on averaged F0 are steeper than classical singers'. Similar trends are also (not very clearly) observed in RWC data base results. One interesting point in this mimicry database analysis is that dependencies on averaged F0 and biases are different between singers' own style and mimicry. This may suggest that they are using vocal tract lengths as a means to mimic other singers' voices and styles.

V. Discussion

Vocal tract length is one of the dominant factors for determining spectral shape. It does not interfere linguistic information for human speech perception but has strong impact on timbre perception especially in singing voices. The proposed method provides a possibility to separate effects of this innate factor in singing performance and elucidate acquired skills by introducing normalization procedure based on the estimated vocal tract lengths.

Figure 6 shows the estimated vocal tract lengths of an inexperienced singer. The inexperienced voices and professional singers' voices are used together to estimate relative vocal tract lengths shown in Fig. 6. The trend of novice's seems close to that of tenor2's. For example, adjusting average vocal tract lengths of each professional singer's to match the average of the novice's reduces innate timber differences. Re-synthesizing such adjusted

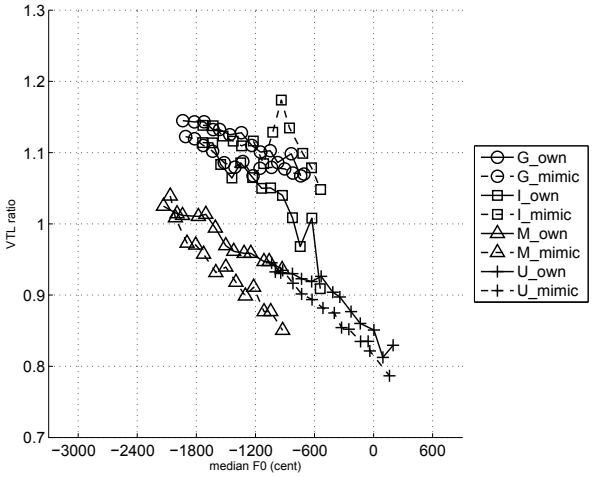


Fig. 5. Estimated relative vocal tract lengths for the original style and mimicry. Each alphabet in the legend represents each singer. Singing samples in singers' own style are represented using solid lines.

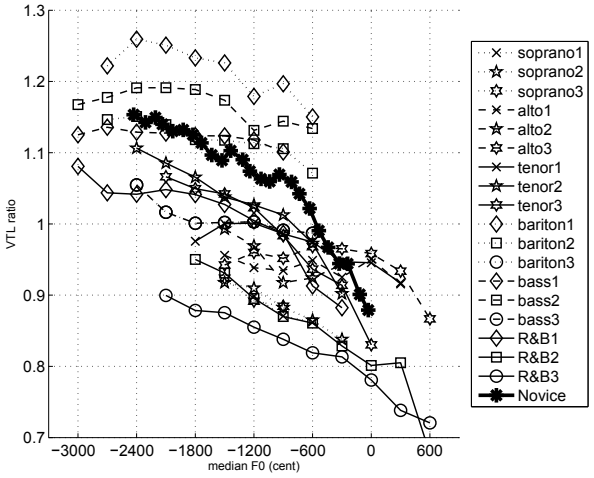


Fig. 6. Estimated relative vocal tract lengths of an inexperienced singer (represented as novice using a solid thick line). Professional singers' voices used in Fig. 1 and inexperienced voice samples were analyzed together.

target voices may clarify differences of voice control between novice and professionals and maybe useful feedback for training.

VI. Conclusion

A new computationally efficient and reproducible method for relative vocal tract length estimation is applied to singing voices stored in the RWC music database for music research and mimicry singing database. The analysis results indicated that variation of the estimated relative vocal tract lengths that vocal tract length variations are dependent on singer, F0 and relatively independent of vocal effort (intensity), lengths (such as *staccato* or *tenuto*) and *vibrato*. It also suggested that the vocal tract length is used as a means to implement mimicry. These findings suggest that the proposed vocal tract length estimation method is useful as an assisting tool for voice training.

Acknowledgement

This work was partly supported by Grants-in-Aid for Scientific Research category (B)24300073, (B)25280063 and Exploratory Research 24650085. It is also by Wakayama University.

References

[1] J. Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, 1987.

[2] S. Maeda and Y. Laprie, "Vowel and prosodic factor dependent variations of vocal-tract length," *Proc. Interspeech 2013*, pp.3196–3200, 2013.

[3] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, 27(3–4), pp.187–207, 1999.

[4] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, Hideki Banno, "TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation," *Proc. ICASSP 2008*, Las Vegas, 2008, pp.3933–3936.

[5] M. Kobayashi, R. Nisimura, T. Irino and H. Kawahara, "Estimated relative vocal tract lengths from vowel spectra based on fundamental frequency adaptive analyses and their relations to relevant physical data of speakers," *Proc. ICA 2013, Montreal Canada*, 5aSCb44, 2013.

[6] S. O. Ternstrom, "Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5 kHz," *Journal of the Acoustical Society of America*, 123(5), pp.3379–3379, 2008 .

[7] M. Goto, "Development of the RWC Music Database," *Proc. ICA 2004, Kyoto Japan*, pp.I-553-556, 2004.

[8] N. Migita, M. Morise and T. Nishiura, "Study of effective features for controlling the differences of vibratos among singers by utilizing singing database," *Trans. Information Processing Society of Japan*, 50(2), pp.1910–1922, 2011.

[9] H. Kawahara, M. Morise, R. Nisimura and T. Irino "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," *Proc. ICASSP 2013*, pp.6797–6801, 2013.

TEMPORALLY FINE F0 EXTRACTOR APPLIED FOR FREQUENCY MODULATION POWER SPECTRAL ANALYSIS OF SINGING VOICES

H. Kawahara[1], M. Morise[2], K. Sakakibara[3]
[1]Department of Design Information Sciences, Wakayama University, Wakayama, Japan
[2]Department of Computer Science and Engineering, University of Yamanashi, Kofu, Japan
[3]Department of Communication Disorders, Health Science University of Hokkaido, Sapporo, Japan
[1]kawahara@sys.wakayama-u.ac.jp, [2]mmorise@yamanashi.ac.jp, [3]kis@hoku-iryo-u.ac.jp

Abstract: **Application of a new fundamental frequency (F0) extractor with temporally fine resolution to singing voice databases revealed details of higher frequency modulation power spectra of F0, which cannot be analyzed by the conventional F0 extractors in such details. Expressive performance found in POP-singing and Japanese traditional performances indicated faster frequency modulation (for example, higher than 50 Hz) sometimes characterizes extreme expressions.**

Keywords:— **Fundamental frequency, fundamental component, instantaneous frequency, voicing, vocal fold vibration**

I. INTRODUCTION

Vocal fold vibration in voicing is not always regular nor stable [1]. Strong singing expression sometimes makes use of this irregularity effectively [2, 3]. Analysis of these irregular vibration *non*-invasively (using speech signal as the source) requires fine temporal resolution and mathematically transparent analysis tools. A new F0 extractor with temporally fine resolution [4] is extended and applied to singing voices stored in the RWC music database [5] and the voice archive of Japanese traditional singing voices [6] for investigating frequency modulation power spectra, which cannot be analyzed in detail by conventional F0 extractors (for example refer to [7, 8]). The proposed method provides a means for substituting EGG measurement when only sound recording is feasible.

II. METHODS

The proposed method is a set of three procedures. The first one is an initial F0 candidate extractor based on higher-order waveform symmetry measure [4]. It is followed by the second procedure based on Kalman smoother [9]. The final stage uses a stabilized representation of instantaneous frequency of periodic signals for refining selected initial estimate of F0 [10, 4]. [1] Following sections introduce these procedures.

A. Initial F0 estimate based on higher-order symmetry

This procedure extracts F0 candidates as the reciprocal of the fundamental interval of waveform repetition of filtered outputs. One of Nuttall windowing functions [11] is used as the impulse response of this bank of low-pass filters because of its low side lobe levels and steep decay. Deviation of the filter output from pure sinusoids is represented as a Minkowski distance consisting of temporal and level asymmetry measures and higher-order (in this case deviation at mid-point of each half cycle) asymmetry measure. Then, this deviation is nonlinearly converted to yield a symmetry measure η_E $(0 \leq \eta_E \leq 1)$, where $\eta_E = 1$ represents pure symmetry.

This symmetry measure is defined by the following equation [4].

$$\eta_E(x,k,f) = \exp\left(-\alpha\left(\sum_{q\in K} w_q \tilde{d}_q^{\beta}(x,k,f)\right)^{\frac{1}{\beta}}\right) \qquad (1)$$

$$\sum_{q\in K} w_q = 1\,, \quad \tilde{d}_q = \frac{d_q[k-1] + d_q[k] + d_q[k+1]}{3\sqrt{V(d_q)}},$$

where x represents output of a filter with the nominal frequency f at k-th extrema of waveform. The element q represents the type of deviation d_q in a set $q \in K = \{AM, FM, SM\}$. (AM: level, FM:temporal and SM:midpoint (higher-order) deviations) Parameters α, β determines the shape of nonlinear mapping.

Since no prior information of F0 is available for designing the best filter to extract the fundamental component, a set of filters processed the input signal simultaneously and the relevant filter candidates were selected based on higher-order waveform symmetry measure η_E. This simple architecture assures fine temporal resolution and efficient computation (and also parallel processing).

[1] All these three procedures are implemented using Matlab. They are computationally efficient and the whole process runs faster than real-time using a note PC (test machine: MacBookPro, with 16 GB memory and 2.6 GHz Intel Core i7.).

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

B. Kalman smoother for candidate selection

This process assumes dynamics of a latent variable ρ_t behind observed F0 $f_{0,t}$, where t represents the time index. Equations and procedures used here are similar to the reference [9]. In this formulation, the latent variable ρ_t represents some underlying un-observable. This latent variable ρ_t should be clearly distinguished from $f_{0,t}$, which represents the observable instance.

A first order dependency of probability $p(\rho_t) \propto p(\rho_t|\rho_{t-1})$ is introduced to model the dynamics of this latent variable. The following relations are assumed for the latent variable ρ_t and observation of F0 $f_{0,t}$.

$$p(\rho_t|\rho_{t-1}) \sim N(\rho_{t-1}, \phi^2) \quad (2)$$

$$p(f_{0,t}|\rho_t) \sim N(\rho_t, \sigma^2), \quad (3)$$

where $\sim$ represents that the lefthand side variable follows the righthand side distribution. $N(\mu, \sigma^2)$ represents a normal distribution with mean μ and variance σ^2. The solution to this process is the Kalman smoother described in the reference [9].

Variance of state transition ϕ^2 in Eq.(2) and Variance of observation σ^2 in Eq.(3) have to be determined to apply the Kalman smoother to higher-order symmetry-based F0 candidates extractor. Variance of observation σ^2 is numerically determined as a function of the symmetry measure η_E based on simulation results using pulse plus noise signals with different signal to noise ratios.

$$\sigma^2(\eta) = \frac{2157^2}{1+\exp(15.3171(\eta-0.16))} + \left(1 - \frac{1}{1+\exp(15.3171(\eta-0.16))}\right) \times(120\cdot(1-\eta))^2, \quad (4)$$

where the variance $\sigma^2(\eta)$ is represented in terms of square of cent.

Variance of state transition ϕ^2 can be determined from the statistics of F0 movement. For example, Fig. 10 of the reference [7] shows distribution of F0 slope in terms of octave/s. By assuming random process, $\phi^2 = 400$ is derived for 1 ms frame rate.

C. Instantaneous frequency-based refinement

Estimated value of the latent variable ρ_t is used to select the best filter. Then, the reciprocal of the repetition interval of the filter output is used as the initial value of the instantaneous frequency-based refinement procedure similar to the post processing procedure used in the reference [4].

Periodic variations of instantaneous frequency is eliminated by using two time windows located half pitch period apart and weighted averaging [10]. The initial estimate of F0 is used to design a set of F0-adaptive time windows and their locations. The first updated F0 is calculated from instantaneous frequencies of the first two harmonic components. This updated F0 is used to select first six harmonic components and their instantaneous frequencies are used to calculate the final revised F0.

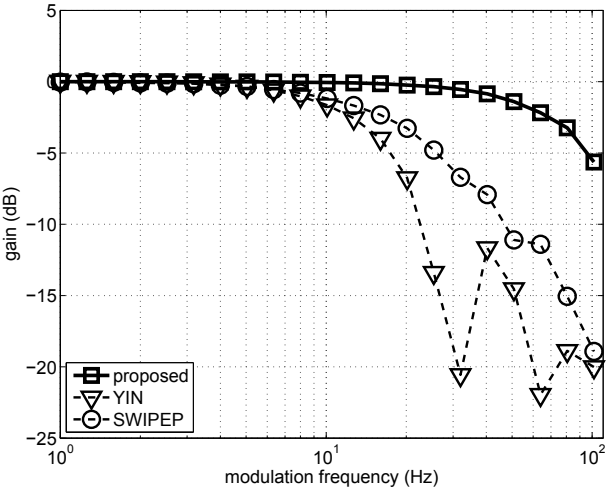


Fig. 1. Modulation transfer functions of F0 extractors to frequency modulated F0.

D. Modulation transfer function and distortion

Response to fast varying F0 was tested using test signals consisting of harmonically related frequency modulated sinusoids. Figure 1 shows modulation transfer functions of several F0 extractors. The modulation depth is one semitone peak-to-peak in this case.

Figure 2 shows power spectra of estimated F0 trajectories. One of Nuttall windows is used in this analysis since the first side lobe level is lower than -90 dB. Distortion of the proposed method is 1/10 to 1/100 of the other methods. The results of modulation transfer function analysis and spectral analysis of F0 trajectories illustrate that the proposed method is suitable for more precise physical measurements of F0.

III. Modulation power spectrum analysis

The procedure was applied to singing voices stored in the databases mentioned below and yielded F0 trajectories calculated in 1ms frame rate. The frequency modulation power spectra were calculated using the differentiated F0 trajectories (represented in terms of cent/s) with one of Nuttall windows. Then, they were converted and calibrated to 1/6-octave band levels in terms of semitone (rms values). Only voiced segments having longer dura-

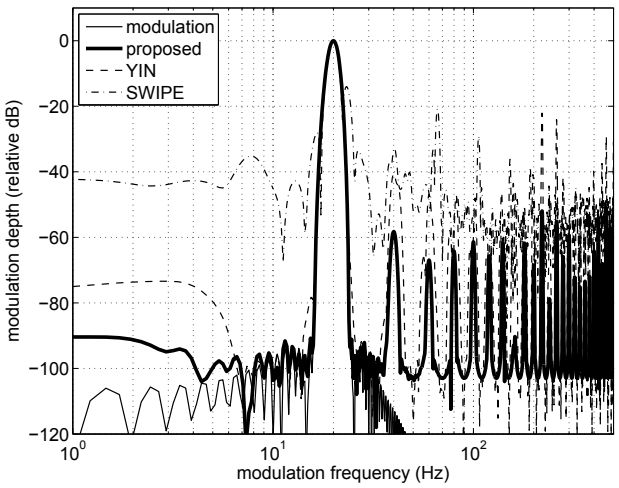


Fig. 2. Distortion of the estimated F0 trajectories using 20 Hz sinusoidal modulation. Modulation depth is a half semitone peak-to-peak. The average F0 is 300 Hz.

tion than 500 ms were analyzed. Prior to these analyses, each long-term power spectrum was inspected to properly remove low-frequency background noise and interference caused by commercial alternating curtent.

A. *Database:*

Two databases were used. The first one is RWC music database for computer music research [5]. The database is a a collection of 100 audio CDs with full of copyright cleared contents for academic research purpose and individual sounds of musical instruments. Singing voices are stored as one category of musical instruments. It consists of three sopranos, three altos, three tenors, three baritones, three basses and three R&B singers' voices in various singing styles. The singers were instructed to sing the full range (as far as the singer can sing) of chromatic scale in Japanese five vowels in designated strength (*piano, mezzo forte* and *forte*), expression and styles (*staccato, vibrato* and *farsetto*).

The other database is a collection of very famous Japanese traditional vocal performance masters' voices [6]. Some of the master performers are designated as the Japanese living national treasure. They are asked to sing a common verse in their traditional singing styles. They are also asked to sing Japanese five vowels. In addition to these compulsory recordings, they recorded traditional songs also. The voices are stored in eighteen audio CDs. It uses the B&K's 1/2" omni-directional condensor microphone (Type 4190), calibrated in pressure field. Both were sampled at 44100 Hz 16 bit format.

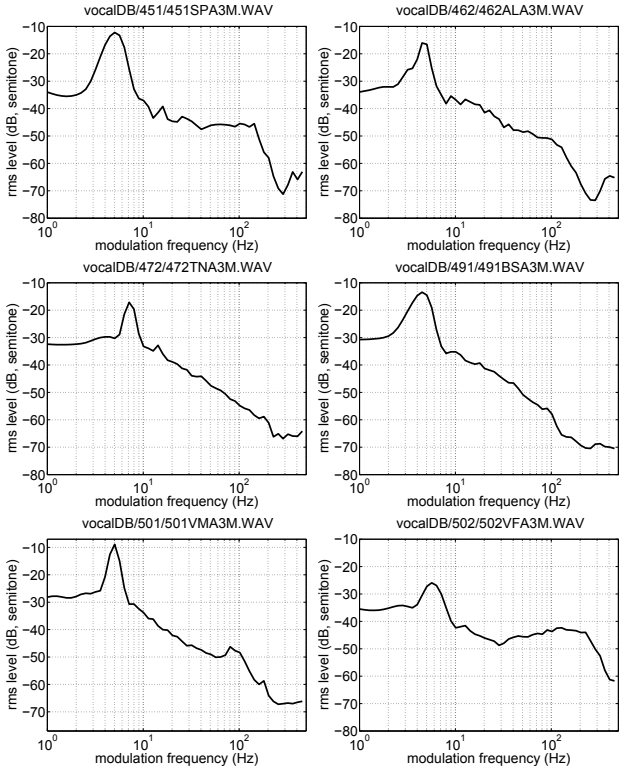


Fig. 3. Frequency modulation power spectra of F0 trajectories for chromatic scale singing of five Japanese vowels. From top left to bottom right, soprano, alto, tenor, bass, male and female R&B singers sang in *mezzo forte* with *vibrato*.

IV. Results

Figure 3 shows some of modulation power spectra of classical singers and R&B singers of RWC database. For classical singers, modulation levels decay in higher modulation frequency range. Peaks from 4 Hz to 7 Hz represent *vibrato*. From 10 Hz to 100 Hz, generally monotonic decaing trend is commonly found. The slope in this range distributes from -5 dB/decade to -20 dB/decade. [2] [3] Other classical singers also showed similar shapes.

For R&B singers, male and female singer's results are shown in the bottom plots. A peak of enhancement in the higher modulation frequency range seems to be common feature of these two singers. However, the third R&B singer showed similar characteristics with classical singers.

Figure 4 shows results excerpted from the traditional Japanese singer's data. In each row, left plot shows modulation power spectrum of sang vowels. The top plots

[2] Note that white noise shows $+10$ dB/decade slope.

[3] Also note that spectral shape in modulation frequency range higher than about 150 Hz is not very reliable due to possible distortion effects.

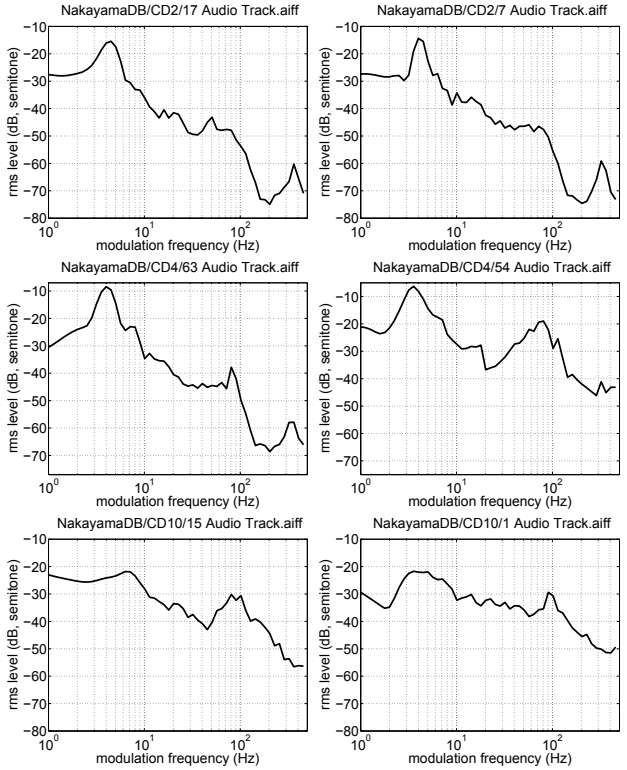


Fig. 4. Frequency modulation power spectra of F0 trajectories for sang five Japanese vowels (left) and common verse (right). Each raw corresponds to each singer.

are results of Japanese Buddist chant (Tendai sect). The top-right plot shows results of common verse sang in the address to the god style. A slight spectral peak (condensation) exists around 50 Hz to 70 Hz, but it is not very clear.

The peak is more salient in the middle and bottom plots. The middle plots show the results by the master player of Kanze school of Noh. The common verse was sang in transition from "tsuyogin" (strong chanting) to "yowagin" (weak changing). The bottom plots show the results by the master player of Kabuki. The common verse was sang in "utaigakari" (chanting in Noh-like style).

Several voices in this traditional Japanese song database were not be able to be analyzed because they are too irregular. Introduction of relevant analysis method for such voices is the next research topic.

V. Conclusion

Application of a new F0 extractor with temporally fine resolution to singing voice databases revealed details of higher frequency modulation power spectra which cannot be analyzed by the conventional F0 extractors. Expressive performance found in pop singing and Japanese traditional performances indicated faster frequency modulation (for example, higher than 50 Hz) sometimes characterize such extreme expressions.

Acknowledgement

This work was partly supported by Grants-in-Aid for Scientific Research category (B)24300073 and Exploratory Research 24650085. It is also by Wakayama University.

References

[1] I. Titze, "Principles of Voice Production," *Prentice Hall*, 1994.

[2] K. Sakakibara, H. Fuks, N. Imagawa, and N. Tayama, "Growl voice in ethnic and Pop styles," *Proc. Int. Symp. on Musical Acoustics, Nara, Japan*, 2004.

[3] H. Kawahara, M. Morise and K. Sakakibara, "Interference-free observation of temporal and spectral features in 'shout' singing voices and their perceptual roles," *Proc. SMAC SMC–2013, Stockholm, Sweden*, pp.256–263, 2013.

[4] H. Kawahara, M. Morise, R. Nisimura and T. Irino "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," *Proc. ICASSP 2013*, pp.6797–6801, 2013.

[5] M. Goto, "Development of the RWC Music Database," *Proc. ICA 2004, Kyoto Japan*, pp.I–553–556, 2004.

[6] I. Nakayama, "Comparative studies on vocal expression in Japanese traditional and western classical-style singing, using a common verse," *Proc. ICA2004, Kyoto Japan*, pp.1295–1296, 2004.

[7] A. de Chevengné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp.1917–1930, 2002.

[8] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp.1638–1652, 2008.

[9] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, 20(1), pp.102–105, 2013.

[10] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," *Proc. ICASSP 2011*, pp.5420–5423, 2011.

[11] A. H. Nuttall, "Some windows with very good side- lobe behavior," *IEEE Trans. Audio Speech and Signal Processing*, vol. 29, no. 1, pp.84–91, 1981.

[12] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, Hideki Banno, "TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation," *Proc. ICASSP 2008*, pp.3933–3936, 2008.

VOCAL TRACT SHAPING AND FORMANT FREQUENCIES IN SOPRANOS WHISTLE REGISTER

M. Echternach[1], P. Birkholz[2], L. Traser [1,3], M. Burdumy [1,4], R. Kammberger [5], B. Richter [1]
[1] Institute of Musicians` Medicine, Freiburg University Medical Center, Germany
[2]Department of Phoniatrics and Pedaudilogy, Aachen University Medical Center, Germany
[3]Department of Otolaryngology, Freiburg University Medical Center, Germany
[4]Department of Radiology, MRI Physics, Freiburg University Medical Center, Germany
[5]Institute of Microsystem Technology, Freiburg University, Germany

Abstract: **The resonatory properties of the vocal tract at very high soprano fundamental frequencies (F0) are not yet understood in detail. We analyzed a single professional soprano subject using 2D dynamic real-time MRI (24fps) and static 3D MRI in the pitch range of B5 to G6. From the 3D MRI data for pitch C6 and G6, formant frequencies were on the one hand measured from physical models created by 3D printing, and on the other hand calculated from area functions obtained from the 3D vocal tract shapes. Our analysis showed that there were only minor modifications of the vocal tract between B5 and G6. Also, the formant frequencies did not exhibit major differences between C6 and G6. Our investigation therefore was not able to confirm that there is still a formant tuning at these high fundamental frequencies.**

Keywords : **Whistle register-Vocal tract-MRI**

I. INTRODUCTION

It has often been assumed that singers produce high F0s above 1000Hz by a separate vocal mechanical principle, leading to the use of special vocal register terminology such as “whistle” or “flageolet” register. There have been different hypotheses concerning the voice production of sopranos’ whistle register. They included voice production by air vortices [1], vortices with vocal tract/voice source interaction [2] or flow modification by vocal fold oscillations with persisting gap [3,4]. However, a recent study in a single professional soprano subject revealed vocal fold oscillations with total glottal closure up to F0s of 1568Hz recorded with transnasal high speed digital imaging with 20.000 frames per second [5]. Here, the oscillatory properties did not change significantly between the pitches of C6, D6, E6, F6 and G6. However, there was a perceivable difference between D6 and E6 which was accompanied by the subject’s feeling of a register change from “head” to “whistle” [5]. The examiner that performed the endoscopy observed a strong narrowing of the pharynx from D6 to E6. The question arises if the perceptual sound changes from D6 to E6 might be explained by changes of vocal tract shape and associated resonances.

II. METHODOS

In this investigation we analyzed the same subject as described before [5] using two MRI protocols. In the first protocol the subject was analyzed using a dynamic real time MRI in coronar and midsagital two-dimensional view with a frame rate of 24 frames per second in supine position. Here, the subject was asked to sing a scale from B5 up to G6. In the second protocol the subject was asked to sustain the pitch of C6 and G6, respectively, for at least 11 seconds. Here the MRI was performed allowing three-dimensional construction of the entire vocal tract. Out of this material the vocal tract was segmented. In the dental clinic a dental laser scan was performed and the data were fusioned with the MRI segmentation. Out of this vocal tract model, including teeth, a three-dimensional rapid prototype was printed.
In order to estimate formant frequencies a micro loud speaker was brought to the vocal tract at the glottal level. White noise were introduced into the vocal tract. At the mouth level the modified signal was recorded.
Furthermore out of the segmented threedimensional MRI Material an area function was built and resulting formant frequencies were calculated.

III. RESULTS

As demonstrated by the dynamic real-time MRI the pharyngeal width was diminished between D6 and E6. This was mainly caused by a narrowing in the lateral/medial dimension, as shown in figure 1. Furthermore, the same effect was visible in the three-dimensional models of the vocal tract between the pitch

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

of C6 and G6 (figure 2). As a consequence, it could be assumed that formant frequencies would differ. In fact, an acoustic analysis of the printed vocal tract showed that F2 was slightly lower for C6. However, F1 and F3 was nearly stable with a frequency of about 1250Hz (figure 3). Also it was visible that for C6, H1 matches F1 and also H2 matches F2. However, for G6 F0 is much higher than F1 but H2 is in the region of F3. Formant estimations from the computed model did agree with the formant frequencies of F1 and F3 by a great amount. However, F2 was lower for this kind of calculation.

IV. Discussion

Our investigation shows that for the analyzed single subject, vocal tract shape differs between C6 and D6 and between E6 and G6. Since there were perceptual differences between D6 and E6 with the subject's feeling of a register change and without major changes in vocal fold oscillation patterns, as described in a previous investigation [5], the authors assumed that the perceptual difference is mainly caused by a crossing of the first formant by F0. Using broadband acoustic excitations, Garnier and coworkers were able to demonstrate different tuning situations for soprano singing at very high pitches [6]. However, the general tuning strategy of F0 matches F2, as described by these authors, could not be verified in our single subject. For high pitches it was found that H2 matches F3.

REFERENCES

1. Van den Berg,J.W. Vocal ligaments versus registers. *NATS Bulletin* **20**, 16-21 (1963).
2. Herzel,H. & Reuter,R. Whistle register and biphonation in a child's voice. *Folia Phoniatr Logop* **49**, 216-224 (1997).
3. Švec,J.G., Sundberg,J., & Hertegard,S. Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography. *J Acoust Soc Am* **123**, 347-353 (2008).
4. Keilmann,A. & Michek,F. Physiologie und akustische Analysen der Pfeifstimme der Frau. *Folia Phoniatr* **45**, 247-255 (1993).
5. Echternach, M., Döllinger, M., Sundberg, J., Traser, L., Richter, B.: Vocal fold vibration at high soprano fundamental frequencies. J Acoust Soc Am 133: EL82-87, 2013
6. Garnier,M., Henrich,N., Smith,J., & Wolfe,J. Vocal tract adjustments in the high soprano range. *J Acoust Soc Am* **127**, 3771-3780 (2010)

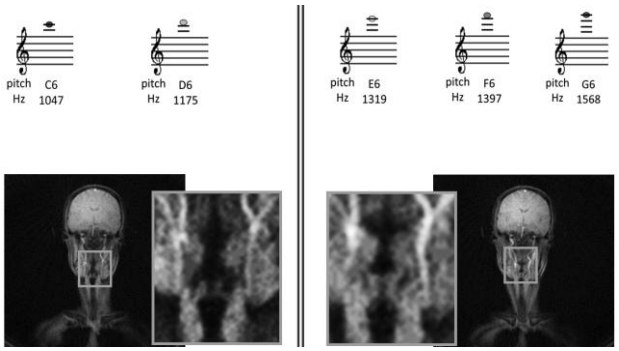


Figure 1: Representative pictures from real time MRI in the coronar plane. The enlarged windows demonstrate the narrowing of the pharynx in latero/medial dimension.

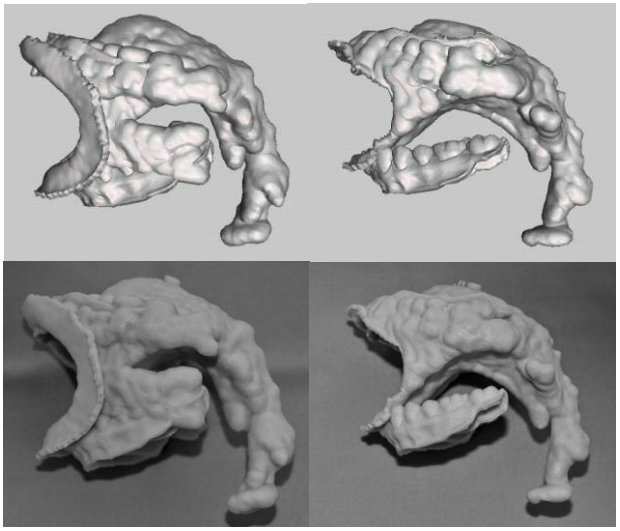


Figure 2: Segmented vocal tract models out of the MRI material (upper raw, left: pitch C6, and right: pitch G6) and photographs of the 3D print outs, (lower raw left: pitch C6, and right: pitch G6, respectively).

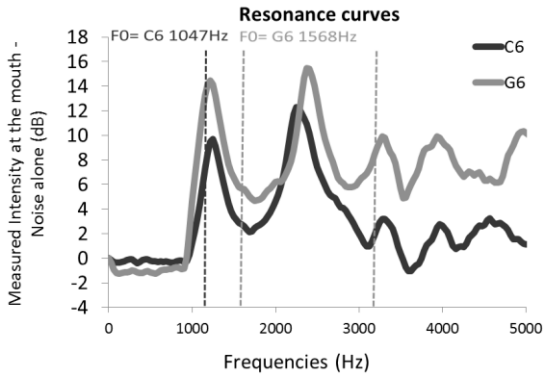


Figure 3: Spectrum of modified white noise brought into the vocal tract at the glottis and measured at the mouth. The black line represents the spectrum as recorded during phonation of C6 and the grey line as recorded during phonation of G6. The dashed line indicate the fundamental frequency.

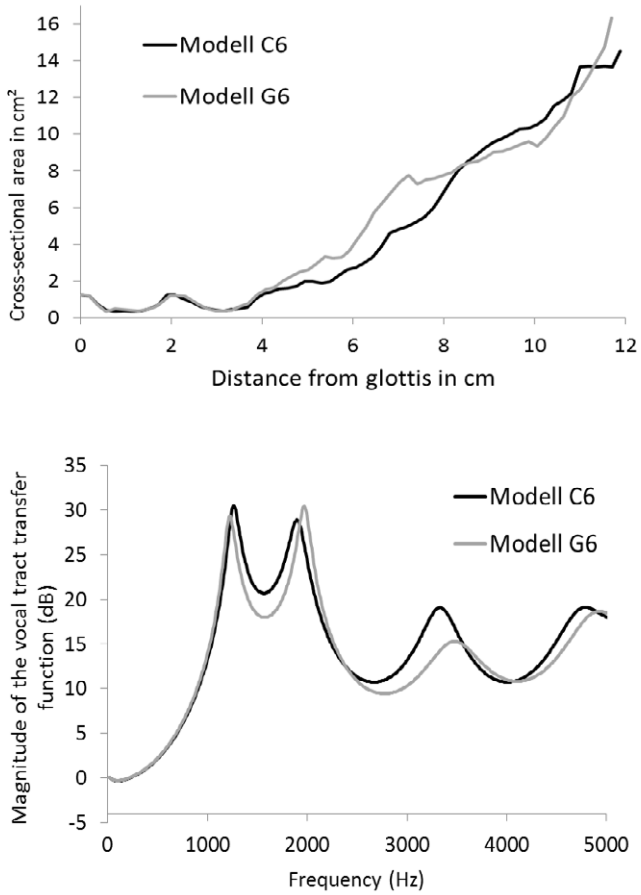


Figure 4: Area functions (top) and volume velocity transfer functions (bottom) calculated for models C6 and G6.

SINGING EXCISED HUMAN LARYNGES: RELATIONSHIP BETWEEN SUBGLOTTAL PRESSURE AND FUNDAMENTAL FREQUENCY

N. Hanna[1,2], N. Henrich[1], A. Mancini[3], T. Legou[4], X. Laval[1], P. Chaffanjon[1,3]

[1] Department of Speech and Cognition, GIPSA-lab, Université de Grenoble, France, noel.hanna@gipsa-lab.grenoble-inp.fr
[2] School of Physics, The University of New South Wales, Australia
[3] Laboratoire d'Anatomie Des Alpes Françaises, Faculté de Médecine de Grenoble, France
[4] Laboratoire Parole et Langage, Aix-en-Provence, France

Abstract: Ex vivo **studies of excised human larynges allow measurements of physical parameters that are important for our understanding of speech and cannot be otherwise determined from *in vivo* studies. This study focuses on the relationship between subglottal pressure and fundamental frequency of phonation, under several experimental conditions: two different air supply (pressure or flow), and three different control settings (pressure or flow, arytenoid compression, vocal-fold extension). A female human larynx was used.**

The results demonstrate a general linear behavior between subglottal pressure and fundamental frequency, with two different gradient regions (50 Hz/hPa or 1 Hz/hPa) depending on the range of downstream pressure. The nonlinear behavior of the human laryngeal system is well known and is illustrated here by pitch jumps and hysteresis cycles. The driving flow source enables the activation of a high-frequency vibratory mode between 600 and 1000 Hz, and greater dynamic ranges of subglottal pressure and sound pressure level.

Keywords : **Excised human larynges, pressure source, flow source, fundamental frequency, subglottal pressure, nonlinear laryngeal behavior**

I. INTRODUCTION

Human voice production is controlled by airflow supply interacting with laryngeal and vocal-tract configurations. Airflow provides an aerodynamic energy to the vocal system, through subglottal pressure and flow. The conversion of aerodynamic energy to mechanical and acoustical energy depends strongly on the biomechanical properties of the vocal folds, which is characterized by the fundamental frequency (f_0) of vibration. Subglottal pressure and f_0 are therefore two key source parameters in voice production.

The relationship between subglottal pressure and f_0 has been studied theoretically, e.g. [1], and experimentally (for a review, see [2]). Fundamental frequency typically rises from 1 to 7 Hz for an increase in pressure of 1 hPa [1,2], but values as high as 20 Hz/hPa can be found [2,3]. The relationship is nonlinear and it depends on the degree of vocal-fold elongation and adduction [2].

Over the past fifteen years, the subglottal pressure-f_0 relationship has been experimentally studied mainly on excised larynges [4], for the sake of a better experimental control of vocal-fold settings and due to the fact that direct measurement of subglottal pressure in humans is very invasive.

For reasons of practicality, excised larynx experiments are often performed on species other than humans [5]. Even if these non-human larynges may be used for assessing aerodynamic and acoustical models, they greatly differ with respect to biomechanical properties of the vocal folds, to laryngeal anatomy and to pressure-frequency behavior [5]. Furthermore, studies of excised dog, pig, sheep and cow larynges have noted that the supra-glottal tissues and structures in some animal models may cause a non-zero supra-glottal pressure, which may have an effect on the glottal resistance [6,7].

Recent measurements of phonation threshold pressure and flow on excised human larynges within 24 hours postmortem reported that phonation threshold pressures were typically lower than those measured in excised canine [8].

The relationship between subglottal pressure and f_0 in human voice production remains an open question. This preliminary study aims first to explore this relationship in excised human larynges. A second aim is to explore the impact of the type of airflow supply on the pressure-frequency relationship. Indeed, a constant flow source is a more physiological boundary condition below the glottis, which may impact voicing [9].

II. METHODS

This paper presents measurements made on a human female larynx, which was harvested and frozen within 48 hours postmortem, and thawed before the experiment. The larynx was made to self-oscillate by application of a source of pressurized air.

The specimen was secured to the experimental bench posteriorly and anteriorly at the cricoid cartilage. An

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

intubation tube was used to supply the source of pressurized air.

Crico-thyroid tension was maintained throughout the experiment by securing the epiglottis. The aryepiglottic folds were held separate throughout by the use of Beckman-Eaton laminectomy retractor.

In order to simulate the lateral crico-arytenoid muscle action used *in vivo*, the level of adduction/abduction of the vocal folds was modified by manual control of metal probes inserted into the muscular processes of the arytenoid cartilages.

The oscillation of the vocal folds was monitored by an Electroglottograph (EGG) (EG2-PCX2, Glottal Enterprises, Syracuse, New York, USA) with electrodes aligned with the plane of the vocal folds. The fundamental frequency of vibration (f_0) was estimated from the EGG signal using the YIN algorithm [10].

Subglottal pressure was measured by means of a specialized aerodynamic workstation (EVA2, S.Q.Lab, Aix-en-Provence, France) with a probe inserted into the crico-thyroid ligament.

A microphone was placed at a distance of 30 cm from the larynx in order to monitor the experiment and measure the sound pressure level (SPL).

The experimental setup is illustrated in Fig. 1.

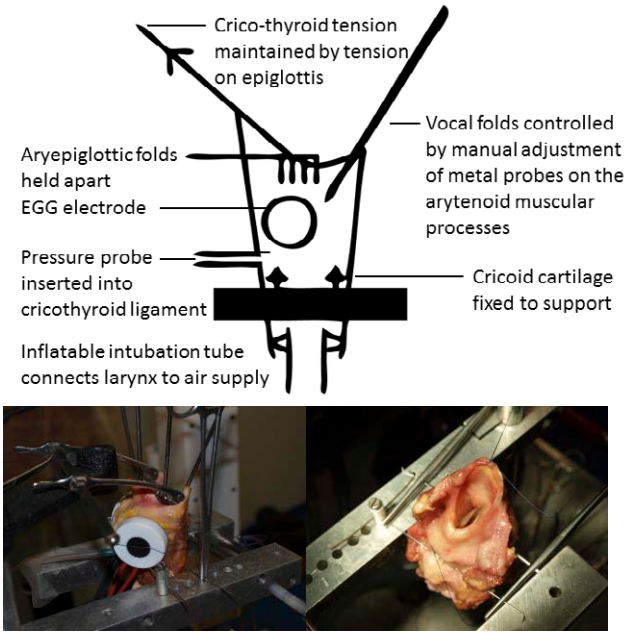


Figure 1. Schematic description of the experimental setup (top) and photographic views (bottom).

The study was conducted in two parts. In the first part (sequence 1) the oscillation in the larynx was driven by a pressure source (Mecafer, LT50 Compressor). The driving pressure was first varied while the larynx geometry was kept fixed (1A). It was then kept constant while either the lateral compression of the arytenoid cartilages was varied (1B), or the vocal fold extension was varied (1C). In the second part of the study (sequence 2), the oscillation was driven by a flow source. The driving flow was first varied (2A). It was then kept constant while either arytenoid compression (2B) or vocal fold extension (2C) were varied. The independent variables are summarized in Table 1.

Table 1. Experimental sequences

Sequence Number	Source Type	Independent Variable
1A	Pressure	Subglottal pressure
1B	Pressure	Compression of arytenoids
1C	Pressure	Vocal fold extension
2A	Flow	Subglottal flow
2B	Flow	Compression of arytenoids
2C	Flow	Vocal fold extension

III. Results

Larynx behavior

As shown in Fig. 2, the human female larynx was able to oscillate over a wide range of fundamental frequency, typically between 300 and 900 Hz. The corresponding dynamic range of SPL was approximately 40 dB, reaching a maximum of 90 dB at 30 cm. Subglottal pressures varied up to 40 hPa.

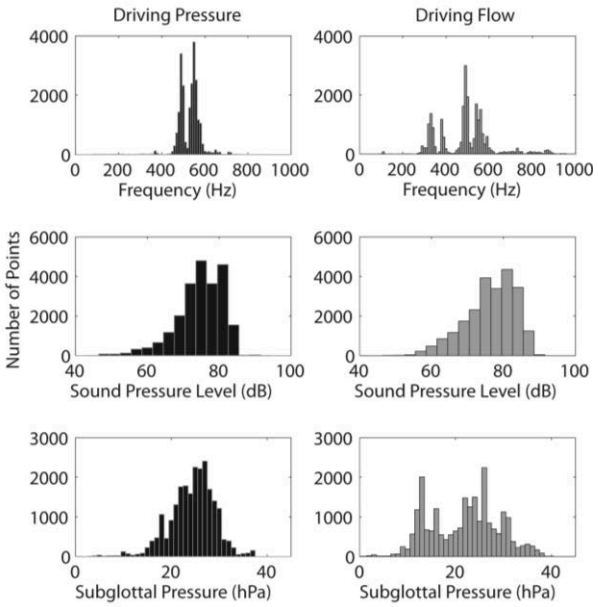


Figure 2. Histograms of fundamental frequency (top panels), SPL (middle panels), and subglottal pressure (bottom panels). Either a driving pressure source (left panels) or a flow source (right panels) were used. The bin sizes are 10 Hz for frequency, 3 dB for SPL and 1 hPa for subglottal pressure.

Fig. 3 plots the relationship between subglottal pressure and f_0 across all experimental conditions listed in Table 1. The results for the driving pressure source are shown in black, and the driving flow source are in grey.

When the larynx configuration is kept fixed (Fig. 3(A)), two distinct regions are highlighted with solid black lines showing the gradients. For subglottal pressure lower than 10-15 hPa, an increase of about 50 Hz/hPa is observed, for oscillatory frequencies f_0 between ~100 and 500 Hz (~A2-B4). For subglottal pressure higher than 10 hPa, the relationship shows a gradient of ~1 Hz/hPa.

A similar gradient of ~1 Hz/hPa is found when the laryngeal configuration is varied (Fig. 3(B) and 3(C)). Differences between pressure source and flow source are observed, which will be discussed later in the paper.

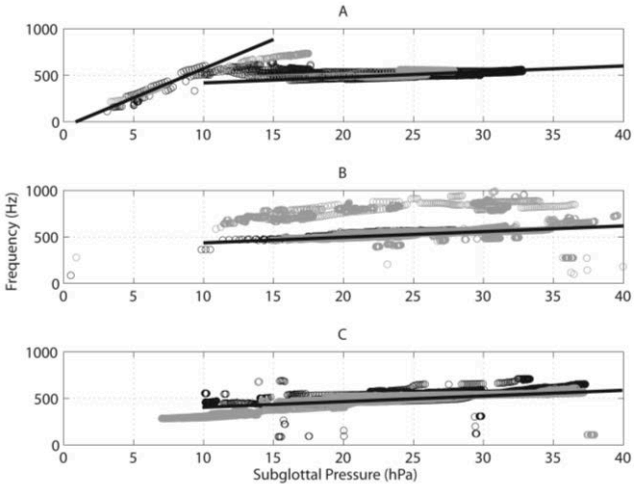


Figure 3. Relationship between subglottal pressure and f_0, with (A) variation of flow (grey) and pressure (black), (B) variation of arytenoid compression under constant flow (grey) and pressure (black), (C) variation of vocal fold extension under constant flow (grey) and pressure (black). Straight lines show approximate gradients.

Vocal fold nonlinearity

The variation of driving source (A), arytenoid compression (B) and vocal fold extension (C) all show a globally linear relationship between subglottal pressure and f_0 as shown by the straight lines in Fig. 3. However, the relationship is locally nonlinear, as it is interrupted by sudden pitch jumps. The pitch jumps were observed under all experimental conditions, as illustrated by samples in Fig. 4.

Pitch jumps occurred between 20 and 30 hPa, with a variability which may reflect small changes in the control parameters. A hysteresis effect is clearly demonstrated in each case, due to the nonlinear behavior of the laryngeal vibratory system. The gradients between subglottal pressure and f_0 calculated on either side of the jumps are of similar values.

In general, upward jumps occurred at higher frequencies than downward jumps. However, in some cases for the driving pressure source condition, the downward pitch jumps occurred at a higher frequency that the upward jump, as shown in Fig. 4(A).

Fig. 4(C) shows two successive pitch jumps under the driving flow source condition. A second pitch jump is observed above 30 hPa, a result which was also found under the constant pressure condition for similar elongation of the vocal folds.

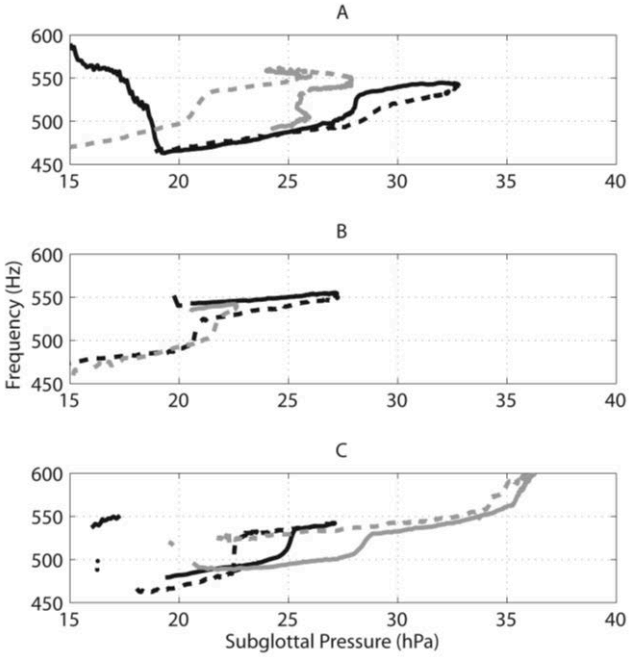


Figure 4. Examples of nonlinear behaviors and pitch jumps with (A) variation of flow (grey) or pressure (black), (B) variation of arytenoid compression under constant flow (grey) and pressure (black), (C) variation of vocal fold extension under constant flow (grey) and pressure (black). Solid lines show increases in the independent variable, dashed lines show decreases.

Driving pressure and driving flow sources

Fig. 2 illustrates two main differences between the experimental conditions. The first is that the flow source allows oscillation at lower subglottal pressures, with greater sound pressure level at frequencies in the normal range of phonation (<500 Hz). Furthermore as shown in Fig. 3(C) there is a slightly steeper gradient in the relationship between subglottal pressure and f_0 at these frequencies.

Secondly, the flow source activates an extreme high frequency mode of oscillation between 600 and 1000 Hz, as can be seen in Fig. 3(B) while the arytenoid compression was varied. In contrast to the behavior under the constant pressure condition (shown in black), two different modes of vibration are found at the same

subglottal pressure. The fundamental frequency of oscillation can occur either between ~500 and 600 Hz comparable with the pressure source case, or ~600-1000 Hz (musically ~D5 – C6), frequencies typically achieved only by sopranos.

This latter 'soprano' mode of oscillation was highly unstable and occurred in 15% of measurements with the flow source. Fundamental frequencies above 600 Hz were also observed in 5% of measurements with the pressure source but oscillation did not occur higher than 800 Hz.

These high fundamental frequencies were only observed towards the end of the experimental session. They are therefore possibly due to dehydration of the tissue over time and as such may not reflect the normal range of oscillation of the vocal folds.

IV. Discussion

In all experimental conditions a similar linear increase of f_0 with subglottal pressure above 10 hPa or 500 Hz with a gradient of ~1 Hz/hPa was observed in line with previous measurements [1,2]. In this region, there were typically one or two pitch jumps of ~40 Hz, which showed hysteresis with increasing and decreasing of the independent control parameter, similar to observations of nonlinearity under variation of vocal fold tension [11].

When the larynx geometry was fixed and the pressure or flow was varied, a steeper gradient of ~50 Hz/hPa was observed below 15 hPa, more than twice the values found in the literature [2,3]. However, the experimental data is scarce in this region, and it appears that there may be a difference in behavior over this range under the driving flow condition. The understanding of the behavior in this region calls for further research.

V. Conclusion

These preliminary data produced from a study of one female human larynx highlight differences in vocal fold behavior due to changes in the subglottal air supply.

Further work aims to improve the experimental setup to allow for quantifiable measurements of arytenoid movements and to extend the process to allow measurement of the downstream air flow.

References

[1] I.R. Titze, "On the relation between subglottal pressure and fundamental frequency in phonation," *J. Acoust. Soc. Am.*, vol. 85(2), pp. 901-906, 1989.

[2] F. Alipour, and R.C. Scherer, "On pressure-frequency relations in the excised larynx," *J. Acoust. Soc. Am.*, vol. 122(4), pp. 2296-2305, 2007.

[3] P. Lieberman, R. Knudson, and J. Mead, "Determination of the rate of change of fundamental frequency with respect to subglottal air pressure during sustained phonation," *J. Acoust. Soc. Am.*, vol. 45(6), pp. 1537–1543, 1969.

[4] J. Van den Berg, and T. Tan, "Results of experiments with human larynxes," *ORL*, vol. 21(6). pp. 425-450, 1959

[5] F. Alipour, and S. Jaiswal, "Phonatory characteristics of excised pig, sheep, and cow larynges,". *J. Acoust. Soc. Am.*, vol. 123 (6), pp. 4572–4581, 2008.

[6] F. Alipour, and S. Jaiswal, and E. Finnegan, "Aerodynamic and acoustic effects of false vocal folds and epiglottis in excised larynx models," *The Annals of otology, rhinology, and laryngology*, vol. 116(2), pp. 135-144, 2007.

[7] Alipour, F. and S. Jaiswal (2009). "Glottal airflow resistance in excised pig, sheep, and cow larynges." *Journal of Voice*, vol. 23(1), pp. 40-50.

[8] T. Mau, J. Muhlestein, S. Callahan, K.T. Weinheimer, and R.W. Chan, "Phonation threshold pressure and flow in excised human larynges," *The Laryngoscope*, vol. 121(8), pp. 1743-1751, 2011.

[9] M.S. Howe, and R.S. McGowan, "Voicing produced by a constant velocity lung source,", *J. Acoust. Soc. Am.*, vol. 133(4), pp. 2340–2349, 2013.

[10] A. De Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917-1930, 2002.

[11] J. G. Švec, H. K. Schutte, and D.G. Miller, "On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges," *J. Acoust. Soc. Am.*, vol. 106(3), pp. 1523-1531, 1999.

BIOMECHANICAL EVALUATION OF THE SINGING VOICE

Pedro Gómez-Vilda[1], Elisa Belmonte-Useros[2], Victoria Rodellar-Biarge[1], Víctor Nieto-Lluis[1], Agustín Álvarez-Marquina[1], Luis M. Mazaira-Fernández[1]
[1]NeuVox Laboratory, Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28223 Pozuelo de Alarcón, Madrid, Spain
[2]Escuela Superior de Canto, C/ San Bernardo 44, 28015, Madrid, Spain
e-mail: pedro@fi.upm.es

Abstract: **Teaching the adequate use of the singing voice conveys a lot of knowledge in musical performance as well as in objective estimation techniques involving the use of air, muscles, room and body acoustics, and the tuning of a fine instrument as the human voice. Although subjective evaluation and training is a very delicate task to be carried out only by expert singers, biomedical engineering may help contributing with well-funded methodologies developed for the study of voice pathology. The present work is a preliminary study of exploratory character describing the performance of a student singer in a regular classroom under the point of view of vocal fold biomechanics. Estimates of biomechanical parameters obtained from singing voice are given and their potential use is discussed.**

Keywords: **vocal fold modeling, singing performance, voice production, vocal effort**

I. INTRODUCTION

The correct use of the singing voice involves good musical knowledge, singing performance and the use of air, larynx characteristics, room, and body acoustics [1]. The subjective evaluation of these factors for advanced singing training is a task for expert singing professors. Unfortunately most of the times, professors have subjective auditory evaluation as the only tool available. No matter how fine the professor's skill is in evaluating the student's performance, it is always exposed to subjective appreciations which may not be as accurate as desirable. Biomedical engineering may help proposing techniques developed for the study of voice pathology [2] which may be adapted to singing voice quality analysis. The work presented here is an exploratory study motivated by the need of objectively estimating objectively aspects of singing expressed subjectively before. A collaboration between the NeuVox Lab and the Superior School of Singing of Madrid allowed the recording of real performances from students and professors of the school both at the study classroom and at the stage with the aim of evaluating the singer's 'stage fright'. The use of specific tools [3] in the estimation of aspects as tone, loudness, vocal fold biomechanics and glottal closure during different scales, has allowed depicting a colourful yet highly semantic picture of what is the singing voice. Estimations of real recordings and their preliminary statistical results are being presented and discussed [4]. This study must be seen as a due sequel of early works conducted in the NeuVox Lab some years ago [5], [6]. The ultimate goal of the study is to provide a methodology for the objective analysis of some characteristics of the singing voice to graduate the vocal effort of the singer, produce objective estimates of singer's performance in real time, and evaluate the emotional overload of the performer (stage fright).

II. BIOMECHANICAL PARAMETER ESTIMATION

The key technique used for the analysis of voice quality is adaptive vocal tract inversion to produce an estimate of the glottal source. Accurate spectral domain techniques [2] allow the estimation of a set of biomechanical parameters associated to a 2-mass model of the vocal folds [3] as the one depicted in Fig. 1.

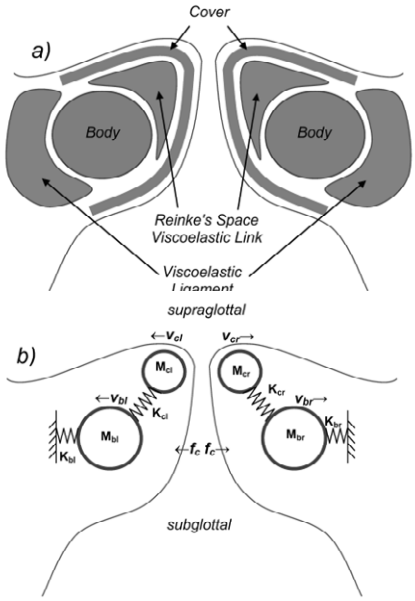


Fig. 1 Vocal fold 2-mass biomechanical model assumed in the study. a) Structural description of vocal folds. b) Equivalent model in masses and viscoelasticities.

The template (a) shows the classical body-cover structure of the vocal folds. The biomechanical model in (b) shows the cover masses M_{cl} and M_{cr} for the left (l) and

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

right (r) vocal folds. Masses M_{bl} and M_{br} are associated to the body. Visco-elastic parameters K_{cl} and K_{cr} explain the relations between tissue strain and acting forces on the cover. K_{bl} and K_{br} are the visco-elastic body parameters. The parameter set is composed of 68 features, including jitter, shimmer, NHR, and mucosal/aaw ratio (aaw: average acoustic wave) composing a total of 68. This is the parameter subset in the present study:

- Parameter 1: Fundamental frequency f_0 (pitch).
- Parameter 2: jitter relative, variation of fundamental frequency between each two near phonation cycles.
- Parameter 3: shimmer relative, variation of the glottal source area between each two near phonation cycles.
- Parameter 5: noise/harmonic ratio, ratio of contents associated to turbulent noise and harmonic lines.
- Parameter 6: mucosal/average acoustic wave ratio between the part of the glottal source contributed by glottal dynamics and the cycle average.
- Parameter 35: Dynamic mass associated to the body, average of M_{bl} and M_{br}.
- Parameter 37: Stiffness parameter associated to the body, average of K_{bl} and K_{br}.
- Parameter 38: Unbalance of dynamic body mass, relative variation between each two neighbor cycles.
- Parameter 40: Unbalance of body stiffness, relative variation between each two neighbor cycles.
- Parameter 41: Dynamic mass associated to the cover, average of M_{cl} and M_{cr}.
- Parameter 43: Stiffness parameter associated to the cover, average of K_{cl} and K_{cr}.
- Parameter 44: Unbalance of dynamic cover masses, relative variation between each two neighbor cycles.
- Parameter 46: Unbalance of cover stiffness between each two neighbor cycles.

The estimation of the biomechanical parameters requires the solution of an inverse problem given the power spectral density (psd) of the glottal source $s_r(t)$ as

$$\|S_r(\omega)\| = \left| \int_{-\pi}^{\pi} s_r(t) e^{-j\omega t} dt \right| \qquad (1)$$

relating glottal source psd with the vocal fold vibration. A cost function between the power spectral density and the transfer function of an electromechanical model of the vocal folds, given as $T(\omega)$ may be defined as

$$L(\omega,\mu,\xi,\sigma) = \int_{2\pi} \left(\|S_r(\omega)\| - \|T_c(\omega,\mu,\xi,\sigma)\| \right) d\omega \qquad (2)$$

where μ, σ and ξ stand for the estimates of the massive, viscous and elastic parameters of the biomechanical model ($R_{bl,r}$, $M_{bl,r}$, $K_{bl,r}$, $R_{cl,r}$, $M_{cl,r}$, $K_{cl,r}$) Assuming a single second-order functional as

$$T_c(\omega,\mu,\xi,\sigma) = \left[\left(\mu_c - \omega^{-1}\xi_c \right)^2 + \sigma_c^2 \right]^{-1} \qquad (3)$$

the process of optimization implies the fulfilling of the following conditions for the model parameters

$$\frac{\partial L}{\partial \mu_c} = 0; \quad \frac{\partial L}{\partial \xi_c} = 0; \quad \frac{\partial L}{\partial \sigma_c} = 0; \qquad (4)$$

Solutions for these conditions may be found either by forcing the derivatives of the functional L to zero deriving expressions for the three fitting parameters, or by adaptive gradient methods (see [2] for a further explanation). Estimates from each parameter on a balanced database of 50 male and 50 female normative speakers collected and evaluated by endoscopy at Hospital Universitario Gregorio Marañón de Madrid (Spain) are used as a normative database. Not all these parameters are equally relevant for this study. The relationship between the most relevant biomechanical parameters (body mass and stiffness, and their unbalances) and the classical acoustic ones (pitch, jitter and shimmer) is given in terms of Pearson's correlation coefficient in Table 1 (for male speakers) and Table 2 (for female speakers).

Table 1. Pearson's Correlation between biomechanical and acoustical parameters (male normative population). x_1: pitch; x_2: jitter; x_3: shimmer; x_{35}: body mass; x_{37}: body stiffness; x_{38}: body mass unbalance; x_{40}: body stiffness unbalance

P_{ij}	x_2	x_3	x_{35}	x_{37}	x_{38}	x_{40}
x_1	0,01	-0,13	**-0,97**	**0,99**	0,01	-0,01
x_2		**0,61**	0,12	0,12	**0,89**	**0,98**
x_3			0,25	-0,04	**0,60**	**0,63**

Table 2. Pearson's Correlation between biomechanical and acoustical parameters (female normative population). x_1: pitch; x_2: jitter; x_3: shimmer; x_{35}: body mass; x_{37}: body stiffness; x_{38}: body mass unbalance; x_{40}: body stiffness unbalance

P_{ij}	x_2	x_3	x_{35}	x_{37}	x_{38}	x_{40}
x_1	0,21	-0,15	**-0,95**	**0,96**	0,16	0,19
x_2		**0,55**	0,03	0,45	**0,88**	**0,97**
x_3			0,32	0,04	**0,59**	**0,59**

It may be seen that in both male and female normative population pitch (x_1) is strongly counter-correlated with body mass (x_{35}) as it should be expected, contrary to body stiffness (x_{37}) which shows large direct correlation with pitch. Jitter (x_2) is moderately correlated with respect to shimmer (x_3), but shows strong correlation regarding body mass unbalance (x_{38}) and body stiffness unbalance (x_{40}). Shimmer (x_3) is moderately correlated with body mass body mass unbalance (x_{38}) and body stiffness unbalance (x_{40}). A moderate level of correlation means that both parameters show some similarity although their information content is not completely redundant.

III. METHODS

Recordings of singing voice were taken in two different scenarios: at the classroom during singing lessons, and in the performing stage in front of the grading jury and general public. To ensure proper quality of voice and reduce interference from piano guidance or

reverberation effects, highly directional wireless chest microphones were used (Sennheiser ME4 clip-on condenser cardioid). Sampling rate was fixed at 96,000 Hz and 32 bits. A MOTU Traveller Firewire Audio Interface Recording System was used. The performers were students of the Superior School of Singing, 7 male and 4 female (2 bass, 3 baritones 2 tenors, 1 mezzo, 4 sopranos, ages ranging from 20-32 years). In the classroom they were asked to produce fifth/ninth scales articulating the five cardinal vowels (/a, e, i, o, u/). In stage auditions they sang classical pieces from repertoir. Recordings from one of the sopranos are used in the present exploratory study to show how biomechanical parameters grade singing effort and performance.

IV. RESULTS

Estimates of the parameters listed in section II over a fifth/ninth span produced by one of the sopranos is given in Fig. 2 (at the end of the paper). All parameter estimates have been normalized to their means from a general normative database of 50 female subjects inspected by Hospital Universitario Gregorio Marañón of Madrid. It may be noticed that some parameters show almost no influence with the tone change, as the *body mass* (x_{35}), whereas others as the *body mass unbalance* (x_{38}) show important changes along the tone scale. As it may be seen in the leftmost column *absolute pitch* (x_1) follows closely the expected tonal scale, first raising, then sloping down along the fifth, thence repeating the same pattern on a larger span for the ninth scale. Pitch accuracy behaves as $(f_0)^2/f_s$, where f_0 is pitch and f_s is sampling frequency. For the larger tone produced (D_5, f_0=1174.66 Hz) the accuracy is 14.37 Hz, whereas for the lowest tone (C_4, f_0=523.25 Hz) it is around 2.85 Hz. In the worst case the accuracy would be equivalent to one eight of tone.

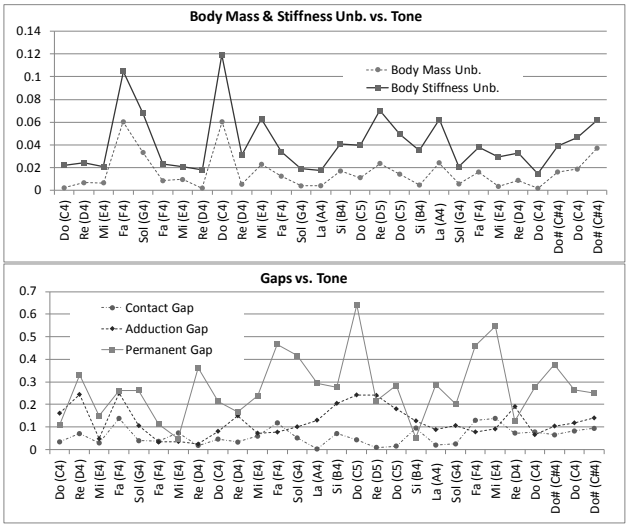


Fig. 3. Top) Details of body mass and stiffness unbalances (x_{38}) and (x_{40}). Bottom) Glottal gap defects (permanent, contact, adduction).

Fig. 3 (top) gives details about two relevant merit factors, which are the unbalances of the *body mass* (x_{38}) and *stiffness* (x_{40}), evaluated as the relative differences between neighbour cycles. Large values of these parameters reveal asymmetric irregularities in tonal passages where performers experience phonation difficulty, as F_4 in the ascending fifth and C_4 in the ligature between both scales. Fig. 3 (bottom) gives other important merit factors as glottal gap defects, defined as the improper opening found during vocal fold contact (contact gap defect), the lack of complete closure all over the phonation cycle (permanent gap defect), and the improper fluctuations during the closing phase (adduction gap defect). Permanent gap maybe very relevant for singing, as it measures the amount of constant opening found in the glottis, giving an estimation of air use efficiency. The larger the permanent gap the larger the air escape and the lower the air use efficiency. It may be seen that permanent gap is especially large in certain tones as C_5 and E_4 (ninth).

V. DISCUSSION

The results of the study avail the definition objective measurements for singing voice performance based on the biomechanical description of the vocal folds. Due to the limitations of the present study based in the description of a single performer, statistical significance cannot be claimed. Nevertheless some interesting important findings may be remarked, as a close following of the performance tuning to be used by the student and professor during the classroom in real time, as well as measurements of vocal effort (not shown), estimates of vocal fold mass and especially stiffness to provide a clear hint to voicing performance, or the biomechanical unbalances, especially those affecting stiffness which can be used as marks to voicing deficiencies to be corrected. Specific relevance should be attributed to glottal gap defects, with special emphasis in the permanent defect, as a mark of improper air usage.

VI. CONCLUSIONS

The results of the study fulfil the objectives stated in section I, as producing objective measurements of singing voice performance based on the biomechanical description of the vocal folds. Some important facts may be remarked from the analysis of the data shown above:

- A close tracking of the performance tuning can be estimated and presented to the student and professor at the classroom in real time.
- Measures of vocal effort can be provided similarly.
- Estimates of vocal fold mass and especially stiffness may provide a clear hint to voicing performance, particularly as statistical dispersion is concerned.

- Biomechanical unbalances, especially those affecting stiffness could be eventually used to marks to voicing deficiencies to be corrected using classical voicing techniques in singing.
- Specific relevance should be attributed to glottal gap defects, with special emphasis in the permanent defect, as a mark of improper air usage.

Many other estimates can be obtained and included in a biomechanical study of singing voice, such as the distribution of the harmonic/noise ratios, the open, close and return quotients, or the parameters of tremor and vibrato. These would be especially relevant to investigate and characterize the stage fright, one of the objectives of the study in the long run. The next steps to be covered are to extend the analysis to the group of singers already recruited in the database to evaluate the statistical significance of this approach.

Acknowledgments: This work is being funded by grants TEC2009-14123-C04-03 and TEC2012-38630-C04-04 from Plan Nacional de I+D+i, Ministry of Economy and Competitiveness of Spain. Special thanks are due to Drs. Ramírez, Scola, and Poletti from Hospital Universitario Gregorio Marañón of Madrid.

REFERENCES

[1] Mürbe, D., Pabst, F., Hofmann, G., and Sundberg, J., "Effects of a professional solo singer education on auditory and kinesthetic feedback—a longitudinal study of singers' pitch control", *Journal of Voice*, Vol. 18, No. 2, 2004, pp. 236-241.

[2] Gómez, P., Fernández, R., Rodellar, V., Nieto, V., Álvarez, A., Mazaira, L. M., Martínez, R, and Godino, J. I., "Glottal Source Biometrical Signature for Voice Pathology Detection", *Speech Comm.*, Vol. 51, 2009, pp. 759-781.

[3] Titze, I. R. and Story, B. H., "Rules for controlling low-dimensional vocal fold models with muscle activation", *J. Acoust. Soc. Am.*, Vol. 112, No. 3, 2002, pp. 1064-1076.

[4] Gómez, P., Rodellar, V., Nieto, V., Martínez, R., Álvarez, A., Scola, B., Ramírez, C., Poletti, D., and Fernández, M., "BioMet®Phon: A System to Monitor Phonation Quality in the Clinics", *Proc. eTELEMED 2013,* Nice, France, 2013, pp. 253-258.

[5] Gómez, P., "Biomechanical Evaluation of Vocal Fold Performance in Singing Voice", Lecture at The Voice Foundation's 37th Annual Symposium 2008: Care of the Professional Voice - The Westin, Philadelphia, PA, May 28 - June 1 (2008)

[6] Murphy, K., "Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice", PhD. Thesis, Universidad Politécnica de Madrid, 2008. http://oa.upm.es/1079/1/katharine_murphy.pdf.

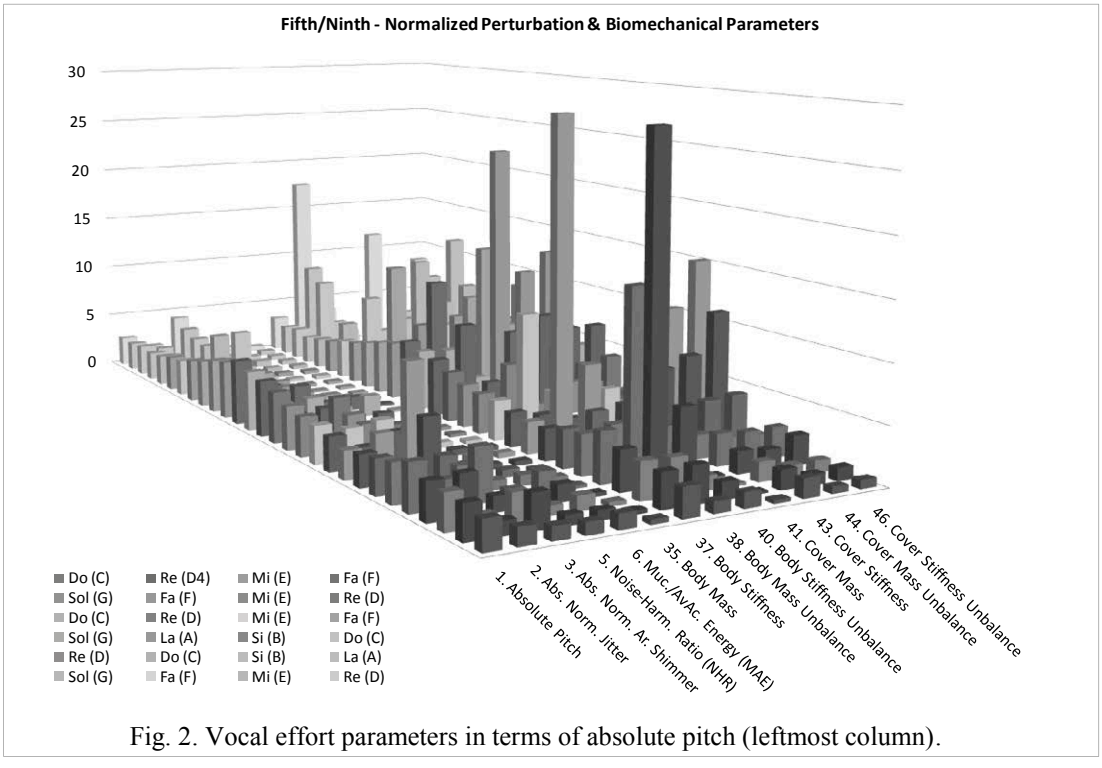


Fig. 2. Vocal effort parameters in terms of absolute pitch (leftmost column).

« The Use of software Overtone Analyzer for analyzing vocal emissions »

TRAN QUANG HAI

CREM – Centre de recherche en ethnomusicologie, CNRS , Paris, France.

tranquanghai@gmail.com

Abstract : **Sygyt Software was founded in Germany in 2003 by Bodo Maass (acoustician engineer) and Wolfgang Saus (famous overtone singer) to explore the creation of tools that help people to become better musician and to realize the full potential of their voice.**

Overtone Analyzer is a software application for the interactive recording and exploration of sounds . The visual display of a sound enables the quick recognition of the fundamental melody, the sound color (timbre) and the overtones. It also makes it easy to visually compare audio files . Overtone analyzer is particularly suited as a feedback tool to practice singing, and to document vocal development over the course of a voice education or therapy .

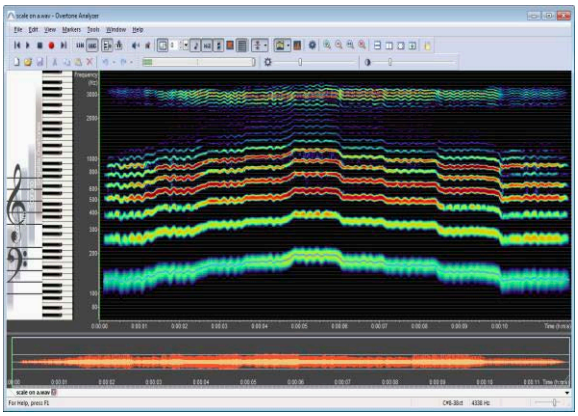
During the presentation, many live examples given by the presenter will show voices in different timbres, registers (Tibetan chanting, Peking Opera, Projected throat voice, Inuit throat game, and Mongolian overtone singing), and some training methods showing how to write words with voice, and also how to help patients to improve vocal illness

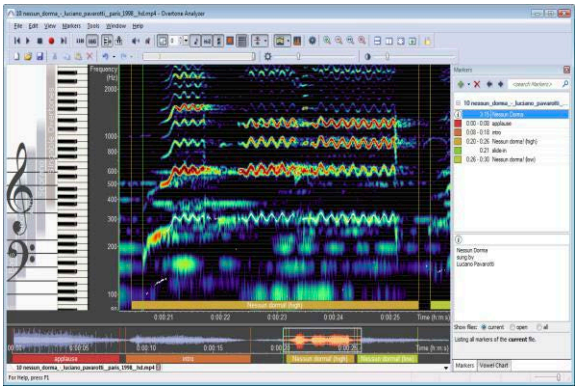
Keywords: **overtone analyzer, throat voice, overtone singing**

INTRODUCTION

Sygyt Software (3) was founded in Germany in 2003 by Bodö Maass (acoustician engineer) and Wolfgang Saus (famous overtone singer) to explore the creation of tools that help people to become better musician and to realize the full potential of their voice.

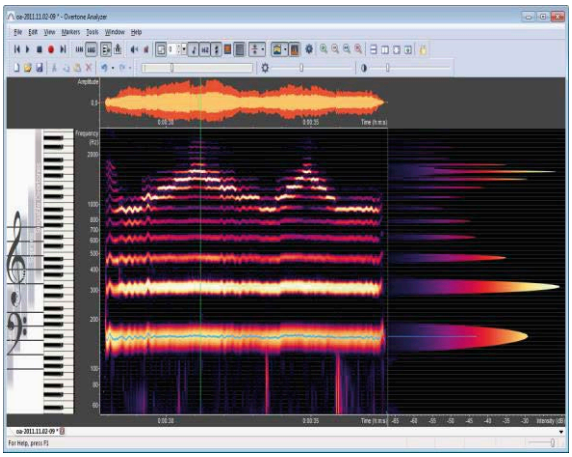
Overtone Analyzer is a software application for the interactive recording and exploration of sounds . The visual display of a sound enables the quick recognition of the *fundamental melody*,





the *sound color* (timbre)

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)



and the overtones. It also makes it easy to visually compare audio files . Overtone analyzer is particularly suited as a feedback tool to practice singing, and to document vocal development over the course of a voice education or therapy .

THE USE OF OVERTONE ANALYZER FOR ANALYZ ÌNG DIFFERENT VOCAL TRADITIONS

PEKING OPERA

In Chinese opera, breath is based in the pubic region and supported by the abdominal muscles. Performers follow the basic principle that "Strong centralized breath moves the melodic-passages" (*zhong qi xing xiang*). Breath is visualized being drawn up through a central breathing cavity extending from the pubic region to the top of the head. This "cavity" must be under the performer's control at all times, and he or she develops special techniques to control both entering and exiting air. The two major methods of taking in breath are known as "exchanging breath" (*huan qi*) and "stealing breath" (*tou qi*). "Exchanging breath" is a slow, unhurried process of breathing out old air and taking in new. It is used at moments when the performer is not under time constraint, such as during a purely instrumental musical passage or when another character is speaking. "Stealing breath" is a sharp intake of air without prior exhalation, and is used during long passages of prose or song when a pause would be undesirable. Both techniques should be invisible to the audience and take in only the precise amount of air required for the intended vocalization. The most important principle in exhalation is "saving the breath" (*cun qi*). Breath should not be expended all at once at the beginning of a spoken or sung passage, but rather expelled slowly and evenly over its length. Most songs and some prose contain precise written intervals for when breath should be "exchanged" or "stolen".

INUIT THROAT GAME

Inuit throat-singing is not singing *per se*. Ethnomusicologists suggest that it should be viewed as vocal games or breathing games more than anything else. Traditionally, they are considered 'games in which one makes noises', as the Inuit would say. Because of the way they use the voice, the throat, deep breathy sounds, rhythms, as well as its similarity to Mongolian and Tuvan throat-singing, it is now called throat-singing. It appears that the main reason why ethnomusicologists suggest to call them vocal games is that they do not use only the throat, they also use regular voice. Traditionally, they are games the women employed during the long winter nights to entertain the children, while the men are away hunting (sometimes for up to a month or more). As already mentioned, they are generally done by two persons, but sometimes we can find four or more performers singing together.

Inuit throat-singing (1) is done the following way: two women face each other; they may be standing or crouching down; one is leading, while the other responds; the leader produces a short rhythmic motif, that she repeats with a short silent gap in-between, while the other is rhythmically filling in the gaps. The game is such that both singers try to show their vocal abilities in competition, by exchanging these vocal motives. The first to run out of breath or be unable to maintain the pace of the other singer will start to laugh or simply stop and will thus loose the game. It generally last between one and three minutes. The winner is the singer who beats the largest number of people.

Originally, the lips of the two women were almost touching, each one using the other's mouth cavity as a resonator* **1** . Today, most singers stand straight, facing one another and holding each other's arms. Sometimes they will do some kind dance movements while singing (e.g., balancing from right to left). The sounds used include voiced sounds as well as unvoiced ones, both through inhalation or exhalation. Because of this, singers develop a breathing technique, somewhat comparable to circular breathing used by some players of wind instruments. In this way, they can go on for hours.

TIBETAN CHANTING

The formantic voice is the voice using specific overtones to create a melody or a fixed pitch upon

the fundamental (the case of Mongolian and Tuvin xöömij singing style and Tibetan chanting). In this aspect noticed in Mongolian and Tuvin throat song xöömij , a singer creates a constant pitched fundamental considered as a drone, and at the same time, modulate the selected overtones to create a formantic melody from Harmonic 4 (H4) till Harmonic 16 (H16), depending the range of the song. For the Tibetan Buddhist chanting, the fixed fundamental and the fixed overtone (especially Harmonic 10 (H10) are the characteristics of one of the prayers recited by monks belonging to the two tantric colleges Gyütö and Gyüme of the Gelugpa monastery .
Until the middle of the 20th century, Western accounts of Tibetan Buddhism often made it seem mystifying or bizarre. The exile of several thousands of Tibetan Buddhist monks after the Chinese invasion in Tibet enabled the Westerners to have a better understanding of Tibetan Buddhism. For the Gyütö and Gyüme tantric universities, the chant master umze, with his deep voice, produces the sound rich in overtones, trying to create the overtone number 10 (H10 for the Gyütö school with the vowel (ô), or the overtone number 12 (H12 for the Gyüme school with the vowel (ö).

TUVIN AND MONGOLIAN THROAT SINGING

MONGOLIA

It is believed the art of overtone singing has originated from south western Mongolia in today's Khovd and Govi-Altai region. Nowadays, overtone singing is found throughout the country and Mongolia is often considered as the most active place of overtone singing in the world.[2] The most commonly practiced style, Khöömii can be divided up into the following categories (2)

- uruulyn / labial khöömii
- tagnain / palatal khöömii
- khamryn / nasal khöömii
- bagalzuuryn, khooloin / glottal, throat khöömii
- tseejiin khondiin, khevliin / chest cavity, stomach khöömii
- turlegt or khosmoljin khöömii / khöömii combined with long song

Mongolians also sing many other styles such as "karkhiraa" (literally "growling") and "isgeree".

Many of these styles are also practiced around neighboring regions such as Tuva and Altai.

TUVA

Tuvan overtone singing is practiced by the Tuva people of southern Siberia. The history of Tuvan overtone singing reaches very far back. There is a wide range of vocalizations, including Sygyt, Kargyraa (which also uses a second sound source), Khoomei, Chylandyk, Dumchuktaar, and Ezengileer. Most of these styles are closely related to the styles and variations in neighboring Mongolia.

WRITING WORDS WITH OVERTONES

It is possible to use overtones to write some words . An example with the word MINIMUM is shown

UNDERTONES

There are two techniques of singing undertones : making ventricular bands vibrate and creating two false muscles on arytenoid cartilage. With this techinique , one can make a deep sound with one octave lower than the fundamental. It is possible to go to F-3 (that means one octave and a fifth below the fundamental). With this technique of undertones , I have used to help the people who had the throat cancer to create a pathological voice without any surgical operation.

During the presentation, with the help of the software Overtone Analyzer many live examples given by the presenter will show voices in different timbres, registers (Tibetan chanting, Peking Opera, Projected throat voice, Inuit throat game, and Mongolian overtone singing), and some training methods showing how to write words with voice, and also how to help patients to improve vocal illness .

CONCLUSION

The use of the software OVERTONE ANALYZER shows convincing spectrograms to help the understanding of different vocal traditions with overtones and experimental researches in voice.

REFERENCES

[1] Q.H. Tran, N. Bannan, *" Vocal Traditions of the World: Towards an Evolutionary Account of Voice Products in Music"*, Oxfore University Press, pp. 142-172, 2012

[2] H.Zemp, Q.H.Tran, *"The Song of the Harmonics"*, CNRS Audio-Visuel, Paris, DVD 38 minutes, 2006.

[3] B.Maass, W.Saus, *"Software of Overtone Analyzer"* , http://sygyt.com, Germany

HEAVY METAL "GROWL" PHONATION. ANALYSIS OF SUPRA-GLOTTIC & GLOTTIC VIBRATORY PATTERNS DERIVED FROM HIGH-SPEED DIGITAL IMAGING

K. Izdebski[1,3], E. Di Lorenzo[2,1] and Y. Yan[3]

[1] Pacific Voice and Speech Foundation, Chairman, San Francisco, CA, USA, kizdebski@pvsf.org
[2] Private Practice, Roma, Italy, enricohdilorenzo@gmail.com
3 Department of Bioengineering, Chair, Santa Clara University, Santa Clara, CA, USA
Institution/Department, yyan1@scu.edu

Abstract: **Heavy Metal Growl (Gr) vocal quality production was investigated using High Speed Digital Imaging (HSDI) and Distant Chip Scope (DCS) with stroboscopic illumination submitted to a visual inspection. Results showed that Gr phonation is produced by supraglottic vibrations with the true vocal fold edges touching only occasionally and with the glottis staying open during most of the Gr production time. Results also showed simultaneous multi-periodic complex vibrations of the many of the supraglottic structures including false vocal folds, arytenoid mucosal caps and aryepiglottic folds. The results revealed that the synchronized absence of the glottic closure with the vibratory action of the supraglottis prevents injuring the mucosa of the true vocal folds and eliminates phonotrauma. Blurry images were obtained via DCS strobe illumination, while superbly clear images were obtained via HSDI.**

Keywords: **Heavy metal growl voice quality, High Speed Digital Imaging, Distal Chip Scope, stroboscopy, supraglottic activity, vocal-fold vibration, aperiodicity, phonotrauma.**

I. INTRODUCTION

Close to 35 millions of heavy metal (HM) music units have been sold world-wide. Of interest this singing style have been criticized by politicians and ignored by researchers. To correct the gap in understanding how this voice is produced, we initiated a comprehensive visual investigation of the supraglottic vocal tract during HM singing. Although we studied many expressions of HM singing, here we limit our discussion to the so-called "Growl" (Gr) vocal quality, as Gr is a staple of HM performance. To investigate Gr physiology, we employed High Speed Digital Imaging (HSDI) and Distant Chip Scope (DCS) with strobe capacity (Olympus) synchronized with acoustic signals (AS). This allowed us to investigate not only the vocal fold (VF) vibratory properties but also the supraglottic activity (SGA). Data acquired, were submitted to visual analysis. Using this paradigm we defined mechanisms responsible for HM vocalization and showed that the acoustically abusive and aggressive Gr voice quality (VQ) is produced without true vocal folds participation. Hence, proper Gr phonation does not introduce phonotrauma to the true vocal folds (TVF) and therefore prevents dysphonia.

II. METHODS

HSDI recordings from an experienced male HM performer were obtained trans-orally using: 1: a 90 degree rigid scope connected to KayPENTAX Color High-Speed Video System (CHSV), Model 9710, and 2: trans-nasally, using a DCS hooked up to a strobe source by Olympus America Inc.3500 Corporate Parkway, Center Valley, PA 18034. The scope we used was an ENF-VH HD Distal chip scope (3.9mm outer diameter). The LED strobe used was Olympus Model # CLL-S1.The Processor we used was a CV-S190 Visera Elite. All acquired signals were analyzed visually. All HSDI signals were processed also by the special software (Yan & Izdebski), but these results will be reported in the different paper.

III. RESULTS

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

HSDI provided much more detailed images of the area of interest (AI) than those displayed via DCC. From these images we were able to distinguish the role of the specific anatomical structures of the glottis and supraglottis (SGA) in the production of Gr.

Glottis:
HSDI showed specifically that during Gr phonation true vocal fold (TVF) edges touch only occasionally with glottis staying open for most of the production time. When present, the TVF vibration was marginal and the glottis was found un-approximated for most of the Gr phonation time (See Figure 1).

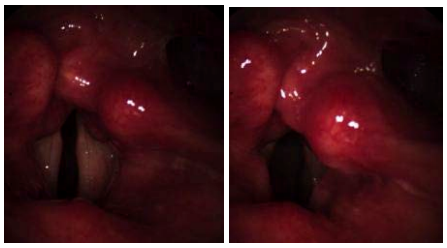


Figure 1

SGA:

SGA Gr activity was expressed by all components of the supraglottis. These included participation of the false vocal folds (FVF), arytenoid mucosal caps, aryepiglottic folds vibratory actions including epiglotttis and all of these appeared to work as a single structure. However, most of the vibratory activity originated from the FVF and propagated onto the arytenoid caps and the aryepiglottic folds.

Both TVF and SGA showed simultaneous different sets of mucosal waves, some characterized by higher frequency, smaller amplitude and shorter propagation and some by a lower frequency, larger amplitude and longer propagation. (See Figure 2)

Figure 2

These combined vibrations are accountable for creating multi-periodic, or in fact very complex aperiodic vibrations that characterize the Gr voice quality.

Kymographic segmentation of this vocal pattern was very difficult to obtain, hence is not reported here. Automatic tracing of the edges of either the TVC or the vibrating SGA area also posed significant difficulties due to this highly aperiodic behavior, hence will be a subject of a different (next) reports.

However, we noticed that the particular vibratory pattern of SGA during Gr phonation is quite different from the sphincter-like activity that we find in dysphonic phonation that employs SGA (Izdebski 2013). In fact we observed different sets of “mucosal waves” generated by the SGA, and we noted essentially no contact of the vibrating edges of the true VC. Also, there were no signs of hard glottic attack in Gr phonation.

The obtained mobile images obtained from HSDI and from DCS are shown in the accompanied videos clips.

IV. Discussion

HM singing is characterized by an aggressive acoustic sound that evokes perception of severe vocal abuse. Yet, when listening to the modal r even high voice of these performers, the vocal acoustic signal is perceived as “normal”. Our study explains this dichotomy, by demonstrating with the help of HSDI and DCC the mechanism of Gr phonation. The Gr phonation is

characterized by multiperiodic or aperiodic mucosal activity derived from multiple supraglottic sources rather than form the activity of the true VFs. This situation can not be assessed adequately via traditional stroboscopy including DCC stroboscopy.

We interpret the chaotic vibratory activity of the supraglottis as reflective of the absence of the hard glottic attack, which is an abusive vocal component to the activity of the TVC. The maintenance of the glottic gap during Gr shifts indicates that the subglottic air pressure from the glottic to supraglottic structures is keep in a safety range and permitting the SGS to vibrate at the high air flow. The results showed that synchronized absence of glottic closure at the TVF level with supraglottic constriction generates vibratory patterns with SGA tissues. This in turns generates this aggressive voice quality without injuring the mucosa of the true VC.

V. CONCLUSION

As far as we are aware, this presentation constitutes the first ever HSDI and DCS based analyses of HM phonation expressed by Gr voice quality. The results showed definitive and specific character of SGA and VF vibration responsible for the production of this unusual VQ referred to as growl (Gr). HSDI showed specifically that during Gr phonation TVF edges touch only occasionally with glottis staying open for most of the production time letting the high airflow that activates vibrations of the compressed SGS. Produced in this way Gr phonation, despite its loud and aggressive perceptive quality, is safe for the TVF. Automatic segmentation of the GW using any of the current known algorithms (Yan & Izdebski, 2013) failed to objectively designate sequential pattern of phonatory tissues due to its extremely chaotic vibratory patterns.

ACKNOWLEGEMENTS

Supported by the intramural funding from the PVSF. We appreciate editorial comments of Dr. R.R. Ward, and the availability of Olympus System provided on pro-bono bases by Mr. Brian Kent, Territory Manger, SF North ENT Business Division of Olympus America Inc. 3500 Corporate Parkway, Center Valley, PA 18034

VI. REFERENCES

Di Lorenzo, E. Heavy Metal. A video presentation of fiberoptic findings. Paper presented at the XX Pacific Voice Conference, SCU, Santa Clara CA, 2010.

Izdebski K,. Non-glottic laryngeal phonation in the traumatically injured larynx, e-Phonoscope, in press, 2013.

Izdebski K. Yan, Y. (Eds.) Visualization of the glottis with HSDI, OCT and NBI. In press an e-Q&A-p, a PVSF Publication, San Francisco, San Jose, CA, USA, 2013.

Yan Y and Izdebski K. Software for processing HSDI. Paper presented at the SPIE – Photonics West, February 2013 Conference, San Francisco, CA, USA

SINGLE LINE SCANNING OF VOCAL FOLDS AS FEEDBACK IN SINGING: THE 'MESSA DI VOCE' EXERCISE

P.H. Dejonckere [1], J. Lebacq [2], L. Bocchi[3], C. Manfredi [3]

[1] University of Leuven, Neurosciences, Exp. ORL, Belgium ; Federal Institute of Occupational Diseases, Av. De l'Astronomie, 1, B-1210 Brussels Belgium.
[2] Institute of Neuroscience CEMO, Université Catholique de Louvain, Avenue Hippocrate 55 bte B1.55.12 B-1200 Brussels, Belgium.
[3] Department of Information Engineering, Università degli Studi di Firenze, Via S. Marta 3, 50139 Firenze, Italy.

Abstract: **This article describes a novel application of the 'single line scanning' of the vocal fold vibrations (kymography) in singing pedagogy, particularly in a specific technical voice exercise: the 'messa di voce'. It aims at giving the singer relevant and valid short-term feedback. An user-friendly automatic analysis program makes possible a precise, immediate quantification of the essential physiological parameters characterizing the changes in glottal impedance, concomitant with the progressive increase and decrease of the lung pressure. The data provided by the program show a strong correlation with the handmade measurements.**
Additional measurements as subglottic pressure and flow glottography by inverse filtering can be meaningfully correlated with the data obtained from the kymographic images.

Keywords: **Messa di voce, single line scan, VKG, inverse filtering, subglottic pressure, singing.**

I. INTRODUCTION

This article describes a novel and original application of the 'single line scanning' of the vocal fold vibrations (kymography) in singing pedagogy. Short-term feedback for relevant physiological parameters of voice production may be very useful for the singer, particularly in acquiring specific technical skills based on motor control, e.g. for producing a 'messa di voce'.

Single line scanning of vocal fold vibrations (kymography, or videokymography: VKG) [1] is an imaging method based on a special digital camera, which can operate in two different modes: standard and high-speed. In the standard mode, the camera provides images displaying the whole vocal folds at standard video frame rate (30/25 frames/s, with 720x486/768x576 pixel resolution). In the high-speed mode, the video camera delivers images from a single line selected from the whole image, at the speed of approximately 7875/7812.5 line-images/s and 720x1/768x1 pixels resolution. The selected line is usually at the level of the midportion of the vibrating folds. The technique allows a clear visualisation of some essential physiological parameters of vocal fold vibration: period, duration of opening, closing and closed phases, amplitude of the vibration and right-left symmetry (Fig.1).

Kymography has been applied successfully to voice pathology [e.g. 2 – 5]. However, another potential field of applications of the technique is singing voice pedagogy, as it provides real time visual feedback for the subject, and allows short term inspection and analysis of the recorded sequence. The method becomes particularly interesting with the input of automatic quantitative analysis of the above-mentioned parameters, as it has already been achieved for pathology, however focusing on asymmetries and irregularities [2,5].

The 'messa di voce' is a gradual crescendo and decrescendo on a sustained (sung) note , and is known as one of the most difficult exercises for singers. This exercise is frequently associated with the 'candle test', a centuries old method used by singers to evaluate airflow. It simply consists of singing a vowel with a candle flame placed about five inches from the mouth. Depending on the exhaled airflow, the flame will waver very little (or not at all) or flutter wildly [6]. During a 'messa di voce' by a trained singer, the flow increase at the *ff* must remain minimal.

'Messa di voce': the underlying physiology

From a physiological point of view, a progressive increase in voice intensity – thus primarily in subglottic pressure - needs several critical requirements [7 ; 8]: (1) a constant adjustment of tension in the intrinsic laryngeal muscles, particularly the m. vocalis and the m. cricothyroideus, in order to keep the fundamental frequency (Fo) constant and to compensate for the enhanced passive strain in the folds, induced by the averaged higher vibration amplitude; (2) a specific mechanical regulation at the transition from a 'one-mass-model' vibration in *pp* (falsetto-like) to a 'two-mass-

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

model' (usual modal register), particularly at the early beginning of the crescendo; (3) a permanent control of the voice quality in order to avoid audible noise appearance due to increase of the transglottic airflow, concomitantly with the increase of the lung pressure. Actually, the performer must succeed in obtaining at each intensity level an exact balance between expiratory pressure and glottal impedance. Ideally, it results in only a slight increase in transglottic flow [9].

II. Material & Methods

(1) Kymography system: The single line scanning system used in these experiments is comprised of a Lambert CCD-Kymocam with technical characteristics corresponding to the above-reported description, a rigid 90° Wolf laryngeal telescope, a JVC-magnetoscope and a monitor. The telescope has a magnifying facility, with narrow depth of field and critical sharpness adjustment.

Vocal material: A trained vocalist (baritone), was asked to produce series of 'messa di voce' utterances on different pitches, avoiding to elicit vibrato. Vocal fold vibrations were recorded using the VKG system and the subject could have a real time visual control on the screen. After some preliminary trials, the subject became able to handle the scope himself and to find the optimal placement, combining comfort and quality of vocal fold image. During four different sessions 62 recordings were achieved. The sound was also recorded by a Sennheiser MD 4210 microphone at 10 cm of the lips, for analysis of SPL and Fo (using PRAAT 5.3.10, 2012 by P. Boersma & D. Weenink: www.praat.org.). One of the recordings, considered as representative and providing a complete visualization of the endolarynx was selected. For demonstration 40 single line scans, each showing two vibratory cycles, were taken at 125 ms intervals in order to cover the complete utterance (5 s). Fig. 1 shows two characteristic kymograms, respectively at the early beginning of the 'messa di voce' and at approximately the maximal SPL of the 'messa di voce' (the closed phase near half of the period).

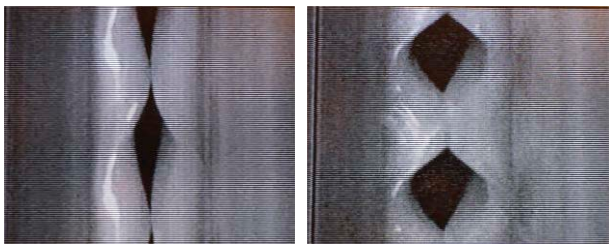


Fig. 1: Two examples of single line scans during a 'messa di voce'. Left at the beginning (*pp*), right at *ff*.

These images were enlarged and printed for manual measurements of the following parameters: right and left period, right and left amplitude, open and closed phase, right and left opening time. All measurements were made manually, independently by two observers, and averaged. The closed quotient is the quotient 'duration of closed phase / duration of cycle'. An adequate calibration was achieved with respect to time and distance. For distance, scale paper was filmed without changing the focusing.

(2) Program for automatic analysis: First attempts to automatically analyze single line scans of vocal fold vibration were achieved by Qiu & al. [10]. The approach used is based on monodimensional active contours, where each line has its own energy to be minimized separately from other lines. With the VKG-Analyser used in the current experiments, a planar active contour is applied, i.e., the set of all points (two for each VKG line) is considered as a pair of lines (corresponding to left and right vocal folds) that contain a surface. During subsequent iterations, the contour is modified in order to adapt to the shape given by the dark pixels of the image. Planar snakes allow the algorithm to find a global minimum of the energy, thus performing a global optimisation instead of a series of local ones. Moreover, in the VKG-Analyser, specific parameters are evaluated for the case of incomplete vocal fold closure. A detailed description of the VKG-Analyser has been provided in a recent publication [5]. Each image in the sequence can be processed using a digital image processing algorithm developed and optimized for the analysis of VKG recordings. It performs intensity adjustment, noise removal and robust techniques for vocal fold edge detection to avoid fluctuations of the grey levels in regions at a distance from the vocal folds.

(3) The vocal folds contour detection algorithm consists of two main steps: the first one defines an initial contour of the glottal area opening, using an adaptive threshold. A refining iterative procedure, based on active contours, is applied to the region, giving the final segmentation. The control parameters which drive both steps are determined automatically by the program. However, the user can manually adjust some of the controls for improving segmentation using a set of controls present in the user interface. The software allows to individually select the desired frame(s) to be processed from the video recording. Once the final contour has been obtained, the parameters of interest are evaluated. The software is designed so as to give a value of each parameter for each video frame by averaging the parameters over all the vibration periods which can be observed on the frame. This reduces the variability of the results smoothing out noise and eases the management of data giving a fixed number of values for a given video sequence,

independently from the acquisition. The program also allows storage and retrieval of subject's data, display of tracked parameters, results and statistics.

(4) Additional measurements

For the purpose of this study, the vocalist – once trained in optimizing his 'messa di voce' - repeated the exercise for separate indirect measurement of subglottic pressure and flow-glottography. For indirect measures of subglottic pressure, the short flow interruption method was used [11]. The glottal volume velocity waveform was recorded with a Rothenberg mask and the MSIF2 inverse filtering system of Glottal Enterprises, Inc.

III. RESULTS

Fig. 2 shows the basic acoustic characteristics of a typical 'messa di voce'. Amplitude increases and decreases (~ 35 dB at 10 cm) while Fo remains stable (165 Hz). Duration of the sequence is about 5 s.

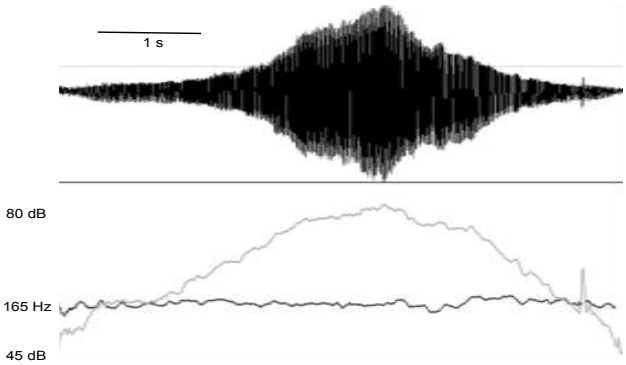


Fig. 2: messa di voce: microphone signal, Fo and intensity (duration about 5s).

Changes in % over time in different physiological parameters can be displayed: amplitude of vibration (Right + Left), duration of the closed phase and closed/open quotient (quotient of duration of the closed phase / duration of open phase), opening and closing speed etc. Fig. 3 shows as an example the evolution of the closed/open quotient, with as well the measures made by hand as those made by the computer program.

The lowest value for each parameter is set as 100%, in order to illustrate the relative changes. The smoothing curves (least squares fit) are shown as solid lines. The correlation coefficient (manual / automatic) is 0.88. Similarly but for amplitude measurements, the correlation coefficient is 0.84 (Fig. 4). Results of the additional instrumental measurements are shown in Figs. 4 and 5: estimates for subglottic pressure vary between about 3 and 20 hPa. There is a satisfactory control of the air flow around approximately 200 – 250 ml./s.

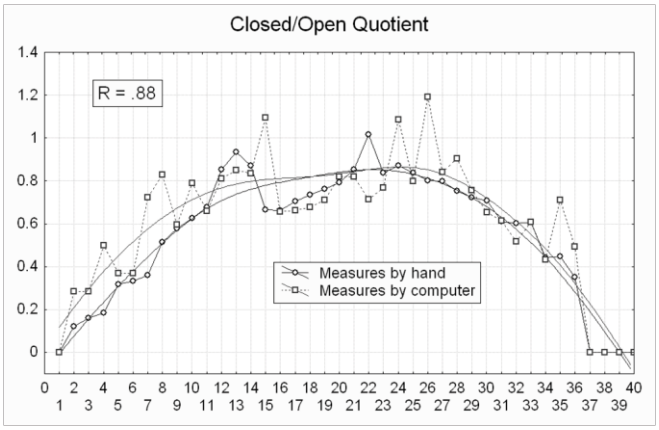


Fig. 3: Evolution over time (5s) of the closed/open quotient. Comparison of measures by hand and by computer

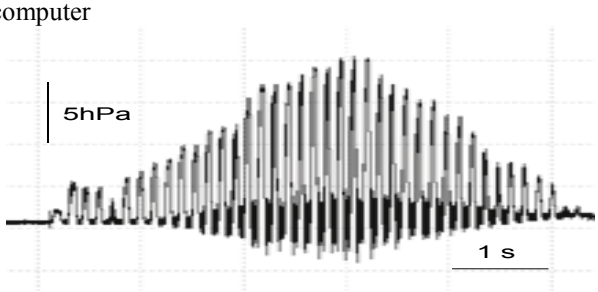


Fig. 4: Estimate of subglottic pressure by measurement of the intraoral pressure (flow interruptions method).

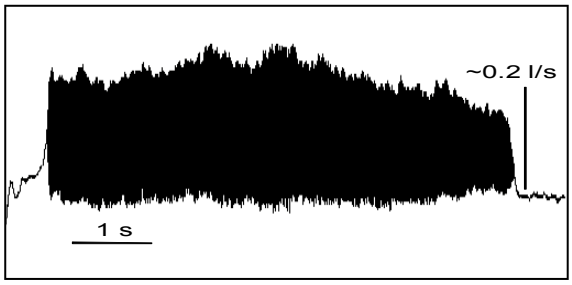


Fig. 5: Glottal volume velocity waveform measured with Rothenberg's mask and inverse filtering.

IV. DISCUSSION

Global feedback on the acoustic phenomenon

During the execution of the 'messa di voce' the changes of intensity are, in this vocalist, clearly correlated with the changes of breath pressure [12], although changes in spectral balance (singer's formant) may contribute to perception of increased loudness. Ideally, one could expect the exercise to be performed as a symmetric triangle: a linear increase in loudness, followed by a linear decrease.

Detailed feedback on vocal fold biomechanics

The cricothyroid and thyroarytenoid muscles are primarily involved in regulating the biomechanical properties of vocal fold vibration. As the vocal intensity

increases, it is expected that their tension will also be increased at a gradual pace to resist the growing breath pressure, but without compromising tonal accuracy. Videostroboscopy of the vocal fold vibration has demonstrated that the degree of glottal closure increases with intensity of phonation in normal subjects [13]. Furthermore, also in normal subjects, the glottal impedance is well reflected by the electroglottographic closed quotient, and the quotient clearly increases with voice SPL; this increase in vocal fold closure is plausibly related to an active thickening of the vocal fold edge (in modal register), and to a longer closed phase of the vibratory cycle [14]. This may account for the limited concomitant increase in transglottal flow [10]. Titze [8] has hypothesized (using computer modeling) that the maximal glottal source power should occur somewhere around a closed/open quotient of 0.5, i.e. when open and closed phases are approximately equal (as observed in the demonstrated case, see Fig. 2, right).

The need for an automatic analysis program

However, these quantitative data are of practical value only if they are available immediately after the 'messa di voce'. Therefore, in this work, a new user-friendly tool, the VKG-analyser [2;5] is used and tried out for the automatic extraction, tracking and computing of quantitative parameters from the VKG images. In the present study, comparative plots of the closed/open quotient measured manually (average of two raters) and by the automatic analysis program confirm the efficacy of the analysis program: the correlation coefficient between the two measurements is 0,88 (Fig. 3). For amplitude measurements, the correlation coefficient is 0,84.

Additional measurements

Both subglottic pressure and flow-glottography are fitted for real-time visual feedback: the direct visualization of the pressure data will help in increasing symmetry (Fig. 4) while monitoring the phonation flow plays the role of the historical candle in front of the mouth (Fig. 5).

V. Conclusion

'Single line scanning' of the vocal fold vibrations appears to be – beside its clinical usefulness - also suitable for investigating a specific technical voice exercise (and musical ornament) as the 'messa di voce'. It makes possible a precise quantification over time of the essential physiological parameters characterizing the changes of glottal impedance concomitant with the progressive increase and decrease of the lung pressure. However, introduction of a valid and user-friendly automatic analysis program of kymography-images appears indispensable for opening such new applications in the field of voice pedagogics by short-term feedback. The data provided by the automatic analysis program show a strong correlation with handmade measurements. Additional measurements (subglottic pressure and phonation flow) can meaningfully be correlated with the data obtained from the kymography-images.

References

[1] Svec J, Schutte H K. Videokymography: high-speed line scanning of vocal fold vibration. J Voice 1996; 10: 201–205.

[2] Manfredi C, Bocchi L, Bianchi S, Migali N, Cantarella G. Objective vocal fold vibration assessment from videokymographic images, Biomed. Signal Process. Control 2006; 1: 129–136.

[3] Svec J, Sram F, Schutte H K. Videokymography in voice disorders: what to look for ? Ann. Otol. Rhinol. Laryngol. 2007; 116: 172–180.

[4] Piazza C, Mangili S, Del Bon F, Gritti F, Manfredi C, Nicolai P, Peretti G. Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study. Eur Arch Otorhinolaryngol 2012; 269: 207-212.

[5] Manfredi C, Bocchi L, Cantarella G, Peretti G. Videokymographic image processing: Objective parameters and user-friendly interface, Biomedical Signal Processing and Control 2012; 7: 192– 201.

[6] Taylor D C. A Rational Method of Voice Culture based on a Scientific Analysis of all Systems, Ancient and Modern. N.Y.,The MacMillan Co., 1922.

[7] Titze I R. Principles of Voice Production. Englewood Cliffs, NJ: Prentice Hall, 1994.

[8] Titze I R. More on Messa di voce; Journal of Singing 1996; 52: 31 – 32.

[9] Smitheran JR, Hixon TJ. A clinical method for estimating laryngeal airway resistance during vowel production. J Speech Hear Disord 1981;46:138-146.

[10] Qiu Q, Schutte H K, Gu L, Yu Q. An automatic method to quantify the vibration properties of human vocal folds via videokymography. Folia Phoniatr Logop 2003; 55: 128-136.

[11] Hertegard S, Gauffin J, Lindestadt P A . A comparison of subglottal and intraoral pressure measurements during phonation. J Voice 1995; 9: 149-155.

[12] Titze I R, Long R, Shirley G I, Stathopoulos E, Ramig L O, Carroll L M, Riley W D . Messa di voce: an investigation of the symmetry of crescendo and decrescendo in a singing exercise. J Acoust Soc Am 1999; 105: 2933 – 2940.

[13] Sodersten M, Lindestadt P A. Glottal closure and perceived breathiness during phonation in normally speaking subjects. J Speech Hear Res 1990; 33: 601-611.

[14] Dejonckere P H. Control of fundamental frequency and glottal impedance with increasing sound pressure in normal and pathological voices. Voice 1994; 3: 10-16.

ANTICIPATION OF A NEUROMUSCULAR TUNING IN M. VOCALIS PERTURBS THE PERIODICITY OF VOCAL FOLD VIBRATION: THE UNEXPEXTED FINDING OF A PITCH-MATCHING EXPERIMENT COMPARING SINGING STUDENTS WITH HIGH-LEVEL PROFESSIONALS

P.H. Dejonckere [1], J. Lebacq [2], C. Manfredi [3]

[1] University of Leuven, Neurosciences, Exp. ORL, Belgium ; Federal Institute of Occupational Diseases, Av. de l'Astronomie, 1, B-1210 Brussels Belgium.
[2] Institute of Neuroscience CEMO, Université Catholique de Louvain, Avenue Hippocrate 55 bte B1.55.12 B-1200 Brussels, Belgium.
[3] Department of Information Engineering, Università degli Studi di Firenze, Via S. Marta 3, 50139 Firenze, Italy.

Abstract: **10 young female singing students were compared to their teachers in a vocal pitch-matching task with three standardized intervals (third, a fifth and an octave) and in two conditions (with / without piano modelling). Period durations at the offset of the base tone and at the onset of the target tone were analyzed in detail. Contrary to untrained individuals, the singing students appear to have already reached a degree of neuromuscular control in pitch matching comparable to that of professionals. There is also no effect of the interval extent on the accuracy of pitch-matching, and the reference of the piano does not play a significant role. In general, a Fo-instability characterizes the first cycles of the target tone; it could be explained by mechanical readjustments. However, the periodicity of the last cycles of the base tone is also significantly perturbed at the moment at which the subject anticipates the pitch jump, and this perturbation increases with the extent of the pitch jump to be achieved. This perturbation plausibly reflects an equivalent of the pre-phonatory burst of asynchronous muscle action potentials observed before the onset of phonation. However the phenomenon is so fast that it is not perceived by the listener.**

Keywords: **pitch- matching, singing, perturbation, anticipation, M Vocalis.**

I. INTRODUCTION

Trained singers perform more accurately than untrained individuals in a pitch-matching task, consisting of matching the vocal fundamental frequency to a preset target tone [1-2]. This ability seems specific to neuromuscular voice regulation. Murry and Caligiuri [3] have compared the performances of adult singers and adult non-singers on an auditory-motor task (pitch matching) and a visuomotor task (turning the wrist to match a visual target). Singers performed more accurately in the pitch-matching task than non-singers, and less experienced singers were less consistent during pitch-matching tasks than more experienced ones, but wrist targets were achieved with comparable efficiency. These results suggest that specific neuro-muscular training and specific experience with motor planning play a key role in technical aspects of singing abilities; however, the importance of their influence is not known. Does a singing student already master this ability quite early, or is it only acquired after years of practice at a professional level? Other relevant aspects that pertain to the accuracy of the neuromuscular control are the influence of modelling the target tone by providing the note with a piano (auditory information), and the effect of the extent of the interval. To address these questions, young conservatory students were compared with their teachers in a pitch matching experiment with three different intervals and two conditions: with or without modelling.

To get the best possible insight in the control mechanism of pitch adjustment, all sound periods in the critical phases (first cycles of target tone and last cycles of base tone) were measured period by period. Voice onset is known to show typically much greater frequency perturbation than the steady state midportion of a sustained vowel [4]. An additional aim of the study was to compare this short-term instability in our two groups of singers.

II. MATERIAL & METHODS

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Subjects were 10 healthy female singing students (ages 18-22) on the one hand, and 10 healthy female professional classic singers (ages 26-43) on the other hand, all of them being also singing teachers in three different conservatories. They were asked to produce 10 times 3 standard intervals (third, fifth and octave) starting from a same base tone d1 (~ 294 Hz) that was given by a piano. In the five first trials, the interval was modelled by the piano, while in the last five ones only the start tone was provided. The subjects were requested to sing on /a/ at comfortable loudness level without vibrato and with a short interruption between the two tones of each interval (thus not legato and without portamento). All recordings were made digitally in a quiet room (conservatory classroom), with a Sennheiser MD 421 N microphone at 30 cm from the mouth and at a sample frequency of 44.100 Hz. The PRAAT program (PRAAT 5.3.10, 2012 by P. Boersma & D. Weenink: www.praat.org.) was used for displaying and stretching the oscillograms. Period duration measurements were made cycle by cycle with cursors on the computer screen on the last ten cycles before the interval, and on the first fifteen cycles after the pitch jump (Fig. 1). This procedure is very easy in normal voice signals, and has been checked in another context [5].

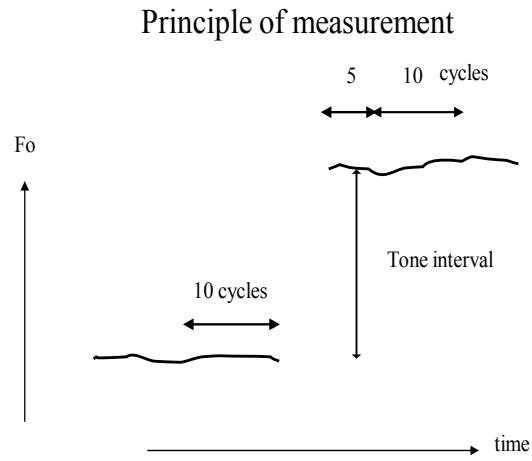


Fig. 1: Pitch-matching task for a given interval. Principle of measurement. X- axis is time, Y-axis is fundamental frequency. At a given moment the singer makes a pitch jump (third / fifth or octave) without legato, thus with a short interruption in voice emission. Cycle duration is measured just before (10 cycles) and just after (5 & 10 cycles) the jump.

III. RESULTS

Fig. 2 shows the Fo quotients for the three intervals (in average 1.2, 1.5 and 2), with their standard deviations, in both professionals and in students, with or without modelling of the interval at the piano.

The standard deviations (SD) of the quotients reflect the accuracy of the interval. A null SD should indicate that the singer always realizes exactly a frequency quotient of 1.2, 1.5 or 2. No systematic deviations were observed in the sense of making the interval smaller or larger. The average standard deviations of the quotients do not significantly differ for the three intervals.

No significant differences were observed between students and teachers, and there is no significant effect of modelling by the piano.

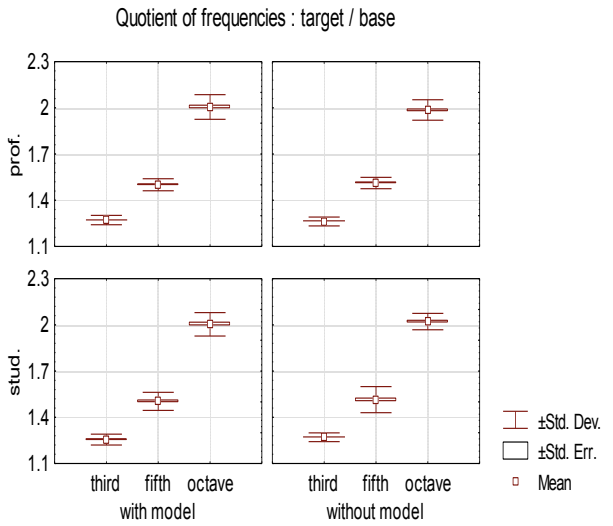


Fig. 2: Fo quotients for the three intervals (in average 1.2, 1.5 and 2) with their standard deviations in professionals and in students, with and without modelling of the interval at the piano.

In both groups of subjects, for all three intervals and for the two modelling conditions, there is a significant increase in Fo-perturbation (variation coefficient of period) in the first five cycles of the target tone when compared with the 10 subsequent cycles (always $p < .001$). However there is no difference between singing students and singing teachers, and no significant effect of the extent of the interval (Figs. 3 & 4).

Interestingly, the last ten cycles before the interval also demonstrate a significant increase in Fo-perturbation compared to the steady state phonation ($p<0.001$). This is observed in both groups of subjects. Furthermore, the variation coefficient of these last ten cycles significantly increases ($p < 0.01$) with the extent of the interval (Figs. 3 & 4).

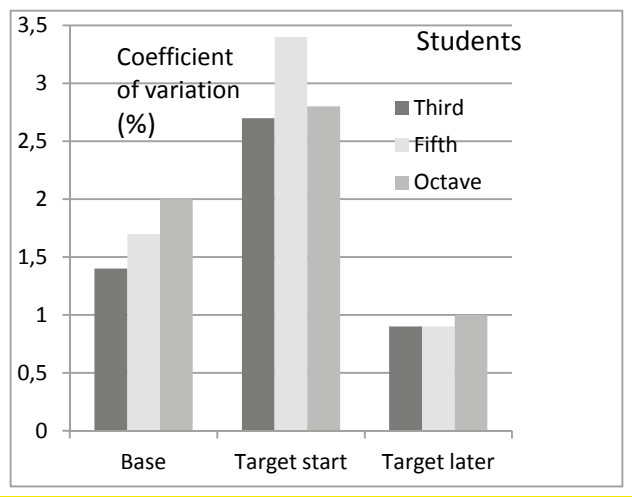


Fig. 3: Fo perturbation in the last 10 cycles of the base tone, in the first 5 cycles of the target tone and in the 10 subsequent cycles of the target tone. (Singing students)

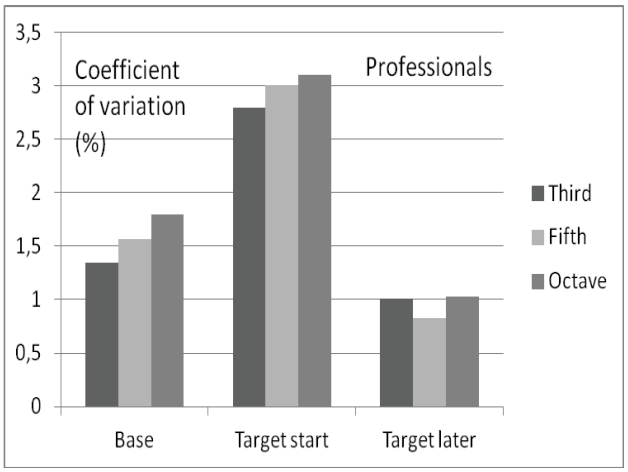


Fig. 4: Fo perturbation in the last 10 cycles of the base tone, in the first 5 cycles of the target tone, and in the 10 subsequent cycles of the target tone. (Professional singers)

IV. Discussion

The hypothesis that professional classical singers achieve a more precise neuromuscular control in pitch tuning than young students is to be rejected. This suggests that this ability does not require a very large amount of training or experience.

In the categories of subjects investigated in this work, modelling of tone intervals by a piano seems ineffective.

As expected, the frequency perturbation is, for all conditions, much larger during the first cycles of the target tone than in the ten subsequent cycles. This phenomenon is plausibly related to a necessary mechanical readjustment and stabilization of muscle tension and to a new balance between glottal resistance and lung pressure. Here again no difference is found, neither between groups nor between intervals.

Most intriguing is the period perturbation induced by the anticipation of the pitch jump. The jitter phenomenon is still incompletely understood, but it is usually explained by mechanical factors within the vocal fold tissues (viscosity, non-linearity), by internal body noise (pulsations in blood vessels) and by asynchronous firing of motor units in the M. vocalis and cricothyroideus. The third factor is probably mainly directly involved in the increased Fo perturbations observed in the present experiment.

The pre-phonatory tuning of the M. vocalis is a well known phenomenon and can be clearly observed using electromyography: the thyroarytenoid muscles show a burst of muscle action potentials preceding the onset of phonation followed by a decreased activity [6]. The time interval between the start of the change in electrical activity and the onset of voice emission is 100 to 200 ms (actually from 50 to 600 ms) [7]. This enhanced pre-phonatory activity consists of firing of motor units within the M. vocalis and these action potentials are asynchronous with the vocal fold vibration frequency. This could explain the perturbation effect on the regularity of this vibration frequency. As there is a clear positive correlation in modal register between voice pitch and muscle tension in the M. vocalis [8], it may be expected that the importance of such a pre-target tuning activity is dependent on the extent of the ascending tone interval. Consequently, it may be hypothesized that it interferes accordingly with the tension regulation during the last ms of the base tone.

V. Conclusion

Young singing students appear to have already reached a degree of neuromuscular control in pitch matching comparable to that of true professionals.

For a mild range of tone intervals (third – fifth – octave) there is no effect on the accuracy of pitch-matching. Reference of the piano does not play a significant role either.

In general, a Fo-instability characterizes the first cycles of the target tone and it could be explained by the mechanical readjustments. However the last cycles of the base tone are also perturbed at the moment the subject anticipates the pitch jump and the perturbation increases with the extent of the pitch jump that will be achieved. This perturbation plausibly reflects an equivalent of the pre-phonatory burst of asynchronous muscle action potentials observed before the onset of phonation.

However the phenomenon is so fast that it not perceived by the listener.

REFERENCES

[1] T. Murry," Pitch-matching accuracy in singers and non-singers." J Voice, vol. 4, pp. 317-321, 1990.

[2] J.M. Estis, A. Dean-Claytor, R.E. Moore and T.L. Rowell, "Pitch-matching accuracy in trained singers and untrained individuals: the impact of musical interference and noise. " JVoice vol. 25, pp. 173-180, 2011.

[3] T. Murry, MP Caliguiri, "Phonatory and nonphonatory motor control in singers." J Voice, vol. 3, pp. 257-263, 1989.

[4] Y. Koike, "Application of some acoustic measures for the evaluation of laryngeal dysfunction", Studia Phonologica, vol. 7, pp. 17-23, 1973.

[5] P.H. Dejonckere, A. Giordano, J. Schoentgen, J. Fraj, L. Bocchi, C Manfredi,, "To what degree of voice perturbation are jitter measurements valid? A novel approach with synthesized vowels and visuo-perceptual pattern recognition", Biomedical Signal Processing and Control, Vol. 7, pp.: 37-42, 2012.

[6] A.D. Hillel, "The study of laryngeal muscle activity in normal human subjects and in patients with laryngeal dystonia using multiple fine-wire electromyography", Laryngoscope, Vol. 111, 2 Suppl., pp. 1- 47, 2001.

[7] P.H. Dejonckere, EMG of the larynx. Marc Pietteur: Liège ISBN 2-87211-000-3, 1987.

[8] P.H. Dejonckere, "Les mécanismes musculaires élémentaires de régulation de la tension de la corde vocale au cours de la phonation", Folia Phon. vol. 32, pp. 1-13, 1980.

VALIDATION OF THE ITALIAN VERSION OF THE SINGING VOICE HANDICAP INDEX

G. Baracca[1], G.Cantarella [1], S. Forti[2], F. Fussi[3]

[1] Otolaryngology Department, Fondazione IRCCS Cà Granda Ospedale Policlinico, Milan, Italy, giovanna.baracca@gmail.com
[2] Audiology Unit, Fondazione IRCCS Cà Granda Ospedale Policlinico, Milan, Italy aut_est@yahoo.it
[3]AUSL di Ravenna, Ravenna, Italy ffussi@libero.it

Abstract: **Singers constitute a specific population sensitive to vocal disability, which may have a higher impact on their quality of life compared to non-singers. A specific questionnaire, the Singing Voice Handicap Index (SVHI) was created and validated aimed to measure the physical, social, emotional and economic impacts of voice problems on the lives of singers. Aim of this study was to validate the Italian version of the SVHI. The validated English version of the SVHI was translated into Italian and then discussed with several voice care professionals. The Italian version of the SVHI was administered to 214 consecutive singers (91 males and 123 females, mean age: 32.62±10.85). Voice problem complaints were expressed by 97 of the singers, while 117 were healthy and had no voice conditions. All subjects underwent a phoniatric consultation with videolaryngostroboscopy to ascertain the condition of the vocal folds. Internal consistency of the Italian version of the SVHI showed a Cronbach's α of 0.97. The test-retest reliability was assessed by comparing the responses obtained by all subjects in two different administrations of the questionnaire; the difference was not significant (p=ns). The SVHI scores in healthy singers was significantly lower than the one obtained in the group of singers with a vocal fold abnormality (29.26±25.72 and 45.62±27.95, p<0.001, respectively). The Italian version of the SVHI was successfully validated as a suitable instrument for the self-evaluation of handicaps related to voice problems in the context of singing.**

Keywords : **Singing, Voice Disturbance, Questionnaire Design, Self Report**

I. INTRODUCTION

Self-administered questionnaires are used to assess the impacts of health problems on the quality of life of patients: a disability in performing a daily task, defined as a handicap, could cause a disadvantage in social, economic or environmental aspects of life [1]. Several questionnaires have been designed to measure the impact of voice problems on the lives of individuals: the most popular is the Voice Handicap Index (VHI) [2], which has been validated and translated into several languages [3-5]. The VHI was developed to assess the subjective perception of disability related to voice disorders in all types of patients [2].

Singers constitute a specific population of professionals particularly at risk for voice problems. The perception of a voice problem in singing is often related to specific symptoms, such as difficulty in the passaggio, vocal endurance and diminished range [6], aspects that are not assessed by the VHI. Furthermore, singers are often more sensitive to vocal disabilities, which may have a higher impact on their quality of life compared to non-singers [7]. Hence, to obtain a self-assessing instrument able to evaluate vocal disability in singers, in 2007, Cohen et al. created and validated a specific questionnaire, the Singing Voice Handicap Index (SVHI), aimed to measure the physical, social, emotional and economic impacts of voice problems on the lives of singers [8]. The SVHI is a 36-item self-administered questionnaire that is able to assess difficulties related to voice health status typical of the singing professional. The items address symptoms frequently reported to laryngologists and speech pathologists by singers. Among singers, the SVHI is also more sensitive to clinical changes than the VHI [9], which proves the validity of the SVHI in measuring treatment outcomes in the singing population. The aim of this study was to validate the Italian version of the VSHI.

II. METHODS

Development of the Italian version of the SVHI: An Italian translation of the validated English version of the SVHI was carried out by a qualified professional translator. The first version of the questionnaire was then discussed by 2 phoniatricians, 2 speech therapists, 2 singing teachers and 2 professional singers to improve the translation and to make it more understandable to singers. After, the new Italian version was to re-translated in English and, finally, re-translated in Italian language. Each of the 36 items of the questionnaire was

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

individually scored on a 5-point Likert scale ranging from "never" (score of 0) to "always" (score of 4). The score was based on how often each statement was experienced by the singer, with higher numbers representing more self-perceived handicaps [8].

Participants: The Italian version of the SVHI was administered to 214 consecutive singers (91 males and 123 females, mean age: 32.62±10.85, range: 14-60 years). The diagnosis of each patient in the study group was determined by the clinical history and by rigid and/or flexible laryngeal videoendoscopy including stroboscopy. The stroboscopic findings were classified into 5 groups: normal, functional (including incomplete closure with a gap along entire length of the vocal folds during phonation and subjects with absence of organic lesions but with perceptual audible voice changes and complaints), inflammatory (including hemorrhage, inflammation of the vocal fold mucosa, posterior laryngitis, and Reinke's edema), mass on vocal fold (including nodules, polyps, and cysts) and high stiffness of the vocal fold (including sulcus, vergeture, and scarring).

Administration of the Italian version of the SVHI: all subjects were asked to fill out the questionnaire following the instructions written at the bottom of the form, prior to a phoniatric consultation and the laryngeal examination. A second copy of the questionnaire was mailed 7 days after the first consultation with a request to fill out the questionnaire based on the perception of the present status and return the form.

Analysis: The statistical tests were performed using SPSS 17.0 for Windows (SPSS Inc., Chicago, IL). The internal consistency of the questionnaire was determined by Cronbach's α coefficient; the item-total correlations were calculated for all items. The test-retest reliability was assessed for the total score of the SVHI. Pearson's product-moment correlation was used to evaluate the test-retest reliability of the SVHI by comparing the first and the second responses. The SVHI total scores of the singers with a vocal fold pathology (functional, inflammatory, mass on the vocal fold, high stiffness of the vocal fold) and of the healthy singers were compared to test the clinical validity of the questionnaire by the use of the nonparametric Mann-Whitney test.

III. Results

The total number of participants was 214. There were 117 healthy singers with no voice complaints (45 males and 72 females, mean age: 31.94 ± 10.95 years, range: 14 – 60 years) confirmed by normal videolaryngostroboscopic findings. There were 97 singers with a voice problem (46 males and 51 females, mean age: 33.39 ± 10.74 years, range: 15 – 60 years) who had a clinical objective diagnosis by videolaryngostroboscopy; 7 were included in the functional group, 31 in the inflammatory group, 52 in the mass on the vocal fold group and 7 in the high stiffness of the vocal fold group.

All participants completed the questionnaire without assistance in less than 10 minutes. The mean score was 45.62 ± 27.95 for the pathological singers and 29.26 ± 25.72 for the healthy singers. As expected, the scores of the control group were significantly lower than the pathological group ($p<0.001$). The retest was completed by 70 subjects (29 pathologic and 41 healthy singers). The mean SVHI scores at the first and second submissions were 36.08 ± 31.37 and 34.33 ± 24.32, respectively. The internal consistency and reliability of the SVHI was very high (Cronbach's $\alpha=0.97$); the correlation between the SVHI scores at the first and second submission (test-retest) was strong ($r=0.98$, $p<0.001$). Both the functional group (mean score: 67.00 ± 34.28, range: 12-109) and the mass on the vocal fold group (mean score: 45.23 ± 25.98, range: 8-126) reported SVHI values significantly higher than the healthy group (mean score: 29.26 ± 25.72, range: 0-137; $p=0.003$ and $p=0.004$, respectively). No significant differences were found among the pathological subgroups, i.e. among functional, mass on the vocal fold, inflammatory and high stiffness (p=ns). Furthermore, the differences between the healthy group and the inflammatory group, (mean score: 41.23 ± 28.51, range: 0-143) and between the healthy and high stiffness of the vocal fold groups (mean score: 46.57 ± 30.74, range: 11-87) were not significant (both p=ns). The ANOVA results were unaffected by age, gender or style of singing.

IV. Discussion

Singers are more sensitive to many early symptoms of voice abnormalities, and they are more likely to seek help and report problems related to their singing voice [8]. Singers represent 11.5% of all patients at voice consultations, while constituting only 0.02% of the general population [10]. This is partly due to the importance they give to their voice status, a critical social and occupational factor that can significantly affect their quality of life. A realistic overview of the singer's condition is critical to facilitate the most suitable management of this unique group of patients. Therefore, Cohen et al., in 2007 [8], created and validated a health status instrument for use in singers, called the SVHI. The use of the SVHI can determine how voice problems impact the quality of life of singers.

The Italian version of the SVHI described in this study supports its important psychometric properties, as the internal consistency and the test-retest reliability were very high. Furthermore, the SVHI was able to discriminate between healthy voice conditions and some pathological voice conditions (functional disturbances

and lesions with mass on the vocal fold), a result that further supports its validity. According to Cohen et al. [17], singers with a mass on the vocal fold had an SVHI score significantly higher than healthy singers. Additionally, in our research singers with a diagnosis of a functional voice disorder obtained a significantly higher SVHI score compared to healthy singers non-singers. Concerning singers with an inflammatory aspect of the vocal folds observed by videolaryngostroboscopy our interpretation is that these types of disturbances are often short in duration, so in many cases, they constitute an occasional cause of dysphonia. Cohen et al. [17] found that the chronicity of a voice problem is a critical factor influencing the SVHI score. The short duration of some inflammatory pathologies of the vocal folds could have caused the lower SVHI scores of this group of patients. The other subgroup of pathological singers that did not show a significant difference of the SVHI scores compared to the healthy singers group are singers with lesions, such as sulcus or scarring, that lead to high stiffness of the vocal folds. For this group of singers a higher SVHI score is expected, along with a strong adverse impact on the mucosal wave amplitude, instead the difference respect of healthy singers was slight, so there are two possible explanations. The first is that self-evaluation is simply another dimension than the biomechanics of vocal fold vibration, and that one need not to expect a clear relation. The second is that to notice the slight differences in SVHI scores between healthy and high stiffness of the vocal fold groups needs to increase the number of pathological subjects.

It must be addressed that none of the singers selected to participate in this study refused to complete the questionnaire. Furthermore, all singers completed the SVHI in no more than ten minutes, without the need for assistance. This is important as it underlines the good compliance of the Italian version of the SVHI, demonstrating that it is acceptable and easy to administer.

Correlations between the Italian version of the SVHI and aspects that were not analyzed; duration of voice complaints, comorbidities and certain voice styles may be interesting areas for future studies.

V. Conclusion

The Italian version of the SVHI is a reliable and valid tool for measuring the level of handicap related to voice problems perceived by singers, as demonstrated by the adequate internal consistency and reliability. The Italian version of the SVHI allowed discrimination between healthy and pathological vocal fold conditions.

References

[1] Organization WH, *International Classification of Impairments, Disabilities, and Handicaps: a manual of classification relating to the consequences of diseases World Health Organization*, Geneva, Switzerland, 1980

[2] B.H. Jacobson, A. Johnson, A. Grywalsky , A. Silbergleit, G. Jacobson, M.S. Benninger, C. Newman, "The voice handicap index: development and validation". Am J Speech Lang Pathol, Vol. 6, pp 66-70, 1997.

[3] A. Bonetti, L. Bonetti, "Cross-cultural adaptation and validation of the Voice Handicap Index into Croatian". J Voice, vol. 27, pp. 130 e137-130 e114, 2013

[4] A.F. Saleem, Y.S. Natour, "Standardization of the Arabic version of the Voice Handicap Index: an investigation of validity and reliability". Logoped Phoniatr Vocol, Vol. 35, pp. 183-188, 2010

[5] M. Behlau, L.M. Alves Dos Santos, G. Oliveira, "Cross-cultural adaptation and validation of the voice handicap index into Brazilian Portuguese". J Voice, Vol. 25, pp. 354-359, 2011

[6] C.A. Rosen, T. Murry T, "Voice handicap index in singers." J Voice, Vol. 14, pp. 370-377, 2000

[7] T. Murry, A. Zschommler, J. Prokop J, "Voice handicap in singers". J Voice, Vol. 23, pp. 376-379, 2009

[8] S.M. Cohen, B.H. Jacobson, C.G. Garrett, J.P. Noordzij, M.G. Stewart, A. Attia, R.H. Ossoff, T.F. Cleveland TF, "Creation and validation of the Singing Voice Handicap Index". Ann Otol Rhinol Laryngol, Vol. 116, pp. 402-406, 2007.

[9] S.M. Cohen, D.L. Witsell, L. Scearce, G. Vess, C. Banka, "Treatment responsiveness of the Singing Voice Handicap Index". Laryngoscope, Vol. 118, pp. 1705-1708, 2008.

[10] I.R. Titze, J. Lemke, D. Montequin, "Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report", J Voice, Vol. 11, pp. 254-259, 1997.

[11] S.M. Cohen, J.P. Noordzij, C.G. Garrett, R.H. Ossoff, "Factors associated with perception of singing voice handicap", Otolaryngol Head Neck Surg, Vol. 138, pp. 430-434, 2008.

Session IV:
VOICE MONITORING

INCORPORATING REAL-TIME BIOFEEDBACK CAPABILITIES INTO A VOICE HEALTH MONITOR

Andrés F. Llico[1], Matías Zañartu[1]*, Daryush D. Mehta[2], Jarrad H. Van Stan[2], Harold A. Cheyne II[3], Agustín J. González[1], Marzyeh Ghassemi[4], George R. Wodicka[5], John V. Guttag[4], and Robert E. Hillman[2]

[1] Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile
[2] Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[3] Bioacoustics Research Program, Laboratory of Ornithology, Cornell University, Ithaca, NY, USA
[4] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[5] Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA

* Corresponding author. Email: matias.zanartu@usm.cl

***Abstract*: Many common voice disorders are chronic or recurring conditions that result from abusive and/or faulty patterns of vocal behavior referred to as vocal hyperfunction. Thus, an ongoing goal is the development of long-term ambulatory monitoring for clinical assessment, prevention and modification of hyperfunctional patterns of vocal behavior. This paper reports our initial efforts toward real-time behavioral biofeedback using the smartphone-based Voice Health Monitor that records the high-bandwidth acceleration from the neck skin above the sternal notch. By incorporating real-time estimation of fundamental frequency and sound pressure level, the monitor can provide various types of biofeedback to the user through a vibrotactile alert. The performance of the monitor is compared with that of the commercially available Ambulatory Phonation Monitor (APM) using a bioacoustic transducer tester that generates repeatable vibratory stimuli recorded from ahuman subject. Ambulatory features computed include phonation time, fundamental frequency, sound pressure level, and compliance to a specified threshold level. The results support the implementation of biofeedback in the smartphone-based system and illustrate that the new platform's technology performs more reliably than the APM. Future work calls for the exploration of more sophisticated algorithms to measure vocal behavior and provide clinically meaningful biofeedback.**

***Keywords:* Voice use, ambulatory voice monitoring, neck accelerometer, vocal hyperfunction, biofeedback.**

I. INTRODUCTION

It is believed that abusive and/or faulty patterns of vocal behavior lead to functional dysphonia or phonotraumatic lesions, such as nodules and polyps, on the vocal folds. It has been suggested that this type of vocal behavior, often referred to as hyperfunction, could be better characterized and treated by incorporating daily long-term ambulatory voice monitoring and biofeedback into the clinical management process. [1]. Cheyne and colleagues developed such an ambulatory monitoring system that employed a neck surface accelerometer as the phonation sensor that provided a number of advantages over microphone-based systems [2]. This device, the Ambulatory Phonation Monitor (APM, Model 3200, KayPENTAX), has been commercially-available for research and clinical use since 2006 and was used in this study as a reference. The APM does not store the raw accelerometer waveform and thus only operates as a data logger of fundamental frequency (F0) and sound pressure level (SPL) every 50 ms for up to a maximum duration of approximately 14 hours. The APM can also provide biofeedback (via a pager vibrator) based on upper or lower thresholds set for F0 or SPL. Ambulatory biofeedback using the APM has been shown in early case studies to have the potential to facilitate vocal behavioral changes targeted in voice therapy [1].

Our group recently developed an enhanced ambulatory system, referred to as the Voice Health Monitor (VHM), employing the same accelerometer sensor coupled to a smartphone platform [3], as shown in Fig. 1a and 1b. The VHM overcomes many technical limitations of the APM, thus providing the capability to acquire and archive raw acceleration data for over 7 days, with an 80 dB dynamic range, 11.025 Hz sample rate, 16-bit quantization, and processing power to run complex algorithms [3]. Prior to this study, the VHM operated only as a waveform acquisition system, without biofeedback capability. This study aims to expand the VHM operation to incorporate real-time biofeedback features and to compare its performance with that of the APM.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

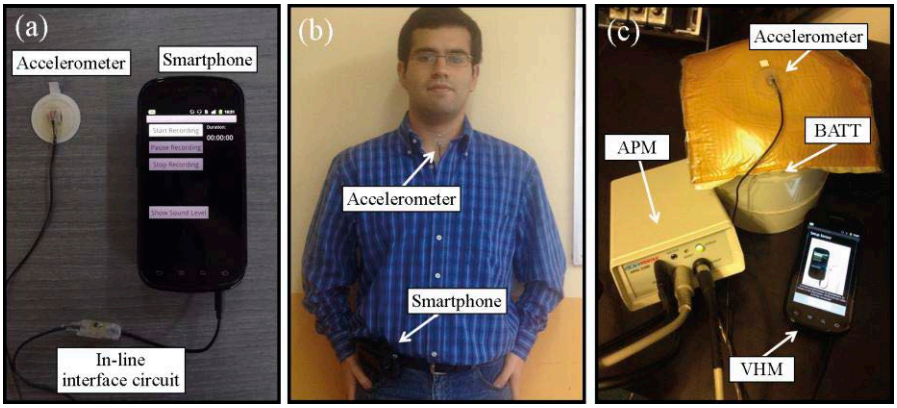


Fig. 1. Mobile Voice Health Monitor: (a) Smartphone, in-line interface circuit, and accelerometer input, (b) Subject wearing the VHM system, and (c) BATT platform for APM and VHM comparison.

II. METHODS

Given the processing capabilities of the smartphone platform, numerous biofeedback targets and approaches can be implemented in the VHM. In this study, we focused on the real-time estimation of F0 and SPL to mimic and contrast the current real-time biofeedback performance of the APM. This comparison was performed using a repeatable excitation signal and biofeedback triggering setup. As shown in Fig. 1c, the same light-weight accelerometer provided the input stimulus to both systems. The accelerometer was mounted on a bioacoustic transducer tester (BATT) [4] that was set to have a flat, band-limited response between 70 Hz and 2 kHz. The BATT was excited with an ambulatory recording previously captured with the VHM from an adult male subject with normal voice (a teacher during a 90-minute lecture), thus providing a signal comparable to that initially obtained with the VHM. Both systems were calibrated with the same subject-specific parameters that related accelerometer level to acoustic SPL.

Each 50 ms frame was divided into two 25 ms subintervals, and the frame was considered voiced if both subintervals exceeded 62 dB SPL. SPL was then re-computed over the entire frame duration. F0 for each voiced frame was equal to the reciprocal of the first peak location in the normalized autocorrelation function if the peak exceeded a threshold of 0.25 [3]. F0 was restricted to the range of 60 to 500 Hz, otherwise, SPL and F0 frame values were set to zero.

A frame counter kept track of the number of frames above-alarm threshold, which is set to 95 dB SPL, increasing when a frame is labeled as voiced, and decreasing when it's not voiced. Biofeedback was triggered when counter reached the equivalent of 300 milliseconds.

III. RESULTS

The summary statistics for the APM and VHM are shown in Table I. With the same conditions and calibration provided to each system, the measured phonation time, percent compliance, and biofeedback time differ slightly. Average differences between F0 and SPL estimates were around 2 Hz and 1 dB, respectively. Although these average measures were similar, differences were observed in the histograms for each parameter (Fig. 2). The slightly greater APM values around the average F0 in Fig. 2a are consistent with the APM labeling more frames as voiced than did the VHM. The SPL histograms in Fig. 2b show that most differences involved the extreme values of the distribution. These findings indicate that the APM labeled more lower-energy and higher-energy frames as voiced and shifted the center of the distribution, thus explaining its increased accumulated phonation time, lower compliance time, and higher biofeedback time. Although the overall differences under the testing conditions are small, the results for the VHM better align with those reported in the literature [1-3,5].

Table I: Summary statistics of ambulatory phonation measures for both systems.

Device	Total Time (hh:mm:ss)	Phonation Time (hh:mm:ss)	Average F0 (Hz)	Average SPL (dB)	% Compliance (SPL <= 95 dB)	Biofeedback Time (hh:mm:ss)
APM	02:08:47	00:38:03 (29.61%)	145.9	82.7	93.5	00:00:46 (2.03 %)
VHM	02:08:52	00:32:45 (25.42%)	148.1	81.6	96.3	00:00:19 (1.01 %)

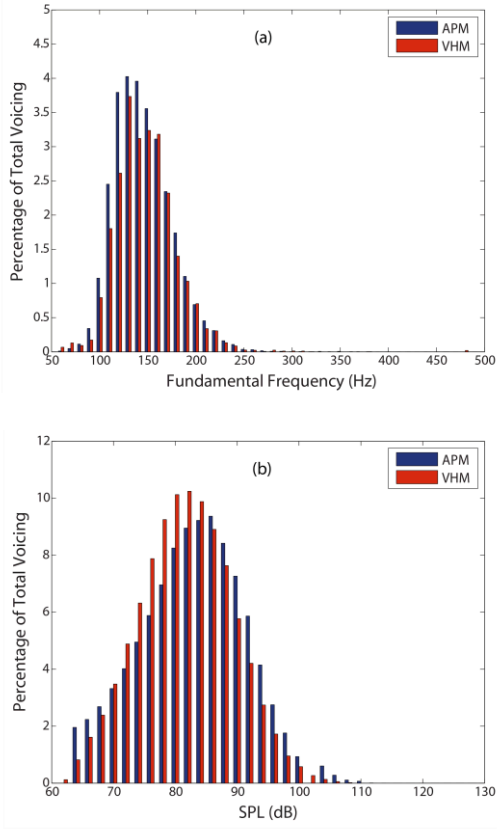


Fig. 2. Histograms for APM (blue) and VHM (red): (a) F0 and (b) SPL

IV. DISCUSSION

Given that the APM is based on technology that is at least 8-10 years old, it is not surprizing that is does not perform as well as the much newer VHM. The VHM provides a 16-bit quantization of the accelerometer signal versus 7-bit quantization employed by the APM. Furthermore, the APM operates with fixed-point arithmetic, where the signal level (in uncalibrated dB units) can only be saved in whole number units that are later converted to whole number SPL values; thus, the level resolution of the APM exhibits round-off error and a potentially coarse representation of SPL. Due to these differences in memory allocation and amplitude quantization, the VHM has approximately 40 dB more in dynamic range than the APM. These factors may explain the differences in resolution for the estimated units of dB SPL and would indicate that the APM is less precise when representing SPL levels for monitoring and biofeedback purposes (in accordance with our clinical observations). The added real-time biofeedback capabilities in the VHM are also more reliable and well suited for professional voice users and patients that have larger vocal ranges.

V. CONCLUSION AND FUTURE WORK

Real-time biofeedback capabilities were added to the VHM based on SPL and F0 thresholds, and its performance was compared with that of the APM. The VHM showed better performance than the APM in terms of its quantization, dynamic range, and computational precision. Subsequent investigations with the VHM in the context of an enhanced real-time biofeedback include the estimation of aerodynamic parameters using impedance-based inverse filtering [6] and z-score assessment [7], as well as wireless connectivity with a server in the clinic. The ability to better facilitate vocal behavioral changes with these new real-time features remains to be tested.

ACKNOWLEDGMENT

This work was supported by NIH-NIDCD grant R33 DC011588, CONICYT grant FONDECYT 11110147, and MIT MISTI grant MIT-Chile 2745333.

REFERENCES

[1] R. E. Hillman and D. D. Mehta, “Ambulatory monitoring of daily voice use,” *Perspectives on Voice and Voice Disorders*, vol. 21, no. 2, pp. 56–61, 2011.

[2] H. A. Cheyne, H. M. Hanson, R. P. Genereux, K. N. Stevens, and R. E. Hillman, “Development and testing of a portable vocal accumulator,” *J. Speech. Lang. Hear. Res.*, vol. 46, no. 6, pp. 1457-1467, 2003.

[3] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, “Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 11, pp. 3090–3096, 2012.

[4] S. S. Kraman, G. A. Pressler, H. Pasterkamp, G. R. Wodicka, “Design, construction, and evaluation of a bioacoustic transducer testing (BATT) system for respiratory sounds,” *IEEE Trans. Biomed. Eng.*, vol. 53, no. 8, pp. 1711–1715, 2006.

[5] I. R. Titze, E. J. Hunter, and J. G. Švec, “Voicing and silence periods in daily and weekly vocalizations of teachers,” *J. Acoust. Soc. Amer.*, vol. 121, no. 1, pp. 469–478, 2007.

[6] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, “Subglottal impedance-based inverse filtering of speech sounds using neck surface acceleration,” *IEEE Trans. Audio Speech Lang. Proc.*, 21(9), pp. 1929-1939, 2013.

[7] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, “Objective assessment of vocal hyperfunction: An experimental framework and initial results,” *J. Speech Hear. Res.*, vol. 32, no. 2, pp. 373–392, 1989.

TOWARD AN OBJECTIVE AERODYNAMIC ASSESSMENT OF VOCAL HYPERFUNCTION USING A VOICE HEALTH MONITOR

Matías Zañartu[1*], Víctor Espinoza[1], Daryush D. Mehta[2], Jarrad H. Van Stan[2], Harold A. Cheyne II[3], Marzyeh Ghassemi[4], John V. Guttag[4], and Robert E. Hillman[2]

[1]Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile
[2]Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[3]Bioacoustics Research Program, Laboratory of Ornithology, Cornell University, Ithaca, NY, USA
[4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

*Corresponding author. Email: matias.zanartu@usm.cl

Abstract: **Vocal hyperfunction is a description of abnormal patterns of vocal behavior that may lead to many common voice disorders. Previous studies demonstrated that disorders associated with hyperfunction could be detected in patients by measuring aerodynamic and acoustic parameters from recordings of a single sustained vowel using a Rothenberg mask setup. Although ambulatory systems have shown the best potential for unobtrusive long-term monitoring of vocal function, their ability to differentiate hyperfunctional from normal patterns of vocal behavior has not been assessed. This study provides an initial quantitative evaluation of the capabilities of a neck surface acceleration signal to objectively detect abnormal vocal behaviors associated with hyperfunctionally-related disorders. The goal is to verify if such detection is possible using a neck accelerometer signal rather than an airflow mask and incorporate vocal gestures from multiple vowels and running speech. An impedance-based inverse filtering algorithm is used to estimate aerodynamic parameters from the neck-surface acceleration signal. The results obtained when contrasting five patients with vocal nodules to five paired normal subjects indicate that the accelerometer-based assessment offers comparable discrimination capabilities as those from the aerodynamic recordings. The results also provide a first indication that this discrimination is possible with an expanded sample that includes other sustained vowels and running speech.**

Keywords: Voice use, ambulatory voice monitoring, neck accelerometer, vocal hyperfunction, inverse filtering.

I. INTRODUCTION

Many common voice disorders are likely to result from faulty and/or abusive patterns of vocal behavior, referred to as *vocal hyperfunction* [1]. These patterns can be difficult to assess accurately in the clinical setting and may be better characterized when individuals wear an ambulatory voice monitor while engaging in their typical daily activities. Current methods for ambulatory assessment of vocal function are based on measurements of neck surface acceleration and constitute a non-invasive, unobtrusive, noise-robust approach that maintains confidentiality [2]. However, there is a lack of statistically robust studies that demonstrate the true diagnostic utility of such systems. Our group strives to advance accelerometer-based ambulatory monitoring of vocal function by validating it as a reliable and cost-effective clinical tool that can be used to accurately identify and differentiate patterns of voice use that are associated with hyperfunctional voice disorders.

Our recently developed system, the Voice Health Monitor (VHM) shown in Fig. 1, is an accelerometer-based system that uses a smartphone platform that takes advantage of technological advances to allow for recording and storing raw acceleration data for at least 7 full days [3]. The acceleration data is subjected to an inverse filtering technique known as sub-glottal impedance-based inverse filtering (IBIF) to provide a

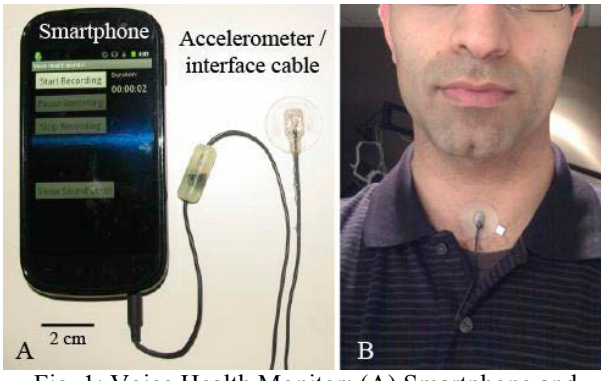


Fig. 1: Voice Health Monitor: (A) Smartphone and accelerometer assembly with (B) illustration of neck-surface sensor position.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

non-invasive estimation of the glottal airflow based on the neck-surface acceleration signal [4]. To achieve that task, the IBIF algorithm uses a transmission line model of the subglottal system, a lumped representation of the mechanical properties of the neck, and subject-specific parameters estimated during a calibration session.

In this study, we provide an initial quantitative evaluation of the capabilities of the accelerometer-based aerodynamic parameters extracted using subglottal IBIF to discriminate disorders associated with vocal hyperfunction. Our goal is to replicate the analysis performed by Hillman *et al.* [1] for selected parameters and conditions of interest. To accomplish this goal, we aim to address the following research questions:

1. Can we discriminate between patients with hyperfunctionally-related disorders and subjects with normal voices using a neck accelerometer signal, comparable with previous results obtained with airflow mask–derived aerodynamic measures?
2. What are the best aerodynamic measures to extract from the acceleration signal?
3. How is discrimination affected by changing the articulatory gesture from a sustained vowel /a/ to other vowels and running speech?

II. METHODS

We recruited five adult female subjects with bilateral vocal fold nodules and five female subjects matched for age and occupation. We recorded the neck-surface acceleration signal using the VHM simultaneously with three physiological signals of interest: oral airflow, electroglottograph (EGG), and sound pressure level (SPL). These signals were used to compute both subject-specific parameters for the IBIF scheme (see [4] for details) and compensate for loudness variation when performing the statistical analyses. The subglottal IBIF algorithm estimated four aerodynamic measures from the neck-surface accelerometer signal windowed into 100 ms non-overlapping frames: maximum flow declination rate (MFDR), amplitude of the modulated flow (AC Flow), open quotient (OQ), and speed quotient (SQ). These measures were selected so that we could compare the results of this study with those from [1].

The aerodynamic measures were computed for each of four vocal gestures performed at a comfortable loudness level (no specific target SPL): sustained vowels /a/, /i/, /u/, and the Rainbow Passage. Normal and regressed Z-scores were obtained for each gesture following the procedure described in [1], but using the matched-normal subject as the reference (rather than using a normative data set). Thus, mean measure values within each gesture for each patient were normalized by the means and standard deviations of the same vocal gesture from the matched subject. The normal (Z_N) and regressed (Z_R) Z-scores are computed, respectively, as

$$z_N = \frac{\bar{x}_P - \bar{x}_N}{\sigma_N} \quad \text{and} \tag{1}$$

$$z_R = \frac{\tilde{x}_P - \bar{x}_N}{\sigma_N}, \tag{2}$$

where $\bar{x}_P$ and $\tilde{x}_P$ are the normal and regressed observations for the patient, respectively, and $\bar{x}_N$ and σ_N are the mean and standard deviation of the matched subject. Given that the comparison is performed against one gesture on a single paired subject, the standard deviation refers in this case to the stability of the signal every 100ms, rather than the variation of a population.

Fig. 2 displays a graphical representation of the method used to compute Z_R. A robust least-square linear regression [5] is calculated for the normal subject data and used to extrapolate each aerodynamic measure to correct for loudness differences. The Z_R score for each patient's measures is given by the distance to the regression line, normalized by the standard deviation measured for the same gesture by the matched subject. Z_R scores are reported when a high Pearson's correlation ($|r| \geq 0.7$) exists between a given measure and SPL.

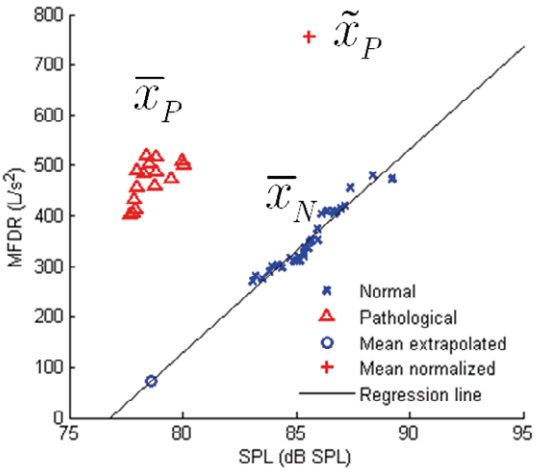


Fig. 2: Computation of regressed Z-scores for MFDR from the sustained vowel /a/ of subject pair 1

We present both Z_N and Z_R scores to evaluate the potential of the accelerometer signal to discriminate normal and pathological cases with and without the correction for the effect of SPL. This information may be used in the future to determine how to implement a real-time ambulatory application of this method for use in biofeedback.

Table I: Discrimination power between each subject pair. Regressed Z-scores scores for sustained vowel /a/ are comparable with those in [1]. Nomenclature: Z-scores ≥ 2 (+), Z-scores ≤ 2 (−), |Z-score| < 2 (no symbol).

Gesture	Subject pair	Normal Z-score				Regressed Z-score			
		MFDR	AC Flow	OQ	SQ	MFDR	AC Flow	OQ	SQ
Sustained vowel /a/	1		+	−		+	+		
	2	+	+				+		
	3	+				+			
	4		+				+		
	5	−	−				+		
Sustained vowel /i/	1	+	+			+	+		
	2		+				−		
	3	+	+			+	+		
	4	+	+			+	+		
	5				−				
Sustained vowel /u/	1	+	+			+	+		
	2	+	+			−	−		
	3	+	+			+	+		
	4	+	+			+	+		
	5	−	−			+	+		
Rainbow Passage	1		+						
	2								
	3								
	4		+						
	5								

III. RESULTS AND DISCUSSION

The summary of the Z-scores from the five subject pairs is shown in Table I. There are fewer Z_R scores Z_R reported because only those with $|r| \geq 0.7$ and $|Z_N| \geq 2$ were considered. For both scores, AC Flow is the most salient measure, followed by MFDR. The features OQ and SQ did not exhibit discriminating power. When comparing the results from vowel /a/ (the only gesture evaluated in previous studies) with the reference framework described in [1] for the nodules patients, the same trend in terms of salient measures is observed.

There were some negative values of Z_N for the /a/ vowel but no negative Z_R scores following adjustments/corrections for SPL (when correlations with SPL were high), which also matched previous results [1]. Our results for different sustained vowels yielded comparable discrimination power using Z_N but a slightly less robust behavior using Z_R; i.e., negative scores were observed for one subject pair (#2). It is possible that the SPL effect was not completely adjusted for due to the fact that only one matched subject was used in the linear regression, rather than a normative dataset. It is also possible that these lower vowels provide a different loading where reduced Z-scores are in fact correct. Given that these vowels have not been tested in previous studies, further investigations are needed.

The mean parameter values extracted from the Rainbow Passage did not correlate highly enough with SPL to allow for the calculation of Z_R scores. This may have been due to the averaging of measurements across many complex/variable phonemic environments. However, normal Z-scores using the unadjusted mean parameter values did provide salient scores in two of the patients for AC Flow. These findings suggest that smaller temporal windows may provide better discrimination and support the potential for computing real-time Z-scores in an ambulatory device.

V. CONCLUSION

This initial evaluation indicates that the accelerometer-based estimates of aerodynamic parameters obtained via subglottal IBIF provide comparable discrimination capabilities between normal and pathological subjects as was observed in previous studies using actual aerodynamic recordings. The results also provide a first indication that this discrimination is possible with other sustained vowels and even with running speech, thus motivating the continued development of these approaches for applications in ambulatory voice monitoring systems.

ACKNOWLEDGMENT

This work was supported by NIH-NIDCD grant R33 DC011588, CONICYT grant FONDECYT 11110147, and MIT MISTI grant MIT-Chile 2745333. V.E. acknowledges scholarships from CONICYT and Universidad de Chile.

REFERENCES

[1] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, "Objective assessment of vocal hyperfunction: An experimental framework and initial results," *J. Speech Hear. Res.,* vol. 32, no. 2, pp. 373–392, 1989.

[2] R. E. Hillman and D. D. Mehta, "Ambulatory monitoring of disordered voices," *Perspectives on Voice and Voice Disorders*, vol. 21, no. 2, pp. 56–61, 2011.

[3] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 11, pp. 3090–3096, 2012.

[4] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, "Subglottal impedance-based inverse filtering of speech sounds using neck surface acceleration," *IEEE Trans. Audio Speech Lang. Proc*., vol. 21, no. 9, pp. 1929–1939, 2013.

[5] R. A. Marona, R. Douglas Martin, V. J. Yohai, "Robust Statistics," John Wiley & Sons, 2006.

VOICE THERAPY ASISSTANT: A USEFUL TOOL TO FACILITATE THERAPY IN DYSPHONIC PATIENTS

I. D. Castro Miller[1], M. Moerman [2]

[1] University of Ghent, Erasmus Mundus CEMACUBE Student: Master in Biomedical Engineering, Belgium.
ivand.castro@hotmail.com
[2] AZ Maria Middelares, ENT/phoniatrics and Head and Neck Surgery department, Belgium.
mieke.moerman@azmmsj.be

Abstract: **A MatLab-based "Voice Therapy Assistant" software is presented, in which a patient is guided in order to perform prescribed speech therapy exercises by himself, and thus increase the frequency and surveillance of therapy and obtain results in less time. The software includes recording of each exercise in the therapy, and an assessment tool in which the speech therapist can evaluate both the performance and the evolution of the patient. The acoustic speech parameters Pitch, Noise to Harmonic Ratio, Relative Jitter, Intensity, and Maximum Phonation Time are calculated for the analysis, and a first set of 4 exercises for the patient is included. Future work includes a software called "Tele-Fon" which comprises the addition of a videoconference in order to allow telepractice of the therapist, and increase even more the frequency and results of the therapy, along with the possibility of treating patients that are located in communities in which no therapist is available.**

Keywords : **MatLab, Telepractice, Voice Therapy.**

I. INTRODUCTION

Voice therapy is used in many situations in which phonation improvement can be achieved by this means. This kind of repetitive "training" of the way a person speaks, is very useful in different pathological and non pathological situations, and in spite of the lack of existence of an unified tool to evaluate its effectiveness, it has shown to have positive results on patient's speech[1].

This paramedical activity has been applied in a variety of patients, including not only functional but also organic1 dysphonia [2] [3], or as a treatment after surgery, when the latter is needed.

The therapy's length has shown an enormous variation among different studies [1], as it varies from patient to patient, and depends on many factors such as the techniques applied. Although there are some recent studies about intensive techniques [4] that could lead to short term effects, most of the therapies require several visits from a speech therapist to the patient (or vice versa) in order to perform the required exercises, which implies mobilization costs, an increase on the time required to perform the therapy, and a reduction in the frequency of the therapy sessions.

These therapy sessions are accompanied by an assessment of progress that is performed by two main means: perceptual assessment and/or acoustic analysis [1] [5]. This fact means that either the speech therapist or a speech analysis system has to be present in order to keep track of the exercises, and make sure that the patient follows them in the right way.

Speech specialized software and hardware (ie. Kay Elemetrics [6], Praat [7], AMPEX [8][9][10]) is nowadays being used in order to do diagnosis, follow up of patients, and research in the phonetics field. In spite of this, these tools have to be properly used by a speech professional, are not meant for patient use, and are limited exclusively for speech analysis and diagnosis, leaving out the assessment of exercises to be performed by the therapist.

In order to increase the frequency of therapy, improve the recovery process and at the same time keep a record of the exercises performed by the patient, a software denominated "Voice Therapy Assistant" was designed. This software guides the patient through different therapy activities that can be done by himself, records the required information of each, and provides the therapist an acoustic analysis tool to evaluate the performed exercises after several sessions performed by the patient on his own.

[1] Abnormal function of voice related to a physical pathology, such as vocal cord nodules or polips [Audrey Millar, et al, 1999]..

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

II. METHODOS

The *"Voice Therapy Assistant"* was designed in MatLab, and included a Graphical User Interface (GUI) in which both the patient and the therapist could go through the different functions.

The software handles the patient information in an Excel file, which is filled when a new patient is added to the system; at the same time, a folder is created with the patient identification number in order to store the audio *.wav* files corresponding to the performed therapy exercises. The therapist has also the option to delete a patient, in which case not only his information will be deleted, but also the folder containing his information.

When the patient is using the assistant, he will be asked for his identification number and then will be guided through the specific exercises that were prescribed to him by the therapist, explaining each of them in a detailed manner. When he is done with the session, a reminder of when to repeat the session again, and when the next meeting with the therapist should be is shown.

The acoustic analysis tool for the therapist shows relevant variables and signals each session, not only for the analysis of the evolution of the patient, but also in order to identify whether the patient is following the exercises as it should be done. The variables and signals included in the analysis tool are:

- Sound signal vs. time graph
- Pitch vs. time graph
- Noise to Harmonic Ratio vs. time graph
- Relative Jitter
- Lowest Intensity
- Highest Intensity
- Lowest pitch
- Highest pitch
- Pitch total range
- Maximum Phonation Time (Specific exercise)

Although the possible exercises for speech therapy include a wide range of activities, this first version of the software copes with 4 main exercises as follows:

- StA1: Posture and Breathing
- StA2: Breathing and MPT measure
- StS3/SlS2: Glottis Adduction
- SlS1: Tongue muscles and strength

In the following sections the calculation method for the parameters will be discussed.

A. Pitch detection algorithm

In order to calculate the natural frequency F0 at which vocal cords are vibrating, several algorithms have been developed through time, and the different approaches have shown a general improvement of the obtained results.

One of the initial approaches [Gold and Rabiner, 1969] [11] works with the speech signal in the time domain, and is a low complexity algorithm based on peak and valley detectors and similarity detectors that does the identification process by looking for similarity patterns in the speech waveform. Although its simplistic approach, it is still useful for applications with low computational resources.

An also computational favorable algorithm was based on the average magnitude difference function of the signal with the time-lag version of itself [12], but its application was only due to restrictions in computational power, which is no longer the case, as the correlation concept can be nowadays applied at ease.

For the rest of methods, most of them are based on the autocorrelation concept [Rabiner, L. 1977] [13], in which the speech signal is multiplied by a shifted version of itself, allowing to identify the points in which the pattern is repeated, and thus calculate the period T of the vocal cord vibration. The autocorrelation is calculated as shown (1).

$$P_{i,k} = \sum_{j=m}^{m+n-k-1} s_j s_{j+k} \qquad (1)$$

s= voice signal
k= autocorrelation lag
m=frame index i frame interval z (samples)*
n=window width (samples)

This generic autocorrelation method, has the specific characteristic of choosing a limited window of analysis *w* (it has to be at least twice the glottal period T0), which introduces a reduced statistical significance as the lag k decreases, due to less samples being correlated., and implies large sizes for *w* in order to have a good performance at long lags (low F0).

Other widely known algorithm in the speech signal processing is the so called *"Power Cepstrum"*, which is based on the squared inverse Fourier transform of the short-time logarithm of the spectrum of the squared signal, and was first defined by Bogert, B.P in 1963 [14], and later applied to speech by Noll, A.M in 1967 [15]. This calculation leads to local maximum values at times kT, being T the period of vocal cord vibration. In spite of

this, the calculation window has the same characteristics of the autocorrelation calculation, and thus related disadvantages to the method.

This disadvantages are covered by the Cross-Correlation method shown in (2), in which the window *w* can be as long as the smaller expected period T0, as it has no theoretical bounds. In spite of this, the same fact that overcomes the disadvantage of the generic autocorrelation, introduces a new one, as the correlation at k=0 might not be the highest peak, and thus simple threshold detection is not enough to generate reliable pitch candidates.

$$X_{i,k} = \sum_{j=m}^{m+n-1} s_j s_{j+k} \tag{2}$$

Improving the mentioned techniques, an algorithm named RAPT is introduced in 1995 by David Talkin [16], and is based on the calculation of the Cross-Correlation which is then normalized by the energy of the signals, as is shown in (3) and (4):

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_j e_{j+k}}} \tag{3}$$

$$e_j = \sum_{i=j}^{j+n-1} s_i^2 \tag{4}$$

This method gives as a result a signal that can be evaluated at its maximum peaks in order to look for pitch candidates, at peaks close to 1, and guarantees a value close to zero (or considerably less than 1 in the worst cases) for lags different than 0.

A comparison study between RAPT algorithm and other 5 state of the art techniques, stated that it is comparable to the widely used PRAAT software [17], and indicated RAPT with the lowest number of gross pitch errors.

Taking in to account the presented perspective, and newly improvements in the RAPT algorithm, a RAPT-based algorithm (IRAPT) was chosen for the *"Voice Therapy Assistant"* software. This algorithm, developed by Azarov et. al [18] in 2012, includes an instantaneous version of the initial RAPT algorithm, having as a result a pitch estimate each 93 ms, and an additional post processing procedure to improve accuracy, obtaining a Gross Pitch Error (GPE) considerably lower than the original RAPT article.

B. Relative Jitter

Jitter is defined as the relative variation of the natural frequency, and its relative measure is defined as the average consecutive difference between consecutive periods, divided by the average period as shown in (5).

$$Jitter\ (absolute) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{5}$$

Where **Ti** are the period lengths and **N** is the number of extracted periods.

This has been used as an indicator of the voice quality [1], being a parameter that evidences the severity of dysphonia of the patient included in most of the clinical speech analysis.

C. Harmonic to Noise Ratio -HNR-

This parameter indicates the relationship between the harmonic signal and the noise, in order to quantify the vocal function, and identify possible hoarseness due to leaks on the glottal closure during phonation.

The HNR is calculated then as the relation in dB between the energy of the harmonic signal and the energy of the noise present in the speech acustic signal, as indicated in (6).

$$HNR = 10 \log \frac{E_s}{E_n} \tag{6}$$

Three main methods were found that were developed in order to fulfill this requirement [19][20][21]. A comparison of these methods by [Severin, F. et al. 2005] [22], determined that *"they are all good indicators of the amount of noise in speech"*, and are thus *"efficient for voice quality analysis"*. Considering this, the algorithm from [G. de Krom 1993] was chosen to be included in the software, only because of availability reasons.

D. Intensity

The intensity of the voice was obtained in dB, by calculating the energy of the signal and its ratio with the energy of the hearing treshold I0 as indicated in (7).

$$I(dB) - 10 \log_{10} \frac{I}{I_0}$$
$$I_0 = 10^{-12} \frac{watts}{m^2} \tag{7}$$
$$I = s^2$$

E. Maximum Phonation Time -MPT-

This parameter is measured by considering the active recording time during the exercise. The details can be confirmed by the therapist as the voice signal of the exercise is available in the analysis tool.

III. Results

The *"Voice Therapy Assistant"* was finally compiled in a windows executable file, in which any user without MatLab installed could use the software. Fig. 5 and Fig. 6 show an example of one exercise indication for the user and the evaluation tool of the speech therapist respectively.

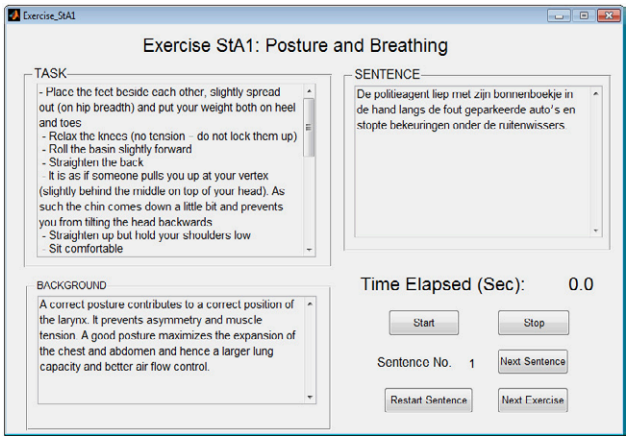


Figure 1. Example of exercise window for patient

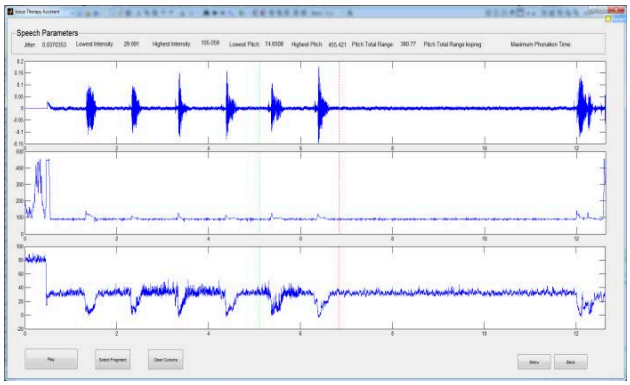


Figure 2. Example of analysis window for therapist

IV. Discussion

A MatLab based software for self Voice Therapy was built, and 4 specific exercises were included. The software guides the patient through the prescribed exercises and records his voice, showing in one of the cases some parameters while the recording is being performed.

The software also allows a speech therapist to review the exercises that have been done by a specific patient on his own, and to assess his performance by means of acoustic parameters extracted from the audio signal.

Although an analysis of different available algorithms for extraction of each of the parameters was done, it has to be considered that these techniques should be ideally tested for the specific application and compared with the software and hardware used in the clinical environment.

It is also a fact, that there are some other pitch extraction techniques available that were not included in the analysis, such as YIN [23] (2002, Autocorrelation based with modifications), and SWIPE [24] (2007, Based on comparison with pitch of saw-tooth waveforms), and that could also be considered for further testing in the application.

According to this, this first version of the software is intended to be tested by therapists in their patients, in order to evaluate its functionality and propose some modifications both in the user interface and in the parameter extraction techniques.

V. Conclusion

The authors identify the potential use of this software in the daily practice of speech therapists, not only by allowing self therapies of the patient himself, but also by including the possibility to the therapist to do a follow up by using the recorded parameters. Although this first version of the software does not include an on-line data sharing for the recorded parameters (follow up is done when the therapist visits the patient), following versions could include the transmission of the parameters via internet.

As future work, the authors also propose a distance-guided therapy, in a new version of the software called *"Tele-Fon"* in which the use of internet allows the therapist to communicate in a videoconference with the patient, and guide him through each of the exercises included in the software.

There is also the need to include a broader range of exercises, so that the therapist can use the software in patients with different pathologies.

References

[1] R Speyer. Effects of voice therapy: A systematic review. Journal of Voice. October, 2006, vol. 22, no. 5, p. 565-580.

[2] Eva B. Holmberg, et al. Efficacy of a behaviorally based voice therapy protocol for vocal nodules. Journal of Voice. 2001, vol. 15, no. 3, p. 395-412.

[3] Janina K. Casper, Thomas Murry. Voice therapy methods in dysphonia. Voice disorders and phonosurgery II, Otolaryngologic Clinics of North America. October 2000, vol. 33, no. 5, p. 983-1002.

[4] Rita R. Patel, et al. Boot Camp: A novel intensive approach to voice therapy. Journal of Voice. January, 2010. vol. 25, no. 5, p. 562-569.

[5] P.H. Dejonckere, et al. Tridimensional assessment of adductor spasmodic dysphonia pre- and post- treatment with Botulinum toxin. European Archives of Oto-Rhino-Laryngology. April, 2012. p 1195-1203.

[6]Kay Pentax [online] Available on: http://www.kayelemetrics.com/ . [Consulted March 1 2013].

[7] Praat: Doing Phonetics by computer. [online]. Amsterdam, the Netherlands. Available on: http://www.fon.hum.uva.nl/praat/ [Consulted March 12 2013].

[8] P. H. Dejonckere, et al. Voicing quantification is more relevant than period perturbation in substitution voices: an advanced acoustical study. European Archives of Oto-Rhino-Laryngology. April, 2012. p 1205-1212.

[9] ELIS DSSP. Disordered Voice Analysis software. Ghent, Belgium. Available on: http://dssp.elis.ugent.be/downloads-software [Consulted March 15 2013].

[10] Moerman M.B. J, Pieters G, Martens JP, Van der Borgt MJ, Dejonckere PH. Objective evaluation of quality of substitution voices. Eur Arch Otorhinolaryngol 2004; 261 (10): 541-7.

[11] B. Gold, L. Rabiner. Parallel processing techniques for estimating Pitch periods of speech in the time domain. J. Acoust. Soc. Am. 1969, vol. 46, no. 2B, p. 442-448.

[12] Ross, M. Shaffer, H. Cohen, A. Freudberg, R. Manely, H. Average magnitude difference function pitch extractor. IEEE Transactions on Acoustics, Speech and Signal Processing. 1974, vol. 22, no. 5, p. 353 - 362.

[13] Rabiner, L. On the use of autocorrelation analysis for pitch detection. IEEE Transactions on Acoustics, Speech and Signal Processing. 1977, vol. 25, no. 1, p. 24-33.

[14] B. P. Bogert, M. J. R. Healy, and J. W. Tukey. The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking. Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15, 209-243. New York: Wiley, 1963.

[15] A. Michael Noll. Cepstrum Pitch Determination. Journal of the Acoustical Society of America. 1967, vol. 41, no. 2, p. 239-309.

[16] Talkin, David. Speech Coding and Synthesis: A robust algorithm for pitch tracking. Elsevier Science B.V. 1995. p. 495-518.

[17] Onur Babacan, Thomas Drugman, Nicolas d' Alessandro, Nathalie Henrich, Thierry Dutoit. A comparative study of pitch extraction algorithms on a large variety of singing sounds. University of Mons, Belgium.

[18] Elias Azarov, Maxim Vashkevich, Alexander Petrovsky. Instantaneous pitch estimation based on RAPT framework. 20th European Signal Processing Conference -EUSIPCO 2012- (August 27-31 2012: Bucharest, Romania).

[19] Guus de Krom. A Cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. Journal of Speech and Hearing Research. 1993, vol. 36, p. 254-266.

[20] D' Alessandro, C. Decomposition of speech signals in to deterministic and stochastic components. International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95. vol. 1. p. 760-763.

[21] Boersma, Paul. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings 1993, vol. 17, p. 97-110.

[22] Severin, François, Baris Bozkurt, and Thierry Dutoit. HNR extraction in voiced speech, oriented towards voice quality analysis. Proc. eUSiPcO, vol. 5. 2005.

[23] Alain de Cheveigne. YIN, a fundamental frequency estiamtor for speech and music. Journal of the Acoustical Society of America. 2002, vol. 111, no. 4, p 1917-1930.

[24] Camacho A, Harris JG. A sawtooth waveform inspired pitch estimator for speech and music. Journal of the Acoustical Society of America. 2008, vol. 124, no. 3, p 1638-1652.

EVALUATION OF SURGICAL TREATMENT OUTCOME IN REAL-TIME CONDITIONS USING A PORTABLE DEVICE: PRELIMINARY DATA.

D. Kiagiadaki[1], A. Cateau [2], M. Remacle [1,3], J. Schoentgen[4], T. Dubuisson[5]

[1] Department of Otolaryngology-Head and Neck Surgery, Louvain University Hospital of Mont Godinne, Yvoir, Belgium,
[2] Faculty of Psychology and Educational Sciences, Université Libre de Bruxelles (ULB), Belgium,
[3] University of Louvain, Saint-Luc University Hospital, Brussels, Belgium
[4]Laboratory of Images, Signals and Telecommunications Devices,
Université Libre de Bruxelles (ULB), Brussels, Belgium
[5] Ir, Phd, xperthis company, Mons, Belgium
E mail addresses: derygr@yahoo.com, acateau@ulb.ac.be, marc.remacle@uclouvain.be, jschoent@ulb.ac.be, thomas.dubuisson@xperthis.be

Abstract: **In the present study, we present the preliminary results of clinical voice testing with the ¨i-Pod¨ device, for the objective assessment of surgical treatment outcome in association with subjective patient's evaluation. Four patients with benign laryngeal pathologies were tested before and after the surgical intervention, as well as eight healthy subjects. The level of vocal discomfort was noted during the testings. Acoustic analysis included fundamental frequency, intensity and Signal-to-Dysperiodicity Ratio. All subjects completed also the VHI questionnaire and the Bipolar Scale of Dejonckere, at the time of each testing.**

Acoustic measurements didn't change significantly neither between the morning and afternoon nor between the recording intervals for the control group (paired t – test, *P*>0.05). The total changes in the patients' group showed no worth-mentioning changes except for F0. SDR was able to discriminate patients and controls both in pre- and post-operative measurements (*P*<0.05).

Interestingly, patients showed an increased total discomfort value post-operatively, fact that could be explained by individualizing the results. Vocal discomfort correlated mostly with the VHI functional sub-scale and variably with the scoring of several items of Dejonckere's scale.

Preliminary data show the sensibility of the ¨i-Pod¨ to assess changes in daily activities and between patients and controls.

Keywords: **real-time voice monitoring, portable device**

I. INTRODUCTION

The idea of a mobile, real-time monitoring of daily voice use has been recently the subject of many research studies. Such devices could be useful in complementary laboratory voice assessment in "real-life conditions", concerning the diagnostic work-up and the evaluation of treatment outcome. Already, portable voice accumulators have been developed, enabling the estimation of F0, SPL and phonation time [1, 2, 3]. The use of accelerometer sensors has both been tested in pilot studies [4-6] and in already commercially available devices. The portable device "i-Pod'', has been designed for the continuous assessment of voice disorders with real-time coupling of acoustic and patient self-evaluation measures. In the present study, we aim to present the preliminary results of clinical voice testing, for the objective assessment of surgical treatment outcome in association with subjective patient's evaluation.

.

II. METHODOS

Technology: ¨I-Pod¨ is a voice recording device, using the iPhone platform combined with a microphone attached to the patient's skin, at the lateral surface of the thyroid cartilage.

Participants: Four patients, (2 males and 2 females, mean age 29 years) with benign laryngeal pathologies (polyps, nodules, sulcus vocalis) were tested before and after the surgical intervention, when the healing period was considered completed. Eight healthy subjects (females, mean age 25.3 years) with important professional voice demands participated, also, tested in 2 different time – intervals. The evaluation included the testing with the portable device in the subject's professional or domicile environment, in the morning and in the afternoon of a single or two consecutive days. Recording settings were adjusted to 2 minutes every 10 minutes. The level of vocal discomfort was noted during the recordings in a visual analogue scale (VAS) with the use of a tactile cursor on the device's screen.

All subjects were asked to complete the VHI [7] questionnaire and the Bipolar Scale of Dejonckere [8], at the time of each testing.

Data extraction and analysis: Sound files (wav.), as well as voice discomfort measurement files, were extracted with the use of a conventional PC. In total, a mean of 15 and 10 voiced recorded files per testing were collected for control group and the patients´ group respectively. Voiced segments were manually selected from every sound file. Acoustic analysis of the voice signals was

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

achieved with the aid of the xperthis® software analysis system, concerning the fundamental frequency (F0), intensity (I) and Signal-to-Dysperiodicity Ratio (SDR). Due to the high variability of the mean values calculated automatically by the system, only the median values were taken into account and the mean values were calculated subsequently for the morning and afternoon of each recording. Statistical analysis was carried out with the SPSS IBM 20.0.

III. Results

A. Acoustic measurements

Acoustic measurements didn't change significantly neither between the morning and afternoon measurements nor between the recording intervals for the control group (paired t-test, $P>0.05$). The total changes in the patients' group showed no worth-mentioning changes except for F0 (Table 1). Mean values of F0 and I were not significantly different between patient and controls, whereas SDR was able to discriminate patients and controls both in pre and post-operative measurements ($P<0.05$).

B. Vocal discomfort

Changes in vocal discomfort are shown in Table 2. None of the changes were statistically significant (paired t-test, $P>0.05$). Interestingly, patients showed an increased total discomfort value post-operatively, fact that could be explained by individualizing the results. Subjects with high discomfort values, had usually decreased intensity and SDR values (indicating highly perturbated voice). Controls reported always less vocal discomfort compares to patients, however these differences were not proven statistically significant.

C. Vocal discomfort and subjective rating questionnaires

Preoperatively, VHI grading (total and subscale) was statistically significant between the 2 groups, as well as the rating of items 7 and 8 of Dejonkere's scale (´´désagréable – agreeable`` and ´´raque – pure``, respectively) ($P<0.05$). No such differences were observed post-operatively.

Correlations of vocal discomfort with the VHI were mostly with the functional sub-scale, when all the participants were analysed in one group as well as for the control group separately, only for the pre-operative measurements (rho= 0.6 and 0.73 respectively, $P<0.05$). Moreover, it correlated strongly with the scoring of several items of Dejonckere's scale, in a variable and not reproducible manner.

D. Vocal discomfort and acoustic measurements

No correlations were observed between vocal discomfort and the acoustic mean values between pre and post-operative measurement, neither for the control or the patient group.

Table 1. Changes in acoustics measurements.

Measurement [Mean (SD)]	Control (N=8)	Patients (N=4)
F0 morning_pre	242.6(30.4)	203.8(40.6)
F0 morning_post	222.2(34.1)	174.9(33.1)
F0 after_pre	235.4(24.6)	190.5(35.5)
F0 after_post	221.5(18.2)	172(20.9)
I morning_pre	74.8 (3.2)	70.9(2.1)
I morning_post	74(2.2)	67.3(4.3)
I after_pre	75.5(2.8)	75(5.2)
I after_post	75.4(2.9)	68.7(4.1)
SDR morning_pre	14.8(1.5)	8.3(5.2)
SDR morning_post	14.4(3.1)	8.9(1.5)
SDR after_pre	15(3.2)	9.8(1.4)
SDR after_post	14(2.2)	10(1.7)

Table 2. Vocal discomfort changes.

Vocal discomfort [Mean (SD)]	Control (N=8)	Patients (N=4)
Discomf. morning_pre	19.7(10.6)	21.2(4.9)
Discomf. morning_post	25(12.5)	50.2(37.8)
Discomf. after_pre	23.5(11.4)	37.2(18.2)
Discomf. after_post	25.7(9.7)	51(38.3)
Discomf.tot_pre	21.6 (9.8)	29.2(10)
Discomf.tot_post	25.4(9)	50(33)

IV. Discussion

Real-time voice monitoring and evaluation with the use of portable devices, has been a challenging field of research, as it can provide detailed information of voice production and behavior during a continuous period of time. The ¨i-Pod¨ device is easy to use and its reliability of recording with the use of a conventional microphone, has been previously proved in a laboratory environment. It is the first time from its development that it is used in ecological conditions with real-time coupling of acoustic measurements and the patient's self-evaluation (vocal discomfort).

The test-retest in the control group proved its reliability in reflecting the subjects' overall state of vocal performance in both recordings. In the patients' group it was able to detect the daily voice fluctuations, giving the opportunity to visualize not only the overall voice characteristics but also the number of fluctuations per day of recording.

The ability to evaluate patients` discomfort fluctuations is a basic advantage of the "i-Pod´´ device. The users have the opportunity to note the level of vocal discomfort at any moment, during or out of voice recordings. This can be useful for the evaluation of vocal performance in relation with the voice loading, both for patients and for healthy individuals (e.g professional voice users), as well as its changes e.g, after an operation or during a working day. In our group of participants, the indices of vocal loading (F0, I) and their changes, could not be correlated with the vocal discomfort evaluation, however, more valuable conclusions could be extracted when each subject was studied individually. Moreover, the fact that among the patients, the overall vocal discomfort took greater values after the surgery, could be attributed partially, to problems in the use of cersor in one patient and a long lasting rehabilitation period for a second patient, suffering also of vocal dysfunction. Finally, the variable and not constant strong correlations of vocal discomfort values with either the VHI or Dejonckere's scale scoring, indicates that the vocal discomfort in real time conditions cannot reliably be demonstrated in the scoring of the widely used self-evaluating questionnaires (e.g. VHI).

Among the acoustic measurements, SDR was proven the most sensitive for the detection of pathological voices (statistically different between controls and patients). As far as the surgical outcome is concerned, although we observed raised SDR values after the surgery, the small number of patients didn't permit to perform a reliable statistical analysis. The SDR [9] has been shown to correlate strongly with the degree of perceived hoarseness and discriminate adequately the patients´ and normophonics´ connected speech samples.

The clinical use of the ¨i-Pod¨ device was proven easy and reliable. Among the technical issues to be improved, is the limited battery capacity, which permitted only 2 hours of autonomous function and consequently a limited recording duration time. Also, the visible microphone, discouraged people -mainly normophonics- to participate in the study, but generally, it didn't pose a problem to the patients, when its usefulness for the assessement of vocal function, was adequately explained.

V. Conclusion

The preliminary results of the ¨i-Pod¨ use in clinical practice have been promising. It is able to provide a real-time voice assessment, coupling acoustic and patient self-evaluation measures. Daily discomfort fluctuations can provide information for the functional element of dysphonia and the state of the healing procedure and rehabilitation post-operatively for the patients, as well as for the effect of vocal loading for the controls, fact that cannot always be adequately estimated with the use of conventional scales and questionnaires (Dejonckere's Bipolar Scale, VHI).

References

[1] AC. Ohlsson, O. Brink, A, "A voice accumulation-validation and application.'' *Speech Hear Res,.* vol. 32(2), pp 451-457, 1989.

[2] R. Buekers, E. Bierens, H. Kingma, EH. Marres, "Vocal load as measured by the voice accumulator.'' *Folia Phoniatr Logop.*, vol 47(5), pp 252-261, 1995.

[3] HA. Cheyne, HM. Hanson, RP. Genereux, KN. Stevens, RE. Hillman, "Development and testing of a portable voice accumulator." *Speech Language and Hearing Research,* vol. 46, pp. 1457–1467, 2003.

[4] DD. Mehta, M. Zañartu, SW. Feng, HA. Cheyne, "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform.'' *IEEE Trans. Biomed Eng.,* vol 59(11), pp. 3090-3096, 2012.

[5] A. Carullo, A. Penna, A. Vallan, A. Astolfi, P. Bottalico, "A portable analyzer for vocal signal monitoring.'' *Instrumentation and Measurement Technology Conference (I2MTC), IEEE International*, pp. 2206-2211, 2012.

[6] PS. Popolo, JG. Scaronvec, IR. Titze, "Adaptation of a pocket PC for use as a wearable voice dosimeter.'' *J Speech Lang Hear Res.,* vol. 48(4), pp. 780-91.

[7] BH. Jacobson, A. Johson, C. Grywalski, A. Silberglent, G. Jacobson, MS. Benninger, C. Newman, "The Voice Handicap Index (VHI): development and validation.'' *Am J Speech Lang Pathol,* vol 6, pp. 66–70, 1997.

[8] PH. Dejonckere, FR. Dejong-Estienne, "Techniques de base d' évaluation de la voix'', Cabay, 1985.

[9] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen, "Multi-band dysperiodicity analyses of disordered connected speech.'' *Speech communication*, vol. 5(1), pp. 131-141, 2011.

A NEW TECHNIQUE TO RECORD A VOICE SOURCE SIGNAL

K.V. Evgrafova[1], V. V. Evdokimova, [2] P. A. Skrelin[3], T. V. Chukaeva, [4] N. V. Shvalev[5]

[1] Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia, evgrafova@phonetics.pu.ru
[2] Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia, evdokimova@phonetics.pu.ru
[3] Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia, skrelin@phonetics.pu.ru
[4] Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia, chukaeva@phonetics.pu.ru
[5] The State Academic Mariinsky Theatre, Saint-Petersburg, Russia, dr-nix99@mail.ru

Abstract: **The given paper presents a new method to record a voice source signal. It allows registering the voice source by a special miniature microphone which is located in the proximity of the vocal folds. Thus there appears an opportunity to record the voice source signal and the output speech signal synchronously. Being an invasive technique, the method does not bring any significant discomfort for a subject. The position of the microphone does not interfere with the naturalness of the speech sounds produced. In the paper the procedure of recording is described and the results of the perceptual tests are presented. The perceptual tests were conducted in order to find out if the voice signals recorded can be identified as any of Russian vowels. The recognition patterns of them are given.**

The analysis and comparison of the synchronous voice source and speech signals can provide a better understanding of the human speech production system. Besides, it allows developing voice source models more accurate compared to the existing ones. This knowledge can be applied to many applications such as speech/speaker recognition, speech synthesis, emotion identification, age identification, speech coding and various medical applications.

Keywords: **voice source, vocal folds, speech signal**

I. INTRODUCTION

Analysis of the voice source is essential to the understanding of the human speech production system and, as a result, developing more accurate source models. However, due to the position of the vocal folds, immediate registering the voice source signal is hampered. Typically, the voice source is estimated by separating the source signal from the speech signal. Traditionally, inverse-filtering or joint estimation techniques have been employed to extract the source signal [1-3], [5]. However, these techniques are based on the assumption that speech production is a linear and time-invariant process [4], which is not. The non-linear interactions between the source and the vocal tract can result in inaccuracies which may be reflected in both the source signal and vocal tract filter estimates. Another noninvasive method of obtaining voice source measurements is through the use of electroglottography (EGG) [6]. It uses a pair of electrodes fixed on the neck skin, next to the larynx, sensitive to the vibrational activity present in this region. However, EGG signals are, still, another form of indirect measurement, and can suffer from attenuation of some signal components by the tissues (where the electrodes are fixed), which results in distortion of the EGG information.

In this paper a new approach to obtaining a voice source signal is proposed. It allows registering directly the voice source signal in the proximity of the vocal folds and the output speech signal synchronously.

II. METHODS

Equipment and Procedure

The recording experiment was conducted with the use of two microphones – the capacitor microphone and the miniature microphone QueAudio (d=2.3 mm, waterproof). The AKG HSC20 microphone was located near the lips of the subject. The miniature one was inserted through the nasal cavity and located in the proximity of subject's vocal folds. This procedure was performed by a phoniatrician who used special medical equipment. The speech signal containing isolated vowels and connected speech was registered synchronously through both microphones. The subject of the experiment, a female native speaker of Russian, was asked to pronounce the 6 isolated Russian vowels /a/, /e/, /i/, /1/, /o/, /u/ for 1 second repeatedly. The overall length of the speech signal was 15 minutes. The recordings were made in the recording studio at the Department of Phonetics, Saint-Petersburg State University. Multichannel recording system Motu Traveler and WaveLab program were used. The recordings had a sample rate of 32000 Hz and a bitrate of 16 bits.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Perceptual tests

Two groups of informants (25 individuals) were involved into perceptual tests. The groups consisted of 5 participants who were experts in phonetics and 20 lay participants.

The stimuli were presented to informants in order to find out the way if a voice source system could be identified a speech sound. The samples were organized on a random basis. The informants were asked to make judgments with respect to each stimulus and decide whether it could be identified as any of 6 Russian vowel phonemes.

The results of the perceptual tests were placed in confusion matrices which showed recognition patterns for each stimulus.

III. RESULTS

Overall, the perceptual test did not confirm the presupposition that all the stimuli should sound similar. It was based on the classic theory of speech production by G. Fant according to which these are formants that define the phonetic quality of a vowel.

However, the group of expert participants 1) distinguished and all the stimuli 2) recognized correctly practically all of them. The Russian vowels /i/ and /y/ which are quite similar were confused by some participants.

The group of lay participants also answered that they could hear different vowels. However, not all of the stimuli sounded intelligible for them.

The vowels /a/. /e/, /i/ stayed most intelligible and were identified correctly in most instances.

However, there were strong confusions of /i/ and /u/, /i/ and /y/ and /u/ and /y/.

Besides, /a/ and /i/ vowels were often perceived as labialized.

IV. DISCUSSION

Thus the used technique allowed obtaining the voice source signal which sounded quite similar to the output speech which had been recorded synchronously.

Being an invasive technique, the method did not bring any significant discomfort for a subject. The position of the microphone did not interfere with the naturalness of the speech sounds produced.

According to the classic theory of speech production by G. Fant, the phonetic quality of a vowel is defined by formant values. However, there exists another approach. V. Galunov claims that the quality of a vowel is formed before a voice signal passes through the filter component [5].

The results of the experiment show the acoustic energy is reflected backwards by the filter component.

V. CONCLUSION

The use of the miniature microphone inserted through the nasal cavity allows recording the voice source signal synchronously with the speech signal. The analysis and comparison of the synchronous voice source and speech signals can provide a better understanding of the human speech production system. Besides, it allows developing voice source models more accurate compared to the existing ones. This knowledge can be applied to many applications such as speech/speaker recognition, speech synthesis, emotion identification, age identification, speech coding and various medical applications.

REFERENCES

[1] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Communication, vol.* 11(2–3), pp. 109–118, 1992.

[2] O. Akane, P. Murphy, “Estimation of the vocal tract transfer function with application to glottal wave analysis,” *Speech Communication*, vol. 46, pp. 15-36, 2005.

[3] W. Ding, N. Campbell, N. Higuchi, and H. Kasuya, “Fast and robust joint estimation of vocal tract and voice source parameters. “ *IEEE International Conference on Acoustics*, *Speech, and Signal Processing, ICASSP-97*, pp. 1291-1294, 1997.

[4] G. Fant, *Acoustic Theory of Speech Production*. Netherlands: Mouton, 1960.

[5] V.I. Galunov, V. I. Garbaruk, “The acoustic theory of speech production and the system of phonetic features.” *The100 years of experimental phonetics in Russia*. The proceedings of international conference. The philological faculty, SPSU, 2001 (in Russian).

[6] M. Rothenberg, “A new inverse-filtering technique for deriving the glottal air flow waveform during voicing,” *J. Acoust. Soc. Am.,* vol. 53(6), pp. 1632–1645, 1973.

[7] M. Rothenberg, “A Multichannel Electroglottograph,” *Journal of Voice*, vol. 6., No. 1, pp. 36-43,1992.

VOICE DOSIMETRY IN 92 CALL CENTER OPERATORS

Giovanna Cantarella [1], Elisabetta Iofrida [1], Paola Boria [2],Simone Giordano [2], Oriana Binatti [2], Lorenzo Pignataro [1,3], Claudia Manfredi [4], Stella Forti [5], Philippe Dejonckere [6]

[1]Otolaryngology Department and [5]Audiology Unit, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
[2]Occupational Medicine, Private Practice, Milan, Italy
[3] Dipartimento di Scienze Chirurgiche Specialistiche, Università degli Studi di Milano, Milan, Italy
[4] Department of Information Engineering, Università degli Studi di Firenze, Florence, Italy
[6] Neurosciences,University of Leuven & Federal Institute for Occupational Diseases, Brussels, Belgium.

Abstract* - **The voice is a primary work tool for call center operators, but the main risk factors for voice disorders in this category have not yet been clarified. This study aimed to analyse the vocal behaviour in call center operators and to search for correlations between the daily voice dose and self-perceived voice-related handicap.**

Ninety-three subjects (25 males, 68 females, aged 24-50) underwent ambulatory phonation monitoring during a working day and were administered a general questionnaire (concerning smoking habits, symptoms, extra-work activities) and the Voice Handicap Index questionnaire (VHI).

The recorded vocal doses showed wide inter-subject variability, both at work and during non-work hours. The mean percentage phonation time (PT) during work was 14.74 and ranged from 4 to 31%. The average voice amplitude was higher in subjects with longer phonation time and higher F0. This finding indicates that "intensive talkers" also tend to use a higher voice volume.

The VHI score (mean 13.6 ± 12.2) was not related to the number of work hours, indicating that work time is not a critical factor in causing the perception of voice fatigue. The mean PT was 87.5 minutes (range 17-186 minutes) and was not correlated with age, gender, number of work hours, symptoms, extra-professional voice use and VHI scores. The mean amplitude was significantly higher in subjects with longer PT ($p < 0.001$). PT during work was related to the number of work hours, but no correlation was found between the PT of the whole recording day and the number of work hours.

In conclusion, our study demonstrates that the number of work hours and the percentage PT are not statistically related to the perception of voice disturbances.

Our data show that "safety" limits of vocal load in the call center setting cannot be clearly defined.

In analogy with previous findings on teachers, we postulate that constitutional and psycho-emotional features might be relevant risk factors for the development of voice pathologies.

Keywords **– voice dosimetry, occupational voice, voice handicap index**

REFERENCES

[1] Hillman RE, Heaton JT, Masaki A, et al. Ambulatory monitoring of disordered voices. *Ann Otol Rhinol Laryngol* 2006;115:795-801.

[2] Cheyne HA, Hanson HM, Genereux RP, et al. Development and testing of a portable vocal accumulator. *J Speech Lang Hear Res* 2003;46:1457-67.

[3] Titze IR, Hunter EJ, Svec JG. Voicing and silence periods in daily and weekly vocalizations of teachers. *J Acoust Soc Am* 2007;121:469-78.

[4] Titze IR, Lemke J, Montequin D. Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report. *J Voice* 1997;11:254-9.

*Full paper withheld by authors' request.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Session V:
MODELS AND ANALYSIS (II)

MULTIBAND VOCAL DYSPERIODICITIES ANALYSIS USING EMPIRICAL MODE DECOMPOSITION IN THE LOG-SPECTRAL DOMAIN

A. Kacha[1], F. Grenez[2], J. Schoentgen[2,3]

[1] Laboratoire de Physique de Rayonnement et Applications, Université de Jijell, Jijel, Algeria
[2] LIST Department, Universite´ Libre de Bruxelles, Brussels, Belgium
[3] National Fund for Scientific Research, Belgium
akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

Abstract: **In this paper, empirical mode decomposition (EMD) is proposed as an alternative to decompose the log magnitude spectrum of the speech signal into its harmonic, envelope and noise components. The EMD-based approach used in this study incorporates an appropriate procedure that estimates automatically the thresholds used by the clustering algorithm without knowledge of the fundamental frequency. The frequency range of the harmonic and noise components is divided into ten equally-spaced intervals and the harmonics-to-noise ratios (HNRs) within each interval are used as independent variables to summarize the amount of perceived hoarseness. The proposed method is evaluated on a corpus comprising 251 normophonic and dysphonic speakers.**
Keywords: **vocal dysperiodicities, empirical mode decomposition, multiband analysis, disordered speech.**

I. INTRODUCTION

Acoustic analyses of disordered speech are of great importance for clinical evaluation of voice disorders because they are noninvasive and enable clinicians to monitor the progress of patients and document quantitatively the perceived degree of hoarseness. Despite the number of acoustic markers that have been proposed in the literature to characterize the speech of dysphonic speakers, finding reliable and accurate descriptors of voice function and voice quality is still an issue.

Recent approaches for vocal dysperiodicities estimation have focused on continuous speech. In [1], the performance of multi-band segmental signal-to-dysperiodicity ratio has been investigated in terms of the correlation with scores of perceived hoarseness. It has been concluded that multi-band segmental signal-to-dysperiodicity ratio correlates more strongly with the perceptual assessment of the degree of hoarseness than the full-band analysis.

In [2], we proposed the empirical mode decomposition (EMD) algorithm [3] as an alternative to decompose the log of the spectrum magnitude of the speech signal into its harmonic, envelope and noise components. The acoustic cue named harmonic-to-noise ratio (HNR) has been used to summarize the degree of disturbance in the speech signal and consequently to evaluate the overall quality of the disordered voices produced by dysphonic speakers. The performance of acoustic analysis of speech by means of spectral acoustic cues obtained via empirical mode decomposition (EMD) of the log of the magnitude spectrum of the speech signal has been investigated in [4]. Experimental results have shown that the EMD-based approach results in a high correlation between HNR estimates and average perceived grade scores. In the method proposed in [2], the thresholds involved in the algorithm for IMF clustering have been fixed empirically. These thresholds are f0-dependent, so that, the method requires the estimation of the average fundamental frequency for each stimulus.

In the present study, the performance of multi-band vocal dysperiodicities analysis based on the empirical mode decomposition in terms of correlation of HNR estimates with the perceived degree of hoarseness is investigated. The empirical mode decomposition is applied to the log of the spectrum magnitude of the normalized speech signal to decompose it into its harmonic, envelope and noise components and multi-band analysis is carried out on the HNR. Compared to the method proposed in [2], the EMD-based approach used in this study incorporates an appropriate procedure that estimates automatically the thresholds used by the EMD algorithm without knowledge of the fundamental frequency.

II. METHODS

A. Speech Components Separation

A voiced speech frame $x(t)$ can be modeled as a periodic source component, $e(t)$ convolved with the impulse response of the vocal tract, $v(t)$ [5]:

$$x(t)=e(t)*v(t) \quad (1)$$

where * denotes the convolution.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Windowing the signal frame $x(t)$ and taking the Fourier transform magnitude gives

$$|X_w(f)|=|E_w(f)\times V(f)| \quad (2)$$

where $X_w(f)$, $E_w(f)$ are short-time magnitude spectra of the windowed speech frame and windowed excitation signal, respectively and $V(f)$ is the frequency response of the vocal tract.

Taking the logarithm changes the multiplicative components into additive components:

$$\log|X_w(f)|=\log|E_w(f)|+\log|V(f)| \quad (3)$$

From (3), it is observed that the log magnitude spectrum is the sum of two spectral components: $\log|E_w(f)|$, the log magnitude spectrum of the windowed excitation signal and $\log|V(f)|$, the spectral envelope due to the filtering characteristic of the vocal tract. Because of the presence of aspiration noise at the glottis, the excitation spectrum itself can be regarded as composed of two parts: the first part is a regularly spaced series of harmonics having a decreasing magnitude with frequency and the second part is an irregularly distributed noise.

The log magnitude spectrum can be considered as composed of a slowly varying (with respect to frequency) contour, noted $V_{db}(f)$, due the contribution of the vocal tract, a series of harmonics characterized by a periodic structure, noted $H_{db}(f)$, and an irregular and rapidly varying part, noted $N_{db}(f)$, due to noise at the glottis. The EMD algorithm yields a tool that enables to separate the three components of the log magnitude spectrum. Indeed, the EMD algorithm acts as a filterbank [6], so that the decomposition of the log magnitude spectrum via the EMD algorithm results into several oscillating components (IMFs) that can be clustered into three classes and each class of components is assigned to some part of the log magnitude spectrum.

In [2] the clustering of IMFs has been accomplished by a simple thresholding operation. Let f_j be the average quefrency of the jth-IMF component of the log magnitude spectrum obtained via the EMD algorithm. The different IMFs have been clustered in terms of their mean quefrencies by comparing their mean quefrencies to fixed thresholds th_1=0.3/f0 and th_2=4/f0. A drawback of this clustering procedure is that it requires the estimation of the average fundamental frequency f0 of the speech signal which is not possible for all speakers. In this presentation, we propose a procedure for IMFs clustering that does not require the estimation of the average fundamental frequency.

Let $f0_{min}$ and $f0_{max}$ be the possible minimal and maximal average fundamental frequencies, respectively. The IMFs belonging to the harmonic component are determined according to the following algorithm:

1. Find the sets of IMFs having average quefrencies within the ranges $(0.3/f0_{min}, 4/f0_{min})$ and $(0.3/f0_{max}, 4/f0_{max})$

$$\frac{0.3}{f0_{\min}} < f_j < \frac{4}{f0_{\min}}, \quad j = p_0, p_0+1, \cdots, p_1 \quad (4\text{ - a})$$

$$\frac{0.3}{f0_{\max}} < f_j < \frac{4}{f0_{\max}}, \quad j = q_0, q_0+1, \cdots, q_1 \quad (4\text{ - b})$$

where p_0 and p_1 denote, respectively, the lowest and highest IMF indices the quefrencies of which are whithin the range $(0.3/f0_{min}, 4/f0_{min})$ while q_0 and q_1 denote, respectively, the lowest and highest indices of IMFs having quefrencies whithin the range $(0.3/f0_{max}, 4/f0_{max})$.

2. Form all possible candidates of the harmonic component by varying the lowest index between p_0 and q_0 and the highest index between p_1 and q_1 and then summing the corresponding IMFs for each combination of the indices

$$H_{dB}^{pq}(f) = \sum_{j=p}^{q} IMF_j, \quad \begin{matrix} p = p_0, p_0+1, \cdots, q_0 \\ q = p_1, p_1+1, \cdots, q_1 \end{matrix} \quad (5)$$

where the superscript pq indicates the lowest and highest indices of the IMFs used to form a candidate of the harmonic component.

3. Compute the normalized autocorrelation sequence of each candidate of the harmonic component $H_{dB}(f)$ and perform an exhaustive search to find the normalized autocorrelation sequence with the most prominent peak at a nonzero delay. A large peak of the normalized autocorrelation sequence states for a high regularity of the harmonic component. The estimated harmonic component is given by

$$H_{dB}(f) = \sum_{j=p_h}^{q_h} IMF_j \quad (6)$$

where p_h and q_h denote, respectively, the lowest and highest indices of the IMFs that give rise to a normalized autocorrelation with the highest peak at a nonzero delay.

Once the lowest and highest indices of the IMFs of the harmonic component have been determined, the spectral envelope $V_{dB}(f)$ and noise $N_{dB}(f)$ are estimated as

$$V_{dB}(f) = \sum_{j=q_{h+1}}^{J} IMF_j + r_J(f) \quad (7)$$

$$N_{dB}(f) = \sum_{j=1}^{p_h-1} IMF_j \quad (8)$$

where r_J is the residue of the decomposition.

As an illustration, Figure 1 shows the estimated components of the log magnitude spectrum of a frame of 200 ms of length taken from a sustained vowel /a/ produced by a normophonic speaker.

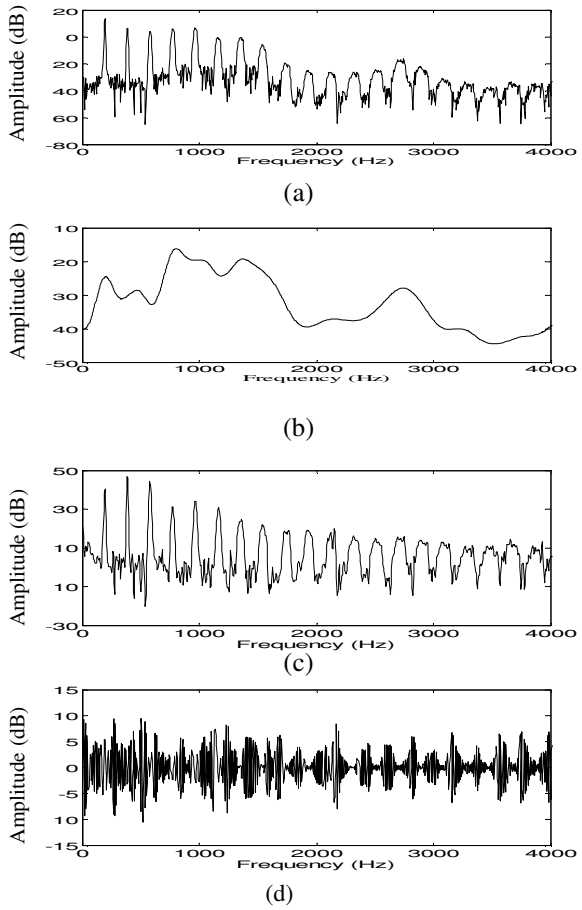


Figure 1: Decomposition of the log magnitude spectrum of a 200 ms speech frame of sustained /a/ into three components via the EMD algorithm. (a) Log magnitude spectrum. (b) Envelope component. (c) Harmonic component. (d) Noise.

B. Baseline Correction

The baseline is the inter-harmonic contour. The baseline correction is necessary because the IMFs are zero-mean oscillating functions so that when the harmonics are large, the inter-harmonics should across zero to take negative values for compensation. Indeed, in the estimated harmonic, the baseline dips slightly towards negative values at low frequencies. The estimated envelope follows the baseline closely at high frequencies and deviates slightly above the baseline at low frequencies. The goal of the baseline correction is to straighten out the baseline.

Each computed candidate of the harmonic component is subject to a baseline correction before computing the normalized autocorrelation sequence. The baseline correction follows that is used in [7] for spectral tilt correction. The correction is carried out in doubly logarithmic coordinates where the envelope of harmonic component is almost a straight line. Firstly, a straight line is fitted to the smallest 60% values of the log harmonic component and secondly, the fitted line is substracted from the harmonic component and added to the spectral envelope to obtain their respective corrected parts.

C. Multi-band Analysis

For a given utterance, the analysis interval is divided into K frames. The frequency band involved in the analysis has been limited to 4 kHz and the analysis is carried out on the harmonics-to-noise ratio (HNR). The frequency range is divided into L=10 equally-spaced intervals and the average HNRs in dB within each frequency interval are used as independent variables to summarize the amount of perceived hoarseness.

Denote by HNR_j the average HNR in the frequency interval j,

$$HNR_j = \frac{10}{K}\sum_{k=1}^{K}\log\frac{\sum_{i=0}^{M-1}H_{jk}^2(i)}{\sum_{i=0}^{M-1}N_{jk}^2(i)}, \quad j=1,\cdots,L \qquad (9)$$

with $H_{jk}(i)$ and $N_{jk}(i)$ denoting, respectively, the magnitude spectrum of the harmonic component and the magnitude spectrum of the noise component in the frequency interval j for the frame k and M is the number of frequency points.

Harmonics-to-noise ratios from the different frequency bands are used as variable predictors of scores of perceived hoarseness. Principal component analysis is performed on the spectral variables and multiple linear regression analysis is carried out to numerically express the correlation between scores of the perceived hoarseness and principal components.

D. Corpus and Perceptual Assessment

The corpus comprises concatenations of two Dutch sentences followed by vowel [a]. Dutch sentences (“Papa en Marloes staan op het station. Ze wachten op de trein.”) have been produced by 28 normophonic and 223 dysphonic speakers with different degrees of dysphonia. Five judges have evaluated the corpus involving the concatenation of the sentences and vowel [a] perceptually. The five judges are professional voice therapists with at least five years of experience in clinical voice quality ratings. Each judge has rated, from 0 to 3, the item “grade” of the (G)RABS scale. “Grade” represents the degree of hoarseness or voice abnormality. The five perceptual scores per stimulus have been averaged [8].

III. RESULTS AND DISCUSSION

The possible minimal and maximal average fundamental frequencies $f0_{min}$ and $f0_{max}$ used by the algorithm have been fixed to 80 Hz and 250 Hz, respectively. Based on our previous investigations, the frame length has been set to 200 ms. Pearson's product moment correlations of the HNRs computed in the different frequency bands with average hoarseness scores of the corpus have been computed and the variation of the correlation coefficient in terms of the frequency band number is shown in Figure 2. The highest correlations are obtained with the third (R=-0.59) and the fourth (R=-0.6) frequency bands which correspond, respectively, to the frequency ranges 800 Hz-1200 Hz and 1200 Hz-1600 Hz. These results are in a good agreement with the values of the cut-off frequency of the filter that gives rise to high correlation between the acoustic marker and the average perceived grade scores in the variogram method [1].

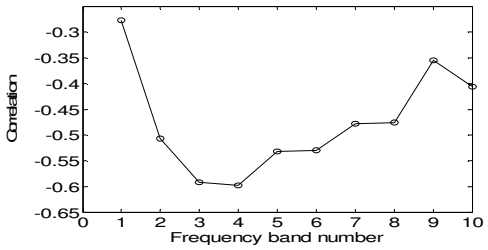


Figure 2: Variation of the correlation coefficient between perceptual scores and HNR values obtained in the different frequency bands.

Linear regression analysis has been carried out on HNRs from the different frequency bands. The multiple correlation coefficient between predicted scores and assigned perceived grade scores is R=0.75. Multiple correlation coefficient is statistically significant (R_{crit} = 0.27, p = 0.05). The first three principal components obtained via principal component analysis of the HNRs in the frequency bands have been used to predict scores of the perceived hoarseness. Figure 3 displays the average perceived grade scores versus the predicted average perceived grade scores for the corpus comprising concatenations of two Dutch sentences followed by vowel [a]. The multiple correlation coefficient between predicted scores and assigned perceived grade scores is R=0.74 indicating the high predictability of hoarseness scores by means of the first three principal components. Multi-band analysis based on empirical mode decomposition results in an improved performance in terms of correlation of predicted scores with scores of perceived hoarseness over full-band analysis the correlation of which is R=0.71. The value of the multicorrelation coefficient is comparable to that obtained in [1] via multi-band analysis of the dysperiodicity in the time domain.

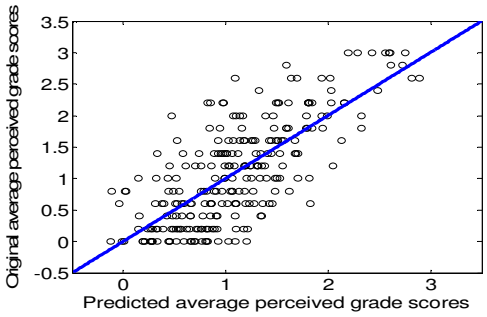


Figure 3: Original average perceived grade scores versus predicted average perceived grade scores via the first three principal components.

IV. CONCLUSION

In this paper, the EMD has been used to estimate the harmonic and noise components of the speech signal by decomposing the log magnitude spectrum the speech signal and the HNRs in different frequency bands have been used as variable predictors of scores of perceived. The proposed approach has been tested on a corpus comprising 251 normophonic and dysphonic speakers. Experimental results have shown that the proposed approach results in a strong correlation between HNR in the different frequency bands and average scores of perceived hoarseness.

REFERENCES

[1] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen, "Multi-band dysperiodicity analyses of disordered connected speech", *Speech Communication*, vol. 53, pp. 131-141, 2011.

[2] A. Kacha, F. Grenez, and J. Schoentgen, "Assessment of disordered voices using empirical mode decomposition in the log- spectral domain", in Proc. Interspeech 2012, Portland (USA), 2012.

[3] N.E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis", Proc. R. Soc. London Ser. A, Vol. 454, pp. 903-995, 1998.

[4] A. Kacha, F. Grenez, and J. Schoentgen, "Empirical mode decomposition-based spectral acoustic cues for disordered voices analysis", in Proc. Interspeech 2013, Lyon (France), 2013, pp. 3632-3636.

[5] G. de Krom, "A Cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", J. Speech and Hearing Res., Vol. 36, pp. 254-266, 1993.

[6] P. Flandrin, G. Rilling, and P. Conçalvès, "Empirical Mode Decomposition as a Filter Bank", *IEEE Signal Proc. Letters*, vol. 11, pp. 112-114, 2004.

[7] J. Schoentgen, M. Bensaid, and F. Bucella, "Multivariate statistical analysis of flat vowel spectra with a view to characterizing dysphonic voices", J. Speech Lang. Hear. Res., vol. 43, pp. 1493-1508, 2000.

[8] Y. Maryn et al., "Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels", J. Voice, 24(5): 540-555, 2010.

Speech representations based on spectral dynamics

Hynek Hermansky

Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, Maryland
hynek@jhu.edu

Abstract: **Short-term spectral envelopes of speech are bases of most conventional speech features. We argue that advantage can be gained by treating speech as sets of roughly syllable-length temporal trajectories of Hilbert envelopes of speech signal in frequency bands. Such a speech representation is consistent with temporal properties of human hearing, and allows for more effective dealing with linear distortions, coarticulation, and frequency-localized noise.**

Keywords : **Temporal representations of speech, robustness of speech processing, consistency with human hearing**

I. INTRODUCTION

Besides carrying a message, speech also carries information about speaker. Experienced listeners can often identify possible health problems just by listening to speech. However, there is no clear agreement how is information coded in the signal, and how it can be accessed by a machine. Since the introduction of Spectrograph™ in the mid of the last century, instantaneous power spectral envelopes are broadly accepted as the main carriers of information in speech. As evidenced by successful spectrum-based speech coding techniques, spectral envelopes do carry information that is sufficient for reconstruction of reasonable-quality speech.

Spectral envelopes can be easily corrupted by many factors that are reasonably well handled by human listeners. Large differences in formant frequencies in phonetically identical sounds produced by different speakers exist., children speech presenting a particular challenge. The ease with which the spectral envelope can be corrupted by relatively benign modifications like linear filtering of the signal is alarming. Thus, e.g., filtering the whole speech utterance by a fixed filter that flattens a particular vowel in the utterance does not prevent a listener to correctly identify the original phonetic value of this vowel in spite of its entirely flat spectrum (Hermansky et al 2013). Other disturbing problems with speech spectra that are known for a long time, persist. Among those, inertia of vocal organs that produces coarticulation among neighboring speech sounds (phones), causes each short-term spectrum to be dependent not only on the current phone but also on the phones that surround it. Well documented are, e.g., substantial differences in short-term spectra of consonants /k/ and /h/ followed by different vowels (Potter et al 1946). In spite of that, human hearing readily and easily accepts these spectrally different sounds as phonetically identical.

We believe that some of these problems might be alleviated by greater emphasis on information carried in frequency-localized spectral dynamics of speech.

II. ALTERNATIVE REPRESENATIONS OF SPEECH

There is no doubt that vertebrate hearing is frequency-selective. The question is why is that so? Is it only to derive spectral envelopes of the acoustic signal? Why would hearing strive only for such unreliable information carriers? Would it be possible that some other elements of the speech signal are also used for decoding the information in speech?

Most natural signals such as speech change over time and the information is carried in these changes. The signal changes are reflected in the dynamics of spectral components. That was eloquently stated by Dudley, who points out that inaudible message in slow motions of vocal tract is made audible by modulating the audible carrier [1].

A. RASTA

Our interest in spectral dynamics started with an engineering problem of dealing with changing communication channel in machine recognition of speech. To deal with the problem we proposed an *ad hoc* but effective RASTA filtering that only passed modulation spectrum components between 1 and 15 Hz to alleviate negative effects of such fixed linear distortions [2]. It is known and many times confirmed by others that human hearing is most sensitive to relatively slow modulations [3][4]. It is then no surprise that the modulation spectrum of speech has most of its energy in the area where hearing is the most sensitive, typically peaking at around 4 Hz, reflecting the syllabic rate of speech [5].

When attempting to derive RASTA filters using Linear Discriminant Analysis on phoneme-labeled speech data, finite impulse response filters with very similar frequency characteristics and rather long (more than 200 ms)

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

impulse responses emerged [6][7], further supporting the need for considering syllable-level spectral dynamics in deriving phonetic values of speech sounds [8][9]. Such a time interval comes as no surprise to any physiologist or psychophysicist, and it is surprising that it for such a long time escaped the attention of most speech engineers. It is found in many psychophysical phenomena and on higher levels of neural processing (see, e.g. [9] for review).

B. TempoRAl Patterns (TRAPS)

A tentative proposal is that the *main reason for frequency-selective hearing of vertebrates is not evaluating the overall shape of the sound spectrum, but that it rather evaluates temporal profiles of signals in individual sub-bands* (Hermansky 1998). The first test of this concept came in the form of a so-called TRAP [10], where 1001 ms long temporal trajectories of spectral power in the individual critical-band sub-bands with their means removed were used to estimate posterior probabilities of phoneme categories in the center of this long temporal pattern at each frequency using nonlinear multilayer perceptron artificial neural net (MLP) based classifier. (Fig. 1).

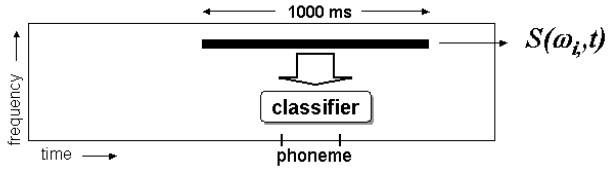


Fig. 1 TRAP approach, where long temporal trajectories of spectral energies in frequency bands are used to estimate phonetic value of speech sound at the centers of the temporal patterns.

Vectors of posterior probabilities from the individual sub-bands are then merged using another MLP classifier (Fig. 2) . In TRAP approach, spectral correlations among the sub-bands are not used. The power in the individual bands merely defines the local signal-to-noise ratio (SNR). The information that TRAP uses is in the local temporal dynamics.

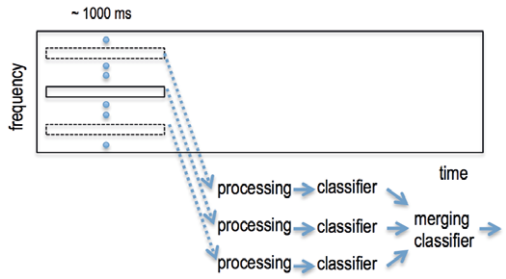


Fig. 2 Estimation of final posterior probabilities of speech sounds in TRAP approach.

Processing of TRAPs prior to the classification can be as simple as computing truncated cosine transform of TRAPS or can be more complex, involving, e.g., emulations of auditory cortical receptive fields in the MRASTA approach [11], where the temporal processing consists of first and second derivatives of Gaussian function of eight different widths and the spectral processing consists of differentiation of three neighboring frequency channels, yielding 448 dimensional vector describing each TRAP at individual carrier frequencies (Fig. 3).

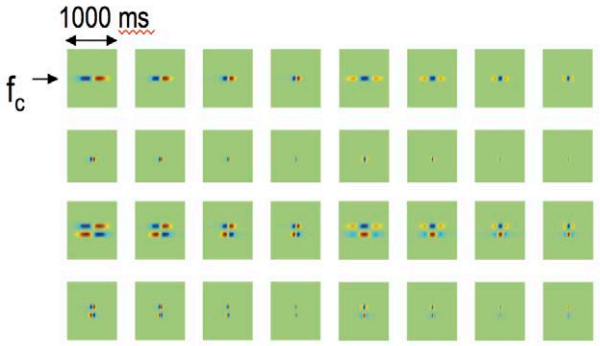


Fig. 3 MRASTA 2-D filters for processing of TRAP at the carrier frequency f_c.

Hilbert envelopes of speech signal in individual frequency bands are filtered by sets of 16 FIR filters illustrated in the upper left corner of the Figure. Local spectral slopes of time-frequency patterns in the respective spectral bands are derived by differentiating over three neigboring bands. This effectively results in a set of 32 2-D time-frequency filters that are applied at each frequency band, that are broadly consistent with known properties of auditory cortical receptive fields.

Actual auditory cortical receptive fields obtained from measurements on auditory cortex of ferrets trained to recognize human speech were applied in a similar manner in [12].

C. Posteriogram

Estimated posterior probabilities of phonemes yield an interesting 2-D representation of speech, which we call posteriogram, illustrated together with the mel spectrum based spectrogram of the same utterance in Fig. 4.

As seen in the figure, phoneme posteriors estimated from TRAPs could handle coarticulation reasonably well since each TRAP spans most of the coarticulation span and can therefore utilize all information about the underlying phoneme. This is consistent with earlier proposals [13] where syllable-length temporal patterns are suggested as information sources for estimating phonetic qualities of underlying speech sounds.

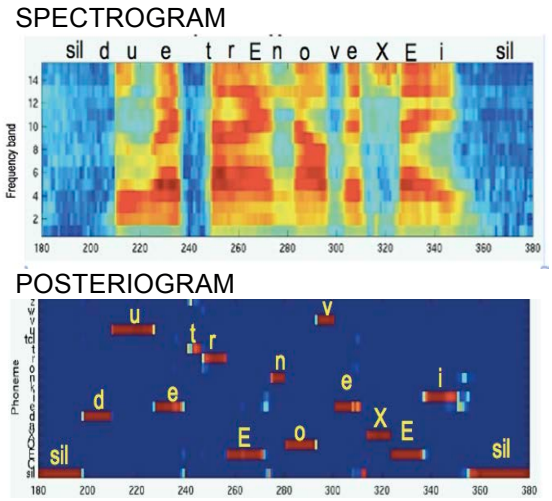


Fig. 4 Auditory-like spectrogram and Posteriogram of a sequence of Italian digits "due-tre-nove-sei". High values are indicated by warm colors.

D. TANDEM

Vectors of phoneme posteriors can be used in automatic speech recognition (ASR) either directly in Viterbi search for the best matching unknown utterance or can be converted to speech representation that can be readily used in current state-of-the-art ASR by so-called TANDEM approach [14] that applies a series of processing steps to estimates of posteriors of speech sounds from the ANN classifier, making them more suitable for the currently dominant HMM/GMM ASR technology. The speech signal is first converted to an auditory-like time-frequency representation. Sufficiently long (typically longer than 200 ms) segments of temporal trajectories of spectral energies in the frequency sub-bands form, after some pre-processing, an input to an estimator of posterior probabilities of speech sounds that has been trained on large amounts of labeled speech data. The final features for an HMM/GMM-based state-of-the-art ASR system are derived from these posteriors by some post-processing that ensures that the features have approximately a Normal distribution and are decorrelated. Such features are currently dominating state-of-the-art experimental systems.

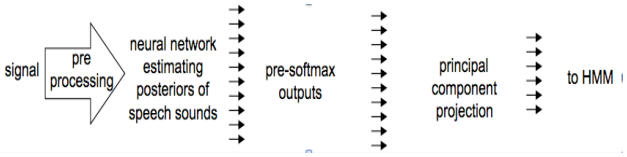


Fig. 4 TANDEM processing of posterior probabilities of phonemes for application in GMM based ASR

E. Frequency Domain Linear Prediction (FDLP)

Since in modulation spectrum-based applications we are primarily interested in temporal trajectories, it is tempting to abandon the short-term analysis altogether. This is possible by using the so-called Frequency Domain Linear Prediction (FDLP) [15][16], where an autoregressive model is not computed from the signal as in the case of the conventional time domain linear prediction (upper part of the Fig. 5) but it computed from a cosine transform of the signal (the lower part of the Fig. 5). To find the autoregressive model of the signal in a restricted frequency range, one can put an appropriate limited-span window on the cosine transformed speech signal. The window span and shape determines the frequency response of the implied frequency filter. Thus, by properly windowing the cosine transform of the signal, one can directly compute autoregressive models of the Hilbert envelopes in the sub-bands over long segments of the speech signal, entirely bypassing any short-term analysis windows. The FDLP model has been shown in dealing with convolutive linear distortions and reverberation is easier [17].

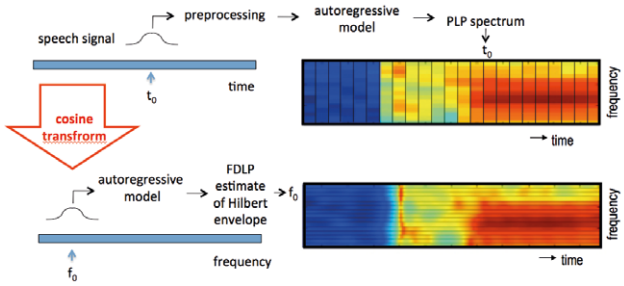


Fig. 5 Conventional linear predictive analysis (the upper part) and FDLP (the lower part).

F. Multistream Processing

Individual frequency channels can be seen as independent processing streams, each attempting to derive information from a particular carrier frequency. That might in principle allow for alleviating streams that are corrupted by frequency-localized corruptions. This ability is quite likely one of reasons for robustness of auditory processing in nature and its emulation would take us a long way in alleviation of fragility of speech technology in noise.

However, such an emulation requires a mechanism for deciding which processing channels are corrupted at any given time. One such technique employs 127 streams, formed by all possible non-empty stream combinations of the 7 band-limited streams with MLP based fusion modules for each of 127 combinations trained on available training data. Several techniques that evaluated performances of the individual streams were investigated

to find the best stream combination [18][19][20][21[22][23]. This problem is still a topic of our current research efforts, and more discussions and details can be found in [24].

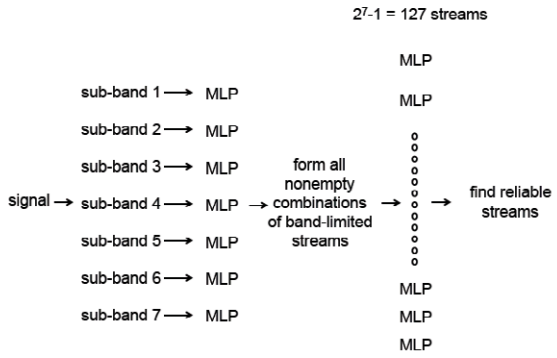


Fig. 4. Multistream classifier for dealing with frequency localized noise

III. CONCLUSIONS

Recent research unambiguously points to the importance of spectral dynamics in coding the information in speech. Working with spectral dynamics is consistent with human processing of speech, allows for alleviation of linear distortions, better handling of coarticulation, and for dealing with noise. More discussion on this topic can be found in [25][26].

REFERENCES

[1] H. Dudley, 1940 "Carrier Nature od Speech," *Bell System Technical Journal,* Vol XIX, No. 4, October 1940

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing* 2(4): 578-589, 1994

[3] R. Riesz, "Differential intensity sensitivity of the ear for pure tones," *Phys. Rev.* 31(5): 867–87, 1928

[4] N. Kanedera, N., T. Arai, H. Hermansky, H., M. Pavel, "On the relative importance of various components of the modulation spectrum of speech," *Speech Commun.* 28(1): 43–55, 1999

[5] T. Houtgast and H.J.M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility, " *Acustica* 28: 66–73, 1973

[6] S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters*," Proc. Eurospeech'97*, Rhodes, Greece, 1997

[7] Valente, F., Hermansky, H., "Data-driven extraction of spectral dynamics based posteriors," in *Handbook of Natural Language Processing and Machine Translation*, J. Olive, C. Christianson, and J. McCary, Eds. Springer-Verlag, 2011.

[8] H. Hermansky, "Speech beyond 10 ms (temporal filtering in feature domain)," *Proc. International Workshop on Human Interface Technology*, Aizu,1994

[9] H. Hermansky, "Should recognizers have ears?," *Speech Communication* 25(1–3): 3–27, 1998

[10] H. Hermansky H and S. Sharma S., "TRAPS: Classifiers of TempoRal PatternS," *Proc. ICSLP'98*, Sydney, 1998

[11] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," *Proc. Interspeech 2005*, Lisbon, 361–364

[12] S. Thomas, K. Patil, S. Ganapathy, N. Mesgarani, H. Hermansky, "A phoneme recognition framework based on auditory spectro-temporal receptive fields," *Proc. Interspeech 2010,* Tokyo, 2458-2461

[13] V. A. Kozhevnikov and L. A. Chistovich, *Speech: Articulation and perception,* Washington, D.C.: Joint Publications Research Service), 250–251

[14] H. Hermansky, D.P.W. Ellis and S. Sharma, "Connectionist feature extraction for conventional HMM systems", *Proc. ICASSP'00*, Istanbul, Turkey, 2000

[15] M. Athineos, H. Hermansky, D.P.W. Ellis, "LP-TRAPS: Linear predictive temporal patterns," *Proc. Interspeech 2004*, Jeju Island, Korea

[16] M. Athineos and D.P.W. Ellis, "Autoregressive modelling of temporal envelopes,", *IEEE Trans. Signal Processing* 55(11): 5237–5245, 2006

[17] S. Ganapathy, S. Thomas and H. Hermansky, "Modulation frequency features for phoneme recognition in noisy speech," *J. Acoust. Soc. Am.* 125(1), 2009

[18] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. ICSLP 1997*

[19] S. Sharma, *Multi-stream approach to robust speech recognition,* OGI Ph.D. dissertation, Portland, OR, 1999.

[20] N. Mesgarani, S. Thomas, and H. Hermansky, "A multistream multiresolution framework for phoneme recognition," in *Proc. Interspeech 2010*, pp. 318–321.

[21] N. Mesgarani, S. Thomas, and H. Hermansky, "Towards optimizing stream fusion," *J. Acoust. Soc. Am.,* vol. 139, no. 1, pp. 14–18, 2011.

[22] T. Ogawa, F. Li, and H. Hermansky, "Stream selection and integration in multi-stream ASR using GMM-based performance monitoring," *Proc. Interspeech 2013*

[23] E. Variani, F. Li, and H. Hermansky, "Multi-stream recogntion of noisy speech with performance monitoring*," Proc. Interspeech 2013*.

[24] H. Hermansky, "Multistream Recogntion of Speech: Dealing with Unknown Unknowns," Invited Paper, *Proc. IEEE,* Vol 101, No 5, May 2013, pp. 1076-1088

[24] Hermansky, H, Cohen, R. J., and Stern, R.M, "Perceptual Properties of Current Speech Recognition Technology," Invited Paper, *Proc. IEEE*, Vol. 101, No 9, 2013

[26] H. Hermansky, "Speech recognition from spectral dynamics," Invited Paper*, Sādhanā,* Vol. 36, no. 5, pp. 729–744, 2011.[1]

ACKNOWLEDGEMENT: Supported by the Intelligence Advanced Research Projects Activity (IARPA) and by Defense Advanced Research Projects Agency (DARPA) via Department of Defense US Army Research Laboratory contract numbers W911NF-12-C-0013 and D10PC20015, and by the JHU Human Language Technology Center of Excellence. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DARPA, DoD/ARL, or the U.S.

MODE-LOCKING OF GLOTTAL JET INSTABILITIES WITH MUCOSA WAVES ON FALSE VOCAL FOLDS

C. Brücker[1], C. Kirmse[2]

[1] Dept. of Fluid Mechanics, TU Bergakademie Freiberg, Freiberg, Germany, bruecker@imfd.tu-freiberg.de
[2] Dept. of Fluid Mechanics, TU Bergakademie Freiberg, Freiberg, Germany, Clemens.kirmse@imfd.tu-freiberg.de

Abstract: **Experimental studies are presented on the control of the glottal jet via varicose waves excited on the surface of the false vocal folds. The results show a strong feedback of jet flow stability with the imposed oscillations of the false vocal folds. As conclusion, shear-layer roll-up is seemingly in lock-in mode with the forced oscillations of the jet for certain excitation frequencies. Therefore, jet stabilization is achieved.**
Keywords : **Glottal jet, mode-locking**

I. INTRODUCTION

The glottal jet flow is highly sensitive to changes of the supraglottal geometry. Earlier studies have shown the reduction of pressure loss by the presence of the false vocal folds. In addition, a possible leakage flow does not change the contribution of harmonic range of vortex-induced source spectrum when the false vocal folds exist. From fluid mechanics of jets it is known that the near-exit area imposes strong feedback on the instabilities of the jet itself and the generation of entrainment and vortex structures in the shear layer. In the present work we investigated the complex fluid structure interaction of the jet flow with the mucosa layer on the false vocal folds. Special focus was laid on the excitation of such waves and possible feedback on the jet shear layer development. A feedback loop is built up and glottal jet instabilities are recorded with High-Speed Particle-Image-Velocimetry (TR-PIV) and flow visualization.

II. METHODS

Experimental methods: flow studies were carried out in a 3:1 enlarged model of the vocal tract with water as the carrier fluid, see Triep & Brücker [1]. All walls are transparent and accessible for light-sheet visualization and Particle-Image-Velocimetry (PIV). Flow is driven through the vocal tract by a constant head imposed on a vertical tank upstream of the vocal tract. A rotating cam model allows the modulation of the glottal gap in a form similar to the convergent/divergent cycle of gottal contour in the motion cycle.

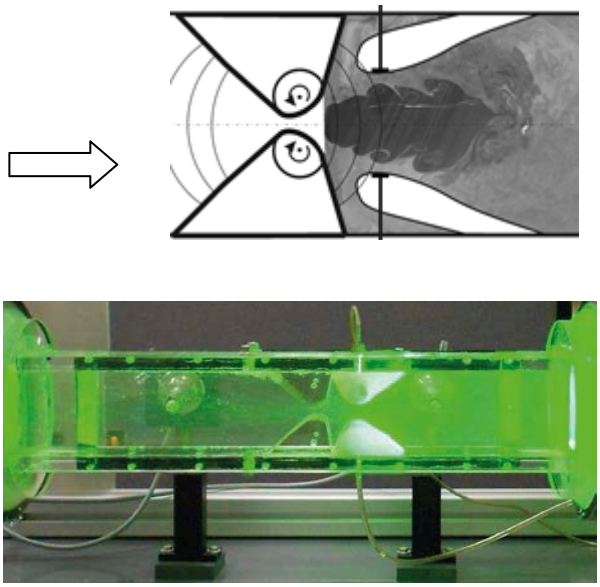


Fig. 1: Experimental setup with excitation of oscillations on the false focal folds.

Experimental procedure: Downstream of the cam model, false vocal folds are placed symmetric on both sides of the channels, see Fig. 1. They are casted from silicone with a hollow air cushion inside. Pressure tubes are connected with the air cushion chambers and are driven with a solenoid valve connected to a pressure line. Thereby, small oscillations of the gap width between the counter-facing folds can be driven in a harmonic manner up to a frequency 20 Hz. This leads do a varicose-type modulation of the gap width as a second constriction of the flow similar as mucosa waves would do when they are excited from the glottal jet flow. Our major interest is a possible feedback of waves and jet instabilities. Therefore we used forced oscillations applied to the false vocal folds with varying frequency and recorded the response of the jet stability and roll-up of the shear layer vortices. The base flow was started by opening the cam model until maximum opening of the glottal gap. Then the cams were stopped in this position. The characteristic

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

time scale of the modulation in the up-scaled model is of order of 4 sec. Oscillations of the false vocal folds were started at the same time as the cams open the gap. Recordings were taken in the major axial plane using light-sheet visualizations and high-speed imaging.

III. Results

A typical flow evolution of the jet is a first straight path of the jet and a later attachment of the jet to one side of the second constriction formed by the false folds. The time required for the jet to attach to one wall was estimated from the flow recordings. Depending on the excitation frequency of the false folds, this time-span after start of the flow can be largely increased up to 5 times the value without any oscillations. So, jet stabilization is achieved by oscillations.

Fig. 2 and Fig. 3 show the different states of the flow after opening of the cam model. Without any oscillations of the false folds, the flow attaches immediately to one side of the false folds and is deflected laterally within the supraglottal space, compare [2]. In contrast, when small oscillations are excited on the surface of the false folds (amplitude 1/20 of maximum glottal gap width), jet flow can be kept in straight path through the center of the second construction for a certain time-span until the jet finally attaches to one side.

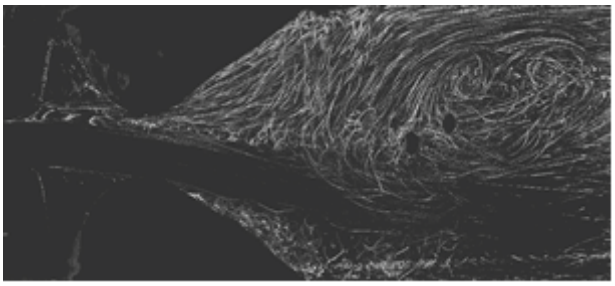


Fig. 2: Attachment of the glottal jet to one side of the false vocal folds after start of the flow. No oscillations on the surface of the false folds.

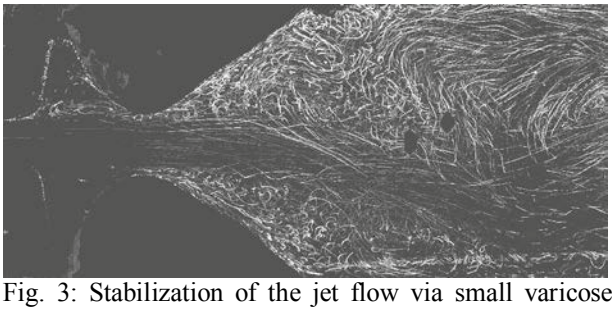


Fig. 3: Stabilization of the jet flow via small varicose oscillations of the false vocal folds at a frequency of 6 Hz (3-times of the fundamental frequency)

It is interesting to note that a stabilization of the jet is possible by these small oscillations such that the jet remains straight and does not attach to the walls. A variation of the frequencies shows that this effect is achieved for a larger range of excitation frequencies. Therefore, attachment of the jet can be delayed to times larger than the characteristic cycle of the glottal gap modulation. Fig. 4 displays the achieved delay times over the excitation frequency. The red bars indicate the rms-value of the values taken from 10 independent experiments. Maximum delay is achieved for a frequency of 3 times the characteristic oscillation frequency of the glottal gap. A small rms value at a frequency of 6Hz indicates a strong feedback with repeatable conditions.

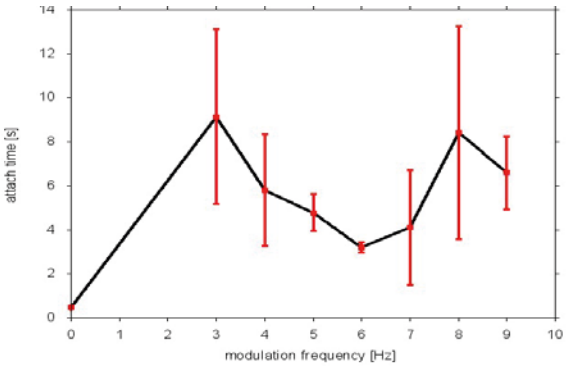


Fig 4: Delay of one-sided jet attachment by oscillations of the fals vocal folds in the enlarged model.

IV. Discussion & Conlusion

The results show a strong feedback of jet flow stability with the imposed oscillations of the false vocal folds. As conclusion, shear-layer roll-up is seemingly in lock-in mode with the forced oscillations of the jet for certain excitation frequencies. This result is important with respect to possible influence of mucosa waves along the false vocal folds for singing voice. As a consequence, higher harmonic parts in the sound spectrum due to vortex motions can be modified and tuned to characteristic frequency bands while random noise due to non-coherent turbulence is diminished.

References

[1] M. Triep, Ch. Brücker, Three-Dimensional Nature of the Glottal Jet. JASA **127**(3), 2010.

[2] W. Mattheus, Ch. Brücker, Mattheus, W., Brücker Ch. Asymmetric glottal jet deflection: Differences of two- and three-dimensional models. JASA **130**(6), pp EL373-EL379.

DIFFERENT TYPES OF PAUSES AS A SOURCE OF INFORMATION FOR BIOMETRY

M. Igras, B. Ziółko

Department of Electronics, AGH University of Science and Technology
{migras,bziolko}@agh.edu.pl www.dsp.agh.edu.pl

Abstract: **Statistics of pauses appearing in Polish as a potential source for biometry information for automatic speaker recognition were described. The frequency of three main types of pauses (silent, filled and breath pauses) usage in monologues, as well as frequency of punctuation (commas and full stops) in their transcriptions were investigated quantitatively. Correlation between temporal structure of speech and syntax structure of the spoken language were examined statistically to verify usefulness of pauses detection for elaborating algorithms of automatic detection of punctuation for spoken Polish.**
Keywords : **pauses, fillers, punctuation, Polish**

I. INTRODUCTION

A set of common disfluencies interferes with sentences borders in spontaneous speech. The most important are: restarts, change of syntax during the utterance and inclusion of intervening sentences. Within words, the most frequent are repetitions, repairs and prolongations of conjunctives, prepositions and final syllables. As far as human perception can focus on the meaning of the utterance and extract the desired information, the automatic speech recognition system literally recognizes whole acoustic content of the speech signal [1]. As a result, the transcription is redundant with notation of disfluencies or slips of the tongue, but diminished of the other types of information present in signal, like punctuation. This information could be also used to differentiate speakers.

The research show three types of acoustic pauses in spoken language. The most intuitive is silence (s_p). Depending on the speaker and situational context, it may be characterized by different length.

Another type are filled pauses (f_p) - pseudo-words, that do not affect sentence meaning, like *yyy, eee, hmm, mmm, ym* (in SAMPA notation: *III*, *eee*, *xmm, mmm*, *Im*,), that perturb utterance fluency. They may often indicate need of insertion of comma or full stop in the adequate position in transcription. The sound of filled pauses are specific for language (in Polish the most common are *yyy/yh* and *mmm*, while for English - *um*) and specific for speaker's habits. The third sort of pauses that we consider are breath pauses (b_p) which strongly indicate insertion of the full stop in transcription.

Breath events [2] and filled pauses [3] can be automatically detected in a speech signal. It allows to apply this methods as a part of biometry systems in speaker recognition task.

Considering the origin of pause usage we marked out 1) regular natural pauses caused by respiration activity (breath pauses), 2) irregular intentional pauses, purposely used as a stylistic form, especially by professional speakers (silent pauses) and 3) irregular, unintentional disfluencies, effect of uncertainty, hesitation or short reflection, in speech of inexperienced speakers even 10-20 per minute (acoustic events like silent pauses or filled pauses).

Information on pauses is used in majority of algorithms of automatic punctuation detection [4], [5]. Some medical aspects of different types of pauses were investigated in context of affective state [6] and mental condition [7] of the speaker.

The obtained knowledge on pauses meaning can be merged with analysis of other temporal features (phoneme length, energy, fundamental frequency [8]) in order to build algorithms for punctuation detection in speech.

II. METHODS

The prepared corpus of spontaneous Polish speech consisted of different types of monologues in formal or half-formal situations. Total duration of recordings is 60 min, including utterances of 24 speakers (13 male, 11 female). Among them, there are both experienced or professional speakers (politicians, professors, professional translators) and inexperienced speakers (students).

The first group of recordings (30 min) are utterances from orations or public presentations: speeches and reports from European Parliament [9], sessions of faculty council, students lectures and reviews. All the speeches, although preceded by preparation of the speakers or supported by slides, were not read and are characterized by all the features typical for spontaneous speech. The second part of the corpus (30 min) consisted of recordings of real time translation of orations during European Parliament sessions [9]. The sort of utterances are specific kind of spontaneous speech, where the speech

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

rate of the translator is determined by the style of the speaker being translated. However, still they are situations of formularization of own utterance, which causes their spontaneous character and induces presence of imperfections characteristic for spontaneous speech.

For comparison with read speech, recordings from audiobooks and AGH Audio-Visual Speech Database [10] were used.

First we transcribed orthographically the content of the recordings to clean (skipping disfluencies, filled pauses or repairs) and syntactically correct texts. On the basis of the observation of the process, the factors affecting the imprecision and ambiguity of inserting punctuation in the transcripts were collected. One of the impediments was ambiguous intonation, especially in case of inexperienced speakers. It manifested by 'enumerating' tone of voice, which caused the same tone in commas and full stops or constructing multiple complex sentences with every clause starting with conjunctive pronounced with extended phonation. In such cases the decision of inserting comma or full stop remained subjective. When a speaker did not signalize the phrases and sentences border with their pronunciation, intonation or pauses, the punctuation was based on the meaning of the utterance. There also often occurred the bonding the last word of preceding sentence with the first in the next one. In translators group we usually observed specific disorder of phonotactics involving artificial prolongations of whole words. Transposals of functional elements of sentences and reorganization of the sentence were also frequent events. It is common for inexperienced speakers to place intervening sentences during the speech or abusing certain words like *let's say, just, simply* (language-specific conversational fillers/discourse makers).

For each transcription, the number of words, full stops and commas were counted. Then the statistics of sentences and phrases lengths were computed: mean length of a sentence and a phrase, as well as mean number of words in sentences and phrases. Then, in the places of punctuations signs, occurrences of pauses were verified. When a full stop were signalized by silent pauses, the time was tagged as s_p. (similarly for commas - s_p,), filled pauses - f_p. (commas - f_p.), b_p. for breath pauses (b_p, for commas). When no type of pause appeared, the place was tagged as n_p. (n_p,).

III. Results

Information of frequency of using punctuation signs in spoken language as phenomena determining speech rhythm were obtained by analyzing the quantity of full stops and commas in transcriptions. Fig. 1 shows meaning of the pauses in determining punctuation in speech. Fig. 2 presents the most frequent types of filled pauses. However, the usage of different types of pauses for signalization of punctuation is strongly individualized between speakers, as presented in Table 1.

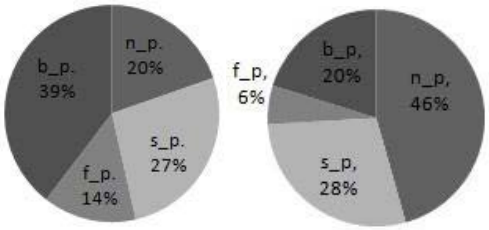


Fig. 1. Different types of pauses determining full stops and commas, and types of filled pauses signalizing punctuation

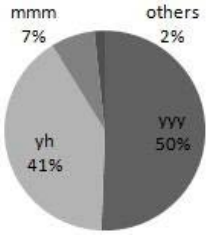


Fig. 2. Different types of filled pauses appearing in spontaneous speech

Table 1. Counts of pauses events denoting full stops and commas (P- results for presentations/orations, T- translation, C- entire corpus, * - lack of audible breaths in recording)

Fullstops

Rec.	#.	n_p.	s_p.	f_p.	b_p.
P1_5	17	2	11	4	*
P2_5	26	1	6	19	*
P3_3	15	1	14	0	*
P4_1	11	2	2	0	7
P5_2	10	4	0	2	4
P6_1	8	0	1	1	6
P7_1	4	0	1	1	2
P8_1	7	0	1	2	4
P9_1	6	0	0	4	2
P10_1	4	1	1	1	1
P11_5	29	1	13	0	15
P12_4	40	13	10	1	16
P	177	25	60	30	57
T1_1	7	4	1	0	2
T2_1	8	2	2	2	2
T3_1	8	3	4	0	1
T4_1	11	0	5	0	6
T5_1	8	4	1	0	3
T6_5	32	6	4	5	17
T7_5	18	12	1	1	4
T8_5	24	0	6	1	17
T9_2	9	0	3	1	5
T10_4	27	8	5	4	10
T11_1	5	2	0	0	3
T12_2	17	3	2	0	12
T	174	44	34	14	82
C	351	69	94	49	139

Commas

Rec.	#,	n_p,	s_p,	f_p,	b_p,
P1_5	71	27	35	13	*
P2_5	80	32	36	12	*
P3_3	24	9	15	0	*
P4_1	12	8	2	0	2
P5_2	22	14	2	0	6
P6_1	10	7	2	0	1
P7_1	14	6	3	1	4
P8_1	6	4	1	0	1
P9_1	11	7	2	2	0
P10_1	21	10	9	1	1
P11_5	51	6	11	0	34
P12_4	99	55	24	1	18
P	421	185	142	30	67
T1_1	7	5	2	0	0
T2_1	7	4	0	1	2
T3_1	4	3	1	0	0
T4_1	8	3	1	0	4
T5_1	8	5	0	1	2
T6_5	57	29	13	1	14
T7_5	41	20	11	1	9
T8_5	25	7	7	1	10
T9_2	13	5	1	0	7
T10_4	49	30	9	1	9
T11_1	13	6	2	0	5
T12_2	19	6	3	3	7
T	251	123	50	9	69
C	672	308	192	39	136

As we estimated, speech rate in spontaneous monologues is about 115 words per minute (with standard deviation between speakers is about 20 words/min). Mean length of sentence (containing average 19 words) was about 10 seconds, while mean length of a speech unit divided by punctuation (average 7 words) - 3.8 s. The results were similar for both orations/ presentations and real time translations.

Among all full stops in transcription, 39% are correlated with occurrences of breath pause, 27% silent pause, 20% filled pause. Among all commas, 28% are pointed by silent pause, 20% breath pause and 6% filled pause. Lack of any kind of pause (words bonding in pronunciation) was registered in 20% occurrences of full stop and 46% commas for spontaneous speech, and only for 1,3% full stops and 42% commas for read speech. Among all occurrences of filled pauses, 8% indicate full stops and 6% indicates commas, among breath pauses the proportions are, respectively, 10 and 11%.

The most commonly used types of filled pauses are: prolonged 'yyy' (a half of the cases), short 'yh' (41%) and 'mmm' (7% of counts). As for acoustically registered breath pauses, average for a speaker was about 11 breaths per minute. In normal physiological condition, at rest, the value of breath per minute is 12-20.

To investigate the influence of experience and oratorical abilities on pauses and speech rate, we divided corpus of spontaneous monologues into recordings of experienced speakers (professors and politicians) and inexperienced speakers (mainly students). Average values of selected temporal features of each group are compared in Table 2.

Table 2. Comparison of selected features for experienced and inexperienced speakers: average values and standard deviation (in brackets)

	Professional speakers	Inexperienced speakers
#words/minute	108 (23)	117 (26)
#words/sentence	17(4)	22(6)
#f_p/minute	4(3)	10(5)
n_p. [%]	12(15)	13(13)
s_p. [%]	26(31)	24(23)
f_p. [%]	10(12)	34(30)
b_p. [%]	50(17)	27(8)

As expected intuitively, professionals speak slowlier, with less disfluencies and formulate shorter sentences, which makes their speech more adjusted for efficient listening and understanding by recipients. Also their dynamic breathing rhytms are much more concordant with sentences boundaries (a half of fullstops were correlated with breath pauses). Such conscious dynamic breathing (taking a breath before beginning of a sentence or phrase) is one of the basic voice emission principles, often emphasized by authors of handbooks on speaking skills and techniques [11],[12].

IV. Discussion

While the full stops can be easily recognized by pauses detection, the commas does not seem to be possible to detect on the basis on pauses alone, without taking into account another parameters.

Both lack of punctuation and occurrence of disfluencies in spontaneous speech transcripts are factors that disturb their processing by natural language processing systems, parsers or information extraction systems, mainly because usually language models do not contain disfluencies and operate on full sentences [13]. Research on punctuation in spoken language can improve ASR systems, increase readability and usefulness of automatic transcripts for human, and adapt them to be processed by language models. Moreover, modeling of pauses in spoken language can be applied to more natural-sounding speech synthesis systems [14].

V. Conclusion

Beyond applications of the research on pauses for speech technology systems, it can be used also directly in the biomedical field.

Connotations between pauses and punctuation, as well as frequency and types of pauses vary between individuals and depend on speaking style of each person, speech quality, culture, experience and preparation for oral presentations. Thereby, the temporal features can be used for speaker biometry or evaluation of speaker oratorical skills.

Further research will cover also other reasons of pauses frequeancy and duration variability. One of them is a type of personality of the speaker or even mental illnesses - quantity and duration of silent pauses can be indicators of emotional state of the speaker or a measurable symptom of psychic disorders like schizophrenia or bipolar affective disorders. Frequency of filled pauses and breath pauses during monologues will be invesitgated as a significant marker of speaker stress and emotional arousal.

ACKNOWLEDGMENTS

The project was supported by The National Research and Development Centre granted by decision 072/R/ID1/2013/03.

REFERENCES

[1] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok and M. Mąsior: *Automatic speech recognition system dedicated for Polish.* Proceedings of Interspeech, Florence (2011)

[2] M. Igras and B. Ziółko: *Wavelet method for breath detection in audio signals*, IEEE International Conference on Multimedia and Expo (ICME 2013) San Jose, California, USA July 15-19, 2013

[3] K. Barczewska and M. Igras: *Detection of disfluencies in speech signal*, Young scientists towards the challenges of modern technology : 7th international PhD students and young scientists conference : Warsaw, 17–20 September 2012, pp. 36

[4] D. Baron, E. Shriberg and A. Stolcke: *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues.* (2002) 949-952

[5] E. Shriberg, A. Stolcke, D. Hakkani-Tur and G. Tur: *Prosody-based automatic segmentation of speech into sentences and topics* (2000)

[6] I. Homma and Y. Masaoka, *Breathing rhythms and emotions*., Experimental physiology, vol. 93, no. 9, pp. 1011–1021, Sept. 2008.

[7] V. Rapcana, S. Darcy, S. Yeap, N. Afzal, J. Thakore, and R.B. Reilly: *Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia*. Medical Engineering & Physics 32 (2010) 1074-1079

[8] M. Igras and B. Ziółko: *The influence of phoneme duration, energy and frequency features on the prominence of accent and sentence boundaries in spoken Polish*, Approaches to Phonology and Phonetics: APAP Lublin, 21-23.06.2013, Book of abstracts, pp. 28

[9] J. Loof, C. Gollan and H. Ney: *Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system.* Proceedings of Interspeech, Brighton (2009) pp. 88-91

[10] M. Igras, B. Ziółko and T. Jadczyk: *Audiovisual database of polish speech recordings*. Studia Informatica 33 2B (2012) 163-172

[11] M. Kotlarczyk, *Sztuka żywego sowa (eng. The art of living word)*, Gaudium Lublin, 2010.

[12] Z. Pawłowski et al., *Emisja głosu - struktura, funkcja, diagnostyka, pedagogizacja (eng. Emission of voice - structure, function, diagnostics, pedagogization)*, Wydawnictwo Salezjańskie Warszawa, 2008.

[13] E. Shriberg: *Spontaneous speech: How people really talk and why engineers should care*. In: in Proc. European Conf. on Speech Communication and Technology (Eurospeech. (2005) 1781-1784

[14] B. Zellner: *Pauses and the temporal structure of speech.* Fundamentals of speech synthesis and speech recognition (1994) 41-62

ACOUSTIC MODEL OF TRACHEAL STOMA NOISE PRODUCTION FOR SPEECH ENHANCEMENT IN POST-LARYNGECTOMIZED PATIENTS

R. Pietruch[1]

[1] Industrial Research Institute for Automation and Measurements PIAP, Warsaw, Poland, rpietruch@piap.pl

Abstract: **This paper focuses on the methods for speech enhancement of people who underwent total laryngectomy. The patients can no longer use the natural sound source and have to breathe through a tracheal stoma. In previous work, it has been shown that silently articulated speech of many subjects is highly distorted by noise. The spectra of isolated vowels had regular characteristics, regardless of the sound being articulated. In the present study, the effect of breathing sounds on speech signal is explored. A model of noise production mechanism during silent articulation considers acoustic propagation in the trachea or tracheostomy tube. A linear prediction of speech signals is applied to estimate the synthesis filter characteristics of noise production. The similarity of artificial and estimated noise spectra is examined in terms of the simplification applied in the model. This work leads to a discussion of how to efficiently reduce stoma noise from speech signal using digital signal processing methods. As a pilot study, acoustic signals of patients' speech is denoised by using a Kalman filter.**

Keywords : **Kalman filter, autoregressive process, voice denoising, total laryngectomy.**

I. INTRODUCTION

The speech processing algorithms are applied to patients who underwent complete removal of the larynx (total laryngectomy). The problem of objective speech factors evaluation in silently articulated speech of laryngectomees was reported in [1]. Isolated vowels could not be distinguished using spectral analysis and formants tracking methods. The authors concluded that noise from tracheostoma plays a significant role in masking speech spectrum. The phenomenon indicates an impact of breathing sounds on voice quality and intelligibility in laryngectomized patients.

Total laryngectomy is usually performed as a treatment for laryngeal carcinoma. After larynx removal, tracheostomy procedure is performed. The trachea is diverted to the outside and sutured to the skin of the sternal notch. This procedure is called tracheostomy. After the trachea is brought to the outside, a small hole is made in the throat. The patient breathes through the opening referred to as a tracheal stoma (TS). After surgery, the tracheostomy tube (TT) is inserted into the stoma for it to remain open whilst it heals. Once the stoma has healed, a patient has made significant progress and the physician deems it safe, the tube can be permanently removed.

There are many significant consequences and complications of the surgical procedure. As an impact on speech, the patients lose the source of vibration (larynx) and air supply (respiration from lungs). Following the loss of vocal cords, patients are not able to phonate adequately. As a consequence, two kinds of speech signal defects can be found: poor quality of articulated sounds and disturbances by breathing noise. The patients need to modify the articulation patterns to compensate for the lack of phonation and pulmonary air. They should also learn how to efficiently synchronize the breathing with speech. There are different ways for articulation among the patients depending on the rehabilitation performed.

The main goal of phoniatric rehabilitation is to pronounce vocalized sounds (that are naturally articulated with the use of vocal cords) using alternative methods. In pharyngeal, esophageal or gastric speech, the patients use other tissues for phonation (e.g. in esophageal speech, alaryngeal voice is articulated using a pharyngo-esophageal segment). By surgical prosthesis the patients can produce tracheoesophageal speech. The patients can also use electromechanical devices (referred to as electrolarynx) to produce electrolaryngeal speech.

The certain percentage of laryngectomees never acquires an alternative voice. They communicate with silently articulated words called pseudo-whisper. The pseudo-whisper involves a compression of the air in the oral cavity acting in combination with articulatory movements of the consonant sounds for the production of voice [2]. Their speech is based mostly on lips or tongue smacking and frication. Many patients that did not acquire a substitute voice try to speak and exhale air simultaneously (as they used to before). Therefore, the air coming out from TT produces unwanted noise, making speech less intelligible.

There were several research on laryngectomees' speech intelligibility enhancement. Most of the work was done for esophageal speech, e. g. [3]. On the other hand, silent speech interfaces have been developed [4].

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

The method for colored noise removal using auto-regressive (AR) model and Kalman filter was described in [5]. There have also been several papers of speech de-noising based on auto-regressive moving-average (ARMA) model [6]. Before applying this method, the spectral characteristics of noise must be known. Thus, an acoustic model of noise production describing sound characteristics can support noise elimination algorithms.

The objective of this study was to develop an acoustic model of noise production mechanism during silent articulation. In this paper the authors described the model and validate it with the speech samples. They experimented in efforts to make the model correspond more closely to physical situations.

The main goal of the study was to suppress noise from TS to enhance the speech to make it more intelligible for listeners.

II. METHODOLOGY

A. Acoustic model of noise production

A standard vocal tract model considers simplified physical approximations for the noise source, acoustical propagation, and radiation. A widely used digital waveguide modeling considers a narrow acoustic tube being a one-dimensional resonator [7]. Only plane-wave propagation is considered. When using speech enhancement procedures [3], the speech signal is usually modeled as AR process. In the case where the acoustic source is not placed at the beginning of an acoustic tube, the model is described by ARMA process.

Model of isolated tracheostomy tube: An overview of applied tracheostomy tubes allows the use of a simple acoustic model of noise production. It was assumed that noise source is located at the entrance of the tube where sounds are generated by turbulent airflow. The vortex is formed by a tube being an obstacle [8], thus theoretical place of noise generation are opposite for inhalation and exhalation. The airflow and the sound waves travel through the tube of the length $L_R \approx 9$cm with constant diameter d_R and cross-sectional area S_R. After assumption for infinite length trachea with its cross-sectional area $S_T >> S_R$, TT can be modeled as an acoustic pipe in which both ends are open. Such a pipe has full harmonic series with the main pitch wavelength twice as long as the pipe. Corresponding digital filter order N_R for sample frequency f_s can be calculated by using equation (1), where sound velocity in air c = 344 m/s.

$$N_R = 2 L_R f_s/c \quad (1)$$

Tracheal stoma alone: In case of tracheal opening alone, noise is generated by the random pressure fluctuations produced at the TS when air is forced through the constriction. As air exits from a constriction it forms a jet which gradually mixes with the surrounding air. According to [9] the spectral characteristics of the sound generated by a jet depends only on the jet velocity and diameter. The total sound power spectrum has a broad peak at about SV/d Hz, where V is the flow velocity in the center of the jet as it exits the constriction, d is the jet diameter, and S, the Strouhal number according to [9]. The sound can be also produced after sternal notch skin excitation by inhalation and exhalation airflow. The acoustic tube of the trachea will introduce the zeros into speech signal frequency characteristics. The lungs have complicated geometry, with successively branching tubes, extending to a quite small scale at the alveoli, which according to [10] produce little reflection in the frequency range up to 3kHz. According to [11] the simplest model of the subglottal airways is a uniform tube with rigid walls and cross-sectional area $S_T = 2.5$ cm open at the inferior end. The effective acoustic length of subglottal airways is $L_T = 20$cm, The frequency spectrum of this acoustic pipe has only odd harmonic series with the main pitch wavelength fourth as long as the pipe.

Tracheostomy tube in trachea joined model: The authors proposed a cross-sectional area model TT inserted into TS. Because the tube is inserted into the trachea the authors suggested a model with three-port junction and following branches:

- TT branch, regular tube with length L_R, diameter d_R and cross sectional area S_R opened to air at the end,
- trachea to lungs direction, regular tube with length $L_T - L_R$, diameter d_T and area S_T opened to lungs at the end,
- trachea to TS direction, regular tube with circular ring sectional area $S_T - S_R$ (inner diameter d_R and outer d_T), and length L_R with rigid wall at the end.

During an exhalation the sound source is most likely placed at the entrance of TT [8] while inhalation noise generator can be placed elsewhere. The presence of side-branching resonator introduces antiresonances in the spectrum (zeros to filter model) [12].

B. Material

Audio recordings of 14 patients after total laryngectomy, speaking with pseudo-whisper, were chosen from material described in [1]. A single microphone audio signal was downsampled from 44.1kHz to 10 kHz using FFmpeg application. The material of this research did not contain recordings of breathing. A middle 20 ms sample of isolated Polish vowel 'e' was used for noise estimation purpose. Audio signal of isolated Polish word 'boso' spoken by one of the patients was used in a denoising experiment. The word is in form of consonant-vowel-consonant-vowel CVCV transition. There was no information if patient wore TT, thus presence of TT was concluded from video material.

In many cases, when the stoma was hidden under some elastic band, the authors assumed the absence of TT.

C. Speech enhancement algorithms

It was assumed that post-laryngectomized patients speech is composed of clean speech signal and the uncorrelated additive noise. In the experiment, the all-pole linear predictive model of noise production model was used.

The methods of adaptive signal processing were applied to find an optimal order of auto-regressive (AR) process of noise production. AR parameters were estimated using Kalman filtering procedure according to [13], implemented as 'aar' function in GNU Octave package 'tsa' ver. 4.2.4. This algorithm also calculates adaptive ARMA estimates. The ratio of mean squared error and mean squared signal (MSE/MSY) was calculated to measure the Goodness-of-fit. The approximation of breathing sound characteristics was further used for noise suppression algorithm.

The authors applied Kalman filter with transition matrix corresponding to AR model according to [5]. Slightly modified Matlab script for colored noise removal implemented and published by [5] was used. The characteristics of noise were calculated for one of the patients, from the same fragment of vowel 'e' as in the previous section. The speech signals were modeled as 10-th -order AR process. According to equation (1), for sample frequency 10 kHz, vocal tract length of about 17.2 cm and up to 5 speech formants can be modeled. The number of noise AR process coefficients were specified after noise estimation algorithm.

IV. Results

A. Noise AR process order estimation

The MSE to MSY ratio was extracted for every patient and number of coefficients from range 1-16. The results are shown on Fig. 1. After visual inspection, the authors chose 6 AR model coefficients for denoising algorithm. The authors also found that moving-average (MA) parameters extension improves the approximation rate much less than AR order.

B. Tracheostomy tube noise reduction

The noise suppression algorithm has been performed for one of the patients wearing TT. The original and processed sound signals in time domain are shown on Fig. 2. The spectrum was not degraded after denoising.

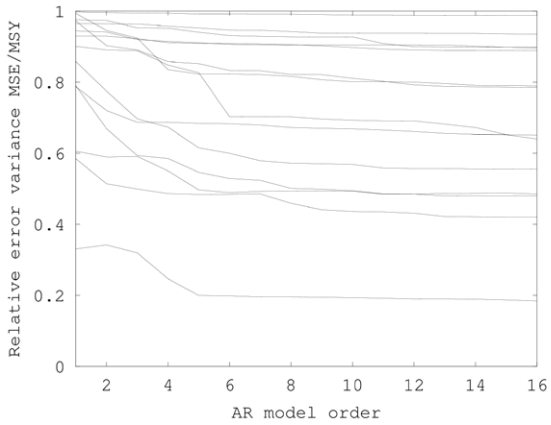


Fig. 1. Relative error variance MSE/MSY of AR process for 14 patients and model order from 1-16.

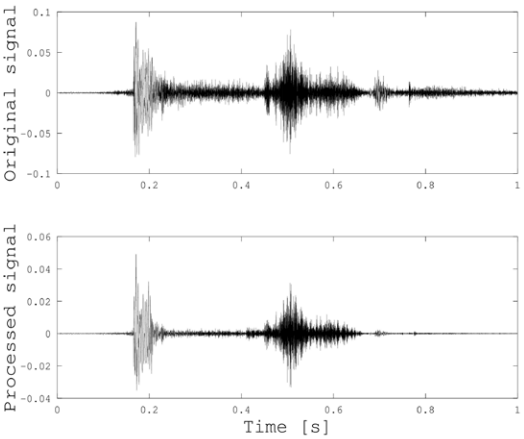


Fig. 2. Original (top) and processed (bottom) sound signals of word 'boso'

V. Discussion

The vowel 'e' phonation was used for noise estimation because there were no audible sounds from oral cavity for isolated vowels; any other vowel than 'e' could be used instead for noise characteristics estimation. Also, the open question that remains is if the noise from inhalation, which takes place before phonation, can be the source of usable noise characteristics evaluation.

With the chosen AR coefficients number up to 3 poles can be modeled in 5 kHz range. From visual examination of original and approximated frequency spectrum, the authors concluded that the formants positions are mostly related to TT length. For resonances related to trachea length, about 2 times greater number of AR parameters should be assumed. Patients speaking with pseudo-

whisper voice cannot articulate the fundamental frequency. Thus, the main problem in this study is to suppress noise in the speech which has also noisy characteristics. Because of this phenomenon, results were somewhat inconclusive. The authors evaluated very high error rate of signal approximation for most of the subjects. Problems with the model can be also due to an inaccurate source representation. The characteristics of the sound radiated from the stoma are determined by the source location. However, a noise source location cannot be precisely determined. Moreover, the authors did not investigate, how the stoma noise interacts with the vocal tract.

The stoma noise was suppressed with little distortions of speech. The effectiveness of the useful speech signal extraction from noisy samples should be measured and compared. The improvement of voice quality can be measured with the help of Multi-Dimensional Voice Program (MDVP) tools. In a future work, the authors should explore another method for speech denoising, e.g. wavelets, non-linear models or neural networks.

A computer program for automatic noise reduction can support patients in speech communication through internet. To enable automatic speech denoising, an acoustic signal should be classified as speech or breath. Some visual parameters like mouth opening extracted from video images can support this decision. An alternative solution for noise and speech signal separation is to use two microphones, one placed at stoma and another at mouth. An adaptive algorithm for noise reduction can be then applied.

In [1] the authors reported another consequence of respiration and articulation patterns change in pseudo-whisper voice. Due to low capacity of air supply, duration of sound-characteristic spectrum is very short compared to noise duration. Moreover, vowel-characteristic formants can be noticed only in the CV (consonant-vowel) transitions. In a future work, the authors should propose some extrapolation algorithm for phonemes sounds prolongation.

VI. Conclusion

The authors did not find any notable visual correlations between sound characteristics and TT presence. From subjective evaluation, the authors concluded that breathing noise from TT was suppressed. Thus, for noise suppression purposes it is sufficient to assume that additive noise is modeled as AR process.

References

[1] R. Pietruch, M. Michalska, W. Konopka, and A. Grzanka, “Methods for formant extraction in speech of patients after total laryngectomy,” *Biomedical Signal Processing and Control*, vol. 1/2, pp. 107–112, 2006.

[2] L. Di Carlo, Speech After Laryngectomy, ser. S*pecial education and rehabilitation monograph series. Syracuse University Press*, 1955.

[3] B. Garcia, I. Ruiz, and A. Mendez, “Oesophageal speech enhancement using poles stabilization and kalman filtering,” *in Acoustics, Speech and Signal Processing IEEE International Conference on*, 2008, pp. 1597–1600.

[4] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, “Silent speech interfaces,” *Speech Communication*, vol. 52, no. 4, pp. 270 – 287, 2010.

[5] J. Kybic, “Kalman filtering and speech enhancement,” Master’s thesis, *Czech Technical University, Prague, Czech Republic,* 1998.

[6] M. Geist, O. Pietquin et al., “Kalman filtering & colored noises: the (autoregressive) moving-average case,” in *Workshop Proceedings of ICMLA 2011*, 2011, pp. 1–4.

[7] J. O. Smith, “A new approach to digital reverberation using closed waveguide networks,” *no. STAN-M-31, Burnaby, B.C., Canada,* 1985, pp. 47–53.

[8] S. W. Rienstra and A. Hirschberg, An introduction to acoustics. Technische Universiteit Eindhoven, 1999.

[9] C. H. Shadle, “The acoustics of fricative consonants,” 1985.

[10] N. H. Fletcher, L. C. L. Hollenberg, J. Smith, A. Z. Tarnopolsky, and J. Wolfe, “Vocal tract resonances and the sound of the Australian didjeridu (yidaki) II. Theory,” *The Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 1205–1213, 2006.

[11] S. M. Lulich, A. Alwan, H. Arsikere, J. R. Morton, and M. S. Sommers, “Resonances and wave propagation velocity in the subglottal airways,” *Acoustical Society of America Journal*, vol. 130, p. 2108, 2011.

[12] Y. Qi and R. A. Fox, “Analysis of nasal consonants using perceptual linear prediction,” *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1718–1726, 1992.

[13] A. Schlögl, The Electroencephalogram and the Adaptive Autoregressive Model: Theory and Applications, *ser. Berichte aus der Medizinischen Informatik und Bioinformatik. Shaker Verlag GmbH*, 2000.

WLP-BASED TV-CAR SPEECH ANALYSIS AND ITS EVALUATION FOR F_0 ESTIMATION

Keiichi Funaki[1] and Keita Higa[2]

[1]Computing & Networking Center, University of the Ryukyus, 903-0213, Okinawa, Japan
[2]Department of Engineering, University of the Ryukyus, 903-0213, Okinawa, Japan
Email: funaki@cc.u-ryukyu.ac.jp

Abstract: **Analytic signal is a complex-valued signal whose real part is observed signal and its imaginary part is Hilbert transformed signal for the real one. Since analytic signal provides the spectrum only on positive frequencies, it can be decimated by a factor of 2. As a result, spectrum estimation accuracy can be improved in low frequencies. The remarkable feature makes it possible to improve the performance of speech processing since speech signal provides a large amount of spectral information in low frequencies. We have proposed Time-Varying Complex AR (TV-CAR) speech analysis, MMSE, GLS, ELS, and so on. We have evaluated them for *F*0 estimation, speech recognition, speech coding, and speech enhancement. On the other hand, P.Alku has proposed Stabilised Weighted Linear Prediction (SWLP) and has applied it to speech recognition and speech synthesis. In SWLP, STE (Short Term Energy) is adopted as a weighting function to realize Weighted Linear Prediction (WLP). Introducing STE makes it possible to realize parameter estimation weighting in glottal closure segment. In this paper, we propose SWLP-based MMSE TV-CAR speech analysis and evaluate it with SRH (Summation Residual Harmonics)-based *F*0 estimation. SRH was proposed by A.Alwan that estimates *F*0 so as to maximize the SRH function.**

Keywords : **Speech analysis, Analytic signal, Complex analysis, WLP, *F*0 estimation**

I. INTRODUCTION

Speech signal uttered by human is generated by vocal tract cavity changing that result in resonance and anti-resonance frequencies. Vocal tract can be modeled by a filter and the filter is driven by glottal excitation generated by glottis vibration. Speech analysis is a signal processing algorithm that estimates spectral parameter from speech signal. Speech spectrum consists of two parts; vocal tract filter and glottal excitation. LPC analysis is commonly used for speech coding such as in ITU-T G.729, G.718 [1] or so on. LPC analysis estimates AR speech spectrum corresponding to vocal tract and glottal excitation. Since LPC analysis can estimate stable AR spectrum envelope from speech signal with low computational amount, LPC analysis is used for low bit rate speech coding. However, LPC analysis suffers from some drawbacks such as sensitive for additive noise, time invariant analysis, or so on. In order to cope with the difficulties, many and many speech analysis methods have been proposed. In order to expand it to time-varying analysis, TVAR analysis has been proposed [2]. In order to improve the spectral resolution, non-linear frequency warping analysis such as PLP [3] or Mel-LPC [4], and complex speech analysis [5][6] have been proposed. In order to make it robust against additive noise, robust criteria have been proposed [7].

We have already proposed time-varying complex AR (TV-CAR) speech analysis for analytic speech signal [8][9][10]. TV-CAR analysis can estimate time-varying spectrum and can improve spectral resolution due to the nature of analytic signal. In addition, we have proposed robust algorithm such as WLS [9], GLS/ELS [10] or so on, besides MMSE algorithm [8].

On the other hand, P.Alku has proposed SWLP (Stabilised Weighted Linear Prediction) [7] that estimates stabilized AR filter from speech signal using Weighted Linear Predictive manner. Short Term Energy (STE) [11] of speech is used as a weighting function. The STE weights criterion in glottal closure and anti-weights in glottal open phase, as a result, the WLP with the STE can estimate more accurate and more robust speech spectrum.

We have already proposed F_0 estimation algorithm based on the TV-CAR speech analysis. In [12], weighted autocorrelation for time-invariant complex AR residual by MMSE-based TV-CAR analysis is used for F_0 estimation. In [13], weighted auto-correlation for time-invariant complex AR residual by ELS-based TV-CAR analysis is used for F_0 estimation. In [14], time-varying analysis is evaluated for [13]. In [15], Zero Frequency Resonance (ZFR) based on the TV-CAR analysis has been proposed and the F_0 estimation using the ZFR has been evaluated. In [16], Summation of Residual Harmonics (SRH)-based F_0 estimation based on the TV-CAR analysis has been proposed and evaluated. SRH (Summation Residual Harmonics) is proposed by A.Alwan,et.al.[17] and it focuses on residual harmonics. Power spectrum of LP residual presents peaks at the harmonics of the F_0. The SRH is calculated by using the power spectrum. The SRH is summation of the harmonics minus the half-harmonics. By peak-picking of the SRH in

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

certain range of F_0, one can estimate the F_0.

In this paper, the WLP using the STE is introduced in the TV-CAR analysis and evaluated it on F_0 estimation. The SRH [16] is used as an F_0 estimation to evaluate the WLP-based TV-CAR speech analysis. The remainder of this paper is organized as follows. In Section 2, WLP-based TV-CAR analysis is explained. In Section 3, speech spectrum estimated by the WLP-based TV-CAR analysis is shown. In Section 4, F_0 estimation based on the WLP-based TV-CAR analysis is evaluated.

II. TV-CAR SPEECH ANALYSIS

A. Analytic Speech Signal

Target signal of the time-varying complex AR (TV-CAR) method is an analytic signal that is complex-valued signal defined by

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

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, respectively. Notice that superscript c denotes complex value in this paper. Since analytic signals provide the spectra only over the range of $(0, \pi)$ analytic signals can be decimated by a factor of two. $2t$ means the decimation. The term of $1/\sqrt{2}$ is multiplied in order to adjust the power of an analytic signal with that of the observed one.

B. Time-Varying Complex AR (TV-CAR) Model

TV-CAR model is defined as

$$Y_{TVCAR}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) z^{-i}} \tag{2}$$

where I is AR order. The input-output relation is defined as

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

where $u^c(t)$ and $y^c(t)$ are taken to be complex-valued input and analytic speech signal, respectively. In the TV-CAR model, the complex AR coefficient is modeled by a finite number of arbitrary complex basis. Note that Eq.(3) parameterizes the AR coefficient trajectories that continuously change as a function of time so that the time-varying analysis is feasible to estimate continuous time-varying speech spectrum. In addition, as mentioned above, the complex-valued analysis facilitates accurate spectral estimation in the low frequencies, as a result, this feature allows for more accurate $F0$ estimation if formant structure is removed by the inverse filtering. Eq.(3) can be represented by vector-matrix notation as

$$\begin{aligned} \overline{y}_f &= -\overline{\Phi}_f \overline{\theta} + \overline{u}_f \\ \overline{\theta}^T &= [\overline{g}_0^T, \overline{g}_1^T, \cdots, \overline{g}_l^T, \cdots, \overline{g}_{L-1}^T] \\ \overline{g}_l^T &= [g_{1,l}^c, g_{2,l}^c, \cdots, g_{i,l}^c, \cdots, g_{I,l}^c] \\ \overline{y}_f^T &= [y^c(I), y^c(I+1), y^c(I+2), \cdots, y^c(N-1)] \\ \overline{u}_f^T &= [u^c(I), u^c(I+1), u^c(I+2), \cdots, u^c(N-1)] \\ \overline{\Phi}_f &= [\overline{D}_0^f, \overline{D}_1^f, \cdots, \overline{D}_l^f, \cdots, \overline{D}_{L-1}^f] \\ \overline{D}_l^f &= [\overline{d}_{1,l}^f, \cdots, \overline{d}_{i,l}^f, \cdots, \overline{d}_{I,l}^f] \\ \overline{d}_{i,l}^f &= [y^c(I-i) f_l^c(I), y^c(I+1-i) f_l^c(I+1), \\ &\qquad \cdots, y^c(N-1-i) f_l^c(N-1)]^T \end{aligned}$$

where N is analysis interval, $\overline{\ },y_f$ is $(N - I, 1)$ column vector whose elements are analytic speech signal, $\overline{\ },\theta$ is (LI, 1) column vector whose elements are complex parameters, $\bar{\Phi}_f$ is $(N–1, LI)$ matrix whose elements are weighted analytic speech signal by the complex basis. Superscript T denotes transposition.

C. MMSE-Based Algorithm [10]

MSE criterion is defined by

$$\begin{aligned} \bar{r}_f &= [r^c(I), r^c(I+1), \cdots, r^c(N-1)]^T \\ &= \bar{y}_f + \bar{\Phi}_f \hat{\theta} \end{aligned} \tag{4}$$

$$r^c(t) = y^c(t) + \sum_{i=1}^{I} \sum_{l=0}^{L-1} \hat{g}_{i,l}^c f_l^c(t) y^c(t-i) \tag{5}$$

$$E = \bar{r}_f^H \bar{r}_f = (\bar{y}_f + \bar{\Phi}\hat{\theta})^H (\bar{y}_f + \bar{\Phi}\hat{\theta}) \tag{6}$$

where $\hat{g}_{i,l}^c$ is the estimated complex parameter, $r^c(t)$ is an equation error, or complex AR residual and E is Mean Squared Error(MSE) for the equation error. To obtain optimal complex AR coefficients, we minimize the MSE criterion. Minimizing the MSE criterion of Eq.(6) with respect to the complex parameter leads to the following MMSE algorithm.

$$(\overline{\Phi}_f^H \overline{\Phi}_f)\hat{\theta} = -\overline{\Phi}_f^H \overline{y}_f \tag{7}$$

Superscript H denotes Hermitian transposition. After solving the linear equation of Eq.(7), we can get the complex AR parameter ($a^c{}_i(t)$) at time t by calculating the Eq.(2) with the estimated complex parameter $\hat{g}_{i,l}^c$.

III. MMSE ANALYSIS BASED ON WLP

P.Alku et.al. have proposed SWLP method[7] as the extension of WLP proposed by C.Ma et.al.[11]. The SWLP offers a guarantee for AR filter stability. Since the MMSE algorithm is an extension of covariance method that has no guarantee for filter stability. The MMSE

algorithm is extended in a framework of WLP. In [7][11], the following STE(Short Term Energy) is used as a weighted function *w(t)*.

$$w(t) = \sum_{i=0}^{M-1} |y(t-i-1)|^2 \tag{8}$$

As shown in [7], the STE weights the criterion in glottal closure instant and less weights in glottal open phase. For this reason, accurate estimation can be realized. Weighting function is defined by

$$\bar{W} = \begin{pmatrix} \sqrt{w(I)} & 0 & \dots & 0 \\ 0 & \sqrt{w(I+1)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{w(N-1).} \end{pmatrix} \tag{9}$$

MSE criterion Eq.(6) for forward Linear Prediction is defined by

$$V_f = \bar{r}_f^H \bar{W}^H \bar{W} \bar{r}_f. \tag{10}$$

Minimizing the MSE criterion with respect to $\hat{\theta}$ leads to the following forward Linear predictive MMSE-based WLP algorithm.

$$(\bar{\Phi}_f^H \bar{W}^H \bar{W} \bar{\Phi}_f)\hat{\theta} = -\bar{\Phi}_f^H \bar{W}^H \bar{W} \bar{y}_f \tag{11}$$

IV. Experiments

We have evaluated the estimated spectra every 2msec by means of the WLP method comparing with conventional methods; the MMSE methods. According to the spectra, it is obvious that WLP-based method can estimate more sharp time-varying speech spectra. In order to evaluate the proposed WLP-based TV-CAR speech analysis, the SRH-based *F*0 estimation [16] based on the method is evaluated. The SRH is based on residual power spectrum, thus, the residual is estimated by the WLP method. The experiments were carried out with Keele Pitch Database[18] corrupted by white Gauss or Pink noise[19] whose noise level was -5, 0, 5, 10, 20, 30[dB]. The noise corrupted speech is filtered by the IRS filter [20] for speech coding application. The proposed method was compared with conventional methods as follows.

(1) **SP**: Weighted auto-correlation for Speech signal [21]
(2) **TVR_WLP**: SRH of WLP-based time-varying real-valued
AR residual
(3) **TVC_WLP**: SRH of WLP-based time-varying complex-valued AR residual (**Proposed**)
(4) **TVC_SRH**: SRH of time-varying complex AR residual [16]

Experimental conditions are as follows. AR order *I* is 14 for real analysis, 7 for complex analysis. Basis expansion order *L* is 2 and first order polynomial (1,*t*) is selected as a basis function. The performance is evaluated by using 10 % of GPE (Gross Pitch Error). Figures 1 show the GPEs corresponding to each method for Female speech in which black line with black square means GPEs for method (1), black line with black diamond means GPEs for method (2), red line means GPEs for proposed method (3) and blue line means GPEs for method (4). Figures 1 demonstrate that the SRH based on WLP time-varying real valued AR residual and complex AR residual perform better than Shimamura method [21]. Moreover, the proposed method can perform slightly better than the SRH method [16] for Female speech corrupted by pink noise (red line is slightly below than blue line at 10, 5, 0 dB) although it does perform equally for the SRH method for additional white Gauss noise.

Table 1: Experimental Conditions

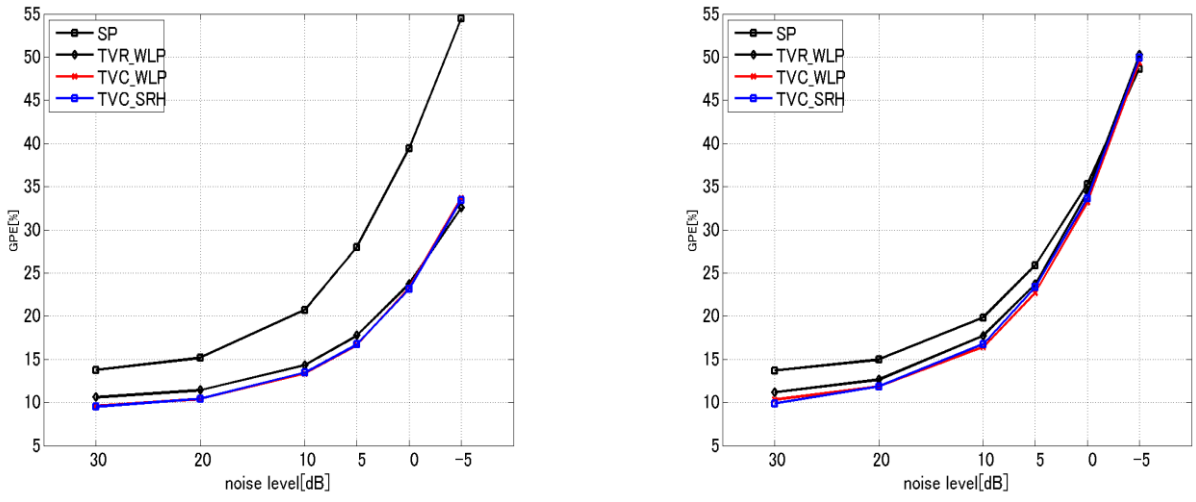
Speech data	Keele Pitch DB.[18]
Sampling	10kHz/16bit
Analysis window	Window Length: 25.6[ms] Shift Length: 10.0[ms]
Complex-valued AR Basis Pre-emphasis	I=7, L=2 (time-varying) $f_l^c(t) = t^l / l!$ No Operation
Real-valued AR Basis Pre-emphasis	I=14, L=2 (time-varying) $f_l^c(t) = t^l / l!$ No Operation
M in Eq.(8)	4
Noise Noise Level	White Gauss noise Pink noise[19] 30,20,10,5,0,-5[dB]

V. Conclusions

We have proposed WLP-based TV-CAR speech analysis that is Weighted LP analysis with weighing function of STE. The STE does not weight for glottal closure phase but also anti-weights for glottal open phase. It enables to estimate more accurate spectrum since assumption of LP does satisfy for glottal closure phase and it does not satisfy for glottal open phase. The estimated spectra indicate that the WLP-based method can estimate more accurate time-varying speech spectrum than the MMSE-based method. Furthermore, we have also evaluated the proposed analysis using the SRH-based *F*0 estimation. The proposed analysis can perform better for female speech corrupted by pink noise. As future works, we are going to proposed WLP-based ELS method and to evaluate it on the *F*0 estimation based on weighted autocorrelation criteria [12] and IRAPT[13]. Furthermore, we are going to evaluate the WLP-based method for Front-End of robust automatic speech recognition (ASR).

References

[1] ITU-T G.718, http://www.itu.int/rec/T-REC-G.718/
[2] A.V.Oppenheim, et.al.,, "Time varying parametric modeling of speech," *Signal Processing,* 1983
[3] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *JASA*.1990.
[4] H.Matsumoto et.al.,"Evaluation of MEL-LPC cepstrum in a large vocabulary continuous speech recognition," *ICASSP-2002*,2002.
[5] S.M.Kay, "Maximum entropy spectral estimation using the analytic signal," *IEEE Trans.* ASSP-26, pp.467-469, 1980.
[6] T.Shimamura,et.al.,"Complex linear prediction method based on positive frequency domain," *IEICE Trans.* Vol.J72-A, pp.1755-1763, 1989.
[7] C. Magi J. Pohjalainen T. Backstrom and P. Alku, "Stabilised Weighted Linear Prediction," *Speech Communication*, 51, pp.401-411,2009.
[8] K.Funaki,et al., "On a Time-varying Complex Speech Analysis," *EUSIPCO-98*, Rhodes, Greece, Sep.9-11,1998.
[9] K.Funaki,et al., "On Robust speech analysis based on time-varying complex AR model," *ICSLP-98*, Sydney, Australia, Nov.30-Dec.4,1998.
[10] K.Funaki, "A Time-Varying Complex AR Speech Analysis Based on GLS and ELS Method," *EUROSPEECH-2001*, Aalborg, Denmark, Sep.7,2001.
[11]C.Ma, Y.Kamp, L.F.Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*,Vol.12, Issue 1, pp.69-81, Mar.1993,
[12] K.Funaki,et al., "Robust F0 Estimation Based on Complex LPC Analysis for IRS Filtered Noisy Speech," *IEICE Trans. on Fundamentals*, Vol. E90-A, No.8.,1579-1586, Aug. 2007.
[13] K.Funaki, "F_0 estimation based on robust ELS complex speech analysis," *EUSIPCO-2008*, Lausanne, Switzerland, Aug.2008.
[14] K.Funaki, "On Evaluation of the F_0 estimation based on time-varying complex speech analysis," *Interspeech2010*,Makuhari, Japan, Sep. 2010.
[15] K.Funaki and T.Higa, "Evaluation of F0 estimation using ZFR based on time-varying speech analysis," *ISCAS2012*, Seoul, May, 2012.
[16] K.Funaki and T.Higa, "F0 Estimation Using SRH Based on TV-CAR Speech Analysis," *EUSIPCO 2012*, Bucharest, Romania, Aug.,2012.
[17] T.Drugman and A.Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," *Interspeech2011*, Firenze, Italy, Sep. 2011.
[18] Keele Pitch Database, University of Liverpool, http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html
[19] NOISE-X92, http://spib.rice.edu/spib/select noise.html
[20] ITU-T Recommendation G.191, Software tools for speech and audio coding standardization, Nov. 2000.
[21]T.Shimamura and H.Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech," *IEEE Trans.* Speech and Audio Processing, vol. 9, no. 7, pp. 727-730, 2001.
[22] IRAPT http://dsp.tut.su/irapt.html



(1) GPEs for additive white Gauss noise

(2) GPEs for additive pink noise

Figure 1 F_0 estimation performances (GPE)

Session VI:
VOICE AND PATHOLOGIES

VOCAL TRACT SETTINGS IN SPEAKERS WITH OBSTRUCTIVE SLEEP APNEA SYNDROME

J. L. Blanco [1], J. Schoentgen [2]

[1] Signal Processing Applications Group, Universidad Politécnica de Madrid
ETSI de Telecomunicación, Avda. Complutense 30, 28040 Madrid, Spain
jlblanco@gaps.ssr.upm.es

[2] National Fund for Scientific Research, Belgium & Laboratories of Images, Signal Processing and Acoustics
CP 165/51, Faculty of Applied Sciences, Université Libre de Bruxelles, 50, Av. F.D. Roosevelt, B-1050, Brussels, Belgium
jschoent@ulb.ac.be

Abstract: **Automatic systems based on speech signal analysis for the early detection of obstructive sleep apnea (OSA) have achieved fairly high performance rates in recent years. However, a satisfactory explanation of these results has not been available. This presentation aims at explaining via an examination of the long-term spectra of OSA patients and normal control speakers these systems' ability to discover OSA speakers on the base of all-purpose cepstral coefficients. An interpretation of the long-term spectra in terms of the underlying tract settings suggests that the speech of OSA patients is characterized by a pharyngeal narrowing that may be captured by acoustic cues of the spectral contour of windowed speech frames. A novel interpretation of long-term spectra in terms of the first principal component of the temporal sequence of short-term amplitude-spectra is also discussed.**

Keywords: obstructive sleep apnea (OSA), spectral analysis of speech, vocal tract settings, long-term average spectrum (LTAS), pharyngeal narrowing.

I. INTRODUCTION

Severe obstructive sleep apnea is characterized by the interruption of breathing during sleep [1], involving episodes that may last for more than 10 seconds and which recurrently occur during the night, with up to more than 30 episodes per hour. This syndrome affects 2 to 4% of the male population between 30 and 60 years of age. A possible effect of OSA is daytime sleepiness, increasing the risk of the patient to get involved in traffic accidents or leading to poor work performance [2].

The discovery via speech analysis of patients that have a propensity to suffer from severe obstructive sleep apnea syndrome (OSA) has become more reliable in recent years. Previous work by the first author has shown that it is possible to discriminate between modal speakers and speakers suffering from severe OSA by means of generic automatic classifiers that rely on a conventional coding of speech frames by means of mel-frequency cepstral coefficients ([3]–[5]). However, no results on speech analysis are available that would enable linking that observation to OSA speaker anatomy, physiology or OSA pathogenesis.

This presentation aims at explaining the ability to discover OSA patients via speech analysis in terms of what is known about speech production in general and vocal tract settings in particular. J. Laver [6] has discussed vocal tract settings extensively. They designate articulatory biases of the neutral tract shape, which are common to all the phonetic segments of the utterances of a speaker. The susceptibility of individual phonetic segments to tract settings is variable depending on the degree by which the properties of a phonetic segment are affected. For instance, one expects +spread sounds to be biased by a lip protrusion setting more than +round sounds that are characterized by lip protrusion anyway.

Easily observable anatomical or physiological OSA cues are patients' weight, height, body mass index and cervical perimeter as well as snoring. The latter together with other symptoms (occasional hypoplasia and/or backward displacement of the maxilla and mandible) suggest that their oro-pharyngeal cavity is narrowed. Magnetic Resonance Imaging of OSA patients has confirmed this and that sole observation can be used to identify OSA cases. The narrowing of the pharynx may be assimilated to a vocal tract setting that is likely to bias a speaker's speech sounds via a shift in vocal tract resonances. Experiments are reported hereafter that have been carried out with a view to testing that hypothesis on speech records.

J. Laver [6] and F. Nolan [7] have suggested tracking vocal tract settings in speech by means of the Long Term Average Spectrum (LTAS). One obtains the LTAS of an utterance by computing the amplitude spectrum for successive frames and averaging the amplitude spectra. Averaging is expected to lessen the contribution of phonetic segment variability and boost what is common to all frames (i.e. the tract setting or phonatory setting, depending on the frame length). In the past, Long Term Average Spectra often have been used to characterize a speaker overall, including information from all the speech frames. Focussing by means of the LTAS on tract settings only has been rare.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

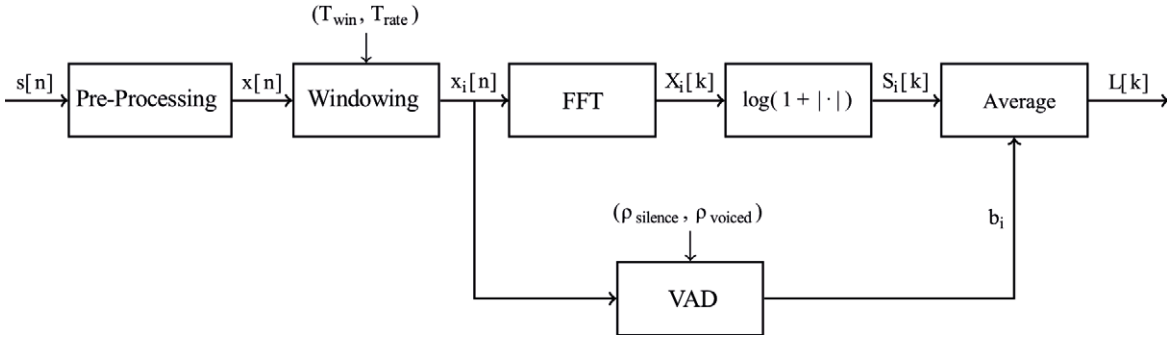


Figure 1: Block diagram of the long-term average spectra estimation.

II. METHODS

Our corpus has comprised speech records from 80 male speakers [8]. Forty of them have been healthy or have suffered from mild OSA (AHI<10), while the remaining forty have been patients suffering from severe OSA (AHI>30). The apnea-hypo-apnea index (AHI) is the number of apnea plus hypoapnea events per hour of sleep, and is often used by clinicians to describe the severity of patients' condition and to adjust their treatment [9]. Hereafter, the AHI<10 speakers are designated as controls and the AHI>30 speakers as patients. The two groups have been balanced for weight, height and age to decrease the influence of secondary speaker characteristics.

Each speaker sustained a complete set of Spanish vowels [i,e,a,o,u] as well as four Spanish sentences. The sentences have been short and phonetically balanced. They have been designed with a fixed number of intonation groups to decrease intra-speaker variability.

The computation of the LTAS has involved (i) pre-emphasizing the speech signal (α=0.99) to remove zero-frequency and ultra-low frequency spectral components, (ii) windowing, (iii) voicing detection, (iv) computing the amplitude spectrum, (v) boosting high-frequency amplitudes and (vi) averaging. The block diagram in Fig. 1 summarizes steps (i) to (vi).

The frame length has been set to T_{win}= 5 ms and the frame hop to T_{rate}= 3 ms to decrease the influence of the voice source harmonics on the amplitude spectrum because the focus is on the spectral contour that reports vocal tract resonances.

Even though other publications have retained vowel nuclei as the sole contributors to LTAS [10], here any voiced frames have been included. Voiced frames have been discovered by means of a voicing detector proposed in [11] and which is easily tunable by weighting each of its two steps that are the following. The first involves a frame-by-frame energy estimation. The purpose is to detect and remove silent intervals. Threshold $\rho_{silence}$ for speech activity detection has been set to 20% of the average energy of the frames of one speech record. The second step involves auto-correlation coefficient ρ_{ss} between an analysis frame and the analysis frame delayed by one sample. The purpose of that step is to detect the frames that are voiced and for which the auto-correlation coefficient is large, i.e. $\rho_{ss}(1) / \rho_{ss}(0) \geq 0.9$. Only frames that have been tagged as voiced have been used to compute the LTAS.

The purpose of boosting is to increase the amplitude of high-frequency components compared to low-frequency components. The reason is that the spectral slope of the voice source causes higher formants to be feeble and barely visible in the amplitude spectrum. The boosting function is the logarithm of the spectral amplitude+1. The +1 guarantees that all amplitudes have positive log-transforms that respect monotony so that the average of the boosted amplitudes can be interpreted meaningfully.

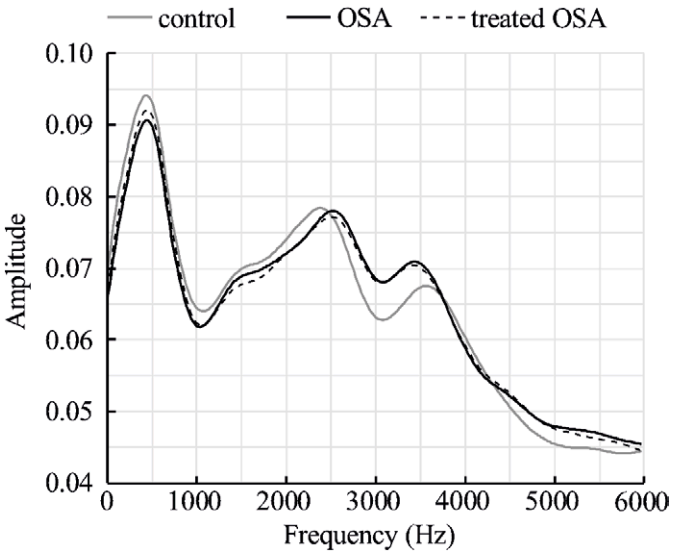


Figure 2: Average LTA spectra for control speakers (continuous, gray line), OSA patients (continuous, black line) and treated OSA patients (dashed, black line) of one Spanish sentence.

Table 1: Sign of the change of the frequencies of formants F1 to F4 when the cross-sectional area of the tube is slightly decreased with regard to the neutral tract shape. The lip and glottal regions are to the left and right respectively.

Upper Airway Regions	bilabial labiodental		dental / alveolar	post-alveolar palatal		velar		oropharyngeal		hypopharyngeal		epiglotto-pharyngeal	epilaryngeal	glottal
Formants	A	B	C	D	**E**	F	G	$\overline{\textbf{G}}$	$\overline{\text{F}}$	$\overline{\text{E}}$	$\overline{\text{D}}$	$\overline{\text{C}}$	$\overline{\textbf{B}}$	$\overline{\text{A}}$
F1	−	−	−	−	−	−	−	+	+	+	+	+	+	+
F2	−	−	−	+	+	+	+	−	−	−	−	+	+	+
F3	−	−	+	+	+	−	−	+	+	−	−	−	+	+
F4	−	+	+	+	−	−	+	−	+	+	−	−	−	+

III. Results

Fig. 2 shows the averages of the LTAS obtained for the control speakers (continuous, gray line), the OSA patients (continuous, black line) and treated OSA patients (dashed, black line). (Palliative) treatment consists in providing continuous positive air pressure during sleep to prevent airway collapse. Treatment does not modify significantly the configurations of the upper airway found in these patients (ca. 1.6 mm increase on average (i.e. 12%) according to [12]).

The average LTAS that are reported have been obtained for one sentence out of four. Results for the other three are similar. The results for sustained vowels are category-dependent and more difficult to interpret.

Fig. 2 evidences differences between OSA and control speakers with regard to the positions of the third and fourth formants, the distance between which is smaller for OSA than for control speakers. Formants F3 and F4 have respectively shifted up and down for the OSA patients. Palliative treatment does not appear to have an influence on the OSA speakers' vocal tract settings, in accordance with its small impact on the anatomy of OSA patients' vocal tracts.

IV. Discussion and Conclusion

The observed differences between OSA and control speakers are best explained in terms of the sensitivity functions of the vocal tract in the vicinity of the neutral vocal tract shape [13]. Table 1 shows the sign of the change of the frequencies of formants F1 to F4 when the area is slightly decreased with regard to the neutral tract shape. Regions labeled A to A-bar within which the signs of formant-specific sensitivity functions are the same actually have unequal lengths. These length differences are not reported in Table 1. One observes three regions the narrowing of which is susceptible to shift formants in agreement with what is observed in Fig. 2.

In Table 1, region B-bar corresponds to the epilarynx the narrowing of which has been linked to the singer's formant. Region G-bar corresponds roughly to the oro-pharynx the narrowing of which is expected for OSA patients. Region E agrees with the palate, to which no role is assigned within the framework of the present study.

One may therefore conclude that the LTAS enables tracking vocal tract settings and that the tract settings and therefore the timbre of OSA speakers tend to differ from the tract settings and timbre of control speakers. The differences are explainable in terms of a pharyngeal narrowing that may be congenital or acquired. Other unexplained evidences reported in the literature for specific speech units extracted from OSA patient records may also be explained by this model (e.g. [4], [14], [15]), reinforcing the interpretation offered here.

V. Tracking settings via principal components analysis (PCA)

An alternative way to track vocal tract settings is by means of a principal components analysis. A spectrogram reports the amplitude of the spectral components as a function of frequency on the vertical axis and time on the horizontal axis. With a view to carrying out principal component analysis and inspecting the first principal component [16], the spectrogram is reinterpreted here as a matrix the rows of which are assigned to indexed

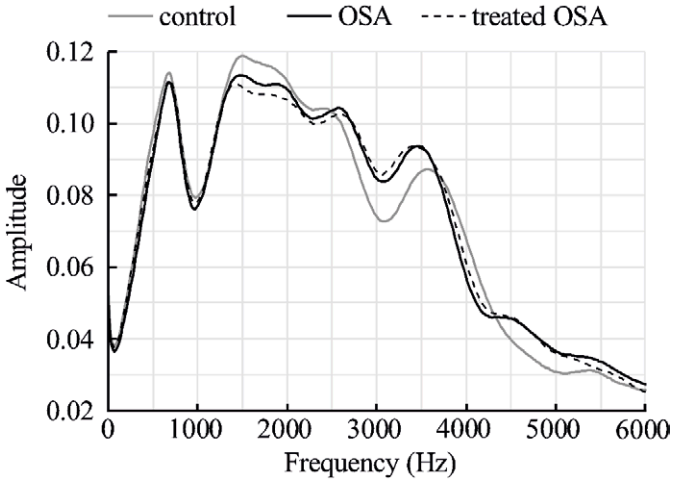


Figure 3: Averaged first principal components for control speakers (continuous, gray line), OSA patients (continuous, black line) and treated OSA patients (dashed, black line) for one single Spanish sentence.

frequency bins and the columns to the index of the analysis frame.

Intuitively speaking, one expects individual amplitude spectra to be a combination of a speaker setting-typical spectrum that is common to all analysis frames, and segment-typical variations that report phonetic identity. When the aggregated segment-typical variations of the spectra are small compared to the setting-typical baseline then the first principal component is expected to report the latter and the higher components the former. The reason is that the mutually uncorrelated principal components are linear combinations of individual amplitude spectra, with the principal components ranked according to the explained variance (i.e. spectral energy).

The feasibility of tract setting analysis via the LTAS suggests that PCA may be suitable for the same task and vice versa. The main difference is that PCA weights analysis frames individually with regard to their phonetic identity, whereas in the LTAS the frame weights are the same.

Fig. 3 shows the averaged first principal components obtained for the control speakers (continuous, gray line), the OSA patients (continuous, black line) and treated OSA patients (dashed, black line). The averaged first principal components are interpretable, similarly to the averages in Figure 2, as spectral contours the F3 and F4 frequencies of which are less distant for OSA speakers than for modal speakers.

ACKNOWLEDGEMENTS

The activities described are partially funded by the Spanish Ministry of Economy and Competitivity as part of the TEC2012-37585-C02 (CMC-V2) Project.

REFERENCES

[1] C. M. Ryan and T. D. Bradley, “Pathogenesis of obstructive sleep apnea,” *J. Appl. Physiol.*, vol. 41, no. 6, pp. 323–330, 2005.

[2] F. J. Puertas, G. Pin, J. M. María, and J. Durán, “Documento de consenso Nacional sobre el síndrome de Apneas-hipopneas del sueño,” *Grupo Español Sueño*, p. 164, 2005.

[3] J. L. Blanco-Murillo, R. Fernández-Pozo, E. López-Gonzalo, and L. A. Hernández-Gómez, “Exploring differences between phonetic classes in Sleep Apnoea Syndrome Patients using automatic speech processing techniques,” *Phon. J. Int. Soc. Phon. Sci.*, vol. 97, pp. 36–55, 2008.

[4] R. Fernández-Pozo, J. L. Blanco-Murillo, L. A. Hernández-Gómez, E. López, J. Alcázar, and D. Torre-Toledano, “Severe Apnoea Detection using Speaker Recognition Techniques.,” in *Proceedings of the BIOSIGNALS Conference*, 2009, pp. 124–130.

[5] J. L. Blanco-Murillo, R. Fernández, D. Díaz, L. Hernández, E. López, and D. Torre, “Apnoea Voice Characterization through Vowel Sounds Analysis using Generative Gaussian Mixture Models,” in *Proceedings of 3rd Advanced Voice Function Assessment International Workshop*, 2009, vol. 1.

[6] J. Laver, *The phonetic description of voice quality*. Cambridge University Press, 1980.

[7] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge University Press, 1983.

[8] R. Fernández-Pozo, L. A. Hernández-Gómez, E. López-Gonzalo, J. Alcázar-Ramírez, G. Portillo, and D. T.Toledano, “Design of a Multimodal Database for Research on Automatic Detection of Severe Apnoea Cases,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, 2008.

[9] M. M. Zhao and X. L. Zhang, “Diagnosis and treatment of obstructive sleep apnea hypopnea syndrome,” *Zhonghua Yi Xue Za Zhi*, vol. 92, no. 18, pp. 1228–1230, May 2012.

[10] E. Keller, “The Analysis of Voice Quality in Speech Processing,” in *Lecture Notes in Computer Science*, 2005, pp. 54–73.

[11] W. J. Hess, “Time-domain digital segmentation of connected natural speech,” in *Proceedings of the 4th International Joint Conference on Artificial intelligence*, 1975, vol. 1, pp. 491–498.

[12] I. L. Mortimore, P. Kochhar, and N. J. Douglas, “Effect of chronic continuous positive airway pressure (CPAP) therapy on upper airway size in patients with sleep apnoea/hypopnoea syndrome,” *Thorax*, vol. 51, no. 2, pp. 190–192, 1996.

[13] M. Mrayati, R. Carré, and B. Guerin, “Distinctive regions and modes: a new theory of speech production,” *Speech Commun.*, vol. 7, no. 3, pp. 257–286, 1988.

[14] J. A. Fiz, J. Morera, J. Abad, A. Belsunces, M. Haro, J. I. Fiz, R. Jane, P. Caminal, and D. Rodenstein, “Acoustic analysis of vowel emission in obstructive sleep apnea,” *Chest*, vol. 104, no. 4, pp. 1093–6, 1993.

[15] M. P. Robb, J. Yates, and E. J. Morgan, “Vocal tract resonance characteristics of adults with obstructive sleep apnea,” *Acta Otolaryngol*, vol. 117, no. 5, pp. 760–3, 1997.

[16] K. Pearson, “On Lines and Planes of Closest Fit to Systems of Points in Space,” *Philos. Mag.*, vol. 1, no. 11, pp. 559–572, 1901.

MODULATIONS OF SPL AND F_0 OCCUR IN SUSPECTED MULTIPLE SCLEROSIS AND INCREASE WITH SEVERITY

E. H. Buder[1], C. Dromey[2], M. Barton[3], M.E. Smith[4], & K. Corbin-Lewis[5]

[1]School of Communication Sciences and Disorders, University of Memphis, Memphis, TN, USA,
[2]Dept. of Communication Disorders, Brigham Young University, Provo, UT, USA,
[3]Dept. of Communication Sciences and Disorders, University of Utah, Salt Lake City, UT, USA,
[4]Dept. of Otolaryngology-Head and Neck Surgery, University of Utah Medical Center, Salt Lake City, UT, USA,
[5]Dept of Communicative Disorders and Deaf Education, University of Utah, Logan, UT, USA
ehbuder@memphis.edu, dromey@byu.edu, michael.barton@hcahealthcare.com,
marshall.smith@hsc.utah.edu, kim.corbin-lewis@usu.edu

Abstract:* **Previous work has shown that the acoustic measures of wow, tremor, and flutter of SPL and F_0 in the sustained vowels of individuals with a diagnosis of *definite* MS can reliably distinguish them from control speakers [1]. The present investigation assesses whether such measures reveal features of phonation in speakers with MS across a range of severities, including *suspected* MS. A total of 79 individuals took part in the study: 35 were diagnosed as having mild MS, 17 as moderate, 15 as severe, and 12 were diagnosed with suspected MS. Participants sustained the vowel /ɑ/. F_0 and amplitude traces extracted from these samples were imported into a Matlab application called the modulogram [2], which produces a graphical display of the extent of modulation of amplitude and frequency as a function of frequency range (wow, tremor, flutter) over time. For each parameter (F_0/SPL) and domain (wow/tremor/flutter), an 'instability' measure is calculated by multiplying the depth of a selected modulation by its duration. Results indicate that, in comparison to normal controls, wow and tremor of SPL were quite sensitive to the MS diagnosis, distinguished levels of severity, and also discriminated suspected MS cases from normal control subjects' phonations.**

Keywords: **Multiple Sclerosis, Vocal Modulations, Fundamental Frequency, Amplitude, Severity**

REFERENCES

[1] Hartelius, L., Buder, E. H., & Strand, E. A. (1997). Long-term phonatory instability in individuals with multiple sclerosis. *Journal of Speech, Language, and Hearing Research, 40,* 1056-1072.

[2] Buder, E. H. & Strand, E. A. (2003). Quantitative and graphic acoustic analysis of phonatory modulations: The modulogram. *Journal of Speech, Language, and Hearing Research, 46,* 475-490.

*Full paper withheld by authors' request.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

DETECTION OF BULBAR ALS USING A COMPREHENSIVE SPEECH ASSESSMENT BATTERY

Y.Yunusova[1], J.S. Rosenthal[2], J.R. Green[3], S. Shellikeri[1], P.Rong[3], J. Wang[4], L. Zinman[5]

[1]Department of Speech-Language Pathology, University of Toronto, Toronto, Canada, yana.yunusova@utoronto.ca
[2]Department of Statistics, University of Toronto, Toronto, Canada, jeff@math.toronto.edu
[3]Department of Communication Sciences and Disorders, MGH Institute for Health Professions, Boston, USA, jgreen2@mghihp.edu
[1]Department of Speech-Language Pathology, University of Toronto, Toronto, Canada, sanjana.shellikeri@mail.utoronto.ca
[3]Department of Communication Sciences and Disorders, MGH Institute for Health Professions, Boston, USA, prong@partners.org
[4]Callier Center for Communication Disorders, University of Texas at Dallas, wangjun@utdallas.edu
[5]ALS/ MN Clinic, Sunnybrook Health Science Centre, Lorne.Zinman@sunnybrook.ca

Abstract: **The study aimed to develop a predictive method that would aid the diagnosis of the bulbar form of Amyotrophic Lateral Sclerosis (ALS) as early as possible, specifically before the onset of obvious clinical signs (e.g., changes in speech intelligibility and speaking rate). Multiple instrumental physiological measures collected across speech subsystems collected longitudinally from over one hundred patients diagnosed with ALS were subjected to multiple analyses. Variable screening was performed using group comparisons with kernel density estimators and with linear regression. Variables identified as showing sensitivity to bulbar ALS onset and progression were used in a linear classifier, which was able to identify individuals who will develop bulbar form of ALS with 80% accuracy. Although preliminary in nature, these results show that instrumental measures might be able to assist in the clinically important early diagnosis of bulbar ALS.**

Keywords: **Bulbar ALS, speech subsystems, speaking rate, intelligibility**

I. INTRODUCTION

Amyotrophic Lateral Sclerosis (ALS) is a devastating neurodegenerative disease with a fast progressing course. There is no biological marker of the condition and the diagnosis is made based on a constellation of clinical observations. As a result, the diagnosis of ALS is significantly delayed. On average, it takes over a year to arrive at the diagnosis [1]. The patients are at risk for multiple referrals to specialists and even for unnecessary surgery [2]. Diagnostic criteria are based on identification of upper and lower motor neuron signs in three regions (i.e., limb, cervical or bulbar muscles) [3]. The bulbar region, defined as musculature involved in speech and swallowing, becomes crucial for timely diagnosis; yet, subtle changes in speech and swallowing might be difficult to detect. The current assessment methods are either perceptual and, thus, characterized by low sensitivity and reliability, or invasive in nature (i.e., needle electromyography is performed on the tongue musculature). There is a clinical need for an assessment protocol that is objective, reliable, and sensitive to early identification of the bulbar form of ALS.

At present, the clinical diagnosis of bulbar ALS is based on the presence of neurological signs of upper and lower motor neuron damage (e.g., atrophy, fasciculations, aberrant reflexes) as well as the system-level measures such as speech intelligibility and speaking rate. Previous research established that speaking rate is more sensitive to disease onset than intelligibility. Speaking rate begins to decline relatively early in the disease until it reaches approximately 120 words per minute (WPM). This cutoff signifies the point in bulbar disease progression when intelligibility begins to decline precipitously [4]. Physiological measures of each bulbar subsystem performance (i.e., respiratory, laryngeal, velopharyngeal, and articulatory) are suggested to show even more sensitivity to disease-related changes in the bulbar mechanism than speaking rate. In the past, each subsystem has typically been studied individually [5,6,7]. Considering substantial heterogeneity of disease presentation and patterns of progression across muscle groups, subsystems, and individuals, the multi-subsystem approach is essential to improve diagnosis.

In this study, we longitudinally assessed the function of each bulbar subsystem with multiple instrumental measures alongside speech intelligibility, speaking rate and ALS-Functional Rating Scale (ALSFRS-R) [8] for a large number of individuals diagnosed with ALS. Based on this dataset, we asked the following questions:

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

(1) Which objective measures within each subsystem distinguish individuals with bulbar ALS from healthy controls?
(2) Which objective measures are sensitive to disease progression over time?
(3) Based on these objective physiological measures, can we predict who will and who will not develop bulbar ALS as disease progresses?

II. Methods

144 individuals (males=89; females=55) diagnosed with ALS took part in the study. The mean age of these participants was 59.6 years (SD=10.3). 53 healthy controls (males=24, females=29) were recruited as well. The mean age was 57.4 (SD=12.6). 34 patients had bulbar onset ALS at the time of diagnosis, the remaining individuals presented with spinal onset ALS, with or without bulbar signs. The participants in the patient group were recorded every three months for the average duration of 22.11 months (SD=16.74). The control participants were recorded once.

The average ALSFRS-R score at the first session was 37.32 (SD=6.32) across all participants. The presence of bulbar ALS was determined by bulbar subscore on ALSFRS-R. At first recording session, the average subscore was 10.42 (SD=1.90). Bulbar performance was also assessed by means of Sentence Intelligibility Test [9] during which individuals were asked to read a series of semantically unpredictable sentences. Speech intelligibility (% words transcribed correctly) and speaking rate (number of words per minute) were determined by a single transcriber unfamiliar with the patients.

The instrumental protocol and measurements are described in detail elsewhere [10, 11]. Briefly, a series of instruments were used to assess the functions of respiratory, laryngeal, velopharyngeal and articulatory subsystems. These instruments included the Phonatory Aerodynamic System (PAS) and Nasometer (KayPentax, MA, USA), acoustic recording equipment (i.e., high fidelity microphone and digital recorder), as well as facial motion (e.g., Optotrak Certus) and tongue motion (Wave) systems (NDI, ON, Canada). Speech tasks included phonation, readings of syllables (e.g., /pa, ma/), words, phrases and paragraphs at normal comfortable speaking rate and loudness.

A large number of measurements per speech subsystem were performed, as those most sensitive to disease onset and progression have not yet been established in the literature. Mean values of multiple repetitions computed by subject/ session composed the final set. The measurements included:

1. Respiratory subsystem – maximum phonation time, minimum and maximum SLP during soft and loud phonation, percent speech and percent pause time, average pause duration, and coefficient of variation of pause duration during paragraph reading, performed using Speech Pause Analysis software (SPA) [12].
2. Laryngeal subsystem – mean fundamental frequency (F0), F0 standard deviation, percent jitter, percent shimmer, F0 maximum, and noise-to-harmonic ratio for a phonated /a/, and laryngeal resistance.
3. Velopharyngeal subsystem – median nasalance scores for a nasal and oral sentence as well as nasalance distance and maximum oral pressure and nasal flow during /pa/ and /ma/.
4. Articulatory subsystem – volume, range, maximum speed and duration of movements of the jaw, lips, and tongue.

For the participants with ALS, all the sessions were subdivided based on the bulbar subscore of ALSFRS-R, speech intelligibility and speaking rate values. The cutoff scores were determined based on the existing literature [4] and the relationship between rate and intelligibility in our sample that showed that the active decline of functional speech was modeled by a linear pattern after speaking rate dropped to 157 and speech intelligibility dropped to 93.

a. The “No-Bulbar” group was composed of sessions without any bulbar signs based on the clinical assessment; they showed ALSFRS-R score =12, rate ≥157, and intelligibility ≥97.
b. The “Bulbar” group was composed of sessions with ALSFRS-R score ≤9, rate≤120, or intelligibility ≤93.
c. The “Early-Bulbar” group was composed of patients with early clinical bulbar signs with criteria outside of those identified in a. and b.

The following analyses were performed. First, each variable was assessed for its sensitivity to the presence of clinically confirmed bulbar disease using kernel density estimators. The kernel density estimators were computed using Gaussian kernel functions, using bandwidths chosen according to Scott's Rule [13]. Such estimators are robust to assumptions about the data such as independence of observations, normality of distribution, etc. [14]. The overlap between density functions served as a metric of similarity between distributions, where the probability of correct guess among observations is equal to one minus half the overlap. We declared the amount of overlap to be significant if it was less than 0.6, corresponding to being able to guess correctly more than 70% of the time.

Second, we examined which variables are sensitive to disease progression using a linear regression of the number of days elapsed from the initial session against the change in each variable. The p values were used to measure the significance of that variable as the bulbar ALS progressed.

Third, a linear classifier was used to derive a predictive score which can be used to determine who will develop bulbar disease over time and who will not, among patients who do not present with the disease initially. Specifically, we used linear regression to find a linear combination of relevant variables which came closest to assigning +1 to patients who will develop bulbar disease, and -1 to patients who will not. This linear combination then gives a predictive score to each new patient, so if their score is positive then we predict that they will develop bulbar disease, while if it is negative then we predict that they will not. Only variables that passed the screening procedures of the two first analyses were used in the classifier.

III. RESULTS

A. Bulbar ALS versus Healthy Controls

In this analysis, we compared healthy controls to ALS data in sessions specified as Bulbar in order to identify measures that are associated with the clinical presentation of bulbar ALS. Fig.1 shows the results of kernel density analysis on the variable Pause Duration. Other physiological measures sensitive to the presence of clinically confirmed bulbar disease included % pause time and number of pauses during paragraph reading, nasal flow during /pi/, laryngeal resistance, maximum speed and duration of opening/ closing cycles of the tongue movement (overlaps ranging between 0.42 and 0.55).

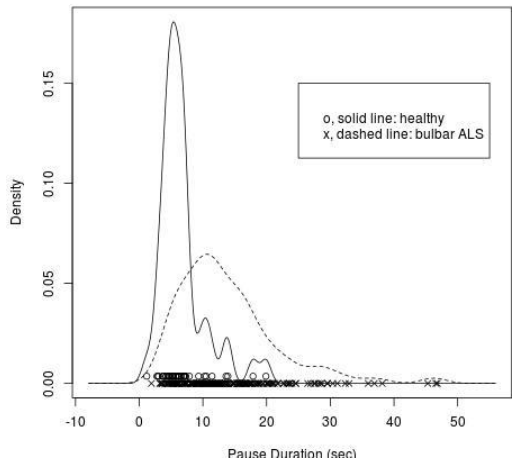


Fig. 1 Kernel density functions for healthy and ALS groups for Pause Duration (overlap=0.46).

B. Change over time

All variables were assessed for their sensitivity to disease progression. For this analysis, only those patients who showed change over time in clinical scores (ALSFRS-R, speaking rate and intelligibility) from the No-Bulbar to either Early-Bulbar or Bulbar groups were selected. 36 out of 67 variables were found to show a statistically significant change with disease progression ($p < 0.05$), including those identified in Step A above.

C. Classification

The five variables with the smallest overlap in step B were selected for the linear classifier to predict who will develop bulbar ALS and who will not, among patients who do not present with the disease. The classifier derived a predictive score given by the following linear combination of the five measures:

$$1.16 + 0.039*\text{PauseDuration} - 22.01*\text{NasalFlowPi} + 0.00316*\text{LarResistance} + 0.00721*\text{TongueMaxSp} - 20.39*\text{DurationOpenJaw} \quad (1)$$

The predictive rule can then be described as follows: Given a new ALS patient who does not currently show clinical bulbar symptoms, if their score of the above linear combination is positive, then we predict that they will develop bulbar symptoms in the future; if it is negative, then we predict that they will not. This predictive rule gives the correct prediction in 80% of the 20 patients for whom we have complete records of all of the required variables.

IV. DISCUSSION

Diagnosing bulbar ALS and predicting disease progression is essential for patient recruitment into clinical trials as well as patient management in a multidisciplinary clinic setting. Approximately 70% of individuals diagnosed with ALS present with spinal signs only (i.e., symptoms associated with arm/ hand/ leg) function. We are exploring the possibility of subclinical bulbar presentation that might be assessed using sensitive instrumental measures of bulbar function. As a result of our preliminary analyses, five variables play an important role in identifying potential early changes associated with bulbar ALS. They include Pause Duration, Nasal Flow during syllable /pi/, laryngeal Resistance, Tongue Maximum Speed and the duration of the Opening-Closing Jaw movement cycle. These variables identified based on their sensitivity to presence of bulbar disease and progression of the disease over time classified individuals into those who will and will not develop bulbar ALS with 80% accuracy.

We are currently continuing to explore other classification approaches (e.g., kernel density, support vector machine [15]) to achieve higher classification accuracy and cross-validating our results on a different data set.

V. Conclusion

Further work is necessary to improve our data reduction and prediction methods, yet current findings provide preliminary indication that we can develop an accurate method for predicting future bulbar symptoms among ALS patients who do not display clinical bulbar signs by virtue of clinical instrumental monitoring.

References

[1] A. Chio, G. Logroscino, and O. Hardiman, "Prognostic factors in ALS: A critical review," *Amyotrophic Lateral Sclerosis,* Vol. 10, pp. 310-323, 2009.

[2] M. Kraemer, M. Buerger, and P. Berlit, "Diagnostic problems and delay of diagnosis in amyotrophic lateral sclerosis," *Clin Neurol Neurosurg,* Vol. 112(2), pp. 103–5, 2010.

[3] B.R. Brooks, R.G. Miller , M. Swash, and T.L. Munsat, "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotroph Lateral Scler Other Motor Neuron Disord,* Vol. 1(5), pp. 293–9, 2010.

[4] L.J. Ball, D. Beukelman, and G.L. Pattee, "Timing of speech deterioration in people with amyotrophic lateral sclerosis," *Journal of Medical Speech-Language Pathology*, Vol. 10, pp. 231–235, 2002.

[5] R. Delorey, H. Leeper, and A. Hudson, "Measures of velopharyngeal functioning in subgroups of individuals with amyotrophic lateral sclerosis," *Journal of Medical Speech-Language Pathology*, Vol. 7, pp. 19-31, 1999.

[6] R.D. Kent, J.F. Kent, G. Weismer, R.L. Sufit, JC Rosenbek, and RE Martin, "Impairment of speech intelligibility in men with amyotrophic lateral sclerosis," *Journal of Speech and Hearing Disorders*, Vol. 55, pp. 721–728, 1990.

[7] L.O. Ramig, R.C. Scherer E.R. Klasner, I.R. Titze, and Y Horii, "Acoustic analysis of voice in amyotrophic lateral sclerosis: A longitudinal case study," *Journal of Speech and Hearing Disorders*, Vol. 55, pp. 2–14, 1990.

[8] J.M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakamishi, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences,* Vol. 2, pp. 13-21, 1999.

[9] K. Yorkston, D. Beukelman, and M. Haken, "Speech Intelligibility Test", 1996.

[10] Y. Yunusova, J.R. Green, J. Wang, G. Pattee, and L. Zinman, "A protocol for comprehensive assessment of bulbar dysfunction in ALS," *J. Vis. Exp*, Vol. 48, e2422, 2011.

[11] J. Green, Y. Yunusova, M.S. Kuruvilla, J. Wang, G. Pattee, L. Synhorst, L. Zinman, and J. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions,",*Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, Vol.* 7, 2013.

[12] J.R. Green, D.R. Beukelman, and L.J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech,", *J. Med. Speech Lang. Pathol*, Vol. 12, pp. 149-154, 2004.

[13] D.W. Scott, "Multivariate Density Estimation: Theory, Practice, and Visualization", John Wiley & Sons, New York, 1992, pp. 1-317.

[14] P. Hall, S.N. Lahiri, and Y.K. Truong, "On bandwidth choice for density estimation with dependent data,". *The Annals of Statistics*, Vol. 23, No. 6, pp. 2241-2263, 1995.

[15] B.E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," *The Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152, 1992.

ACOUSTIC AND ARTICULATORY VARIATION IN THE MID-CENTRAL VOWEL IN APRAXIC AND NORMAL SPEECH

C. M. Menezes

University of Toledo/Dept. of Health and Rehabilitation Sciences, Toledo, Ohio, U.S.A, caroline.menezes@utoledo.edu

Abstract: **This study looks at the production of the mid-central vowel produced in different prosodic conditions by a speaker with apraxia and compares it to a normal speaker matched for age and gender. Results reveal that this vowel is emphasized in apraxic speech consistent with it being longer in duration and produced with an inferior-posterior tongue position when compared to normal speech. However, the jaw movement is reduced when compared to the normal speaker.**

Keywords : **apraxia, schwa, articulation, vowel quality**

I. INTRODUCTION

Apraxia of speech is a neurological condition caused by bilateral lesions to the frontal lobe in the vicinity of the perisylvian region of the brain affecting the programming and sequencing of speech movements [1]. The speech of a person with apraxia is marked by hesitations, and groping behaviors not resulting from muscle weakness, slowness and incoordination [2]. The salient speech errors include slow rate of speech, incoordination of laryngeal and supralaryngeal articulators [3]; co-articulation and sound metathesis [4], predominance of articulatory substitutions, and difficulty initiating speech [5]. Apraxic speech is however, best characterized by inconsistencies in error production, for example, they might say the word "bad" as bad but at other times it might be "tad" or "sad".

The research on vowel acoustics of speech apraxia have revealed mixed results. Some studies show minimal deviations in vowel quality [3; 6; 7] while another reveals a reduction in the vowel space [8]. Reduction of the vowel space reflects a centralization of the vocal articulation [9]. On the other hand, phonological differences in duration such as intrinsic and contrastive durations are maintained in apraxic speech [10]. Research on formant frequencies and durations of vowels tend to indicate that poor intelligibility of apraxic speech is not directly related to vowel acoustics.

Studies on the effect of vowel space on speech intelligibility focus on the corner vowels of the vowel quadrilateral. However, no studies have actually reported on variations of the unstressed vowel occupying the center of the vowel quadrilateral; the "schwa". The mid-central "schwa" in English is a short neutral vowel whose quality varies greatly depending on the phonetic environment in which it exists. It can be an epenthetic vowel, a reduced vowel, a rhotic vowel and an unstressed vowel. The rhythmic quality of English is also determined by the alternation of strong and weak syllables. As a reduced/unstressed vowel the schwa plays a crucial role in the actuation of English rhythm. Therefore, errors in articulating the schwa can affect speech intelligibility both at the segmental and suprasegmental levels. Moreover, if the vowel space is reduced in speech apraxia it is not clear if the reduction is towards the central vowel as would be expected if the central vowel as represented by the uniform tube closed at one end and open at the other that has been used to model the vocal tract.

This paper is interested in studying the articulatory variability in the production of the mid-central schwa in apraxic speech in the English definite article "the" produced in sentence initial and medial positions. Vowel duration will be compared across apraxic and normal speech. We will also study vowel quality differences by studying the variation in first and second formants. Furthermore, we will study the vertical displacement of the jaw since it has already been reported that the first formant (F1) values correlate with jaw height [11].

II. METHODOS

Subjects: Two subjects participated in this study. Subject one a 60-year old female was diagnosed with severe apraxia five years ago following an aneurysm in the frontal lobe of the left hemisphere. The subject with apraxia has been receiving speech therapy for approximately four years. She was selected from a small pool of clients with brain injuries that receive speech therapy at the University of Toledo. Another 60-year old female with no neurological diagnosis was selected as the control for this study. Both subjects come from the same region and speak standard mid-western American English.

Stimuli: The study focuses on the mid-central American English vowel commonly occurring in the English article "the". The vowel in this word is sometimes emphasized in American English as /i/ but in general it is pronounced with the central neutral vowel /ə/. The allophonic variations of this vowel are studied by inserting the target word in different phrase positions. Research has shown that sentence position affects both

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

the production and perception of stress/emphasis syllables [11; 12]. Syllables in phrase initial position are produced with greater magnitude when compared to other syllables in the utterance if spoken as a neutral declarative utterance without emphasis [13; 14; 15]. This phenomenon is known as "phrase initial strengthening". Syllables at the end of an utterance on the other hand experience the phenomenon of "phrase final lengthening" [15]. By using the principles of phrasal prosody we can affect allophonic variations in the quality of this central vowel. An example of the stimuli set is provided below:

The sad American story. Phrase-initial-pre-target-boundary

The SAD American story. Phrase-initial-post-target-boundary

AMERICA the sad story. Phrase-medial-pre-target-boundary

America, the SAD story. Phrase-medial-post-target-boundary

Words indicated in capitals were emphasized to change the phrasing pattern of the utterances. Speech of people with neurological lesion (reported in dysarthria) have been found to contain a high percentage of spirantization of stops and weak consonant constrictions [16] and therefore, we use utterances comprising mostly of +sibilants. There were two types of sentences produced in four different phrasing pattern (as listed above) repeated five times each by each subject (80 items in total).

Procedure: Articulatory data was collected using the AG500 3-D ElectroMagnetoArticulatograph (3-D EMA) at a sampling rate of 200Hz. Acoustic data was recorded directly to the computer at a sampling rate of 16KHz. A computer placed directly in the view of the subjects presented the speech stimuli. Subjects read the list of utterances five times in a random order.

Analysis: To test the variations in the acoustic quality of the schwa the vowel in the target word "the" was analyzed. Duration was measured from the second glottal pulse at the beginning of voiced segment to the penultimate glottal pulse at the end of voicing. There were instances especially in normal speech when the vowel in the target word was extremely coarticulated and devoiced. Target words that did not maintain the phonemic quality of the mid-central vowel were excluded from the analysis. Errors also occurred in the articulatory data collection of the apraxia subject which resulted in us attaching the coil on the mandible after the experiment had started. This also reduced the number of items that were finally analyzed. In total there were 31 utterances in apraxia speech and 27 utterances in normal speech were analyzed. Since the aim of the study was to understand the quality of central vowel the first and second formant values were also measured at the midpoint of the vocalic segment where the formants were relatively static. All acoustic measurements were made using the PRAAT software for phonetic analysis. Articulatory measurements were obtained using the software program Visartico (v 0.9.1). In a preliminary analysis the maximum deviation of the mandible from the occlusal plane was measured. Tongue dynamics are not reported here because all coils placed on the tongue of the apraxic subject fell off at different stages of the analysis due to extreme salivary secretion.

III. RESULTS

A. Vowel Duration

Fig. 1 plots the difference in mean vowel durations for *apraxia* and *normal* speech for the different emphasis condition. For simplicity the different emphasis is labeled by the sample sentences (see above). In this graph we see that the *apraxia* vowel was longer in duration when

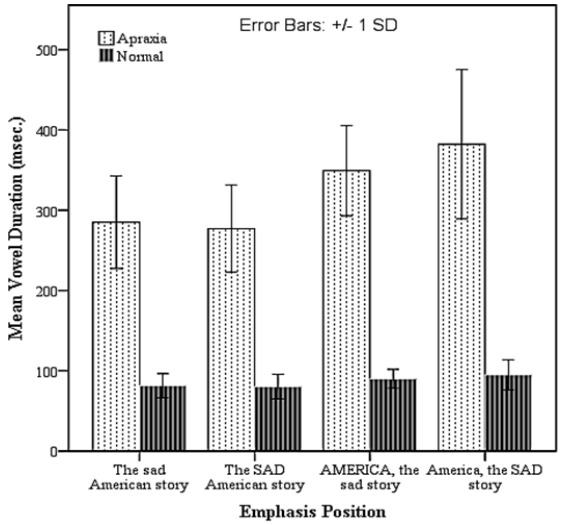


Figure 1: Mean vowel duration (msec.) separated for *apraxia* and *normal* speech for different emphasis types.

compared to *normal* speech. An independent t-Test analysis showed vowel duration between *apraxia* (N = 31, M = 313.81, SD = 74.57) and *normal* speech (N = 27, M = 76.96, SD = 31.64) was statistically significant t (41) = 16.1, p = 0.001, 95% CI [207.15, 266.55], d = 236.85. These results confirm the findings of earlier studies [10, 3, 17]. No significant differences were found among the different emphasis conditions in *normal* speech but in apraxia we see two homogenous subsets. Post-hoc Tukey test reveal significant differences between the penultimate emphasis position and the medial (p = .020) and neutral emphasis (p = .043) position. Looking at Fig. 1, we see that the vowel is longer in the more complex utterance structure (non-restrictive clause). And it is significantly longer when the target is followed by emphasis. Therefore, the apraxia subject elongates the vowel in complex utterance conditions which appears to be uncharacteristic for the *normal* speaker.

B. Vowel Quality

Fig. 2 plots F1 values, with *apraxia* values represented at the top of the graph and *normal* values at the bottom. The terms "lower" and "higher" indicate jaw/tongue position.

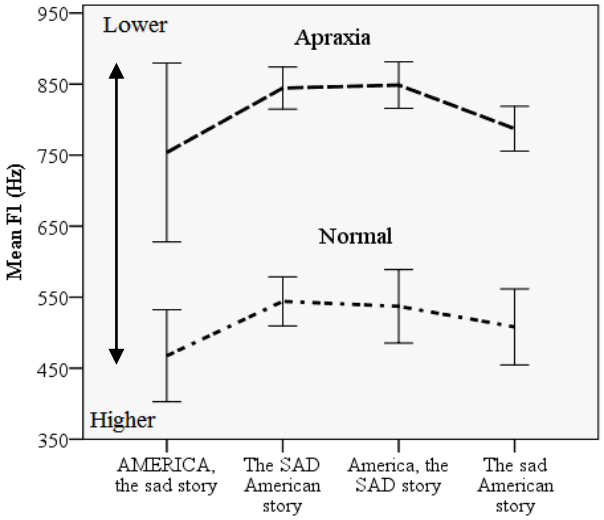


Figure 2: Mean F1 values (Hz) separated for *apraxia* and *normal* speech for different emphasis types.

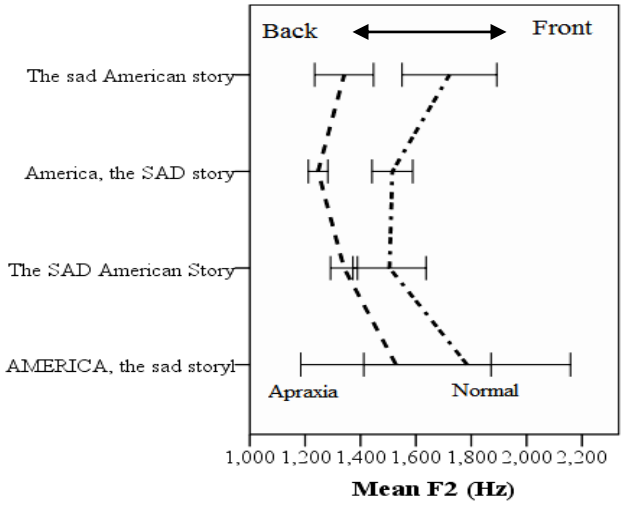


Figure 3: Mean F2 values (Hz) separated for apraxia and normal speech for different emphasis types.

The two speaker conditions were significantly different in F1 values with *apraxia* revealing higher values than *normal*. Independent t-Test analysis showed F1 values between *apraxia* (N = 31, M = 806.74, SD = 103) and *normal* speech (N = 27, M = 511.3, SD = 67) was statistically significant t (52) = 13.1, p = 0.001, 95% CI [250, 340.7], d = 295. High F1 values correlate with lower jaw/tongue positions [see 11 relative to jaw movements]. There were no significant differences between emphasis conditions for both *apraxia* and *normal*. However, in both speakers F1 values where higher for the central vowel when the word following it was emphasized indicating an anticipatory spreading of emphasis to the unstressed vowel for both the *apraxia* and *normal* subjects.

Fig. 3 plots F2 values for *apraxia* and *normal* speech. It is flipped to represent the front-back movement of the tongue as it is now undisputed that the forward-backward movement of the tongue influences F2 values. Again we see that the two speakers have clearly distinct articulatory position. Independent t-Test analysis showed F2 values between *apraxia* (N = 31, M = 1376.9, SD = 271.3) and *normal* speech (N = 27, M = 1648.56, SD = 248.2) was statistically significant t (55) = -3.98, p = 0.001, 95% CI [-408, -134.96], d = -271.65. No significant difference was seen between emphasis conditions for both speakers. However, similar to F1, F2 values in the target vowel were influenced by emphasis. F2 values were higher for both speakers when the schwa was followed by emphasis. In general, when compared to *normal* speech, *apraxic* speech was marked by a more backward directed tongue, which interestingly was the position of the tongue when the vowel was followed by emphasis for both speakers.

C. Jaw Displacement

Fig. 4 plots the displacement of the mandible from the occlusal plane (z-axis of 3-D EMA). Large values here indicate larger vertical displacement of the jaw or lower

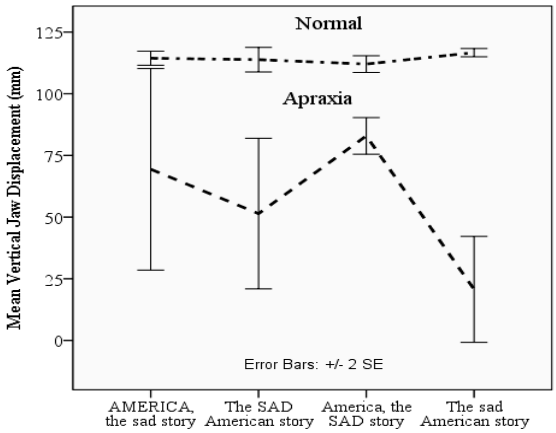


Figure 4: Mean jaw displacement (z-axis) separated for apraxia and normal speech for different emphasis types.

jaw position. *Normal* speech has significantly lower jaw position when compared to *apraxia* speech. Independent t-Test analysis showed jaw displacement values between *apraxia* (N = 31, M = 49.65, SD = 8) and *normal* speech (N = 27, M = 113.58, SD = 4.69) was statistically significant t (30.8) = -7.84, p = 0.001, 95% CI [-80.57, -47.28], d = -63.93. While the jaw is relatively stable for the normal speaker through all emphasis conditions, we see large variations for apraxia within each emphasis

condition and across emphasis condition. The larger jaw movements again occur on the more complex phrasing structure. No significant variations were seen for the different conditions in apraxia due to the larger variation within conditions.

IV. Discussion and Conclusion

The unstressed schwa vowel in this study exhibited several variations depending on the clinical condition of the speaker and utterance phrasing (manipulated here through emphasis) while still maintaining it's phonemic identity. Vowel duration was longer in apraxia than normal speaker consistent with previous findings [3, 10, 17], however, we find effects of phrasing only for apraxic speech.

Vowel quality of the mid-central vowel is also significantly variable in the F1 and F2 values. The apraxic vowel is higher in F1 and lower in F2 than normal schwa. This would translate to a lower jaw and a posterior tongue position in apraxia which is not an open/neutral vocal tract nor is it indicative of vowel centralization. However, in this study the jaw kinematic data does not support traditional hypothesis. The normal vowel has significantly lower jaw position than the apraxic vowel. This result is not consistent with prior research results [11]. Since we did not analyze the tongue movement it is not clear if there is an articulator to articulator compensatory movement. It is quite possible that the jaw is held relatively stable to support a rather ballistic tongue. F1 and F2 values are more exaggerated in apraxia but they follow the normal phrasing pattern. Generally, in complex utterances like the non-restrictive clause structures in English the unstressed vowel is emphasized both in duration and vowel quality. Furthermore, there is an anticipatory co-articulation of emphasis in the conditions where the emphasis is contiguous with the unstressed "the" in these sentences.

This preliminary study reveals that the mid-central vowel schwa has a stressed cognate which is statistically significant in apraxia but also exists in normal speech. Further analysis is required to test if these results are true for different levels of apraxia and for larger numbers of normal speakers.

References

[1] Darley, F. L. (1969). Aphasia: Input and output disturbances in speech and language processing. Paper presented at the annual meeting of the American Speech-Language-Hearing Association, Chicago.

[2] Darley, F.L., Aronson, A.E., & Brown, J.R. (1975). Motor Speech disorders. Philadelphia, W.B. Saunders.

[3] Kent,R . D. and Rosenbeck J. C. (1983). Acoustic patterns of apraxia of speech. *Journal of Speech and Hearing Research*, 26,23 1-248.

[4] MlCoch, A. G., & Noll, J. D. (1980). Speech production models as related to the concept of apraxia of speech. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 4). New York: Academic Press.

[5] Johns, D. F., & Lapointe, L. L. (1976) Neurogenic disorders of output processing: Apraxia of speech. In H. Whitaker & H. A.Whitaker (Eds.), *Studies in neurolinguistics* (Vol. 1). NewYork: Academic Press.

[6] Keller, E. (1975). Vowel errors in aphasia. *Dissertation Abstracts International*, 38(10), 6100A. (UMI No. NK32923)

[7] Jacks, A., Mathes, K. A. & Marquardt, T. (2010). Vowel acoustics in adults with apraxia of speech. *J. of Speech, Language, and Hearing Research*, v. 53, 61-74.

[8] Haley, K. L., Ohde, R. N., & Wertz, R. T. (2001). Vowel quality in aphasia and apraxia speech: Phonetic transcription and formant analyses. *Aphasiology*, 15, 1107–1123.

[9] Weismer, G. and Martin, R. E. (1992) Acoustic and perceptual approaches to the study of intelligibility. In Kent. R. D. (Ed.), *Intelligibility with Speech Disorders: Theory, Measurement and Management* (Amsterdam: John Benjamins).

[10] Baum, S., Blumstein, S., Naeser, M., & Palumbo, C. (1990). Temporal dimensions of consonant and vowel production: An acoustic and CT scan analysis of aphasic speech. *Brain and Language*, 39, 33–56.

[11] Menezes, C. (2003). *Rhythmic pattern of American English: An articulatory & Acoustic study*. PhD. Dissertation, Dept. of Speech and Hearing Sciences,The Ohio State University. (Dissertation)

[12] Zheng, X. and Pierrehumber, J. B. (2010). Effects of prosodic prominence and serial position on duration perception. *JASA* 128 (2), pp. 851-859.

[13] C. Fougeron and P. A. Keating (1997). Articulatory strengthening at edges of prosodic domains. *JASA* 101, 3728-3740.

[14] P. Keating, T. Cho, C. Fougeron, and C. Hsu. (2003). Domain-initial articulatory strengthening in four languages. *In Phonetic Interpretation*. Papers in Laboratory Phonology 6, edited J. Local, R. Ogden, R. Temple, Cambridge University Press, pp. 143-161.

[15] Byrd, D.(2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, 57, 3–16.

[16] Polczynska, M., Tobin, Y. and Sapir, S. (2009). Correlation between vowel centralization and incomplete stop articulation in individuals with traumatic brain injury. *Poznan Studies in Contemporary Linguistics* 45(2), pp. 281-299.

[17] Rogers, M. A. (1997). The vowel lengthening exaggeration effect in speakers with apraxia of speech: compensation, artifact or primary deficit. *Aphasiology*, 1997, vol. 11, n 4/5, pp 433-445.

Session VII:
VOICE AND STRESS/DEPRESSION

EXAMINATION OF SEGMENTAL AND SUPRA-SEGMENTAL PARAMETERS OF DEPRESSED SPEECH

Klára Vicsi, Dávid Sztahó, Fodor Tamás

Department of Telecommunication and Media Informatics of Budapest University of Technology and Economics
vicsi@tmit.bme.hu, sztaho@tmit.bme.hu, fota6666@gmail.com

Abstract: **In this paper acoustic-phonetic analysis was done in order to identify differences in speech production of healthy and depressed people. Read and continuous speech material was gathered from patients diagnosed with different degree of depression. In this study only the read speech was analysed. A reference speech database was used as healthy speech for comparison. It was found that segmental parameters: fundamental frequency, F1, F2 formant frequencies, jitter, shimmer; and supra-segmental parameters: number of phonemes of a fix text, speech rate, length of pauses, intensity and fundamental frequency dynamics and spectral slope in the speech of depressed people shows significant changes compared to a healthy reference group.**

***Keywords*: speech analysis, depressed speech, pathological speech production**

I. INTRODUCTION

According to psychology, depression is a mental state caused by one or more experience of failure, which state has not only emotional symptoms but cognitive, physical and motivational symptoms too. These symptoms can also be observed in speech.

Physicians often use faded, slow, monotonous, lifeless and metallic words as properties of depressed speech. Researches assume that these perceptual properties can be linked to acoustic parameters, such as fundamental frequency, amplitude modulation, formant structure, energy distribution, etc. Proper parameters of speech can indicate depression and suicidal tendency. The topic has been examined for decades and numerous studies have identified acoustic features that can be linked to depression. Some of these studies measure differences between voice of healthy and depressed people, others investigate follow-up monitoring to gather features with high classification performance.

It is proven that psycho motoric disorders are the earliest and most stable indicators of mood disorders. Differences can be observed in motoric behaviours, body movements, speech and delay in motoric answers. These studies examined parameters of speech as psycho motoric symptoms of depression and suicidal tendency.

Determined parameters can be used as tools of differential diagnosis [1]. An early study measured fundamental frequency change in patients [7]. This early study contains experiments with three patients only but it had already identified one of the most important acoustic features of depressed speech.

Nowadays many acoustical, phonetical parameters are investigated, at different levels of speech production: fundamental frequency, variation of fundamental frequencies, formants, power spectral density [1], MFC coefficients [4] or cepstrum [5], speech rate [2], amplitude modulation and other different prosodic parameters [9].

Glottal features were examined in [8]. Glottal features could be very important, because they give information about speech production of voiced sounds, but their measurement method is complex. Exact values are only predicted based on resonance cavity models.

In [2] patients under treatment global features are examined in order to observe correlation between the improvement of the patients' mental state and the measured acoustic parameters. Researchers started treatment of thirty five patients parallel to the work, and analysed the change in their speech production according how they responded to the treatment. In the case of patients with successful therapy, a higher variation of pitch was found along with shortening of pauses in speech, and a higher speech rate. On the other side in the case of patients with no improvement, these changings of parameters were not observed.

In [1] study was reported from another viewpoint. Patients were categorized into healthy, depressed and highly suicidal categories. The following acoustic parameters were measured: fundamental frequency, amplitude modulation, formants, power spectral density. It was found that in the case of women amplitude modulation and fundamental frequency did not show any results, but formant frequencies enabled successful categorization between the former classes. In the case of men, a lower pitch and higher formants were found at depressed and highly suicidal patients. At highly suicidal patients a rise in amplitude modulation was also observed. These parameters could classify sound samples into the three proper categories.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

Lower first formant frequencies were measured in speech of depressed patients in [3]. These seem to contradict to [1], where broadly higher formants were found. This implies that there are many more researches to be done in the topic in order to clarify the results that somehow did not follow a consistent tendency.
Our goal is to identify those acoustic-phonetic parameters, separately in segmental and supra-segmental level, that can characterize the speech of depressed people. This paper is only a preliminary report of a long-term project. We could analyse only the voice of 21 patients, with different degree of depression, but we obtained very promising results.

II. METHODS

Database: The patients were selected with the help of a psychiatrist from the Neurology Department of Semmelweis University, Hungary. 21 native Hungarian patients diagnosed with depression were involved in the research, 11 women and 10 men. Beck Depression Inventory (BDI) score was used to classify the recordings into the following categories: 0-13: minimal degree of depression; 14-19: mild depression; 20-28: moderate depression; 29-63: severe depression.

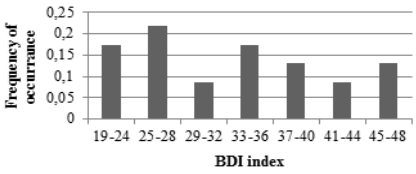


Figure 1: *Distribution of BDI indices of patients in the database.*

The distribution of BDI indices of patients are shown in Fig. 1. The recordings were done at the psychiatry and the text consists of two parts. In the first part the patients were having a free conversation with the therapist. The length of this conversation was about five minutes. In the second part patients read a standard phonetically balanced short folk tale (about 6 sentences all together), frequently used in the phoniatry practice for all European languages, "The North Wind and the Sun".
Recording conditions were the followings. clip-on microphones were used (Audio-Technica ATR3350), with external USB sound card with 44100Hz, at a 16 kHz sampling rate, quantized at 16 bits. The intensity of the recordings was normalized according to maximum amplitude peek.
For reference database 21 healthy speakers (10 male and 11 female) were asked to read also the tale "The North Wind and the Sun".
The recordings were annotated and segmented on phoneme level, using SAMPA phonetic alphabet, with the help of automatic phoneme segmentator developed in our Laboratory. Then manual correction was done.
Due to the restriction of the read speech, from the depression database only the second part was chosen for the analysis. This allowed the examination of the prosodic features, such as length of pauses, length of total speech and number of phonemes.
For each patient the age and gender were noted with additional information about smoking status, illnesses and taken medicine.
Acoustic-phonetic features: The acoustic-phonetic features were examined in two groups: segmental and supra-segmental (prosodic) features. The segmental features were calculated at the middle of all 'E' vowels of the read part of the recording (folk tale). The segmental features were the followings: fundamental frequency of all 'E' vowels (F0),first and second formant frequency of all 'E' vowels (F, F2),jitter (J),shimmer (Sh). For the calculation of formants, fundamental frequency and spectral values a Hamming window was used with 25 ms frame size and these features were computed from the middle of the vowels. The spectral slope was calculated from the mel-scale domain using linear regression.
The supra-segmental (prosodic) features were computed from the total length of the read part of the recording (folk tale). The following features were calculated as prosodic features:
volume dynamics of speech (range of intensity) (Ra_I), fundamental frequency dynamics of speech (range of fundamental frequency) (Ra_F0), total length of pauses (P_length), total length of recording (R_length), ratio of total length of pauses and the total length of recording (R_p_r), number of phonemes (#P), speech rate (SR), articulation rate (AR), ratio of pauses and voiced phonemes (R_p_vph), ratio of voiced and unvoiced phonemes (R_v_uvp).
For the calculation of intensity and fundamental frequency a 100 ms frame size was selected. For the computation of parameters listed the Praat [6] software was used.

III. RESULTS

The values of all the segmental parameters are shown in Table 1 and the values of the supra-segmental parameters are presented in Table 2. The lower fundamental frequency is a common phenomenon in the earlier studies [10], [11], [12]. In our work significantly lower average pitch was found in depressed speech, congruent to the earlier studies.
A lower pitch was reported in [2] in the case of men, however not in the speech of women. In Table 1 it is clearly shown, that both depressed women and men have lower pitch average values.

Table 1: *Results of the measured segmental acoustic features end the significance level.*

Feature	Gender	Group	Mean	Standard deviation	Significance level
F0 [Hz]	Women	Depressed	154	24	99.9%
		Normal	199	33	
	Men	Depressed	101	14	95%
		Normal	116	19	
F1 [Hz]	Women	Depressed	613	74	99.5%
		Normal	695	47	
	Men	Depressed	512	56	<90%
		Normal	531	46	
F2 [Hz]	Women	Depressed	1764	112	99.9%
		Normal	1955	87	
	Men	Depressed	1565	83	99%
		Normal	1672	80	
J [%]	Women	Depressed	3.4	4.2	95%
		Normal	1.9	2.3	
	Men	Depressed	5.3	4.7	95%
		Normal	1.7	2.7	
Sh [%]	Women	Depressed	14.1	8.8	95%
		Normal	8.6	7.7	
	Men	Depressed	17	9.3	95%
		Normal	9.6	6.2	

In [2] a difference in second formant frequencies were also reported, but the tendency was that depressed people had higher formant frequencies. In our work we experienced the contrary. Significantly lower F1 and F2 formant frequencies were observed congruent to the results published in [1]. The contradiction between the two results could be due to language characteristics or the wrong conclusion because of the low sample number in the database.

In the case of measured jitter and shimmer of all 'E' vowels, a tendency of higher values were observed in the voice of depressed patients. This can be due to the uncertainty in speech production. However the standard deviations also increased.

The range of the intensity in the recordings shows that depressed patients speak with less variance in their loudness. This can be caused by the monotonous, faded voice of the depressed people, which is usually reported by physicians.

The same tendency occurs in the case of fundamental frequency range. These parameters reflect that these patients speak with less emphasis, and can be important features in a classification task.

In accordance with other studies in the topic, we have found that the length of the pauses shows a large difference between the reference group and depressed patients. During the reading of the tale the patients kept more pauses, which can be caused by psycho-motoric disorders. The longer pauses resulted longer recordings, which can also be seen from the data.

It is interesting to see that the number of phonemes is higher in the case of depressed speech. At first this can be confusing due to read speech, but the hesitations, the restarts in the speech not only cause longer total recording time, but a higher phoneme number.

Table 2: *Results of the measured prosodic features and the results of T-probe, and significance level*

Feature	Gender	Group	Mean	Standard deviation	Significance level
Ra_I [dB]	Women	Depressed	32.6	4.1	99.9%
		Normal	44.6	6.3	
	Men	Depressed	34.5	5.2	99.9%
		Normal	44.9	3.6	
Ra_F0 [Hz]	Women	Depressed	49.8	12.5	95%
		Normal	61.4	14.3	
	Men	Depressed	32.5	11.4	97.5%
		Normal	47.6	17.8	
P_length [sec]	Women	Depressed	9.2	3.3	95%
		Normal	6.6	3.1	
	Men	Depressed	12.2	5.6	95%
		Normal	7.6	4.2	
R_length [sec]	Women	Depressed	50.2	7.9	97.5%
		Normal	43.5	6.8	
	Men	Depressed	53.5	12.0	95%
		Normal	44.0	10.1	
R_p_r [%]	Women	Depressed	19.0	7.5	90%
		Normal	14.7	5.0	
	Men	Depressed	22.1	7.4	95%
		Normal	16.5	4.5	
#P	Women	Depressed	464.4	24.1	99.9%
		Normal	434.1	7.4	
	Men	Depressed	473.6	20.2	99.5%
		Normal	444.7	13.4	
SR [unit/sec]	Women	Depressed	9.5	1.7	<90%
		Normal	10.2	1.7	
	Men	Depressed	9.2	1.9	90%
		Normal	10.5	1.9	
AR [unit/sec]	Women	Depressed	11.9	3.0	<90%
		Normal	12.0	1.7	
	Men	Depressed	11.8	2.0	<90%
		Normal	12.5	1.9	
R_p_vph [%]	Women	Depressed	34.5	12.7	<90%
		Normal	29.3	16.5%	
	Men	Depressed	56.0	33.9	90%
		Normal	39.2	16.7	
R_v_uvp [%]	Women	Depressed	261.9	155.2	<90%
		Normal	211.2	89.8	
	Men	Depressed	149.9	82.8	<90%
		Normal	126.6	46.6	

Significance tests were carried out in order to examine the effectiveness of the separation according to the computed features. The results of T-probe are marked on the Table 1 and Table 2. The significance levels show, that among the segmental parameters the decrease of the fundamental frequency and the F1, F2 formant frequencies are the most characteristics and among the supra-segmental parameters, the decrease of the volume dynamics of speech (range of intensity) and the increasing number of phonemes in a fix read text are the most important changes in the case of depressed speech.

In Table 3 a summary can be found, where the direction of the changes of the acoustical parameters are presented. It shows the tendency how values are changed in the case of depressed speech.

IV. Discussion and Conclusion

The reviewed studies represent preliminary investigations of the acoustic-phonetic properties of

speech collected from 21 native patients having different degree of depression.

Table 3: *Tendency of change in values in case of depressed speech.*

Segmental features	Tendency in depressed speech	Supra-segmental features	Tendency in depressed speech
F0	-	Ra_I	-
F1	-	Ra_F0	-
F2	-	P_length	+
J	+	R_length	+
Sh	+	R_p_r	+
SS	-	#P	+
		SR	-
		AR	-
		R_p_vph	+
		R_v_uvp	+

These depressed speech samples and speech of healthy people of the Hungarian Reference Database were compared. We successfully identified some segmental and supra-segmental features from continuously read speech that can show significant differences between the speech of depressed patients and a healthy reference group. We have found that segmental parameters: fundamental frequency, F1, F2 formants frequencies, jitter, shimmer; and supra-segmental parameters: number of phonemes, speech rate, length of pauses, intensity and fundamental frequency dynamics in the speech of depressed people shows significant changes compared to a healthy reference group.

The examined number of depressed patients is small according to the wide range of degree of the depression. But the database is under continuous expansion with more and more recordings with further patients. A proper number of sound samples will allow us to perform a full analysis, and thus we can select a complete set of acoustic features that enables more precise conclusions to deduct.

The ultimate goal would be to found a clear correlation between the severity of depression and the change of the acoustic-phonetic parameters.

Until now standard read speech (the folk tale) was used for the comparison, but a free conversation with the therapist have also been recorded and stored in the depressed speech database for further analysis. It is assumed that some acoustic-phonetic parameters will behave differently than in read speech. It remains to be examined too.

V. ACKNOWLEDGEMENTS

Persons for the speech database were selected by Dr. Lajos Simon psychiatrist from the Neurology Department of Semmelweis University, Hungary. We want to tell many thanks for his help.

The authors would like to thank the COALA project: Psychological Status Monitoring by Computerised Analysis of Language phenomena (COALA) (AO-11-Concordia). Moreover this work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013) organized by VIKING Zrt, Balatonfüred.

REFERENCES

[1] Daniel J. at all: “Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk”, *IEEE Transactions On Biomedical Engineering,* VOL. 47, NO. 7, 2000.

[2] James C. at all:“Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology”, *J Neurolinguistics*, 2007.

[3] Nicholas C. at all:“An Investigation of Depressed Speech Detection: Features and Normalization”, *INTERSPEECH 2011*, 2997-3000. 2011.

[4] Terapong B. at all:“Assessment of Vocal Correlates of Clinical Depression in Female Subjects with Probabilistic Mixture Modeling of Speech Cepstrum”, *2011 11th International Conference on Control, Automation and Systems* Oct. 26-29, 2011.

[5] Thaweesak Y. at all:“Characterizing Sub-Band Spectral Entropy Based Acoustics as Assessment of Vocal Correlate of Depression”, *International Conference on Control, Automation and Systems* 2010 Oct. 27-30, 2010.

[6] Boersma, P:“Praat, a system for doing phonetics by computer”. *Glot International* 5:9/10,341-345. 2001.

[7] Askenfelt, A.at all:“Voice analysis in depressed patients: Rate of change of fundamental frequency related to mental state”, *Dept. for Speech, Music and Hearing, Quarterly Progress and Status Report* Vol. 21, No. 2. pp. 71-84. 1980.

[8] Elliot M.at all:“Investigating the Role of Glottal Features in Classifying Clinical Depression”. *Proceedings of the 25th Annual International Conference of the IEEE*. pp. 2849-2852. 2003.

[9] Michelle Hewlett Sanchez, at all:“Using Prosodic and Spectral Features in Detecting Depression in Elderly Males”, *INTERSPEECH 2011,* Florence, Italy, August 27-31, 2011.

[10] Kuny, S. at all:“Speaking behavior and voice sound characteristics in depressive patients during recovery,” *J. Psych. Res.*, vol. 27, pp.289–307, 1993.

[11] Darby, J. at all:“Speech and voice parameters of depression: A pilot study,” *J. Commun. Disorders*, vol. 17, pp. 75–85,1984.

[12] J. Leff and E. Abberton, “Voice pitch measurements in schizophrenia and depression,” *Psychological Med.*, vol. 11, pp. 849–852, 1981.

AN AUTOMATIC METHOD FOR THE ANALYSIS OF PITCH PROFILE IN BIPOLAR PATIENTS

A. Guidi[1,2], N. Vanello[1,2], G. Bertschy[3], C. Gentili[4], L. Landini[1,2], E. P. Scilingo[1,2]

[1]Univ. of Pisa, Dept. of Information Engineering, Pisa, Italy
[2]Univ. of Pisa, Research Center "E. Piaggio", Pisa, Italy
[3]Univ. Hospital and Univ. of Strasbourg, INSERM u666, Dept. of Psychiatry, Strasbourg, France
[4]Univ. of Pisa, Dept. of Surgical, Medical, Molecular Pathology and Critical Care, Pisa, Italy
andrea.guidi@for.unipi.it

Abstract: **Psychiatric patients affected by bipolar disorder experience a mood swing, often ranging from mania to depression. Investigating biomedical signals to detect physiological correlates of mood changes is a more and more debated issue. In this work we describe an automatic method to analyse prosodic features estimated from speech signals. In particular we explore pitch dynamics in voiced part of syllables using some features borrowed for Taylor's tilt intonational model. However, the approach here proposed differs substantially from Taylor's one in that the features are estimated from all voiced segments without performing any analysis of intonation. This method results in features that acquire a different meaning and can be estimated automatically without any labelling step. The suggested approach has been tested firstly on an emotional speech database. Then an analysis on speech samples acquired on psychiatric patients in different mood states is introduces and the results are discussed.**

Keywords : **bipolar disorders, mood state, voice analysis, voice pitch, swipe', tilt**

I. INTRODUCTION

Bipolar disorder is an increasingly widespread pathology and it is characterized by a mood swing passing through hypomania, euthymia and depression. It is important to endeavour to find out tools aiming to support the physicians in formulating diagnoses. Several studies are concerned with the analysis of biomedical signals to detect physiological correlates of mood changes. Relevant information can be drawn by the analysis of speech signals, whose characteristics have been shown to vary in patients affected by psychiatric disorders with respect to healthy subjects. Prosodic and spectral features have been found to vary in patients with respect to healthy subjects [1, 2, 3]. The analysis of pitch changes has also been proposed with the aim of identifying different emotions and mood changes associated with mental disorders [4, 5]. In this work we describe an automatic method for the analysis of pitch contour. In particular this approach performs a segmentation of running speech and identifies voiced parts of syllables. Descriptive statistics of the pitch profile within each voiced segment are suggested. The results obtained from an emotional speech database are shown. Preliminary results on bipolar patients recorded in different mood states are introduced and discussed. This is a study-part of the European project PSYCHE (Personalised monitoring SYstems for Care in mental HEalth) funded by the Seventh Framework Programme.

II. METHODS

A. Algorithm

The proposed approach consists in a three-step process. In the first step, speech signal is analysed to detect voiced part of syllables by using information about signal intensity and zero crossing rate [6]. The signal intensity is estimated using the autocorrelation method applied to sliding windows. A window width equal to 32 ms and a window step equal to 8 ms are used. The intensity value is obtained by retaining only the frequencies between 5 Hz and 5 kHz. The threshold used to discard unvoiced segments is the median of the speech intensity calculated on the whole signal. In order to detect syllables nuclei, local maxima and local dips of the intensity contour are analysed [7]. Each syllable nucleus is considered centred on a local maximum whose intensity is 1 dB higher than the intensity of a preceding local dip. High intensity unvoiced sounds are differentiated from voiced segments by estimating zero crossing rate: only the segments having high intensity and low zero crossing rate are labelled as voiced. In a second step the pitch contour pertaining each segment is estimated using Camacho's swipe' algorithm [8], based on a spectral matching approach. In each voiced segment, pitch is estimated by means a double procedure and using a sliding window approach. First an early estimate of pitch (f_0) on the whole segment is performed, afterwards the effective pitch contour is extracted by using a time window width of $T=4/f_0$ and a time step of $dt=T/4$ [9]. In a third step the final features, that allow to describe specific

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

characteristics of the pitch profile within each voiced segment, are estimated. In particular the extracted features have been borrowed from Taylor's Tilt Model [10] and are related to "relative sizes of the amplitude and durations of rises and falls for an event". Within each voiced segment an eventual local maximum is detected and the features are estimated as in (1-3)

$$Amplitude^* = \frac{|A_{rise}|-|A_{fall}|}{|A_{rise}|+|A_{fall}|} \quad (1)$$

$$Duration^* = \frac{D_{rise}-D_{fall}}{D_{rise}+D_{fall}} \quad (2)$$

$$Tilt^* = \frac{Amplitude^*+Duration^*}{2} \quad (3)$$

where A_{rise} and A_{fall} are the pitch change during the rising and falling section within a segment respectively, D_{rise} and D_{fall} are the duration of the rising and falling sections. Given the voiced segment and identification processes we adopt, features we are estimating are different by those proposed by Taylor (see Discussion section), even if functionally equivalent.

B. Experimental protocol

To verify voiced segments detection performances, the segmentation step was applied on a database consisting of audio and concurrent electroglottographic (EGG) recordings [11]. Voiced segments as revealed by EGG were considered as the ground truth. In particular we evaluated the percentage of voiced speech detected by the proposed methods, in terms of specificity and sensitivity.

The described features were firstly estimated from an emotional speech database [12]. Ten different sentences, acted by ten different actors (5 female) playing different emotions (anger, boredom, happiness and neutral), have been selected. Both intra- and inter-subject statistical analyses were performed.

Six psychiatric patients (1 female) were recruited for this study. All subjects had a clinical diagnosis of bipolar disorder. Every subject was recorded twice a day (in the morning and in the afternoon) in two different days, corresponding to two different mood states. A physician labelled patient's mood status before each acquisition, using clinician administered rating scales. In this study three different states were identified, namely depressed, euthymic, and hypomanic states. The experimental protocol, approved by the clinical ethical committee, consisted of a neutral text reading: subjects were asked to read a text that was supposed not to elicit a strong emotional reaction. The signals were acquired with a sample frequency equal to 48 KHz and a resolution of 32 bits by means of a high quality directional microphone. Two intra-subject statistical analyses were performed to find out feature changes between records of the same day, i.e. with the same mood label, and between different mood states.

III. RESULTS

All features were shown not to be normally distributed, so non parametric tests were adopted. In particular the Mann-Whitney U-test has been used for the intra-subject analysis and the Kruskal-Wallis test for the inter-subjects analysis.

Segmentation results: the 94% of audio signal labelled as voiced by the proposed method was found to be voiced according to EGG signal segmentation. On the other side, the 77% of EGG signal labelled as voiced, obtained the same classification by using the proposed method on audio records. In conclusion, the proposed approach resulted in a specificity of 90% and a sensitivity of 81%. Since the percentage of detected vowels by our approach was equal to 95.3, the above-described results can be partially explained by an underestimation of voiced segments length.

Emotional Database results: intra-subject analysis revealed that amplitude* (ampl*) and duration* (dur*) showed statistically significant differences among emotions characterized by high arousal with respect to

Table 1: median and median absolute deviation of ampl* parameters extracted from emotional database.

subj.	anger	neutral	boredom	happiness
1	**0,49 ± 0,51*+**	**-0,49 ± 0,50 *∂**	**-0,46 ± 0,54 +•**	**0,48 ± 0,52 ∂•**
2	0,38 ± 0,61	0,46 ± 0,54	0,27 ± 0,73	0,40 ± 0,59
3	**0,30 ± 0,69 ***	**-0,01 ± 0,97 *∂**	**-0,07 ± 0,67 •**	**0,58 ± 0,42 ∂•**
4	**0,51 ± 0,49 *+**	**-0,52 ± 0,48 *◊∂**	**0,33 ± 0,67 +◊**	0,27 ± 0,73 ∂
5	**0,46 ± 0,54 *+**	**-0,51 ± 0,48 *∂**	**-0,36 ± 0,64 +•**	**0,20 ± 0,79 ∂•**
6	0,52 ± 0,48	0,54 ± 0,46	0,07 ± 0,93	0,75 ± 0,25
7	**0,47 ± 0,53 +**	0,26 ± 0,74	**0,08 ± 0,90 +•**	**0,62 ± 0,38 •**
8	**0,72 ± 0,27 *+**	**0,34 ± 0,65 *◊**	**-0,37 ± 0,63 +◊•**	**0,60 ± 0,40 •**
9	**0,46 ± 0,53 *+**	**-0,28 ± 0,72 *∂**	**-0,69 ± 0,31 +•**	**0,83 ± 0,16 ∂•**
10	**0,44 ± 0,55 *+**	**-0,01 ± 0,83 *∂**	**-0,25 ± 0,75 +•**	**0,53 ± 0,46 ∂•**

Table 2: median and median absolute deviation of dur* parameters extracted from emotional database.

subj.	anger	neutral	boredom	happiness
1	**0,08 ± 0,77 *+**	**-0,29 ± 0,52 ***	**-0,33 ± 0,42 +**	-0,40 ± 0,49
2	-0,13 ± 0,69	-0,10 ± 0,66	-0,03 ± 0,78	0,06 ± 0,68
3	-0,14 ± 0,64	**-0,39 ± 0,41 ∂**	**-0,57 ± 0,25 •**	**0,14 ± 0,68 ∂•**
4	**0,11 ± 0,67 ***	**-0,20 ± 0,40 *◊**	**0,08 ± 0,58 ◊**	0,00 ± 0,71
5	**0,00 ± 0,67 *+**	**-0,25 ± 0,46 ***	**-0,25 ± 0,45 +**	-0,08 ± 0,66
6	0,14 ± 0,86	0,33 ± 0,58	0,00 ± 0,78	0,11 ± 0,80
7	**0,05 ± 0,71 +**	-0,09 ± 0,63	**-0,33 ± 0,51 +•**	**0,17 ± 0,75 •**
8	**0,23 ± 0,72 +**	**0,07 ± 0,78 ◊**	**-0,33 ± 0,48 +◊•**	**0,24 ± 0,70 •**
9	**0,13 ± 0,74 +**	-0,25 ± 0,46	**-0,33 ± 0,47 +•**	**0,43 ± 0,57 •**
10	**0,00 ± 0,81 +**	**-0,41 ± 0,44 ◊**	**-0,60 ± 0,28 +◊•**	**0,04 ± 0,88 •**

Table 3: median and median absolute deviation of tilt* parameters extracted from emotional database.

subj.	anger	neutral	boredom	happiness
1	**0,19 ± 0,81 *+**	**-0,38 ± 0,53 ***	**-0,43 ± 0,46 +**	-0,02 ± 0,80
2	0,09 ± 0,77	0,10 ± 0,70	0,01 ± 0,83	0,15 ± 0,65
3	0,04 ± 0,66	**-0,18 ± 0,69 ∂**	**-0,29 ± 0,49 •**	**0,37 ± 0,59 ∂•**
4	**0,30 ± 0,71 ***	**-0,38 ± 0,46 *◊**	**0,15 ± 0,75 ◊**	0,11 ± 0,76
5	**0,18 ± 0,80 *+**	**-0,37 ± 0,49 *∂**	**-0,26 ± 0,58 +•**	**0,10 ± 0,73 ∂•**
6	0,29 ± 0,72	0,31 ± 0,69	0,21 ± 0,79	0,43 ± 0,57
7	**0,24 ± 0,68 +**	0,07 ± 0,70	**-0,13 ± 0,76 +•**	**0,40 ± 0,60 •**
8	**0,52 ± 0,48 +**	**0,11 ± 0,79 ◊**	**-0,27 ± 0,62 +◊•**	**0,39 ± 0,60 •**
9	**0,27 ± 0,73 +**	**-0,25 ± 0,61 ◊∂**	**-0,43 ±0,47 +◊•**	**0,61 ± 0,39 ∂•**
10	**0,17 ± 0,80 +**	**-0,20 ± 0,66 ◊**	**-0,43 ± 0,49 +◊•**	**0,23 ± 0,74 •**

lower arousal states (happiness and anger vs. boredom and neutral). In tables 1 and 2 results are shown. The different symbols highlight statistically significant differences. Concerning tilt* parameter anger and happiness proved to be statistically different from boredom (Table 3). In some subjects differences were seen between neutral and boredom. No differences were displayed between anger and happiness.

An inter-subject analysis revealed that ampl* and tilt* allow to separate anger and happiness from boredom and neutral. Dur* instead allows to separate boredom from happiness and anger. In figures 1-3 the results of the Kruskal-Wallis test are shown. In the graphs each group mean- ranks have been represented by a symbol and an interval around the symbol. If two intervals are disjoint, the groups are significantly different. If their intervals overlap, they are not significantly different. In table 4 the p-values resulting from the test and the median of each group are reported.

Experimental protocol results: an intra-subject analysis on bipolar patients revealed that all subjects, but one, did not show any statistically significant differences between the features related to the same day recording (Table 5-6). An asterisk highlights the p-values that are lower than *0.05*.

All the subjects showed a different mood state in the second acquisition day with respect to the first one (Table 7). In all the subjects but one, the ampl* features highlighted significant differences. Three out of six subjects showed no significant changes between both the dur* and tilt* features (Table 8). Inter-subjects analysis was not performed here, given the small number of subjects.

Table 4: P-values of Kruskal-Wallis test performed on each feature and median of each group.

feature	p-value	anger	neutral	boredom	happiness
ampl*	6,0658E-05	0,47	-0,01	-0,16	0,56
dur*	2,8661E-03	0,06	-0,23	-0,33	0,08
tilt*	2,4856E-04	0,22	-0,19	-0,27	0,30

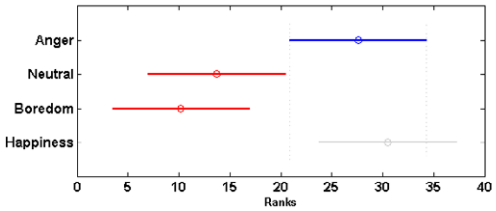


Figure 1: Graphs of Kruskal-Wallis test of amp*.

IV. Discussion

In this work we investigated possible changes in speech related features, on varying mood states in bipolar subjects. In particular we propose the use of parameters describing pitch changes in voiced part of syllables.

The tests about the voiced segments identification revealed that specificity of the proposed approach is good. In our opinion, this is an important result since we have to keep low the probability of labelling unvoiced segments as voiced.

The method here proposed exploit swipe' algorithm for pitch estimation, even if other pitch estimation algorithms could be used. In [8] Camacho has compared his method with some of them, highlighting good performances.

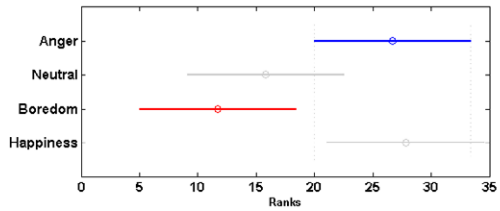


Figure 2: Graphs of Kruskal-Wallis test of dur*.

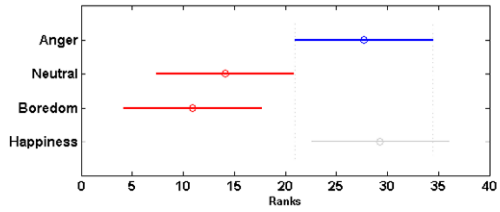


Figure 3: Graphs of Kruskal-Wallis test of tilt*.

Table 5: The symbol * indicates p-values < 0.05 in Mann-Whitney U-test related to patients' features: day 1

	ampl*		dur*		tilt*	
A	0,05 ± 0,86	*0,12 ± 0,84*	-0,33 ± 0,38	*-0,33 ± 0,49*	-0,09 ± 0,73	*-0,06 ± 0,80*
B	-0,57 ± 0,43	*-0,34 ± 0,66*	*-0,45 ± 0,37*	-0,42 ± 0.40	*-0,34 ± 0,55*	-0,33 ± 0,57
C	*-0,45 ± 0,54*	-0,40 ± 0,25	*-0,42 ± 0,40*	-0,44 ± 0,38	*-0,34 ± 0,56*	-0,31 ± 0,58
D	*-0,50 ± 0,49*	-0,59 ± 0,40	*-0,40 ± 0,41*	-0,47 ± 0,33	*-0,38 ± 0,52*	-0,50 ± 0,42
E	*-0,43 ± 0,57*	-0,40 ± 0,95	*-0,50 ± 0,31*	-0,52 ± 0,52	*-0,41 ± 0,49*	-0,38 ± 0,77
F	0,40 ± 0,54	*0,32 ± 0,62*	-0,28 ± 0,54	*-0,28 ± 0,54*	0,04 ± 0,62	*0,03 ± 0,59*

Table 6: The symbol * indicates p-values < 0.05 in Mann-Whitney U-test related to patients' features: day 2

	ampl*		dur*		tilt*	
A	-0,10 ± 0,86	-0,36 ± 0,63	-0,40 ± 0,43	*-0,42 ± 0,40*	-0,20 ± 0,69	*-0,28 ± 0,61*
B	**-0,33 ± 0,66***	***-0,57 ± 0,42****	***-0,40 ± 0,41****	**-0,50 ± 0,33***	***-0,29 ± 0,60****	**-0,45 ± 0,45***
C	*-0,20 ± 0,80*	-0,31 ± 0,68	*-0,29 ± 0,50*	-0,40 ± 0,40	*-0,15 ±0,80*	-0,25 ± 0,64
D	*-0,59 ± 0,40*	-0,83 ± 0,16	*-0.45 ±0,37*	-0.58 ±0,25	*-0,51 ± 0,40*	-0,63 ± 0,28
E	*-0,04 ± 0,95*	-0,09 ± 0,91	*-0,25 ± 0,58*	-0,40 ± 0,40	*-0,08 ± 0,77*	-0,12 ±0,72
F	0,45 ± 0,49	*0,47 ± 0,50*	-0,14 ±0,63	*-0,16 ± 0,65*	0,14 ± 0,55	*0,11 ±0,62*

Table 7: Patients and mood status in each day

	day 1	day 2
A	Hypomania	Euthymia
B	Hypomania	Euthymia
C	Hypomania	Euthymia
D	Depression	Euthymia
E	Depression	Euthymia
F	Depression	Hypomania

Table 8: The symbol * indicate p-values < 0.05 in Mann-Whitney U-test related to patients' features.

	Amplitude*		Duration*		Tilt*	
	day 1	day 2	day 1	day 2	day 1	day 2
A	**0,12 ± 0,84 ***	**-0,36 ± 0,63 ***	-0,33 ± 0,49	-0,42 ± 0,40	-0,06 ± 0,80	-0,28 ± 0,61
B	**-0,34 ± 0,66 ***	**-0,57 ± 0,42 ***	-0,45 ± 0,37	-0,40 ± 0,41	-0,34 ± 0,55	-0,29 ± 0,60
C	**-0,45 ± 0,54 ***	**-0,20 ± 0,80 ***	**-0,42 ± 0,40 ***	**-0,29 ±0,50 ***	**-0,34 ± 0,56 ***	**-0,15 ± 0,80 ***
D	-0,50 ± 0,49	-0,59 ± 0,40	-0,40 ± 0,41	-0,45 ± 0,37	-0,38 ± 0,52	-0,51 ± 0,40
E	**-0,43 ± 0,57 ***	**-0,04 ± 0,95 ***	**-0,5 ± 0,31 ***	**-0,25 ± 0,58 ***	**-0,41 ±0,49 ***	**-0,08 ± 0,77 ***
F	**0,32 ± 0,62***	**0,47 ± 0,5 ***	**-0,28 ± 0,54 ***	**-0,16 ± 0,65 ***	**0,03 ± 0,59 ***	**0,11 ± 0,62 ***

Evanini has reported similar results in [13]. In [14] Swipe' algorithm was used to estimate pitch and jitter on voiced segments, and its performances were compared with those achievable with SIFT algorithm. The two performances of the two approaches were similar as concern average pitch on each voiced segment. On the other hand, Swipe' was found to outperform SIFT as regards jitter estimation.

The proposed parameters are inspired by those introduce by Taylor's Tilt intonational model. However, the parameters here described, although functionally equivalent, are substantially different from Taylor's and do not carry the same information. In fact, in the tilt model intonational events are taken into account, while in this work those parameters were estimated from all voiced segments of syllables. The detection of intonational events relies on the ability of the human labeller and requires training a classifier starting from hand labelled sentences. Our approach is simpler and completely automatic because it bypasses this heavy computational task and studies every voiced segment performing an easier and quicker analysis

Results obtained on the emotional speech database, demonstrated that the proposed parameters highlight significant differences among different emotional speech recordings. This was observed both in intra and in inter subject analysis. In particular, concerning inter subject analysis, some features have been shown to be capable to group emotions by excitation level. More the patients are aroused, and more their speech features show different trend. However, it is important to stress that the emotional database we took into account is a collection of sentences spoken by actors who were "playing" different emotions, while the actors' actual mood is unknown.

Finally intra-subject analyses on bipolar patients have shown that the proposed features have a good specificity. In fact, no statistically significant differences were found among features obtained from acquisitions labelled with the same mood state. The only statistically significant differences that have been found are related to mood changes. Amplitude* has shown a good capability in discriminating different moods, while duration* and tilt* have shown lower performances. The features here proposed could be used along with other features, like pitch and jitter [9], to improve the performances of speech-based mood classifiers.

V. CONCLUSION

In this work we propose a method to estimate prosodic features of voiced segments. The proposed features although borrowed from Taylor's model, were applied applied in a different context, thus achieving a different meaning.

The results on an emotional database have showed that such parameters could be able to find out statistically significant differences in emotional speech. In particular emotions are distinguishable by excitation level.

The analyses on bipolar patients have highlighted that the proposed parameters have a good specificity. In fact, statistically significant differences have not been detected in all couples of records with the same label, but one. On the contrary such statistical differences have been found out comparing speech records corresponding to different emotional states.

In conclusion the proposed method may provide a useful tool to find out statistically significant differences among emotional and mood speeches.

REFERENCES

[1] C. Sobin and H. Sackeim, "Psychomotor symptoms of depression." Am J Psychiat, 154, 14–17, 1997.
[2] A. Nilsonne, et al., "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression." JASA, 1988.
[3] M. Cannizzaro, et al., "Voice acoustical measurement of the severity of major depression," Brain Cogn, 56,(1), 30–35, 2004.
[4] S. Koolagudi and K. Rao, "Emotion recognition from speech: a review," Int J Speech Tech, 15,(2), 99–117, 2012.
[5] M. Bulut and S. Narayanan, "On the robustness of overall f0-only modifications to the perception of emotions in speech," JASA, vol. 123, p. 4547, 2008.
[6] B. Atal and L. Rabiner, "A pattern recognition approach to voiced unvoiced- silence classification with applications to speech recognition," ITASS, 24, 3, 201–212, 1976.
[7] N. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," Behavior Research Methods, 41, 2, 385–390, 2009.
[8] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," JASA, 124, (3), 1638–1652, 2008.
[9] N. Vanello, et al., "Speech analysis for mood state characterization in bipolar patients". In EMBS (EMBC), 2012 Annual International Conference of the IEEE (2104-2107). IEEE.
[10] P. Taylor, "Analysis and synthesis of intonation using the tilt model," JASA, 107, 1697, 2000.
[11] J. Kominek and A. Black, "Cmu arctic databases for speech synthesis cmu language technologies institute," Language Technologies Institute, CMU, Pittsburgh PA, Tech Report CMU-LTI-03-177, 2003
[12] F. Burkhardt, et al., "A database of german emotional speech," in Proc. Interspeech, 2005, 2005.
[13] K. Evanini, C. Lai, and K. Zechner, "The importance of optimal parameter setting for pitch extraction", Proceedings of Meetings on Acoustics. vol. 11, 2011.
[14] N. Vanello, et al., "Evaluation of a pitch estimation algorithm for speech emotion recognition," in Proc. 6th MAVEBA, 29–32, 2009.

ACOUSTIC ANALYSIS OF SPANISH VOWELS IN EMOTIONAL SPEECH

F. M. Martinez-Licona[1], J. Goddard[1], A. E. Martínez-Licona[1]M. Coto Jiménez[1,2]
[1] Universidad Autonoma Metropolitana/Department of Electrical Engineering, Mexico City, Mexico,
{fmml, jgc, aaml}@xanum.uam.mx
[2] Universidad de Costa Rica/Electrical Engineering School, San José, Costa Rica, marvin.coto@ucr.ac.cr

***Abstract*: In this paper an initial acoustic analysis of the five Spanish vowels in emotional speech is given, for both Spanish as spoken in Spain and Spanish as spoken in Mexico. More precisely, two emotional speech databases, which were recorded by professional actors from both Spain and Mexico, were processed to obtain information about the vowels concerning their durations, fundamental frequencies and the first two formant frequencies, as a function of the emotion type. This quantitative information is analyzed to provide some conclusions, such as the possible correlations with the level of arousal of the speaker's emotional state and the similarities and differences between speakers in Spain and Mexico.**
***Keywords :* Emotional Spanish speech, acoustic vowel analysis**

I. INTRODUCTION

How can we differentiate between Spanish vowels in emotional speech? Is there any difference between the vowels articulated in emotional speech in Spanish, as spoken in Spain, compared to those for Mexico's Spanish? As Martín-Butragueño mentions in [1], there are not many papers dealing with acoustic aspects, such as the formant structure, of Mexican Spanish vowels in general, let alone in emotional speech.

Some work has been done for other languages. For example, in [2,3] different acoustic studies were conducted for German and English, respectively, using utterances spoken by actors for a variety of emotions. In a recent paper [4], it was found that the positions of the average $F1/F2$ formant values extracted on a vowel level from a German dataset were strongly correlated with the level of arousal of the speaker's emotional state.

It should be noted that the study of emotional speech has a very close relationship to that relating to the study of stress in human beings. Also, stress is often used to describe negative situations (e.g. fear, anger, anxiety), however there is also eustress (c.f. [5]), which is the term for a positive stress, where one can feel motivated and improve performance. This means that other types of emotion, such as joy, can have an important bearing on positive stress and are worth studying in this context.

In the present paper, we conduct an initial preliminary acoustic study into the above two questions. To this end, we use two speech databases, the first a well-known Spanish emotional speech database [6] developed by the Center for Language and Speech Technologies and Applications (TALP) of the Polytechnic University of Catalonia (UPC) for the purpose of emotional speech research, and the second developed by us. We present the results obtained using these databases for the five Spanish vowels pertaining to their duration, fundamental frequency and the first and second formant frequencies as a function of the emotion type.

The paper is organized as follows: In the next section the emotional speech databases used, and the methods applied to process them are explained. The results obtained are then presented, and finally a discussion and some conclusions follow.

II. METHODOS

One of the emotional speech databases was created by the Center for Language and Speech Technologies and Applications (TALP) of the Polytechnic University of Catalonia (UPC) for the purpose of emotional speech research. The database was part of a larger project, INTERFACE, involving four languages, English, French, Slovene, and Spanish. In the case of Spanish, two professional actors, a male and a female, were used to record the corpus. The speech corpus consists of repeating 184 sentences with the so-called big six emotions of joy, surprise, anger, fear, disgust and sadness, together with several neutral styles. The orthographic transcriptions are also supplied together with speech files stored as sequences of 16-bit, 16kHz speech files without headers, nor compression (Linear PCM, Intel byte format). The second database was essentially a duplicate of the first recorded using two Mexican professional actors, again one male and the other female, and just one neutral style.

The 184 sentences included isolated words, sentences, which can also be in the affirmative and interrogative forms. The distribution is shown in Table 1.

Table 1: *Spanish Corpus Contents*

Identifier	Corpus contents
1 – 100	Affirmative
101 – 134	Interrogative
135 – 150	Paragraphs
151 – 160	Digits
161 – 184	Isolated words

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

It is interesting to note that in a subjective test of the original database with 16 non-professional listeners (UPC engineering students), it was found that over 80% of the sentences were correctly classified initially, and given a second choice, more than 90%. Each emotion was correctly classified by at least half of the listeners.

Although the orthographic transcriptions were available, the speech databases were not initially segmented, and this was required in order to conduct the proposed analyses. To this end, the *EasyAlign* tool [7] was employed. *EasyAlign* is an automatic phonetic alignment tool for continuous speech in several languages, including Spanish, under *Praat* [8].

Once the speech databases had been segmented, several *Praat* scripts from *Spect* [9], as well as others developed by us, were used to obtain vowel information for their time durations, fundamental frequencies and the first two formant frequencies, all as a function of the emotion type as well as the type of Spanish spoken and gender.

III. Results

The analysis of the speech signals included the duration of the vowels and analysis of the formants, the pitch and the tones.

A. Vowel duration

Fig. 1 & 2 show the vowel behavior according the duration of the sound for the case /a/ and /e/ respectively. In the case of vowel /a/ it can be seen that the Mexican male speaker tends to enlarge its duration in the emotions of joy, anger and disgust, compared to the Castilian (or Spanish) speaker. For vowel /e/ the same effect seems to happen in the emotions of joy, surprise and sadness in the Mexican female speaker.

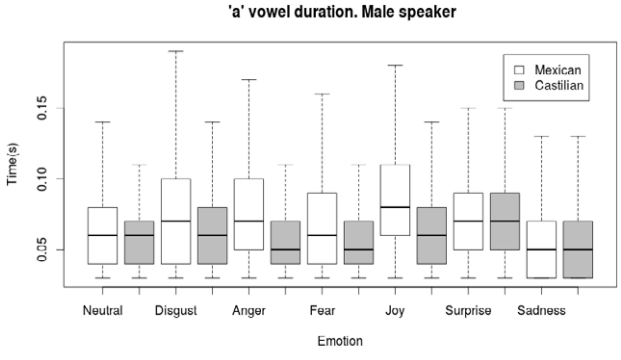


Fig. 1: Duration of vowel /a/ for the male speakers in all the emotions considered.

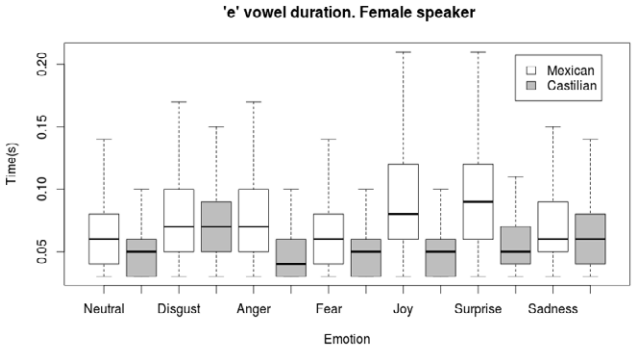


Fig. 2: Duration of vowel /e/ for the female speakers in all the emotions considered.

B. Formant analysis

Fig. 3-6 show the relation between the first and the second formants, F_1 and F_2, of the five Spanish vowels for the seven emotions in the shape of a triangle. The four cases of male-female and Castilian-Mexican, are considered.

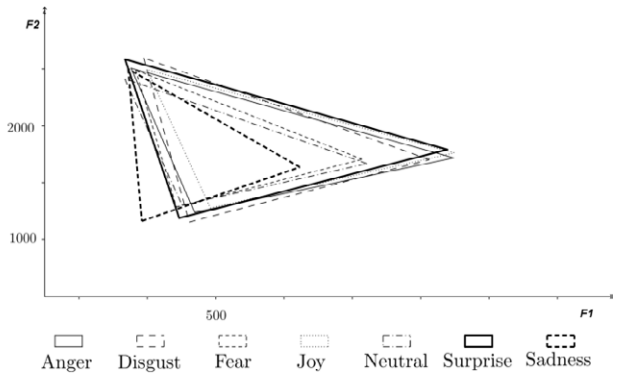


Fig. 3: The triangle of the emotional vowel formants (F_1 and F_2) for the Castilian female speaker.

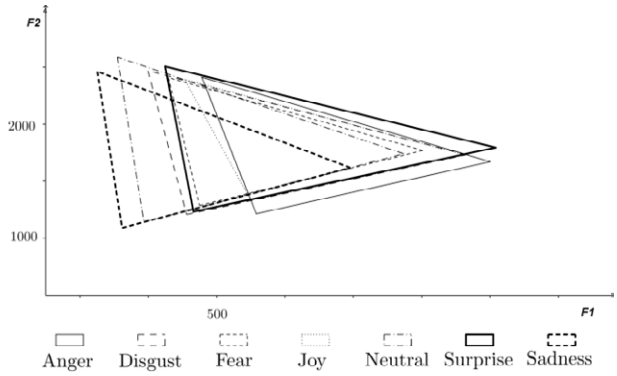


Fig. 4: The triangle of the emotional vowel formants (F_1 and F_2) for the Mexican female speaker.

It is remarkable that in both, male and female Mexican speakers, the triangles are more spread along the emotions than the Castilian ones.

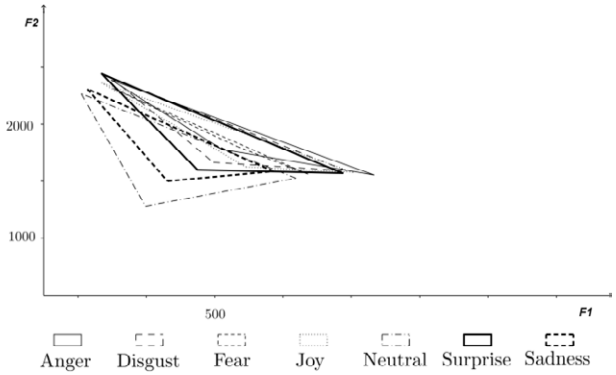


Fig. 5: The triangle of the emotional vowel formants (F_1 and F_2) for the Castilian male speaker.

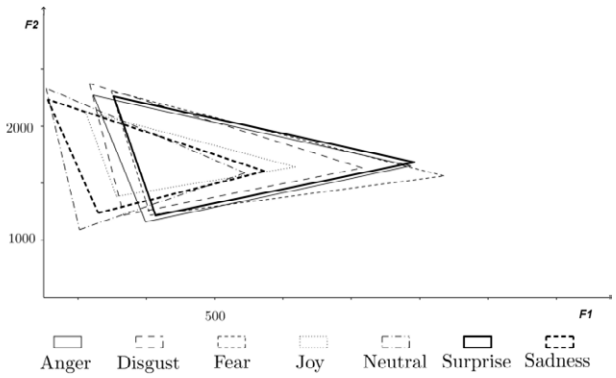


Fig. 6: The triangle of the emotional vowel formants (F_1 and F_2) for the Mexican male speaker.

C. Vowel pitch

Fig. 7 & 8 show the maximum value of the pitch for the five vowels in two different emotions for the male speakers.

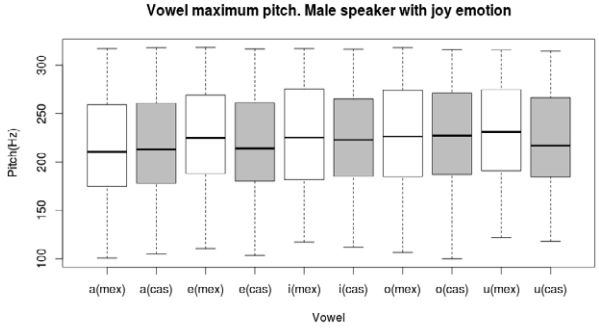


Fig. 7: Maximum pitch of the vowels for the emotion of joy in the male speakers (mex: Mexican, cas: Castilian).

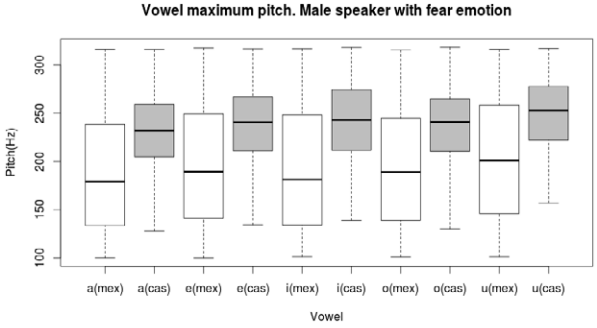


Fig. 8: Maximum pitch of the vowels for the emotion of fear in the male speakers (mex: Mexican, cas: Castilian).

It can be seen that the emotion of joy presents less variations compared to the emotion of fear; it is also interesting to note that this last emotion presents lower mean values for the Mexican speaker while for the first emotion the mean values range is short in both speakers.

D. Tonal analysis

For the tonal analysis, tables 2 & 3 show the mean value, in hertz, and standard deviation for the male speakers in the case of vowel /a/. Excluding the neutral case, the maximum is found in the emotion of surprise for the Mexican speaker, fear for the Castilian speaker, and the minimum is located in the emotion of sadness for both cases.

Table 2: Mean tone and standard deviation for Mexican male speaker, vowel /a/

Emotion	Mean tone (Hz)	σ
Neutral	105.300	8.1351
Anger	177.202	38.958
Joy	170.422	39.678
Disgust	124.780	35.687
Surprise	182.045	42.913
Sadness	115.460	22.549
Fear	155.018	45.370

Table 3: Mean tone and standard deviation for Castilian male speaker, vowel /a/

Emotion	Mean tone (Hz)	σ
Neutral	115.578	16.402
Anger	164.827	32.498
Joy	180.446	42.926
Disgust	139.365	29.771
Surprise	175.945	43.284
Sadness	116.705	16.640
Fear	201.396	36.103

Tables 4 & 5 show the mean value, in hertz, and standard deviation for the female speakers in the case of vowel /e/. Excluding the neutral case, the maximum is found in the emotion of joy for the Castilian speaker, surprise for the Mexican speaker, and the minimum is located in the emotion of sadness for both cases as it happened with the male speakers.

Table 4: Mean and standard deviation for the maximum tone in the Castilian female speaker, vowel /e/

Emotion	Mean tone (Hz)	σ
Neutral	143.687	22.258
Anger	194.657	36.361
Joy	196.657	39.179
Disgust	181.733	36.255
Surprise	189.436	38.572
Sadness	170.059	27.637
Fear	179.236	27.402

Table 5: Mean and standard deviation for the maximum tone in the Mexican female speaker, vowel /e

Emotion	Mean tone (Hz)	σ
Neutral	164.285	39.679
Anger	192.569	46.106
Joy	193.059	44.080
Disgust	182.624	40.979
Surprise	194.884	44.357
Sadness	152.195	32.846
Fear	182.509	40.041

IV. Discussion

Some issues can be discussed about the obtained results. The use of some vowels in both Spanish variations is relevant since /a/ and /e/ are the most frequently present classes and /u/ the least one. The triangle graphics are useful in order to appreciate the effect of the voice characteristics that can differentiate between Castilian and Mexican Spanish. F_1 right shifting is the most remarkable effect in the Mexican speaker along the seven emotions; this can be an interesting point to carry out some more experiments. Using solely the tone analysis the emotions can be separated into two big groups: joy, fear and surprise in one hand, and neutral, disgust and sadness in the other, according to the theory, but this feature is not enough to separate the Spanish variants in none of the cases.

Mexican and Castilian Spanish present differences on prosodic features as rhythm and melodic aspects, they can be used to improve their differences.

The recording conditions were totally controlled so we can consider the signals as acted speech; it has consequences on the naturally of a spoken emotion. A comparison of this database with natural data is desirable. Also it would be important to increase the amount of data for all the emotional states in acted and natural speech.

V. Conclusion

An analysis of the acoustic characteristics on Mexican and Castilian speech signals was performed on the five vowels of the Spanish language. Duration, formants and tone features show differences between them, but it has to be considered the "acted" nature of the data. This is a preliminary study that is being continued in order to deal with the aspects mentioned in the discussion. One of the objectives of this analysis is referred to the modeling of the emotion speech classes for synthesis of emotions.

Acknowledgements

This work was supported by the SEP and CONACyT under the Program SEP-CONACyT, CB-2012-01, No.182432, in Mexico, as well as the Costa Rica University. We want to thank ELRA for supplying the, Emotional speech synthesis database, catalogue reference: ELRA-S0329 (http://catalog.elra.info).

References

[1] P. Martín-Butragueño, "Vocales en contexto," in *Homenaje a Thomas C. Smith-Stark*, Ed. E. Herrera & R. Barriga, México: El Colegio de México, in press.

[2] R. Banse, K.R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, Volume 70, Issue 3, pp. 614-636, 1996.

[3] S. Yildirim, M. Bulut, C. Min Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, "An acoustic study of emotions expressed in speech," *Proceedings of InterSpeech*, pp. 2193-2196, 2004.

[4] B. Vlasenko, D. Philippou-Hubner, D. Prylipko, R. Bock, I. Siegert, A. Wendemuth, "Vowels Formants Analysis allows straightforward detection of High Arousal Emotions," *Multimedia and Expo (ICME), IEEE International Conference on Multimedia and Expo*, pp.1-6, 2011.

[5] M. Sigmund, "Influence of Psychological Stress on Formant Structure of Vowels", Elektronika *IR Elektrotechnika*, Vol. 18, No. 10, pp.45-48, 2012.

[6] Emotional speech synthesis database, catalogue reference: ELRA-S0329. (http://catalog.elra.info),

[7] J. Goldman, "EasyAlign: an automatic phonetic alignment tool under Praat", *Proceedings of InterSpeech*, Firenze, Italy, September 2011.

[8] P. Boersma & D. Weenink, "Praat: doing phonetics by computer," Version 5.3.52, 2013. (http://www.praat.org/)

[9] M. Lennes, "*SpeCT* - The Speech Corpus Toolkit for Praat", http://www.helsinki.fi/~lennes/praat-scripts/

ASSESSING STRESS IN MEXICAN SPANISH FROM EMOTION SPEECH SIGNALS

F. M. Martinez-Licona[1], J. Goddard[1], A. E. Martínez-Licona[1]M. Coto Jiménez[1,2]
[1] Universidad Autonoma Metropolitana/Department of Electrical Engineering, Mexico City, Mexico, {fmml, jgc, aaml}@xanum.uam.mx
[2] Universidad de Costa Rica/Electrical Engineering School, San José, Costa Rica, marvin.coto@ucr.ac.cr

Abstract: **Stress is considered a very serious condition that is a consequence of several alterations. In Mexico it affects about half of the adult population and is considered as the most important issue that young working people and women have to deal. The stressful condition is a complex process that can be partially evaluated through the speech since they can be obtained non-invasively and the signals can be analyzed form acoustic features as well as prosodic and other non-linguistic elements. In this paper, an evaluation of stress from an analysis of Mexican Spanish emotional speech signals is presented. Anger, disgust, fear and a normal state are obtained from a two speakers database and pitch, jitter and shimmer features were extracted. Results show that anger and fear are the emotions that can present useful information for the stress evaluation.**
Keywords: **Stress, emotions, acoustic features**

I. INTRODUCTION

Stress is considered a very serious condition that is a consequence of several physiological alterations; in Mexico 43% of adults suffer from stress, as estimated by the Mexican Institute of Social Security, and according to the National Institute of Psychiatry, Mexico is at the top with patients suffering from work stress affecting mainly young people and women [1]. As a psychological state, stress is a response to a perceived threat or task demand and it usually comes with certain specific emotions like fear, anxiety or anger. Speech may be one of the sources where stress is clearly identified; speech under stress implies that the subject speaks under some form of pressure that results in an alteration of the speech production process [2]. Although speech is primary a vocal activity, there are several human vocalizations that are essentially non-linguistic like voice quality, prosody, rhythm and pausing. The non-verbal content of the voice carries, among other things, information about the physiological and psychological state of the speaker, and it can be obtained from the selection of the appropriate acoustic features [3,4]. Due to the importance that the stressful condition represents, efforts are focused on taking preventive measurements in order to control the effects of stressful situations on people with noninvasive methods, if possible. This paper presents an evaluation of stress from an analysis of Mexican Spanish emotional speech signals, focused in the differences between male and female speakers.

II. METHODOS

Two Mexican speakers, professional actor/actress, recorded three sets of 184 speech tracks each. The tracks included 24 isolated words, 10 digits, 100 affirmative sentences, 34 interrogative sentences and 16 paragraphs. In total there were 1104 tracks that covered the emotions of anger, disgust, fear and a normal state as a reference. The selection of the words, sentences and paragraphs were the same used in [5]. The records were carried out in a professional studio where all the technical aspects and the recording conditions were completely controlled, and the speakers were free to manage themselves for the production of the emotion.

Acoustic features were subtracted from the speech signal using Praat [6]; the features selected were pitch, jitter and shimmer, those which were obtained in order to find the configuration that reveals the most useful information about the differences of the emotions.

A. Pitch

The average pitch over all the utterance was extracted in each track of the database. The mean value was obtained from the pitch contour computed through the autocorrelation method. Since the database contains short and long utterances, this feature was computed from the whole set and the long and short tracks separately.

B. Jitter

Jitter is a measure of period-to-period fluctuations in the fundamental frequency. In general it is calculated between consecutive periods of voiced speech as follows:

$$Jt = \frac{\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \quad (1)$$

where T_i, T_{i+1} are the present and posterior periods of speech and N the total number of intervals. The jitter reported is the local jitter, which is used as a voice quality feature and is defined as the rate between the computed jitter and the mean value of the periods of voiced signal found in the utterance.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

B. Shimmer

Shimmer is a measure of the period-to-period variability of the amplitude value as follows:

$$Shm = \frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (2)$$

where A_i, A_{i+1} are the present and posterior periods amplitude of speech and N the total number of voiced periods. The shimmer reported is the local shimmer, which is defined as the average absolute difference between the amplitudes of consecutive periods divided by the average amplitude.

III. RESULTS

Pitch, jitter and shimmer are features that can analyze tonal behaviors as well as the trends in tone and intensity. As mentioned before, the dataset contains long and short utterances, Table 1 shows the mean values of these features in the complete dataset, Table 2 presents the results on the long utterances and Table 3 does it on the short utterances. It is remarkable that in both speakers the fear value of the mean pitch and the shimmer percentage were higher than the values of the other emotions, and the highest jitter percentage was found in the disgust emotion.

Table 1: *Mean pitch value, jitter and shimmer for the complete dataset. M: male speaker, F: female speaker*

Emotion	Mean Pitch (Hz)		Jitter (%)		Shimmer (%)	
	M	F	M	F	M	F
Normal	101.78	175.50	2.83	2.47	12.34	10.29
Anger	136.61	229.10	3.15	2.15	11.93	10.50
Disgust	217.73	220.24	3.47	2.72	13.93	11.76
Fear	220.79	293.99	3.07	2.16	12.89	12.60

Table 2: *Mean pitch value, jitter and shimmer for the long utterances subset. M: male speaker, F: female speaker*

Emotion	Mean Pitch (Hz)		Jitter (%)		Shimmer (%)	
	M	F	M	F	M	F
Normal	103.28	179.52	2.87	2.43	12.68	10.45
Anger	136.61	229.10	3.15	2.15	11.93	10.50
Disgust	217.73	220.24	3.47	2.72	13.93	11.76
Fear	220.79	293.99	3.07	2.16	12.89	12.60

Table 3: *Mean pitch value, jitter and shimmer for the short utterances subset. M: male speaker, F: female speaker*

Emotion	Mean Pitch (Hz)		Jitter (%)		Shimmer (%)	
	M	F	M	F	M	F
Normal	95.86	159.64	2.70	2.65	11.01	9.69
Anger	136.61	229.10	3.15	2.15	11.93	10.50
Disgust	217.73	220.24	3.47	2.72	13.93	11.76
Fear	220.79	293.99	3.07	2.16	12.89	12.60

Figures 1-3 show the distribution of the features for both male and female speakers, and the complete dataset. It can be seen that the normal male pitch frequency range is very compact compared to the normal female case and that the emotion of fear presents similar mean values and ranges in the two cases.

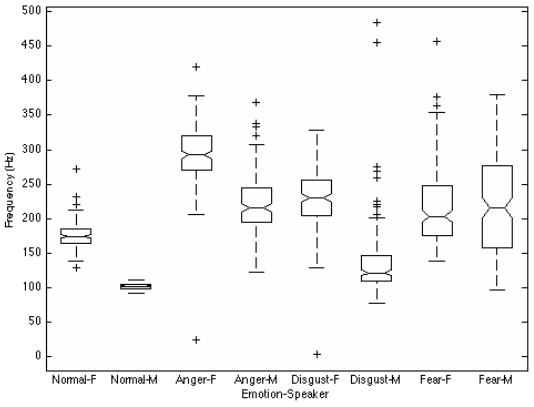


Fig. 1: Mean pitch frequency of the emotion dataset M: male speaker, F: female speaker

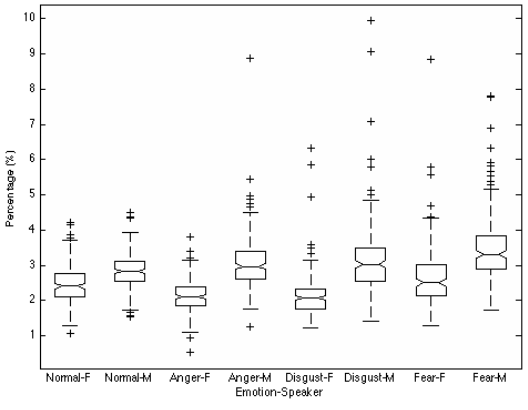


Fig. 2: Jitter percentage of the emotion dataset M: male speaker, F: female speaker

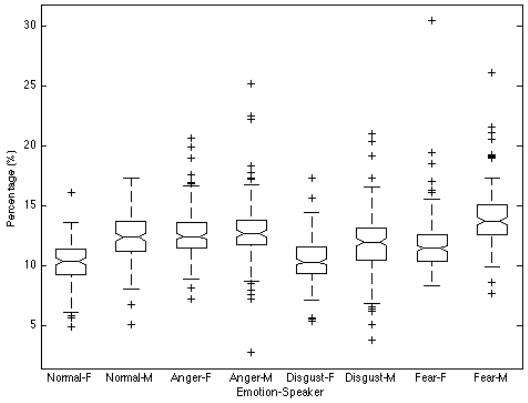


Fig. 3: Shimmer percentage of the emotion dataset M: male speaker, F: female speaker

From Figure 3 it can be noticed that the emotion of anger in the female speaker and the emotion of fear in the male speaker present most outliers than the other emotions. In order to appreciate the effect of the emotional state in the pitch feature, Figure 4 show the pitch contour of the same utterance for each emotion in the male speaker and Figure 5 does it for the female speaker. In this case, anger and fear are the emotions that present an increased value of pitch frequency compared to the reference (normal) and the female speaker differences in the pitch frequencies across all the emotions considered are less evident than the pitch frequencies in the male speaker. The figures also show that, in average, the utterance durations for the female speaker is shorter than the durations of the other one.

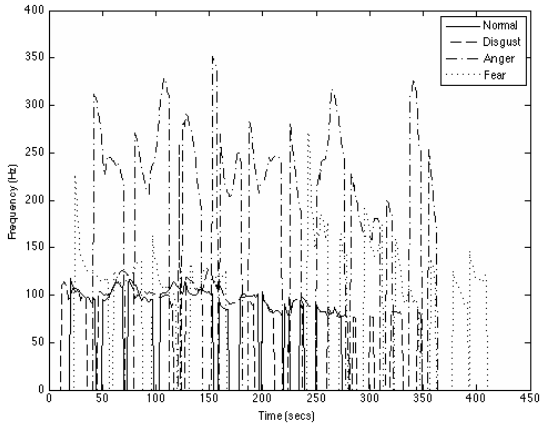


Fig. 4: Pitch frequency contour for the same track in the male speaker record

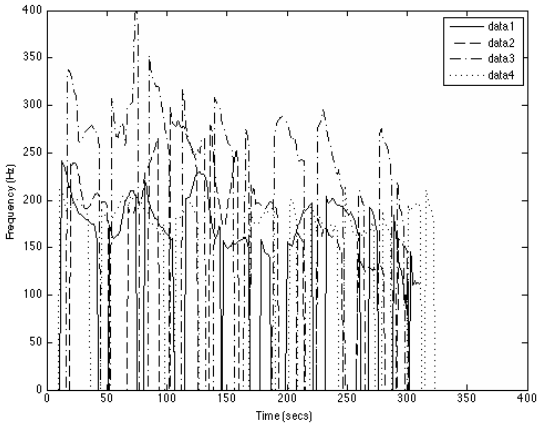


Fig. 5: Pitch frequency contour for the same track in the female speaker record

IV. DISCUSSION

To evaluate stress from speech signals, we have to consider more than one feature and find the best feature configuration that provides useful information. Jitter represents the perturbation in the vibration rate of the vocal cords and shimmer is the analogous in intensity, since there is a glottal component involved and that it is related to the airflow passing though the vocal fold, we might think that they can be important for the analysis. From the results we can see that there is a number of outliers in jitter for male fear and disgust and in shimmer for male fear, disgust and anger; this can mean that these emotional states present noticeable alterations in the production of the speech compared to a normal condition

In this analysis the pitch frequency and the pitch contour provided information that can be used for distinguish among the emotions. Although this is an interesting issue, it is important to annotate that the speech tracks were recorded by professionals; they managed their voices as to find the most accurate approximation to a natural emotion, and in a controlled environment.

During the experimentation we detected some improvements that can be performed. Since Praat was used to obtain the feature values, the calculation of them was restricted to their methods. Pitch frequency can be obtained from different methods such autocorrelation, cross correlation or even cepstral coefficients and the same happens to jitter and shimmer so it is convenient to experiment with them and study their interpretation in order to obtain more information about the relationship to the stressful conditions.

In adition to the features considered in this paper, some other acoustic and prosodic characteristics may be used for the evaluation of emotions and for the assessment of the stress. From the shown results it is possible to

consider the experimentation with more combination of features including the time duration of voiced parts of the records, the short and long pauses during the pronunciation, the change in the time frame for the calculation of the pitch, jitter and shimmer, etc. These are currently under development

The original database contains 16 paragraphs an in the study they were removed form the dataset; this first approach considered short and long tracks as isolated words and simple sentences to have initial clues about the changes along the utterances, the experimentation on the paragraphs as well as the extension to a more natural emotional voices is being carried out.

Finally, it is important to keep present that the evaluation of the stress condition through emotional speech signals can be considered as a tool for the evaluation of the psychological level, according to the stress taxonomy [7], that completes the physical, physiological and perceptual levels; it is evident that the more integral is the analysis, the better will be the understanding of this severe condition.

V. Conclusion

An evaluation of the stress condition from an analysis of Mexican Spanish emotional speech signals is presented. The features of pitch frequency, jitter and shimmer were obtained from isolated words and sentences, recorded in the emotions of anger, disgust, fear and a normal state. Results show that these emotions can present useful information for the stress evaluation, but it is necessary perform more experimentation, obtain more data and analyze in a more integral way.

Acknowledgements

This work was supported by the SEP and CONACyT under the Program SEP-CONACyT, CB-2012-01, No.182432, in Mexico, as well as the Costa Rica University.

References

[1] José Juan Reyes, Mexicans are normal to live with stress, El Economista, 04-11-2012, http://eleconomista.com.mx/sociedad/2012/11/04/mexicanos-ven-normal-vivir-estres-laboral

[2] John H.L., Hansen & Sanjay Patil, Speech Under Stress; Analysis, Modeling and Recognition, in Christian Müller, (Ed), Speaker Classification I, Fundamentals, Features, and Methods Part II, pp. 108-137, 2007.

[3] Rothkrantz, L. J., Wiggers, P., van Wees, J. W. A., & van Vark, R. J. (2004, January). Voice stress analysis. In *Text, Speech and Dialogue* (pp. 449-456). Springer Berlin Heidelberg.

[4] Xi Li, Jidong Tao, Michael T, Johnson, et al, Stress and Emotion Classification using Jitter and Shimmer Features, Proceedings of ICASSP, Vol. IV, pp. 1081-1084, 2007.

[5] ELRA catalogue (http://catalog.elra.info), Emotional speech synthesis database, catalogue reference: ELRA-S0329.

[6] Boersma P, Weenink D., Praat: doing phonetics by computer, V5.3.23, from http://www.praat.org/ , 2012.

[7] Hansen, J.H.L., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, E.J., Anderson, T., Vloeberghs, C.R.A., Trancoso, I., Verlinde, P.: The Impact of Speech Under 'Stress' on Military Speech Technology. In: NATO RTO-TR-10, AC/323(IST)TP/5 IST/TG-01 (2000).

Session VIII:
VOICE AND GENDER-SIBLINGS

WOMENS' VOICE DURING IN-VITRO FERTILIZATION TREATMENT

O. Amir[1], N. Lebi-Jacob[1], O. Harari[2]
[1]Department of Communication Disorders, Sackler Faculty of Medicine, Tel-Aviv University, Israel.
[2]Fertility and IVF Clinic, Haifa, Israel.

Abstract: **This study was aimed to explore the effect of In-Vitro Fertilization treatment on acoustic properties of women's voice. To that end, ten women undergoing In-vitro fertilization treatment volunteered to participate in the study. All women were recorded repeatedly in three successive sessions, before and during treatment. In addition, hormonal assays, endometrial thickness measurements and follicular growth data were collected. Recordings were performed during sustained vowel phonation, and during a reading task. Acoustic analyses included fundamental-frequency measures, as well as frequency- and amplitude-perturbation measures. Repeated-measure analyses of variance were performed to test for treatment effect, and the correlations between the acoustic measures and the hormonal as well as endometrial thickness data were examined. Results revealed a significant reduction in the two fundamental-frequency measures and in the amplitude-perturbation measure along treatment ($P<.05$). In addition, prior to treatment, a negative correlation was found between F0 and estrogen levels. During treatment, however, a negative correlation was found between F0 and endometrial thickness. These findings suggest an association between In-vitro fertilization treatment and specific voice properties, supporting the effect of sex hormones on the larynx and on voice. In addition, the possibility of a ceiling effect for the influence of estrogens on female vocal folds was introduced.**

Keywords: **IVF; voice; hormones; vocal folds; acoustic analysis.**

I. INTRODUCTION

The vocal folds, the larynx and the entire voice mechanism are affected by the female hormonal system [1]. This has been confirmed by cytological smears [2] and by the discovery of hormonal receptors in the laryngeal mucosa and epithelium [1,3]. Behavioral support and acoustic evidence for this relationship was found in various conditions, such as pregnancy [4,5], and while using birth control pills [6,7]. Nonetheless, despite the growing body of research documenting the apparent relationship between the female hormonal system and the voice mechanism, these reported vocal changes have often been subclinical, inconsistent or controversial [4,7].

Women undergoing In-Vitro Fertilization (IVF) treatment are exposed to substantially higher levels of estrogen, compared to women experiencing natural hormonal cycles [8] or those using birth control pills [9]. Therefore, we hypothesized that a pronounced voice effect would be found in women undergoing IVF, more than in previously examined conditions. A single recent study has examined subjective self-reports of women undergoing IVF, and failed to identify changes in vocal symptoms during the ovarian stimulation stage of IVF [10]. Because of the inherent limitations and limited validity of the subjective self-evaluation of voice quality, we designed this study to examine possible acoustic changes in voice characteristics among women undergoing IVF treatment.

II. METHODOS

Ten women, (mean age: 31.1 years, range: 25-45) who enrolled for IVF treatment, volunteered to participate and completed a preliminary anamnesis questionnaire. Exclusion criteria included a history speech or voice problems or therapy, formal singing or voice training, hearing loss, smoking, routine alcohol consumption or substance abuse. All women were healthy, with no remarkable medical history and no routine medication.

All women were evaluated three times within the ovarian stimulation phase of the IVF treatment. Session I was performed prior to the beginning of the treatment. During this session, the participants' voice was recorded and blood tests were taken for hormonal measurements. Session II was performed on day five of the ovarian stimulation. During this session, participants were recorded, blood tests were taken and sonographic examination was performed vaginally to document endometrial thickness and to evaluate number and size of follicles. Session III was performed approximately three days after session II, based on follicle size measurement, and replicated all measurements taken in Session II. The

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

three sessions were performed over a period of nine days, on the average (SD=3).

Voice recordings were performed in a quiet room at the fertility clinic, during routine visits. Participants were recorded while sustaining the vowel /a/ for three seconds, eight times repeatedly, in addition to reading a passage aloud. Acoustic analyses were performed using Praat, and a set of eleven acoustic measures was obtained. These measures included: (a) Mean Fundamental Frequency (F0), (b) Minimum Fundamental Frequency (min-F0), (c) Maximum Fundamental Frequency (max-F0), (d) Fundamental Frequency Range (F0-range), (e) Standard Deviation of Fundamental Frequency (F0-SD), (f) Jitter, (g) RAP5, (h) Shimmer, (i) APQ11, (j) NHR, and (k) Autocorrelation.

III. Results

Group means for the acoustic measures obtained from the participants' recordings during the vowel task are presented in Table 1. Following, group means for the reading task are presented in Table 2.

Table 1. Mean Values and Standard Deviations (in parentheses) of Acoustic Measures Obtained for the Sustained Vowel /a/ at the Three Recording Sessions.

Acoustic Measure	Vowel /a/		
	Session 1	Session 2	Session 3
F0 (Hz)	192.61 (18.27)	191.01 (20.10)	189.32 (19.31)
F0-range (Hz)†	13.89 (20.51)	9.77 (17.13)	7.61 (13.59)
RAP (%)	0.24 (0.17)	0.23 (0.32)	0.20 (0.14)
APQ (%)†	1.78 (0.90)	1.77 (0.87)	1.47 (0.48)
NHR	0.007 (0.011)	0.006 (0.010)	0.005 (0.004)
Auto-correlation	0.993 (0.007)	0.993 (0.009)	0.995 (0.003)

† Statistically significant treatment effect (p<.05)

In the vowel task, a significant main effect for Treatment was found for F0-range and APQ11 [(F2,9=3.69, P=.04) and (F2,9=4.15, P<.05), respectively]. Contrast analyses revealed a significant reduction in both measures (corrected P<.05) from session I to session III, but not between adjacent sessions (I vs. II, or II vs. III). Similarly, a consistent reduction in values of F0, F0-SD and max-F0 was observed, as treatment progressed, but these differences failed to reach statistical significance. All other measures did not reveal a consistent change associated with treatment progress (P>.05).

Table 2. Mean Values and Standard Deviations (in parentheses) of Acoustic Measures Obtained for the Reading task at the Three Recording Sessions

Acoustic Measure	Reading		
	Session 1	Session 2	Session 3
F0 (Hz)*	197.37 (19.38)	192.94 (18.81)	190.31 (19.20)
F0-range (Hz)	146.28 (39.93)	140.28 (40.30)	139.16 (48.44)
RAP (%)	0.69 (0.23)	0.71 (0.18)	0.70 (0.16)
APQ (%)	5.98 (1.53)	6.11 (1.48)	6.33 (1.46)
NHR	0.064 (0.022)	0.599 (0.016)	0.063 (0.021)
Auto-correlation	0.953 (0.014)	0.956 (0.010)	0.953 (0.013)

*Statistically significant treatment effect (p<.05)

In the reading task, a significant main effect for treatment was found for F0 (F2,9=5.28, P=.05). Contrast analysis revealed a significant reduction in F0 (corrected P<.05) from session I to session III, but not between adjacent sessions. Similar to the results of F0, a consistent reduction, though not statistically significant, was observed for F0-SD and F0-range, as treatment progressed. All other measures did not reveal a consistent change associated with treatment.

Further analysis was performed by assigning the ten participants to two groups, based on final endometrial thickness measured at session III, as a general predictor of treatment success. Four women were assigned to a "Medium Endometrial" subgroup, while the other six women were assigned to a "Thick Endometrial" subgroup. A statistically significant difference between the two groups was found only for F0 (F1,8=8.82, P=.018). In addition, this measure yielded a significant difference between sessions (F1,8=8.68, P=.019).

Finally, statistically significant linear correlations were found, during Session I, between estrogen levels and three F0 measures [F0: (r=-0.68, P=.03), min-F0: (r=-0.67, P=.03), and max-F0: (r=-0.62, P<.05)]. No significant correlations were found between endometrial thickness and any of the acoustic measures in this session. During Session II, no significant correlations were found between estrogen levels or endometrial thickness and any of the acoustic measures. During Session III, no significant correlations were found between estrogen levels and any of the acoustic measures. However, statistically significant correlations were found

between endometrial thickness and three F0 measures [F0: (r=-0.64, P=.04), F0-SD: (r=-0.68, P=.03) and max-F0: (r=-0.68, P=.03)].

IV. Discussion

This study presents a novel perspective on the effect of IVF treatment on women's voice, using acoustic analyses. First, as treatment progressed and estrogen levels were raised, women exhibited a reduction in F0, F0-range and APQ11 values. Second, when women were under natural hormonal conditions (i.e., Session I), significant correlations were found between acoustic measures related to F0 and estrogen levels. However, when estrogen levels peaked (i.e., Session III), significant correlations were found between these acoustic measures and endometrial thickness.

Estrogens have a hypertrophic and proliferative effect on mucosa, in addition to increasing capillary permeability [1]. These effects were shown in female vocal folds, similar to other target organs. Therefore, an increase in estrogen levels is expected to lead to an increase in vocal folds' mass due to edema, and to a lowering in F0. This was supported by our findings, showing a lowering in F0, as estrogen levels increased along the IVF treatment.

The fact that no increase in perturbation measures was observed in our study suggests that the increased mass of the vocal folds, caused by estrogens, does not follow a typical pattern of edema or extracellular fluid accumulation. Instead, this could imply that fluids are accumulated *intra*-cellularly, such that normal vibratory pattern is not disturbed and voice quality remains relatively unaffected. This possibility, however, should be further supported histologically.

The second major finding of our study is the negative correlations between estrogen levels and F0 measures at Session I. By itself, this result provides added support to the known effect of increased estrogen levels on lowering F0. However, the fact that this was found only for the natural hormonal climate (i.e., Session I), but not during the other two sessions, could be explained by the large difference in estrogen levels between the natural condition and the later conditions. We, therefore, hypothesize that there is a "ceiling effect" for estrogen levels on the vocal folds, similar to that reported for other target organs [11]. Accordingly, the vocal folds have a maximal thickness capacity, which cannot be transcended. This enforces a limit on the maximal impact of estrogen levels on acoustic properties of the female voice.

The negative correlation between F0 measures and endometrial thickness at Session III supports previous studies showing that increased estrogen levels affect both endometrium and vocal folds similarly [1,2]. The fact that this correlation was only found close to the end of the ovarian stimulation phase (i.e., Session III), suggests a difference in the time required for reaching a clinical effect on the endometrium, in comparison to the vocal folds.

Dividing the participants to two subgroups, based on final endometrial thickness, revealed consistent and significant voice differences, shown even prior to the IVF treatment. This suggests that women's voices can provide indications associated with endometrium characteristics. Since final endometrial thickness is predictive of final treatment outcome [12], it is conceivable that specific acoustic measures obtained prior to treatment, could provide a supplementary indication of the expected treatment outcome. Clearly, such clinical indication should be replicated and confirmed.

V. Conclusion

In this preliminary study, specific vocal changes were documented for the first time in women undergoing IVF treatment. In addition, the possibility of a ceiling effect for the influence of estrogens on female vocal folds was suggested.

References

[1] J. Abitbol, P. Abitbol, and B. Abitbol, "Sex hormones and the female voice," *J Voice*, vol.13, pp. 424-446, 1999.

[2] J. Abitbol, J. de Brux, G. Millot, M.F. Masson, O.L. Momoun, H. Pau, and B. Abitbol, "Does a hormonal vocal cord cycle exist in women? Study of vocal premenstrual syndrome in voice performers by videostroboscopy-glotography and cytology on 38 women," *J Voice*, vol. 2, pp. 157-162, 1989.

[3] S.R. Newman, J. Bulter, 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] A. Raj, B. Gupta, A. Chowdhury, and S.A. Chadha, "Study of voice changes in various phases of menstrual cycle and in postmenopausal women," *J Voice*, vol. 24, pp. 363-368, 2010.

[5] F.M.B.La, and J. Sundberg, "Pregnancy and the singing voice: Reports from a case study," *J Voice*, vol. 26, pp. 431-439, 2012.

[6] O. Amir, T. Shental, C. Muchnik, and L. Kishon-Rabin, "Do oral contraceptives improve voice quality?" *Obstet Gynecol*, vol. 101, pp. 773-777, 2003.

[7] O. Amir, T. Biron-Shental, T. Barer, and O. Tzenker, "Different oral contraceptives and voice quality – An

observational study," *Contraception*, vol. 71, pp. 348-352, 2005.
[8] O. Moraloglu, E.A. Tonquc, M. Ozel, G. Ozaksit, T. Var, and E. Srikaya, "The effects of peak and mid-luteal estradiol levels on in vitro fertilization outcome," *Arch Gynecol Obstet*, vol. 285, pp.857-862, 2012
[9] F.M.B. La, W. Ledger, J.W. Davidson, D.M. Howard, and G. Jones, "The effects of a third generation combined oral contraceptive pill on the classical singing voice," *J Voice*, vol. 21, pp. 754-761, 2007.
[10] A.L. Hamdan, R.A. Barazi, A. Kanaan, S. Sinno, and A. Soubra, "Vocal symptoms in women undergoing in vitro fertilization," *Am J Otolaryng*, vol. 33, pp.239-243, 2012.
[11] N.J. Raine-Fenning, B.K. Campbell, J.S. Clewes, N.R. Kendall, and I.R. Johnson, "Defining endometrial growth during the menstrual cycle with three-dimensional ultrasound," *BJOG-Int J Obstet Gy*, vol. 111, pp. 944-994, 2004.
[12] A. Palatnik, E. Strawn, A. Szabo, and P. Robb, "What is the optimal follicular size before triggering ovulation in intrauterine insemination cycles with clomiphene citrate or letrozole? An analysis of 988 cycles," *Fertil Steril*, vol. 97, pp. 15-28, 2012.

SEX-DEPENDENT AUTOMATIC DETECTION OF VOICE PATHOLOGIES

J.A. Gómez-García[1], J.I. Godino-Llorente[1], G. Castellanos-Domínguez[2]
[1]Bioengineering and Optoelectronics group (ByO). Universidad Politécnica de Madrid, Spain.
[2]Control and Signal Processing group (GC&PDS). Universidad Nacional de Colombia, Manizales, Colombia.
jorge.gomez.garcia@upm.es; igodino@ics.upm.es; cgcastellanosd@unal.edu.co

Abstract: **The automatic detection of pathologies using the speech provides certain advantages, such as non-invasiveness and low cost of implementation, compared to traditional diagnostic approaches. However, the reliability of these automatic detection systems is affected by the underlying variability of speech, and where a major source is introduced due to differences in sex. Having this in mind, the present paper deals with such speech variability by designing sex-dependent pathology detection systems. In this manner, a gender detector identifies the sex of the speaker, and according to the decision taken, feeds a male or female pathology detector which takes a decision on the speaker's condition. Mel frequency cepstral coefficients are employed for the characterization of raw speech signals, as well as glottal components extracted from speech by using inverse filtering. Experiments are performed on the Saarbrücken Voice Disorders Database using recordings of the sustained phonation of vowel /a/.**

Keywords: **sex detection, automatic pathology detection**

I. INTRODUCTION

The automatic detection of voice pathologies enables an objective assessment of the presence of certain disorders, reducing the evaluation time and, subsequently improving the diagnosis and clinical treatment given to patients [1]. In this regard, the traditional detection systems, mainly, employ either linear or nonlinear features for discriminating between normal and pathological conditions.

However, there exist dissimilarities between male and female voices due to physiological, acoustic, and psychophysical factors, producing in turn differences in acoustic parameters [2]. These differences alter the performance of automatic detection systems. As a matter of example, authors in [3] report evidences indicating that sex is relevant to evaluate the presence of laryngeal pathologies, when analyzing speech recordings of sustained vowels. Having those precedents, it is reasonable to consider the design of sex-dependent recognition systems such that the sex dissimilarities are faced separately. To this end, the information obtained from glottal components is considered, since there have been found significant differences in glottal waveform parameters between sexes [2]. Also, since the glottal components constitute the excitation source of speech, it is responsible of many underlying vocal features, such as fundamental frequency or a variety of quality parameters [6]; which are also be of great relevance in automatic detection of voice pathologies [7]. These glottal components are typically extracted from voice recordings, by using a procedure termed *inverse filtering*, which decomposes speech into its glottal and vocal tract components.

With the aforementioned precedents, this work discusses the usefulness of an automatic sex-dependent pathology detection system, which is grounded on glottal and speech characterization through Mel frequency cepstral coefficients (MFCC). The proposed methodology is composed by a sex identification stage, which feeds sex-dependent pathology detectors (either male or female). Afterwards, the classification between normal and pathological speech is carried out by Gaussian Mixture Models (GMM). Experiments are performed using the Saarbrücken voice disorders database [4], employing a subset of recordings containing the sustained phonation of vowel /a/. For extracting glottal components from speech, the Iterative Adaptive Inverse Filtering technique [5] is employed.

This paper is organized as follows: Section II presents the database, the sex-dependent detector and all its stages. Section III presents the obtained results. Finally Section IV presents the discussions and conclusions of the work.

II. METHODS

A. Database

The Saarbrücken Voice Disorder Database [4] holds a collection of voice signals from more than 2000 normal and pathological German speakers. It contains the recordings of the sustained phonation of vowels /i/, /a/, and /u/ produced at normal, high and low pitch, as well as with rising-falling pitch. Voice recordings were obtained using the Computerized Speech Lab (CSL) station 4300B, using a sampling frequency of 50 kHz and 16-bits of resolution. For the purpose of this work, only the /a/ vowel at normal pitch is considered. Additionally, a

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

subset of the database was segmented by a speech therapist, removing those recordings with a low dynamic range or interferences, and selecting registers according to an age balance. After this selection a total of 737 male patients (229 normal and 508 pathological), and 1011 female patients (396 normal and 615 pathological), were chosen.

B. sex-dependent pathology detector

The general scheme followed in this paper is presented in Fig 1. The aim is to include information about the sex of the speaker in the design of the automatic pathology detectors. Specifically, the input speech signal passes through a gender detector which determines the sex of the speaker. Depending on the decision taken, a male or female pathology detector determines if the speaker has a pathological or normal condition.

Therefore, this procedure implies the definition of two *subsystems*: A ***gender detector*** and two ***sex-dependent pathology detectors*** (male and female).

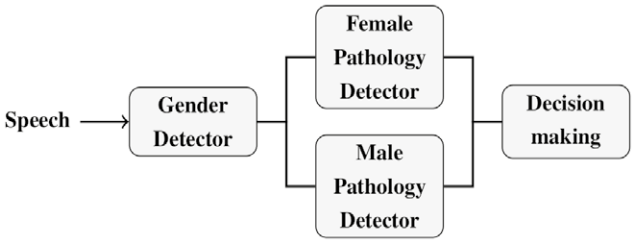


Fig 1. General scheme of the proposed system, composed by a gender detector and two sex-dependent pathology detectors.

C. Methodology

For simplicity, the sex detector and sex-dependent pathology detectors are designed in the same manner. Fig. 2. presents the methodological stages followed by both subsystems, while a detailed explanation of those stages is presented next:

In the ***preprocessing*** stage, all speech signals are [-1,1] normalized and down-sampled to 25 kHz. Next, a short time analysis is carried out using 50% overlapped Hamming windows 40 ms long.

In the ***decomposition*** stage, and from the resulting speech frames, the glottal waveform and the vocal tract model are further extracted via inverse filtering, for which (F_s/1000)+2 coefficients are used for modeling the vocal tract, and 4 coefficients are used for modeling the glottal waveform, being F_s the sampling frequency of the recording (25kHz after the resampling). The iterative Adaptive Inverse Filtering technique [5] is employed, which aims at iteratively refining the vocal tract model and the glottal signal to produce a good estimate of the glottal waveform.

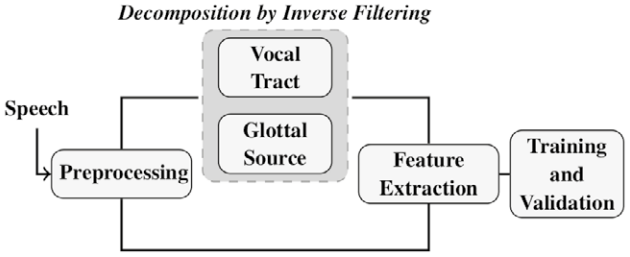


Fig 2. Detailed schema of the methodological stages followed by both detectors (sex and pathology detectors adopt the same stages).

Next, the glottal waveform and the speech are used as inputs of the ***feature extraction*** stage, where both signals are characterized by means of MFCC coefficients, varying its number in the interval [12:2:20]. Additionally, the vocal tract model obtained in the inverse filtering is considered as a feature vector by itself.

In this manner, three experiments are defined:

- *Characterization of raw speech frames*,
- *Fusion of the parameters extracted from the glottal waveform and the vocal tract model*
- *Fusion of the parameters extracted from the speech and glottal waveform, and the vocal tract model*.

Finally, in the ***training and validation*** stage, a 7-fold cross-validation is considered, calculating the classifier accuracy, α, within a confidence interval q. The q range is estimated as $q = \pm 1.96\ \alpha(1-\alpha)/N$, where N is the total number of classified patterns.

Specificity (s_p), sensitivity (s_e), receiver operating characteristic curves (ROC) and area under ROC (AUC) are calculated as well.

The classification scheme is based on simple Gaussian mixture models (GMM), tuned separately for each one of the subsystems: For the gender detector, the number of gaussians is varied in the following set: {3,5,7}. Similarly, for the sex-dependent pathology detection systems, the number of gaussians is varied in the set: {14,21,28,35,42,48}.

Finally, and for the sake of comparison, a *baseline* system which does not employ sex information and which resembles the traditional approach is utilized. The same parameters employed in the sex-dependent pathology detector are used in this baseline system.

III. Results

A. Baseline

The ROC curve for the three features which achieved the best performance for the baseline system are shown in Fig. 3. The best classification accuracy is 85.23±0.6 % and its AUC=0.8.

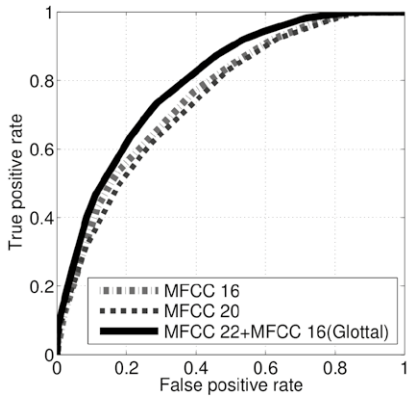


Fig 3. ROC curve for the three features that achieved the best performance in the sex detector.

B. Proposed system

Since the gender detector constitutes the first stage in the design of the proposed system, the ROC curves of the three features which achieved the best performance, are shown in Fig. 4.

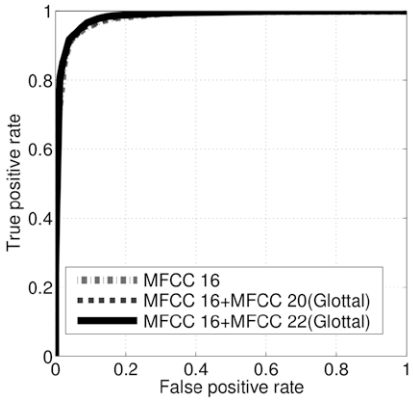


Fig 4. ROC curve for the three features that achieved the best performance in the sex detector.

Despite the performance between the three feature sets is almost indistinguishable, the best result is obtained when fusing 22 MFCC extracted from speech and 16 MFCC extracted from the glottal components. In this particular case, the classification accuracy is 94.01±1.1%, the AUC=0.98 and s_p=s_e=0.94. This operation point is, from now on, used for the gender recognizer which feeds the sex-dependent pathology detectors.

Now the results for the proposed female and male pathology detection systems are presented. The ROC curve for the three features which achieved the best performance for the female detection system are shown in Fig. 5, while the ones for the male detection system are show in Fig. 6.

The best results for the female detection system are when fusing 14 MFCC extracted from speech and 14 MFCC extracted from the glottal source, achieving an accuracy of 86.3±0.9% and AUC 0.80.

On the other hand, and when considering the male pathological detection system, the best results are obtained when fusing 18 MFCC features extracted from the voice signal and 14 MFCC extracted from the glottal waveform. In this case the accuracy is 89.45±0.9% and AUC 0.80.

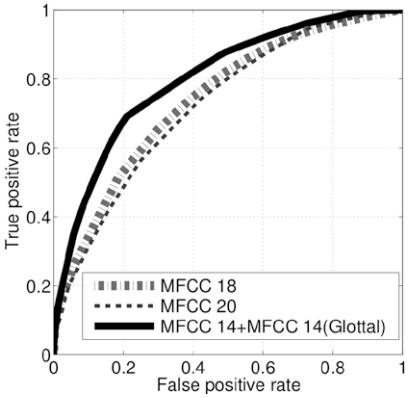


Fig 5. ROC curve for the three features that achieved the best performance in the female pathology detection system.

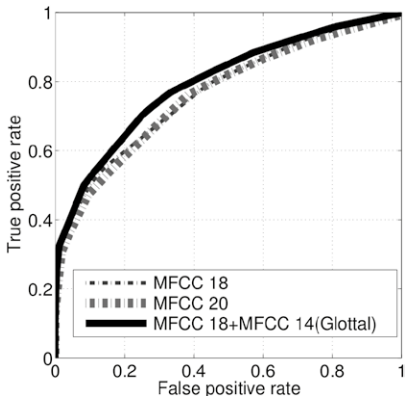


Fig 6. ROC curve for the three features that achieved the best performance in the male pathology detection system.

IV. DISCUSSIONS AND CONCLUSION

This paper presents preliminary results of a sex-dependent automatic pathology detection system, composed by a gender recognizer and two sex-dependent pathology detection systems. The methodology relies on features extracted from both speech and glottal waveforms.

Regarding the gender recognizer, the achieved classifier performance (up to 95% classification

accuracy) suggests the usefulness of the proposed system for automatic sex recognition.

Regarding the sex-dependent recognizer, results suggest that the inclusion of the sex detector in the design of automatic pathology detection systems improves performance, compared to the baseline system. In particular, those improvements were between 1% and 3% in classification accuracy.

It is observed from ROC curves of all experiments, that a lightly better curve is obtained when fusing features extracted from glottal and vocal tract components, rather than using features extracted from the raw speech signal. However, those improvements were not significant and do not provide conclusive evidence on whether or not a characterization of such a kind provides a better performance.

It was also found that the female recognizer provided a lower classification accuracy compared to its male counterpart. Aforementioned finding is in line to that obtained in [3], where the recognition of female speakers seemed to be "harder" than in male ones.

Finally, the best results for the female, male, as well as for the baseline system, are summarized in Table 1.

TABLE 1. Best results obtained among all tested experiments.

System	Set of Features	α	AUC	s_p	s_e
Baseline	22(Voice) +16(Glottal)	85.23 $\pm$ 0.6	0.80	0.84	0.86
Female	14(Voice) +14(Glottal)	86.30 $\pm$ 0.9	0.80	0.87	0.83
Male	18(Voice) +14(Glottal)	89.45 $\pm$ 0.9	0.80	0.93	0.80

This paper has presented a sex-dependent automatic pathology detection system, composed by a gender recognizer and two sex-dependent pathology detection system. The methodology employs MFCC features extracted from both speech and glottal waveforms. It was found that the usage of features extracted from the glottal source, fused with the vocal tract model, provided a slight but not significant improvement in the classification accuracy for both gender and sex-dependent pathology detection tasks, compared to the use of raw speech. However, the differences with the raw speech signal are minimal, and therefore are not conclusive on if they performed better than using the voice signal alone. This is evident, for instance, in Fig. 4, where the ROC curves for features extracted from raw speech, are almost indistinguishable, compared to those curves calculated fusing glottal features and vocal tract components.

Results suggest that the proposed sex-dependent pathology detection system performs better than the traditional sex-independent pathology detection system. Those results are in line with the ones presented in [3] where, with a different database, an improved performance in the detection of pathologies was found, when considering a differentiation between male and female subjects.

It is also important to remark that this is a first approach in the design of sex-specific pathology detectors, and thus the selection, of either the number of gaussians and the number of MFCC coefficients, has been guided by previous works which have used different databases. Therefore, care has to be taken with the results found so far, since the optimal operational point is still under research, and further evidences are necessary to provide extra insight on the validity of the findings, and on the usefulness of the methodology.

With that in mind, a wider range of MFCC coefficients for characterization, as well as a wider range of gaussians components for classification, have to be considered. Additionally, some experimentation with other linear and nonlinear features is of interest, as well as the analysis with feature extraction and selection techniques. However, that remains as future work.

ACKNOWLEDGEMENT

This research was carried out under grants: *Ayudas para la realización del doctorado* (RR01/2011) from Universidad Politécnica de Madrid, TEC2009-14123-C04 and TEC2012-38630-C04-01 from the Spanish Ministry of Education.

REFERENCES

[1] J. I. Godino-Llorente, et al. "An integrated tool for the diagnosis of voice disorders." Medical engineering & physics, vol. 28, no. 3, pp. 276–89, May 2006.

[2] D. Childers and K. Wu, "sex recognition from speech. part ii: Fine analysis," The Journal of the Acoustical society of America, vol. 90, p. 1841, 1991.

[3] R. Fraile, et al, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of MFCC parameters and differentiation of patients by sex." Folia phoniatrica et logopaedica, vol. 61, no. 3, pp. 146–52, 2009.

[4] "Saarbruecken voice database." [Online]. http://www.stimmdatenbank.coli.uni-saarland.de/index.php4

[5] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," Speech Communication, vol. 11, no. 2-3, pp. 109–118, Jun. 1992.

[6] M. Airas, "TKK Aparat: an environment for voice inverse filtering and parameterization." Logopedics, phoniatrics, vocology, vol. 33, no. 1, pp. 49–64, Jan. 2008.

[7] J. Walker and P. Murphy, "A review of glottal waveform analysis," Progress in nonlinear speech processing, pp. 1–21, 2007.

VOICE BIOMETRICAL MATCH OF TWIN AND NON-TWIN SIBLINGS

Eugenia SanSegundo[1], Pedro Gómez-Vilda[2]
[1]Phonetics Lab., Institute of Language, Literature and Anthropology, Spanish National Research Council (CSIC)
C/ Albasanz 26-28, 28037 Madrid, Spain
[2]NeuVox Laboratory, Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28223 Pozuelo de Alarcón, Madrid, Spain
e-mails: eugenia.sansegundo@cchs.csic.es, pedro@fi.upm.es

Abstract: **The similarity in twins' voices has always been an intriguing issue in forensic speaker matching, and has become an important research matter recently. The present work is a preliminary study of exploratory character diving into the similarities of monozygotic (MZ) and dizygotic (DZ) twins' phonation under the point of view of vocal fold biomechanics. The study extends to other siblings' and unrelated speakers' phonation. Estimates of biomechanical parameters obtained from vowel fillers are used to produce bilateral matches between MZ and DZ twins and siblings, and unrelated speakers. These results show interesting relationships regarding genetic load and ambient factors in the adoption of phonation styles.**

Keywords: **voice production, forensic pattern matching, phonation styles, glottal source features, twins' voice.**

I. INTRODUCTION

Recent studies in voice quality are conducted towards the evaluation of phonation performance in relation to either professional voice care, or in meta-acoustic knowledge (neurological deterioration, emotion detection, forensic applications, etc.) These fields of study are becoming more and more demanded nowadays. The aim of the present work is to study the similarities and differences of phonation characteristics in twins' voices, including monozygotic (MZ) as well as dizygotic (DZ) twins for specific forensic use, not disregarding other fields of application, as the clinical one, although this is not the main aim of the paper. A reference to previous work on twin voice quality analysis and vocal performance of interest is that of Van Lierde et al. [1]. The quality measurements used were perceptual GRBAS, breathing performance, fundamental frequency, jitter and shimmer, and the Dysphonia Severity Index. However, the study focused only on monozygotic siblings (MZ). Another relevant reference is that of Cielo et al. [2], although the twin sample used was quite small (2 MZ pairs, one per gender). Their analysis is interesting as far as they use some features not been considered in twins' voice studies before, namely vocal onset and harmonic characterization. The work of Fuchs et al. [3] found that the voices of MZ twins showed more similarity among themselves than those of non-similar speakers regarding vocal range, highest and lowest fundamental frequency, prosodic pitch line, maximum intensity, number of overtones and intensity vibrato.

The study of twins' voices can be approached from many perspectives. Stemming from a typical phonetic division, they may be classified into perception, acoustics or articulation. Some of the acoustic-related studies dealing with voice-quality or glottal parameters have been reviewed in [[4]]. Since perceptual or articulation-based approaches are less relevant for the purpose of this work, we will consider those studying twins' voices from an automatic perspective. The system by Scheffer et al. [[5]] was able to identify twins with a good performance (85% of correct identifications) using MFCC (Mel Frequency Cepstrum Coefficients). The residual error (speakers who were not correctly detected as twins of their actual twins) would suggest that "the twin of a speaker is not necessarily the most difficult impostor for an automatic speaker recognition system" ([[5]]: 2). The automatic system by Ariyaeeinia et al. [[6]] used LPCC (Linear Predictive Coding-Derived Cepstral) parameters, and the speaker representation was based on adapted Gaussian Mixture Models (GMMs). The results showed that the use of long test utterances led to smaller error rates than short ones. Both KyungWha [[7]] and Künzel [[8]] used *Batvox*; the former studied Korean female twin pairs (17 MZ, including 1 triplet and 5 DZ) and the latter studied German male and female twin pairs. The results in [[7]] showed that every twin speaker was correctly identified in the same speaking style condition (reading speech). The performance of the system in [[8]] was better for male than for female voices.

The present work focuses on studying phonation marks (including biomechanical parameters) of relevance in the biometrical description of phonation [[10], [11]]. The working hypothesis is that phonation cycle quotients and biomechanics may offer differentiation capabilities among MZ, DZ and control speakers not explored already. The paper is organized as follows: A description of the materials and methods used in the study is given in section II. In section III results obtained from the bilateral tests and matches of 16 male speakers are discussed. Conclusions are presented in section IV.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)

II. METHODS

Recordings from 40 male native speakers of Spanish (spontaneous conversation) were taken at a sampling rate of 44,100 Hz and 16 bits using HQ microphones in an isolated room. The distribution of speakers was: 7 MZ pairs, 5 DZ pairs, 4 pairs of non-twin siblings and 4 pairs of controls (non-relatives). Spontaneous fillers (long [ε] vowels maintained during more than 200 ms produced by speakers in words like "que", "de", or in hesitation marks like "eeh..." etc.) were used in the study. Recordings from two sessions separated by a 3-week interval were taken per speaker. Speech recordings were around 10 min long, an average of 8-10 fillers found in each recording.

A set of biomechanical parameters as body and cover dynamic mass and stiffness was estimated from the glottal source by inverse filtering [9]. The inter-cycle unbalances of these parameters were also used. Open, Close and Return Quotients were added to the parameter set as well as Contact Gap Defects. The parameter set was completed with jitter, shimmer and NHR ratio to produce a feature vector of 65 parameters given as x_{sij}, where s refers to the subject, i is for the session, and j for the filler. Pair-wise parameter matching experiments were carried out by likelihood ratio contrasts used in forensic voice matching [11]. The test is based on two-hypotheses contrasts: that the conditional probability between voice samples $Z_a=\{x_{aij}\}$ and $Z_b=\{x_{bij}\}$ (from the two subjects under test, a and b) is larger than the conditional probability of each subject relative to a Reference Speaker's Model Γ_R in terms of logarithmic likelihood

$$\lambda_{ab} = \log\left[\frac{p(\mathbf{Z}_b \mid \Gamma_a)}{\sqrt{p(\mathbf{Z}_a \mid \Gamma_R)p(\mathbf{Z}_b \mid \Gamma_R)}}\right] \quad (1)$$

where conditional probabilities have been evaluated using Gaussian Mixture Models (Γ_a, Γ_b, Γ_R) as

$$\begin{aligned} p(\mathbf{Z}_b \mid \Gamma_a) &= \Gamma_a(\mathbf{Z}_b), \\ p(\mathbf{Z}_a \mid \Gamma_R) &= \Gamma_R(\mathbf{Z}_a), \\ p(\mathbf{Z}_b \mid \Gamma_R) &= \Gamma_R(\mathbf{Z}_b) \end{aligned} \quad (2)$$

Following this background, the Forensic Voice Evidence Evaluation Framework is a two-step process:

- Step 1. Model Generation. A model representative of the normative population set considered (male subjects between 18-52 years-old) was created on recordings $\mathbf{Z}_R=\{\mathbf{x}_{Rjk}\}$, as a Gaussian Mixture Model $\Gamma_R=\{\mathbf{w}_R, \boldsymbol{\mu}_R, \mathbf{C}_R\}$, $\mathbf{w}_R$, $\boldsymbol{\mu}_R$ and $\mathbf{C}_R$ being the set of weights, averages and covariance matrices associated to each Gaussian Probability Distribution in the set.
- Step 2. Score Evaluation. The material under evaluation will be composed of different parameterized voice samples in matrix form $\mathbf{Z_a}=\{\mathbf{x}_{aj}\}$, where $1\leq j\leq J_a$ is the sample index, each sample being a vector $\mathbf{x}_{aj}=\{x_{aj1}\ldots x_{ajM}\}$ from vowel-like segments conveniently parameterized. Similarly, the set of the correspondent speaker to be matched will be given as $\mathbf{Z_b}=\{\mathbf{x}_{bj}\}$, where $1\leq j\leq J_b$ will be the sample index, each sample being a vector $\mathbf{x}_{bj}=\{x_{bj1}\ldots x_{bjM}\}$.

The conditioned probability of a sample from speaker *a* x_{aj} matching speaker *b* will be estimated as

$$\Pr(\mathbf{x_{bj}} \mid \Gamma_a) = \frac{1}{(2\pi)^{M/2}|\mathbf{C_a}|^{Q}} e^{-1/2(\mathbf{x_{bj}}-\boldsymbol{\mu_a})^{T}\mathbf{C_s^{-1}}(\mathbf{x_{bj}}-\boldsymbol{\mu_a})} \quad (3)$$

Similarly the conditioned probability of a sample from speaker *a* matching the Reference Model will be

$$\Pr(\mathbf{x_{aj}} \mid \Gamma_R) = \frac{1}{(2\pi)^{M/2}|\mathbf{C_R}|^{Q}} e^{-1/2(\mathbf{x_{aj}}-\boldsymbol{\mu_R})^{T}\mathbf{C_s^{-1}}(\mathbf{x_{aj}}-\boldsymbol{\mu_R})} \quad (4)$$

Finally the conditioned probability of a sample from speaker *b* matching the Reference Model will be

$$\Pr(\mathbf{x_{bj}} \mid \Gamma_R) = \frac{1}{(2\pi)^{M/2}|\mathbf{C_R}|^{Q}} e^{-1/2(\mathbf{x_{bj}}-\boldsymbol{\mu_R})^{T}\mathbf{C_s^{-1}}(\mathbf{x_{bj}}-\boldsymbol{\mu_R})} \quad (5)$$

A full description of this methodology is given in [12].

III. RESULTS AND DISCUSSION

The composition of the sample was the following: 14 subjects are MZ siblings in 7 pairs (numbered as 01-02, 03-04, 05-06, 07-08, 09-10, 11-12 and 33-34), 10 subjects are DZ siblings in 5 pairs (corresponding to speakers numbered as 13-14, 15-16, 17-18, 19-29 and 45-46), 8 subjects are non-twin brothers (BS) in 4 pairs (numbered as 21-22, 23-24, 47-48 and 49-50) and 8 subjects are not known to have any familiar relationship (US), grouped also as 4 pairs (25-26, 27-28, 29-30 and 31-32). Speakers were matched in: a) different-session intra-speaker tests (I: intra-speakers), b) inter-speaker tests (O: inter-speakers). A priori expectations assume that MZ should show the largest LLRs, followed by DZ, then by non-twin siblings; non-related speakers are expected to show the lowest LLRs. The baseline is defined by a reference background set composed of 20 speakers (set B). Scores are qualified as Strong Likeness if above 1, Weak Likeness if between 1 and -1 and Unlikeness if below -1. The hypotheses tested were the following:

H1. Intra-speaker tests should show large LLRs.
H2. MZ inter-speaker tests should show large LLRs.
H3. DZ inter-speaker tests should show also large LLRs although not that large as H1 or H2.
H4. BS inter-speaker tests should show LLRs at least over the background baseline (fixed at $\lambda = -10$).
H5. US inter-speaker tests should show LLR's aligned with the background baseline.

The results of the matching tests are summarized in Table I (see end of paper). The results contradicting the strongest hypotheses (H1 and H2) are marked in bold. Four speakers out of the total of 40 appear to be in the

limit of H1 (03, 48, 49 and 50), five others show strong intra-speaker dissimilarity (04, 09, 15, 20 and 33), and one shows very strong self-dissimilarity (25), therefore 10 out of 40 do not fulfil H1. The rest of the speakers show weak or strong self-similarity in inter-session tests, fulfilling H1. Regarding H2 we find only one out of seven pairs not fulfilling it (11 vs 12). Hypothesis 3 is not fulfilled in one out of five pairs (17 vs 18). H4 is fulfilled in all four cases. Only one pair of unrelated subjects is slightly over the baseline (27 vs 28) out of 4 cases fulfilling H5. The cases affecting only to MZ siblings have been depicted in Fig. 1 for special discussion.

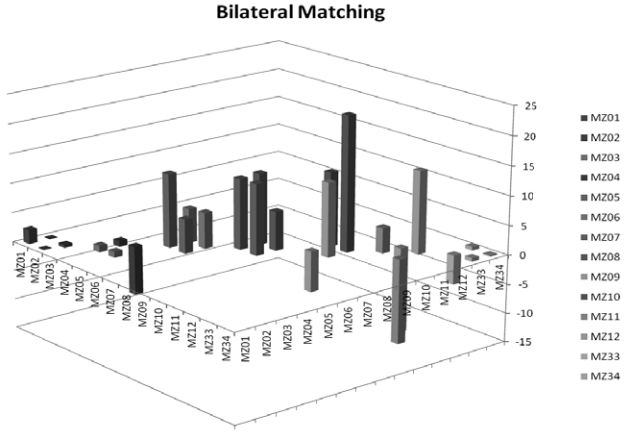


Fig. 1 Summary of the results for the MZ tests.

Three intra-speaker tests out of 13 which do not fulfil H1 correspond to relatively large negative column values (04, 09 and 33) as well as one inter-speaker test not fulfilling H2 (11 vs 12). Two twin pairs show good fulfilment of H1 and H2 (05, 06, 07 and 08), another twin pair do show a weak fulfilment of H1 and H2 (01 and 02), two twin pairs show weak fulfilment of H2, and irregular fulfilment of H1 (03, 04, 33 and 34). Another twin pair shows strong fulfilment of H2 and irregular fulfilment of H1 (09 and 10) and another pair shows good fulfilment of H1 but irregular fulfilment of H2 (11 and 12). Some words have to be said about intra-speaker fulfilment of H1: it is unclear why 10 out of 40 speakers do show self-unlikeness in a larger or smaller extent when one session phonation is tested against another. Several reasons have been considered, as changes in phonation due to emotional stress or even temporary pathological conditions. Excluding weak self-unlikeness the number of cases would be 6 out of 40, which is still a large figure. Possibly some normalization on the selection of the speaker's most characteristic phonation patterns could help in reducing this apparently large value. Regarding H2 the number of non-fulfilments seems smaller (1 out of 7 pairs). Reasons for dissimilarities in MZ within-pair comparisons seem somehow different. The most plausible reason that we can pinpoint is the nature-nurture dichotomy: in other words, the behavioural component of phonation as opposed to genetic reasons (phonation characteristics may be due to learned styles as much as to biological imprinted patterns).

IV. Conclusions

The results of the study show some interesting considerations. Regarding H1 it seems that there are certain speakers who do not show strong intra-speaker similarity (6 out of 40 are in this situation). The immediate reflection is if these could be labelled as "goats" in Doddington's Zoo [[13]]. As far as H2 is concerned it seems that most MZ twins show reasonable inter-speaker (within-pair) similarity except in one pair out of 7. Whether this could be due to behavioural rather than to genetic factors is an open question. In DZ twins (H3) the situation is similar (only 1 out of 5 pairs show low inter-speaker scores). Non-twin brothers fulfil H4 relatively well, since all 4 pairs considered showed scores over the background baseline. Finally non-relative subjects showed scores well around the background baseline giving a good description of what would be considered the normal situation in unrelated speakers. A possible complementary explanation involves the 65 parameter set in such comparisons where some of them may show a greater influence from both genetic and environmental factors. If only the comparisons of MZ twin pairs had yielded large matches, the only explanation possible would be genetic influence. However, the fact that similar values are obtained for MZ and DZ twins cannot lead to that conclusion. The impact of external factors (like a similar living and educational environment, same age, etc.) may be more relevant than it may be thought a priori in this kind of voice studies. Further research would be necessary in order to study the role of each specific parameter intervening in the results, and to extend the study to more speakers.

Acknowledgments: This work is supported by an FPU grant from the Ministry of Education, a grant from the International Association for Forensic Phonetics and Acoustics, and by grants TEC2009-14123-C04-03 and TEC2012-38630-C04-04 from *Plan Nacional de I+D+i*, Ministry of Economy and Competitiveness of Spain.

References

[1] Van Lierde, K. M., Vinck, B., De Ley, S., Clement, G., and Van Cauwenberge, P. "Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach", *Journal of Voice*, Vol. 19, No. 4, 2005, pp. 511-518.

[2] Cielo, C. A., Agustini, R. and Finger, L. S., "Características vocais de gêmeos monozigóticos", *Revista CEFAC*, Vol. 14, No 6, 2012, pp. 1234-1241 (in Portuguese, summary in English).

[3] Fuchs, M., Oeken, J., Hotopp, T., Täschner, R., Hentschel, B. and Behrendt, W., "Die Ähnlichkeit

monozygoter Zwillinge hinsichtlich Stimmleistungen und akustischer Merkmale und ihre mögliche klinische Bedeutung", *HNO*, Vol. 48, No. 6, 2000, pp. 462-469.

[4] San Segundo, E. and Gómez-Vilda, P., "Voice Biometrical Match of Twin and non-Twin Siblings", *Proc of the 1st Multidisciplinary Conf. of Users of Voice, Speech and Singing*, Las Palmas de Gran Canaria, 27-28 June 2013, pp. 132-136.

[5] Scheffer, N., Bonastre, J-F., Ghio, A. and Teston, B. "Gémellité et reconnaissance automatique du locuteur", *Actes, Journées d'Etude sur la Parole (JEP)*, 2004, pp. 445-448.

[6] Ariyaeeinia, A., Morrison, C., Malegaonkar, A., Black, S., "A test of the effectiveness of speaker verification for differentiating between identical twins", *Science and Justice*, 48, 2008, pp. 182-186.

[7] KyungWha, K., "Automatic speaker identification of Korean female twins", *Proc. 19th Ann. Conf. of the Int. Assoc. for Forensic Phon. and Acoust. (IAFPA)*, Trier, Germany, 18-21 July 2010, p. 21.

[8] Künzel, H., "Automatic speaker recognition of identical twins", *Int. Journal of Speech Language and the Law*, Vol. 17, No. 2, 2010, pp. 251-277.

[9] Gómez, P., Fernández, R., Rodellar, V., Nieto, V., Álvarez, A., Mazaira, L. M., Martínez, R, and Godino, J. I., "Glottal Source Biometrical Signature for Voice Pathology Detection", *Speech Comm.*, Vol. 51, 2009, pp. 759-781.

[10] Gómez, P., Rodellar, V., Nieto, V., Martínez, R., Álvarez, A., Scola, B., Ramírez, C., Poletti, D., and Fernández, M., "BioMet®Phon: A System to Monitor Phonation Quality in the Clinics", *Proc. eTELEMED 2013: The Fifth Int. Conf. on e-Health, Telemedicine and Social Medicine*, Nice, France, 2013, pp. 253-258.

[11] González, J., Rose, P., Ramos, D., Toledano, D. T. and Ortega, J., "Emulating DNA: Rigurous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", *IEEE Trans. On Audio, Speech and Lang. Proc.*, Vol. 15, No. 7, 2007, pp. 2104-2115.

[12] Gómez, P., Mazaira, L. M., Hierro, J. A. and Nieto, R., "Distance Metrics in Voice Forensic Evidence Evaluation using Dysphonia-relevant Parameters", *Proc of the VI Jornadas de Reconocimiento Biométrico de Personas*, Las Palmas de Gran Canaria, January 26-27, 2012, pp. 169-178.

[13] Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D., "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation", NIST, Gaithersburg, MD, 1998.

Table 1. Summary of the results for the different tests. MZ: Monozygotics; DZ: Dizygotics; RS: Related Siblings; US: Unrelated Speakers; (I): intra-speaker tests; (O): inter-speaker tests. Divided columns are used for each pair member. Cases: xxvyy means speaker xx versus speaker yy. Matches: Strong Likeness (SL): $\lambda \geq 1$; Weak Likeness (WL): $-1 \leq \lambda < 1$; Unlikeness (UL): $\lambda < -1$. In bold: results contrary to hypotheses H1 and H2 (MZ should be SL or WL, Intraspeaker's should be SL or WL). Hypothesis visual code: H1; H2; H3; H4; *H5*; ~H1-5.

	MZ (I)		MZ(O)	DZ(I)		DZ(O)	RS(I)		RS(O)	US(I)		US(O)
Cases	01v01/02v02		01v02	13v13/14v14		13v14	21v21/22v22		21v22	25v25/26v26		25v26
LLR	2.4	-0.5	-0.0	6.4	-0.7	1.7	0.3	5.9	-3.5	**-42.2**	-0.7	*-11.2*
Match	SL	WL	WL	SL	WL	SL	WL	SL	UL	**UL**	WL	UL
Cases	03v03/04v04		03v04	15v15/16v16		15v16	23v23/24v24		23v24	27v27/28v28		27v28
LLR	**-1.1**	**-8.3**	-1.0	**-8.7**	5.2	-3.2	6.4	-0.3	0.7	10.2	11.9	-9.7
Match	**UL**	**UL**	WL	**UL**	SL	UL	SL	WL	WL	SL	SL	UL
Cases	05v05/06v06		05v06	17v17/18v18		17v18	47v47/48v48		47v48	29v29/30v30		29v30
LLR	12.5	6.1	5.8	1.6	4.3	-10.1	2.9	**-1.2**	-5.5	-0.2	7.5	*-13.2*
Match	SL	SL	SL	SL	SL	UL	SL	**UL**	UL	WL	SL	UL
Cases	07v07/08v08		07v08	19v19/20v20		19v20	49v49/50v50		49v50	31v31/32v32		31v32
LLR	12.0	6.6	12.1	0.6	**-7.7**	-0.4	**-1.3**	**-2.5**	1.6	6.1	5.2	*-12.7*
Match	SL	SL	SL	WL	**UL**	WL	**UL**	**UL**	SL	SL	SL	UL
Cases	09v09/10v10		09v10	45v45/46v46		45v46						
LLR	**-7.0**	23.0	12.6	-1.0	0.0	3.4						
Match	**UL**	SL	SL	WL	WL	SL						
Cases	11v11/12v12		11v12									
LLR	4.3	14.1	**-14.6**									
Match	SL	SL	**UL**									
Cases	33v33/34v34		33v34									
LLR	**-5.0**	0.2	0.6									
Match	**UL**	WL	WL									

AUTHOR INDEX

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 8 th international workshop : December 16-18, 2013*
ISBN 978-88-6655-469-1 (print) ISBN 978-88-6655-470-7 (online)