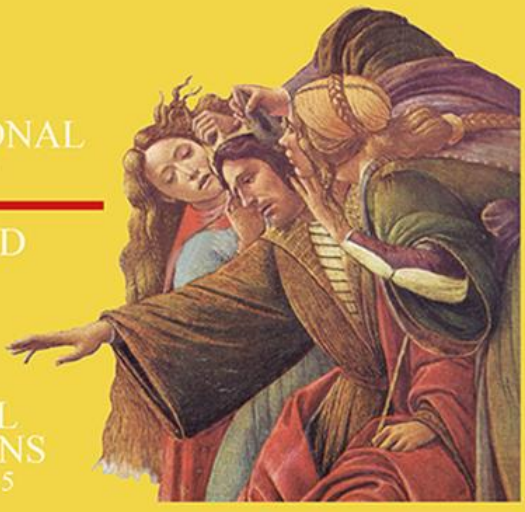
UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

9th
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

September, 2-4, 2015
Firenze, Italy



PROCEEDINGS

FIRENZE
UNIVERSITY
PRESS

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

9th INTERNATIONAL WORKSHOP

September 2-4, 2015
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2015

Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2015.
(Proceedings and report ; 105)

http://digital.casalini.it/9788866557937

ISBN 978-88-6655-792-0 (print)
ISBN 978-88-6655-793-7 (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 (www.fupress.com).

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

MAVEBA 2015

Firenze, Italy

The MAVEBA 2015 Workshop is sponsored by:

Università degli Studi di Firenze
Department of Information Engineering - DINFO

and is supported by:

XION GmbH Pankstrasse 8-10, 13127 Berlin, Germany
www.xion-medical.com

Merz Voice Business, Merz Pharmaceuticals GmbH, Eckenheimer Landstrasse 100, 60318 Frankfurt am Main, Germany www.merztraining.com

Plural Publishing, Inc. 5521 Ruffin Road, San Diego, California, 92123 USA
www.PluralPublishing.com

CONTENTS

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

MAVEBA
2015
Firenze, Italy

FOREWORD

As organizer and chairperson of this conference, I would like to express to all the participants my warmest welcome at the 9th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA2015, which takes place once again in Firenze, Italy, after 16 years since its first edition.

The MAVEBA workshop, never discontinued over the years, came into being in 1999 from the need, particularly felt, of sharing know-how, objectives and results between areas that until then seemed quite distinct such as bioengineering and medicine. It deals with all aspects concerning the study of the human voice with applications ranging from the neonate to adult and elderly.

Over the years MAVEBA has reached full maturity and the initial issues of the workshop have grown and spread also in other aspects of research such as occupational voice disorders, neurology, rehabilitation, image and video analysis. Nevertheless - and even more - it still expresses what was the original aim, namely collecting 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 papers presented at this Workshop.

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

This year and for the first time I have the honour and pleasure to chair the ninth edition of MAVEBA in conjunction with the 11th PEVOC Conference, the largest European conference for voice professionals, taking place just before MAVEBA.

MAVEBA 2015 is characterized by two workshops, concerning mechanical replicas and large-scale simulations of the vocal apparatus. Other equally important subjects are exploited in seven sessions of free papers, three of which are dedicated to the basic theme of the workshop, namely modelling and analysis of the voice signal, the other deal with the following topics: electroglottographic signals and endoscopic video images at high speed, high-pitched signals such as singing, children voice and newborn cry, mechanical models, marker vs. marker less approaches for voice analysis, emotional and therapeutical aspects.

Finally on Saturday September 5th MAVEBA 2015 hosts a one-day summer school on physics-based voice simulation hosted by the EU project EUNISON and coordinated by Sten Ternström, KTH, Stockholm, Sweden.

Therefore MAVEBA 2015 faces a broad range of issues but at the same time, and as always, with great rigor. This has allowed in the past to publish extended versions of some papers presented at MAVEBA in a Special Issue of the prestigious scientific journal Biomedical Signal Processing and Control (Elsevier Ltd.). I am confident that this tradition will continue this year too.

As always, the three intensive days of the workshop (2-4 September, 2015) will be also an opportunity for participants to visit places of Firenze not included in the traditional touristic routes. During the conference participants could visit the museum of the Military Health located inside the Convent and Cloister of San Domenico del Maglio, hosting the Congress. A welcome cocktail will be offered after the visit of the Palazzo Vecchio museum and Arnolfo's tower. A relaxing gala dinner will be given on the Arno banks, enjoying a unique view of Ponte Vecchio.

To conclude I wish to express my deepest thanks to all participants for the high level of contributions that again make this 9th MAVEBA Workshop an event of great scientific relevance worldwide. My thanks also go to the anonymous reviewers of the papers and to the Committee for workshops selection (Giovanna Cantarella Milano, Italy, Ruth Epstein London, United Kingdom, Franco Fussi Ravenna, Italy, Markus Hess Hamburg, Germany, Outi Kähkönen EVTA President, John Rubin London, United Kingdom, Johan Sundberg Stockholm, Sweden), flawlessly coordinated by Prof. Dejonckere, who have freely devoted part of their valuable time to the success of the Workshop.

Last but not least a special thank goes to my closest collaborators Andrea Bandini and Silvia Orlandi without whom this event could not have been organized and carried out with, I hope, the satisfaction of everyone.

Therefore I wish that Firenze, in addition to its artistic, cultural and landscape heritage will remain a pleasant memory of a MAVEBA 2015 inspiring and challenging meeting.

I hope to see you again in two years, still in Firenze!

Claudia Manfredi
Workshop Chair

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

This Volume of Proceedings collects all contributions presented at MAVEBA 2015 organized by date and by session according to the codes listed below.

LEGENDA

Free papers MAVEBA
M1-M3 = Models
SI = Singing; Infants
EI = EEG; Imaging
MM = Mechanical; Markerless
ET = Emotions; Therapy

Workshops MAVEBA

W1 = Workshop n.1
W2 = Workshop n.2

September 2

FP – Models 1

VOCAL FOLD COLLISION MODELING

A. Granados[1], J. Brunskog[1] , M. K. Misztal[2]

[1] Acoustic Technology, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark
[2] Niels Bohr Institute, University of Copenhagen, Copenhagen DK-2100, Denmark
algra@elektro.dtu.dk

Abstract: **When vocal folds vibrate at normal speaking frequencies, collisions occurs. The numerics and formulations behind a position-based continuum model of contact is an active field of research in the contact mechanics community. In this paper, a frictionless three-dimensional finite element model of the vocal fold collision is proposed, which incorporates different procedures used in contact mechanics and mathematical optimization theories. The penalty approach and the Lagrange multiplier method are investigated. The contact force solution obtained by the penalty formulation is highly dependent on the penalty parameter value. Furthermore, the Lagrange approach shows poor results with regard to instantaneous contact force estimation. This motivates the use of an Augmented Lagrange approach to regularize the Lagrange contact force solution. Finally, the effect of the interpenetration volume on contact force and contact area computations is illustrated.**

Keywords : **Vocal folds, collision, constrained optimization, finite element method, contact detection.**

I. INTRODUCTION

Mathematical descriptions of self-oscillating finite element models of the vocal folds have been reported in the literature (e.g., see [1]). A continuum model of the airflow coupled to a deformable three-dimensional body have been one of the main focuses. For purpose of clinical research, investigations on the mechanical conditions that arise during phonation are of special interest. At normal speaking frequencies, vocal fold collision occurs, and the tissue is affected by specific stresses and reaction forces [2]. Hence, a detailed mathematical study of the collision process is expected to contribute to a better understanding of vocal fold mechanics.

In the context of continuum mechanics, the vocal fold contact can be modeled by enforcing position-based constraints to the minimization of the total potential energy of the mechanical system. Methodologies from mathematical optimization theory can be applied in order to solve the contact constrained problem [3]. In this paper, a Penalty method and a Lagrange multiplier approach are investigated for the the case of frictionless vocal fold collision. Furthermore, a penalty regularization of the Lagrange multiplier method is carried out by the Augmented Lagrange technique applied together with an Uzawa type algorithm [3]. Finite element contact discretization and contact detection mechanism that allows for asymmetric collision are presented.

II. METHODOLOGY

A three-dimensional deformable viscoelastic model of the vocal folds driven by a Bernoulli glottal airflow is described. At each time step the new equilibrium position is found as the minimum of the total potential energy by the variational formulation. When collision occurs, the contact constrained minimization problem is solved by different methods.

A. Governing equations

In the absence of volume forces, the vocal fold deformation is described by the equation of balance

$$\nabla \cdot \boldsymbol{\sigma} = \rho \frac{\partial^2 \boldsymbol{x}}{\partial t^2} \qquad \boldsymbol{x} \in v_{\text{solid}}, \tag{1}$$

in the deformed state $v_{\text{solid}} \subset \mathbb{R}^3$, the constitutive equation for a transversely isotropic linear viscoelastic solid as in [4], and the Dirichlet and Neumann boundary conditions

$$\boldsymbol{x} - \boldsymbol{X} = 0 \qquad \text{in } \Gamma_D \subset \partial v_{\text{solid}} \tag{2.1}$$

$$\boldsymbol{\sigma} \cdot \boldsymbol{n} = \boldsymbol{p} \qquad \text{on } \Gamma_N \subset \partial v_{\text{solid}} \tag{2.2}$$

respectively, where $\boldsymbol{\sigma}$ is the stress tensor, $\boldsymbol{X}$ represents the material coordinates, $\boldsymbol{n}$ is the outward normal, and $\boldsymbol{p}$ is the aerodynamic pressure derived from Bernoulli's principle. The Dirichlet boundary where the displacement is zero is placed in the anteroposterior glottal regions; see [3] for further details. The equilibrium position can be found as the minimum of the total potential energy Π. Hence, for

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

admissible displacement variations or test functions $\boldsymbol{w}$ that vanish in the Dirichlet boundary, the weak formulation of the problem takes the form

$$\delta\Pi=\int_{v_{solid}}\rho\,\boldsymbol{w}^t\cdot\frac{\partial^2\boldsymbol{x}}{\partial t^2}\,\mathrm{d}v+\int_{v_{solid}}\nabla\boldsymbol{w}^t:\boldsymbol{\sigma}\,\mathrm{d}v-\int_{\Gamma_N}\boldsymbol{w}^t\cdot\boldsymbol{p}\,\mathrm{d}\Gamma=0, \quad (3)$$

where $\delta\Pi$ indicates the variation of the energy.

When collision between the vocal folds occurs, additional constraints may be activated on the contact area $\Gamma_C\subset\partial v_{solid}$. In order to avoid unphysical body interpenetration for a frictionless contact, non-negativeness of the normal gap between a superficial slave node $\boldsymbol{x}^s$ and a master surface placed at the opposite vocal fold may be enforced by the position-based constraint

$$g_N=(\boldsymbol{x}^s-\overline{\boldsymbol{x}}^m)\cdot\overline{\boldsymbol{n}}^m\geq 0 \qquad \text{on } \Gamma_C\subset\partial v_{\text{solid}}, \quad (4)$$

where $\overline{\boldsymbol{x}}$ is the projection of the slave node onto the master surface; see Fig. 1.

B. Contact constraint enforcement

A penalty, a Lagrangian and an augmented Lagrangian [3] methods are here studied to enforce the inequality constraint in Eq. (4). Only the Lagrangian solution enforces the collision constraint in exact form.

The penalty method consists of a minimization problem where the objective function involves the collision-free potential energy and a term which penalizes infeasible positions on Γ_C as

$$\Pi+\int_{\Gamma_c}\frac{1}{2}\kappa|g_N(\boldsymbol{x})|^2\mathrm{d}\Gamma, \quad (5)$$

where $\kappa>0$ is a penalty parameter. Optimality conditions lead to the variational formulation

$$\delta\Pi+\int_{\Gamma_c}\kappa(\boldsymbol{w}^s-\overline{\boldsymbol{w}}^m)\cdot(\boldsymbol{x}^s-\overline{\boldsymbol{x}}^m)\,\mathrm{d}\Gamma=0 \quad (6)$$

to be combined with Eq. (3). The second term above can be interpreted as minus the reaction force to avoid interpenetration. For non-adhesion contact, the reaction force must be compressive. Hence, it can be seen that as the penalty parameter tends to infinity the normal gap tends to zero, and the optimal of the new minimization problem approaches the exact equilibrium solution at collision. However, large penalty parameters may lead to ill-conditioning of the global matrix.

The Lagrangian method solves the inequality constrained problem by solving the optimization problem with objective function

$$L(\boldsymbol{x},\boldsymbol{\Lambda})=\Pi+\int_{\Gamma_c}\boldsymbol{\Lambda}\,g_N(\boldsymbol{x})\,\mathrm{d}\Gamma, \quad (7)$$

called the Lagrangian function, where $\boldsymbol{\Lambda}$ is the Lagrange multiplier vector, also called dual variables.

Optimality conditions to the problem are

$$\delta\Pi+\int_{\Gamma_c}\boldsymbol{\Lambda}(\boldsymbol{w}^s-\overline{\boldsymbol{w}}^m)\cdot\overline{\boldsymbol{n}}^m\,\mathrm{d}\Gamma=0$$
$$\boldsymbol{\Lambda}\leq 0$$
$$\boldsymbol{\Lambda}\cdot g_N(\boldsymbol{x})=0, \quad (8)$$

which are known as the Karush-Kuhn-Tucker conditions for optimality. Note that the Lagrange multiplier vector can be seen as the compressive reaction forces. From a physical point of view, the last condition indicates that no contact forces are active when the normal gap is positive, and the non-penetration constraint is fulfilled in exact form whenever collision occurs. However, this approach introduces additional unknowns in the form of Lagrange multipliers. Furthermore, the Lagrangian approach in Eq. (8) is a non-smooth contact formulation, and regularization techniques may be used to improve results.

The Augmented Lagrange formulation combines the Lagrange and the penalty approaches, without additional unknowns. A simplified version is the Uzawa algorithm [3] which may be summarized as follows. For an initial Lagrange multiplier vector $\boldsymbol{\Lambda}_k$, a new equilibrium is found by minimization of

$$\Pi+\int_{\Gamma_c}\boldsymbol{\Lambda}_k\,g_N(\boldsymbol{x})\,\mathrm{d}\Gamma+\int_{\Gamma_c}\frac{1}{2}\kappa|g_N(\boldsymbol{x})|^2\mathrm{d}\Gamma, \quad (9)$$

where the last penalty term can be seen as a regularization term for non-smoothness. The Lagrange multiplier vector is updated in an augmentation iteration as

$$\boldsymbol{\Lambda}_{k+1}=\boldsymbol{\Lambda}_k+\min\{\kappa\,g_N(\boldsymbol{x}_{k+1}),\boldsymbol{\Lambda}_k\}, \quad (10)$$

where $\boldsymbol{x}_{k+1}$ is the solution of the minimization problem. The update in Eq. (10) can be seen as a gradient ascent algorithm, as the critical point of the Lagrangian in Eq. (7) occurs at a maximum over the multipliers [5]. As the contact constraint is not solved in an exact form, the augmentation procedure in Eq. (10) continues until a convergence criterion for $g_N(\boldsymbol{x}_{k+1})$ is fulfilled. Furthermore, the penalty parameter can be increased at each augmentation step to speed up the convergence rate. However, to avoid ill-conditioning of the system matrix due to a large

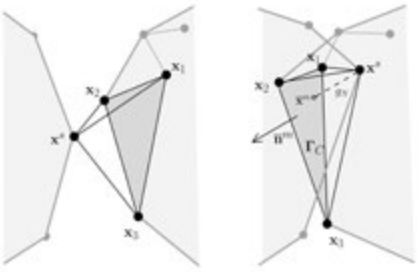


FIG. 1. Conforming interface mesh for collision detection; before (left) and after collision (right).

penalty parameter value, a maximum number of augmentations must be requested.

C. Spatial and temporal discretization

The spatial finite element discretization is based on a tetrahedral mesh. Hence, the interface domain is formed by triangular elements. As Fig. 1 illustrates, a coarse conforming interface tetrahedral mesh may be defined to detect contact; whenever an oriented interface element volume is inverted, collision occurs. Once the slave node $\boldsymbol{x}^s$ and the master surface with vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are detected, by means of an isoparameteric transformation with linear basis functions $N_i(\xi,\zeta)$ defined on a reference triangular element, the projection $\overline{\boldsymbol{x}}^m$ corresponds to the local coordinates $(\overline{\xi},\overline{\zeta})$, and a contact element matrix

$$\boldsymbol{g}^e=\left(\overline{\boldsymbol{n}}^m-N_1(\overline{\xi},\overline{\zeta})\overline{\boldsymbol{n}}^m-N_2(\overline{\xi},\overline{\zeta})\overline{\boldsymbol{n}}^m-N_3(\overline{\xi},\overline{\zeta})\overline{\boldsymbol{n}}^m\right) \quad (11)$$

with $\boldsymbol{g}^e\cdot(\boldsymbol{x}^s\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3)\geq 0$ contributes to the assembled global constraint contact matrix $\boldsymbol{G}$.

The penalty approach in Eq. (6) can be simplified further in the way that follows. Once a negative oriented element volume V^e is found, the compressive reaction force on a colliding element may be approximated numerically as

$$-(\boldsymbol{g}^e)^t\,\kappa\,\boldsymbol{g}^e\cdot(\boldsymbol{x}^s\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3)^t\approx\overline{\boldsymbol{n}}^m(-1\ \ 1\ \ 1\ \ 1)^t\,\frac{\kappa V^e}{4} \quad (12)$$

Hence, a matrix $\boldsymbol{F}_c$ can be assembled. For global mass, damping, and stiffness matrices $\boldsymbol{M}$, $\boldsymbol{C}$, and $\boldsymbol{K}$, respectively, and $\boldsymbol{F}$ a vector of applied aerodynamic forces, the finite element system of the penalty approach is

$$\boldsymbol{M}\ddot{\boldsymbol{x}}+\boldsymbol{C}\dot{\boldsymbol{x}}+\boldsymbol{K}(\boldsymbol{x}-\boldsymbol{X})=\boldsymbol{F}+\boldsymbol{F}_c \quad (13)$$

The optimality condition for a Lagrangian approach in Eq. (9) consists of the equations

$$\begin{aligned}\boldsymbol{M}\ddot{\boldsymbol{x}}+\boldsymbol{C}\dot{\boldsymbol{x}}+\boldsymbol{K}(\boldsymbol{x}-\boldsymbol{X})+\boldsymbol{G}^t\boldsymbol{\Lambda}&=\boldsymbol{F}\\ \boldsymbol{G}\boldsymbol{x}&=\boldsymbol{0}\end{aligned} \quad (14)$$

When the second condition in Eq. (8) is not satisfied for all contact elements, the contact constraint is no longer, and a collision-free finite element system must be solved. The finite element discretization of the variation of the Augmented Lagrange formulation in Eq. (9) yields

$$\begin{aligned}\boldsymbol{M}\ddot{\boldsymbol{x}}_{k+1}+\boldsymbol{C}\dot{\boldsymbol{x}}_{k+1}+\boldsymbol{K}(\boldsymbol{x}_{k+1}-\boldsymbol{X})+\boldsymbol{G}^t\boldsymbol{\Lambda}_k&\\ +\kappa(\boldsymbol{G}^t\boldsymbol{G}\,\boldsymbol{x}_{k+1})=\boldsymbol{F},&\end{aligned} \quad (15)$$

where use is made of Eq. (14) and Eq. (6).

The temporal discretization scheme implemented for calculations is the Hilbert-Hughes-Tavlor $\alpha-$method. The parameters employed are $\alpha=-0.3$ and a time step increment $h=50\ \mu s$. These values give good accuracy and introduce advantageous numerical damping. Further details can be found in [4].

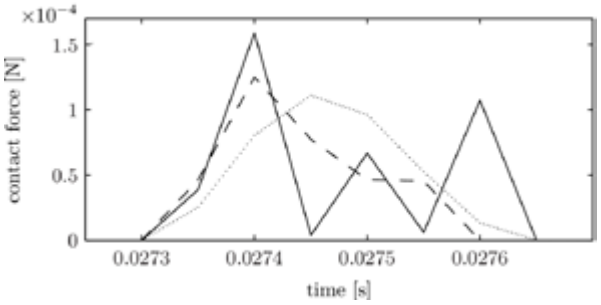


FIG. 2. Mediolateral coordinate of the contact force applied to an interface node. Solid line indicates the results for a Lagrange formulation; dotted line indicates the results for a penalty formulation with $\kappa=10^7$; dashed line indicates an Augmented Lagrangian formulation with 4 augmentations.

III. Results and Discussion

For all simulation, the tissue, geometry and initial conditions can be found in [4]. With regard to the augmented Lagrange technique, augmentations of a Lagrange multiplier associated to a slave node stop when the corresponding normal gap is less than 10^{-5}. The initial Lagrange multiplier vector is set to zero. The initial penalty parameter is 1, which increases by a factor of 10 when the total intersection volume is reduced by less than a 75% at each augmentation step.

The performance of different methods for contact constraint enforcement with regard to contact force estimations is illustrated in Fig. 2. The mediolateral component of the contact force applied to the interface node at initial position $(0.024,-0.136,-0.037)$ as a function of time is shown, for a subglottal pressure of 0.8 kPa. The results obtained by a penalty method with $\kappa=10^7$ are shown in dotted line; the Lagrange multiplier method, in solid line; the Augmented Lagrange formulation with a maximum of 4 augmentations, in dashed line. Comparison between the penalty and Lagrange reaction force solution, makes apparent a spurious non-smooth behavior of the Lagrange multiplier solution. From physical considerations, a smooth transition at each time step may be expected. Consequently, the Lagrange approach may lead to wrong estimations of the instantaneous contact force. In an effort to improve this unsatisfactory behavior, the Augmented Lagrange approach seems to have a regularization effect with 4 augmentations per time step.

Fig. 3 shows the maximum mediolateral component of the total contact force calculated from the summation over all nodal contact forces. Circles

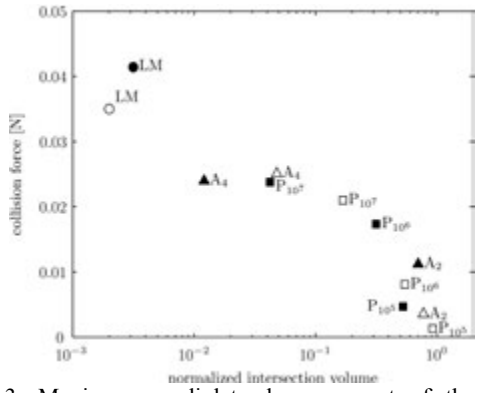


FIG. 3. Maximum mediolateral component of the total contact force as a function of the normalized intersection volume. The selected collision time interval is [0.036, 0.037], and the right vocal fold (x > 0) has been used for calculations.

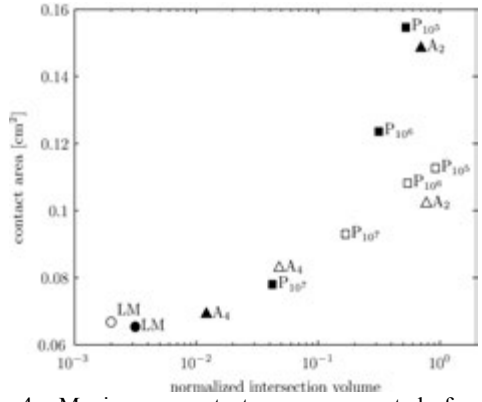


FIG. 4. Maximum contact area computed from the summation of interface triangles with any vertex loaded by a collision force. Symbols are the same as in Fig. 3.

indicate results for the Lagrange multiplier method (LM); squares correspond to a penalty formulation (P), where the subscript corresponds to the value of the penalty parameter; triangles correspond to the Augmented Lagrange approach (A), where the subscript indicates the maximum number of augmentations. Black and white marks show the results for a subglottal pressures of 0.8 kPa and 0.6 kPa, respectively. The horizontal axis corresponds to the interpenetration volume normalized to the maximum intersection volume when the effect of contact forces is neglected. The graph shows a clear effect of the violation of position-based contact constraint on contact force estimations. The Lagrange approach gives the smallest intersection volume, although, theoretically, the intersection volume should be zero. This small error is probably due to the contact detection algorithm. Furthermore, contact force computations with a penalty approach are highly dependent on the value of the penalty parameter. When the subglottal pressure is modified, the Lagrange approach shows robustness in comparison with the penalty results. With regard to the Augmented Lagrange procedure, ideally the augmented multipliers are not dependent on the penalty parameter [3]. However, the numerical solution behaves differently, which may be due to the contact finite element computations. Robustness in the method may be introduced by enlarging the maximum number of augmentations. Nevertheless, exact contact force solution cannot be assured as the penalty parameter tends to infinity, and ill-conditioning of the system matrix may occur.

Fig. 4 shows the maximum contact area computed from the summation of interface triangles with any vertex loaded by a collision force. An influence of the interpenetration volume is apparent from the results. Again, the Lagrange multiplier method seems to be robust for subglottal pressure variations.

IV. Conclusions

Position-based contact constraints of vocal fold collision have been shown to have a clear effect on collision force and contact area estimations. The Lagrange multiplier method for contact constraint enforcement appears to be robust for pressure variations, but poor with regard to instantaneous contact force solution. An Augmented Lagrange approach with an Uzawa algorithm has a smoothing effect by introducing a penalty regularization term. However, the Penalty and the Augmented Lagrange results show strong dependency on penalty parameter choice. Alternative formulations of contact constraint may further improve contact force estimations.

REFERENCES

[1] F. Alipour, D. A. Berry, and I. R. Titze, “A finite-element model of vocal-fold vibration.”, *J. Acoust. Soc. Am.* , vol. 108, pp. 3003–3012, 2000.

[2] H. E. Gunter, “A mechanical model of vocal-fold collision with high spatial and temporal resolution,” *J. Acoust. Soc. Am.*, vol. 113, pp. 994–1000, 2003.

[3] P. Wriggers, *Computational contact mechanics*, John Wiley & Sons, 2002.

[4] A. Granados, J. Brunskog , M. K. Misztal , V. Visseq, and K. Erleben “Finite element modeling of the vocal folds with deformable interface tracking,” in *Proc. Forum Acust.* (Krakow, Poland), 2014.

[5] N. Andreasson, and A. Evgrafov, and M. Patriksson, *Introduction to Continuous Optimization: Foundations and Fundamental Algorithms,* Studentlitteratur AB, 2006.

ASPECTS OF THE GLOTTAL SOURCE CHARACTERISTICS AND CONSEQUENCES FOR THE ACOUSTICS OF THE VOCAL TRACT

M. Fleischer[1] and D. Mürbe[2,3]

[1]Department of Otorhinolaryngology, Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[2]Division of Phoniatrics and Audiology, Department of Otorhinolaryngology, Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[3]Voice Research Laboratory, University of Music Carl Maria von Weber, Dresden, Germany
mario.fleischer@uniklinikum-dresden.de
dirk.muerbe@uniklinikum-dresden.de

Abstract: **In this article, common procedures to apply an equivalent source term are compared. First, an acoustic monopol at the glottal region of models of the vocal tract, and second, a more physiological excitation, that is movement of the vocal folds are applied. It can be shown that both results in a plausible frequency behaviour of the pressure at the lips. But in contrast, the ratio of the pressure to the glottal flow shows significant differences for these different boundary conditions. This behaviour is also observable in case of changed shape of the vocal tract.**

Keywords: **vocal tract transfer function, glottal source, finite-element-modelling**

I. INTRODUCTION

To characterize the acoustic transfer characteristics of the vocal tract (VT) regarding its ability to shape the glottal source spectra, application of appropriate source conditions on numerical models are always needed.

Whereas the movement of the vocal folds is of a very complex nature, in models often an acoustic equivalent such as a monopol is used. This modelling strategy results from several thoughts, e.g. to circumvent the complexity of the formulation of a fluid-structure-interaction taking into account the material properties of the solid structures or the uncertainty of the exact movement of the vocal folds.

In this study, influences of different source conditions on the VT transfer function and the resulting pressure fields inside the VT are investigated and consequences regarding the underlying mechanisms are drawn.

II. METHODS

In this study, VT-models (including the vocal folds) of different complexity, starting from an ordinary pipe up to extra-considering the changed laryngeal area measures (Fig. 1A,C) were analyzed with respect to the impact of the glottal movement on the glottal source and, additionally, of the produced sound at the lips. Here, numerical procedures as implemented in Ansys V14 (ANSYS, INC., Canonsburg, PA) were used for computation of the acoustics. In brief, the Helmholtz equation

$$(-\kappa - \nabla^2)\tilde{p} = 0 \qquad \text{in V} \tag{1}$$

by considering two different types of boundary conditions at the glottis, namely

$$\nabla\tilde{p} \cdot \boldsymbol{n_0} = -\mathrm{j}\omega\varrho\tilde{f}_0 \qquad \text{on } \Gamma_{ga} \qquad \text{(Type I)} \tag{2}$$

$$\nabla\tilde{p} \cdot \boldsymbol{n_1} = -\mathrm{j}\omega\varrho\tilde{f}_1 \qquad \text{on } \Gamma_{gw} \qquad \text{(Type II)} \tag{3}$$

was solved.

Herein, $\tilde{p}$ is the acoustic pressure in the volume V, ∇ is the nabla operator, $\kappa = \omega/c$ is the wave number, ω is the angular frequency and c is the constant speed of sound. $\boldsymbol{n_0}$ is the outward normal vector at the glottal cross-sectional area Γ_{ga}, $\boldsymbol{n_1}$ is the outward normal vector at the wall close to the glottis Γ_{gw} and ϱ is the constant density. Further, $\tilde{f}_0$ is the normal particle velocity, $\tilde{f}_1$ represents the normal wall velocity. At the lips, the impedance was equivalent to that seen by an rigid pistion and sub-laryngeal (in case of Type II), the outgoing plane waves were fully absorbed. It should be noted, that the second option (Type II) is equivalent to the radial movement of the vocal folds as presented in the past ([1], there case I).

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

III. RESULTS

As a result, simplification of the glottal movement to an acoustic equivalent source changes the spectral characteristics of the vocal tract depending on its geometrical shape as shown in Fig. 1B,C. In case of the axisymmetric model with varying diameter of the lower vocal tract small differences in the frequency of the first and the second formant – depending on the used source – were identified (Fig. 1C; dotted vs. solid line). According to [2], adding a constriction into the lower VT, the third and fourth formant were clustered. But in contrast, the ratio of the pressure at the lips to the **resulting** glottal flow (Type II) shows a spectrum which differs from the other two spectra (Fig. 1B,C; dashed line), for example missing formants. Among others, this is caused by the resulting pressure field inside the VT (Fig. 2A&B) that depends strongly on the choosen boundary conditions.

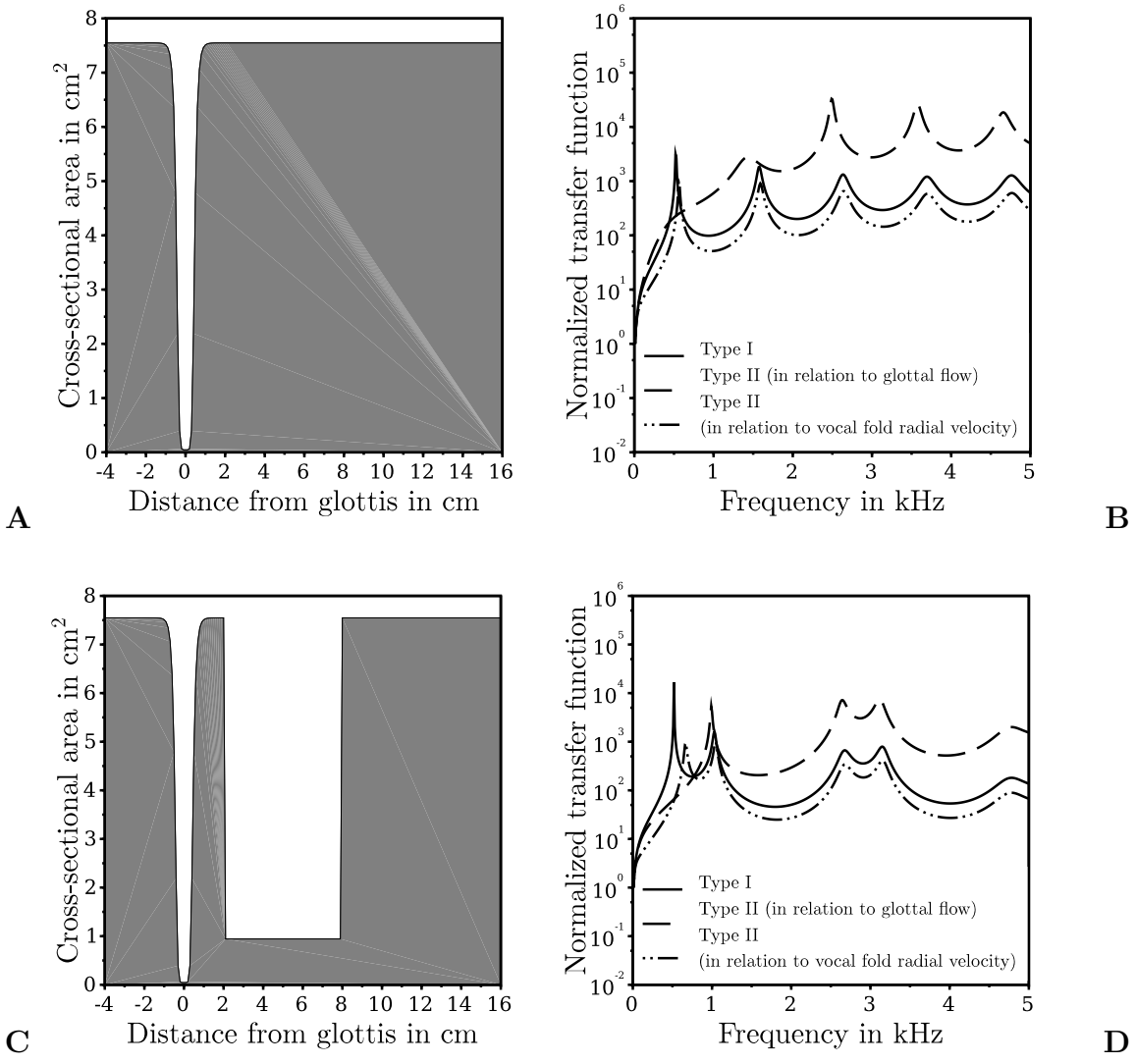


Fig. 1: **Sketch of the simple models used and associated transfer characteristics.** Sketch of an axisymmetric pipe model without (A) and with (C) varying geometrical properties of the lower larynx to investigate principle features of source production (see [1]). (B,D) Ratio of the pressure at the lips to glottal flow (Type I/solid line, Type II/dashed line) and to the vocal fold velocity (Type II/dotted line) (B corresponds to A and D to C, respectively).

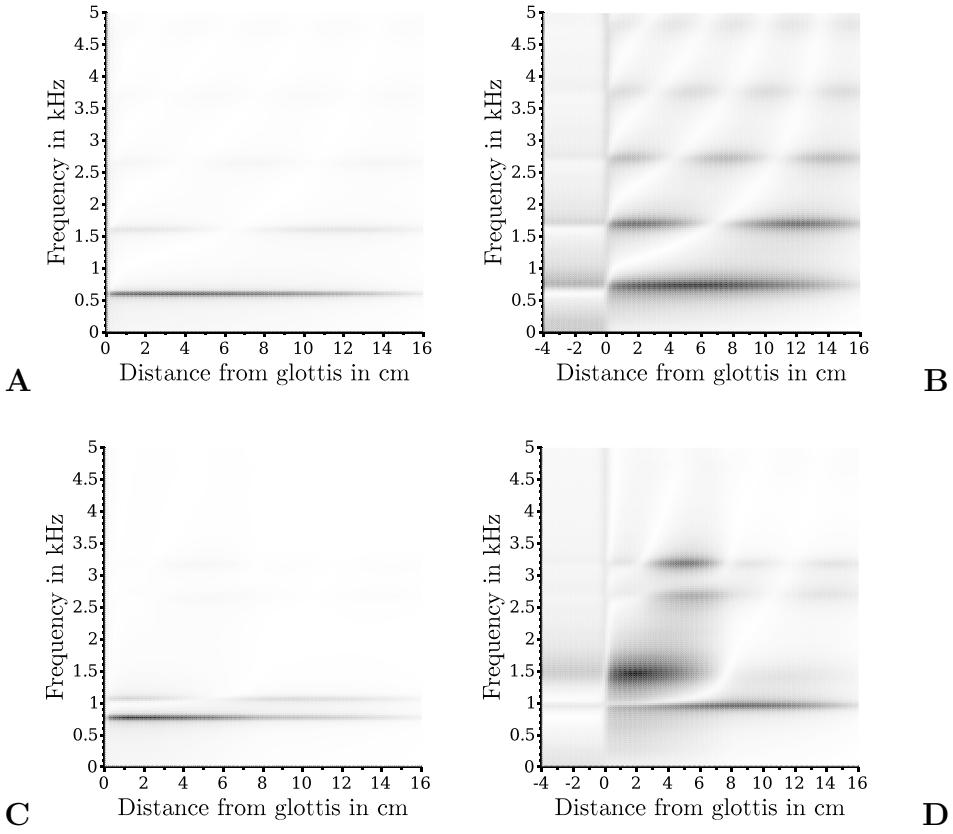


Fig. 2: **Amplitude of the normalized acoustic pressure as a function of space (along the centerline of the models) and frequency.** (A,C) Results from the model with particle velocity applicated at the glottal area and (B,D) with application of radial contraction of the vocal folds and considering the subglottal structures. (A,B) are results for the model without, and (C,D) with narrowing of the lower larynx (see Fig. 1C).

IV. DISCUSSION

As briefly summarized in the section above, the transfer characteristics of the vocal tract depend on its geometry and the source which generates and drives the acoustic waves through the vocal tract. This is caused by the impedances which are 'seen' by the source and the driving mechanisms, and more important, the resulting pressure fields inside the VT. Further work is needed to qualify the statements above by means of a more realistic three-dimensional VT-model (Fig. 3) and to identify the local pressure characteristics within the VT.

Fig. 3: **Three-dimensional model.** Three-dimensional model of a sung german vowel /a/ including a spherical lip region that represents an opened mouth.

REFERENCES

[1] W. Zhao, C. Zhang, S.H. Frankel, L. Mongeau, "Computational aeroacoustics of phonation, Part I: Computational methods and sound generation mechanisms", J. Acoust. Soc. Am., vol. 112, pp. 2134–2146, 2002.

[2] J. Sundberg, "Articulary interpretation of the 'singing formant'", J. Acoust. Soc. Am., vol. 55, pp. 838–844, 1974.

INTERPRETATION OF THE THREE-DIMENSIONAL DYNAMICS OF THE SUPERIOR SURFACE OF THE VOCAL FOLDS

D.A. Berry[1], M. Döllinger[2], F. Alipour[3]

[1]Department of Head & Neck Surgery, University of California, Los Angeles, California, United States
[2]Universitätsklinikum, Abteilung für Phoniatrie und Pädaudiologie, Erlangen, Germany
[3]Speech Pathology & Audiology, The University of Iowa, Iowa City, Iowa, United States
daberry@ucla.edu, michael.doellinger@phoni.imed.uni-erlangen.de, f-alipour-haghighi@uiowa.edu

I. INTRODUCTION

Physical mechanisms of regular and irregular vocal fold vibration were first studied using the method of empirical eigenfunctions and a computational model of vocal fold vibration [1,4]. Later, the same method was used to study the physical mechanisms of vocal fold vibration in laboratory hemilarynx studies, in which the medial surface of the vocal fold was imaged [2,3]. While the method was later extended to clinical studies of vocal vibration [5], the method exhibited significantly less interpretive power in the clinical applications in which the superior surface of the vocal folds was imaged. Our hypothesis is that the interpretive power of the method of empirical eigenfunctions in studying the superior surface dynamics of the vocal folds can be significantly increased if one performs 3D imaging of the dynamics instead of the standard 2D imaging.

II. METHODS

To test this hypothesis, the method of empirical eigenfunctions was employed on the same finite element model used in previous computational experiments [1,4]. The hypothesis was also tested on the 3D superior surface dynamics of the vocal folds from excised larynx experiments, in which the 3D imaging was performed using a laser projection system [6].

III. RESULTS

Our results confirm that the interpretive power of the method of empirical eigenfunctions in studying the superior surface dynamics of the vocal folds is significantly increased when 3D imaging is employed.

IV. DISCUSSION

These results suggest that physical mechanisms of regular and irregular vocal fold vibration may be utilized more fruitfully in clinical studies through the use of 3D, highspeed imaging.

REFERENCES

[1] Berry DA, Herzel H, Titze IR, Krischer K (1994). Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions, J Acoust Soc Am 95, 3595-3604.
[2] Berry DA, Montequin DW, Tayama N (2001). High-speed, digital imaging of the medial surface of the vocal folds, J Acoust Soc Am 110, 2539-2547.
[3] Berry DA, Zhang Z, Neubauer J (2006). Mechanisms of irregular vibration in a physical model of the vocal folds, J Acoust Soc Am 120, EL36-42.
[4] Alipour F, Berry DA, Titze IR (2000). A finite element model of vocal fold vibration, J. Acoust. Soc. Am. 108, 3003-3012.
[5] Neubauer J, Mergell P, Eysholdt U, Herzel H (2001). Spatio-temporal analysis of irregular vocal fold oscillations: biphonation due to desynchronization of spatial modes, J Acoust Soc Am 110, 3179-3192.
[6] Luegmair G, Kniesburges S, Zimmermann M, Sutor A, Eysholdt U, Döllinger M (2010). Optical reconstruction of high-speed surface dynamics in an uncontrollable environment, IEEE Trans Med Imaging 29, 1979-1991.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

FORMULATION OF A STOCHASTIC GLOTTAL SOURCE MODEL INSPIRED ON DETERMINISTIC LILJENCRANTS-FANT MODEL

G. A. Alzamendi[1], G. Schlotthauer[1] and M. E. Torres[1]

[1] Laboratorio de Señales y Dinámicas no Lineales - CONICET, Universidad Nacional de Entre Ríos, Argentina
{galzamendi, gschlott, metorres}@bioingenieria.edu.ar

Abstract: **Accurate estimation of the glottal source is a difficult task in voice signal processing. In the past, several deterministic formulations were proposed to simulate glottal information. However, they are not able to accurately represent any perturbation or aperiodicity occurring in real voices. In this work, a glottal source model inspired on deterministic Liljencrants-Fant model and ruled by a stochastic difference equation is proposed. Following the source-filter theory, a pitch synchronous state-space voice model is formulated combining the proposed glottal model and a time-varying autoregressive vocal tract filter. State-space methods are applied for the estimation of both the glottal source and the vocal tract filter. The method here proposed proved to be useful for voice signal decomposition. Simulations with artificial voice signals demonstrated that a set of parameters characterizing the glottal source can be accurately computed. For real voices, preliminary results suggest that glottal source behavior can be suitable represented using this alternative model.**

Keywords: **Stochastic glottal source, voice decomposition, state-space voice model, glottal source estimation.**

I. INTRODUCTION

Recently, several methods inspired by source-filter theory have been proposed for the decomposition of voice signals into vocal tract and glottal source components [1–3]. Vocal tract behavior is usually modeled by a time-varying autoregressive filter. Whereas different strategies are available to represent glottal information, most of them are based on deterministic glottal functions [4]. However, due to the deterministic formulation, these glottal functions are not able to represent any perturbation or aperiodicity occurring at the glottal level in real voices [5]. Therefore, alternative models are required for a richer representation of glottal source information.

State-space framework has proved to be very powerful for model-based processing of non-stationary stochastic signals. Its most important characteristics are the following [6]: (i) model formulation is straightforward and intuitive, (ii) analytical tools exist for extracting estimates -meaningful statistics- of unobserved processes, (iii) state-space formulation takes uncertainties and errors into account, and (iv) algorithms are available for optimal calculation of unknown parameters. Therefore, in this work we propose an alternative glottal source model, ruled by a stochastic difference equation, suitable for state-space based voice signal decomposition methods.

II. METHODS

A. Stochastic glottal source model

The Liljencrants-Fant (LF) model [4] is one of the most popular parametric representations of the glottal source. It provides a good fit to commonly encountered glottal source waveforms. According to LF model, a glottal source pulse is specified with two interrelated set of parameters (cf. [4]), named the direct synthesis parameters $\{\mathrm{E}_0,\alpha,\omega_g,\varepsilon\}$ and the timing parameters $\{N_p,N_e,N_a,N_0\}$. N_0 is the fundamental period and, therefore, $f_0 = 1/N_0$ is the fundamental frequency. Analytically, the LF model is defined as follows:

$$g_{LF}[n] = \begin{cases} \mathrm{E}_0 e^{\alpha n}\sin(\omega_g n), & 0 \le n \le N_e, \\ -\dfrac{E_e}{\varepsilon N_a}\left(e^{-\varepsilon(n-N_e)} - e^{-\varepsilon(N_0-N_e)}\right), & N_e < n < N_0. \end{cases} \quad (1)$$

Inspired in Eq. (1), we propose a linear time-varying stochastic difference equation for modeling the glottal source. First row in Eq. (1) can be written as follows:

$$\begin{aligned} & E_0 e^{\alpha n}\sin(\omega_g n) \\ & \quad = E_0 e^{\alpha} e^{\alpha(n-1)}\sin(\omega_g[(n-1)+1]) \\ & \quad = e^{\alpha}\cos(\omega_g)\left[E_0 e^{\alpha(n-1)}\sin(\omega_g(n-1))\right] \\ & \qquad + e^{\alpha}\sin(\omega_g)\left[E_0 e^{\alpha(n-1)}\cos(\omega_g(n-1))\right]. \end{aligned} \quad (2)$$

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

Similarly, the second row in Eq. (1) becomes:

$$\begin{aligned}&\frac{E_e}{\varepsilon N_a}\left(e^{-\varepsilon(n-N_e)}-e^{-\varepsilon(N_c-N_e)}\right)\\&=-\frac{E_e}{\varepsilon N_a}\left(e^{-\varepsilon(n-1+1-N_e)}-e^{-\varepsilon(N_c-1+1-N_e)}\right)\\&\approx e^{-\varepsilon}\left[-\frac{E_e}{\varepsilon N_a}\left(e^{-\varepsilon(n-1-N_e)}-e^{-\varepsilon(N_c-N_e)}\right)\right].\end{aligned} \quad (3)$$

In the last expression we assumed that $N_c - N_e >> 1$. Therefore, $e^{-\varepsilon(N_c-N_e)} \approx e^{-\varepsilon(N_c-1-N_e)}$.

Combining the two expressions introduced above and assuming that the glottal source behaves as a stochastic process, the stochastic glottal source (SGS) model is defined by:

$$g_{SGS}[n]=\begin{cases}\mathrm{A}\,g_{SGS}[n-1]+\mathrm{B}u_{SGS}[n-1]+\eta[n], & 0\le n\le N_e,\\ \mathrm{C}\,g_{SGS}[n-1]+\eta[n], & N_e<n<N_0,\end{cases} \quad (4)$$

with parameters $\mathrm{A}=e^{\alpha}\cos(\omega_g)$, $\mathrm{B}=e^{\alpha}\sin(\omega_g)$ and $\mathrm{C}=e^{-\varepsilon}$, auxiliary input $u_{SGS}[n]=E_0\,e^{\alpha n}\cos(\omega_g n)$ and the perturbation $\eta[n]\sim\mathcal{N}(0,\sigma_\eta^2)$.

The SGS model (4) introduced here has three important advantages: (i) it can be studied under the state-space framework, (ii) glottal source shape is characterized by parameters A, B and C, and (iii) any error or misspecification in SGS model can be modeled by the random variable η.

B. State-space voice model

In accordance with the source-filter theory, it is considered that the voice signal is produced by the modulation of the glottal source in the vocal tract [1–3]. In this work, only sustained vowels are considered. Moreover, we assumed that closure and opening instants are known in advance for every glottal pulse. The phonation process was represented by a time-varying autoregressive filter with exogenous input defined by:

$$s[n]=-\sum_{l=1}^{\rho}a_l[n]\,s[n-l]+g_{SGS}[n]+\varepsilon[n], \quad (5)$$

where s is the voice signal, a_l are the vocal tract coefficients, ρ is model order and $\varepsilon\sim\mathcal{N}(0,\sigma_\varepsilon^2)$. Minus sign in Eq. (5) is for mathematical convenience.

As in [1], a state-space voice model (SSM) was formulated combining SGS, Eq. (4), and phonation, Eq. (5), models. Therefore, state-space methods were applied for voice signal processing [6]. Expectation and maximization (E-M) optimization method was applied for the calculation of optimal model parameters [7]. Glottal source g_{SGS} and vocal tract coefficient a_l for $n=1,2,\ldots,N$ were estimated applying Kalman filtering and smoothing methods [6]. Also, 95 % confidence intervals were computed, describing the amount of uncertainty in these estimates.

III. RESULTS

A. Synthetic voices

Sustained vowels /a/ were synthesized from a LF glottal pulse train modulated by an autoregressive filter representing the vocal tract. LF pulses were generated using random time parameters, as in [2]. Vowels with different values of signal-to-noise ratio (SNR), glottal-to-noise ratio (GNR), and f_0 were considered. Only one of them was modified at a time, starting from initial values SNR=60 dB, GNR=60 dB, and f_0=108 Hz. For each setting, 100 signals were synthesized with formant frequencies {800, 1200, 2600, 3200} Hz and bandwidths {60, 50, 105, 110} Hz. Notice that, for these signals, both vocal tract spectrum and LF glottal source were known in advance.

For each signal, optimal model parameters were calculated using state-space and E-M methods, and the mean relative error E_{SGS} between real LF parameters and those estimated considering the proposed method was calculated. Here, we considered parameters $\{\alpha,\omega_g,\varepsilon\}$ because their estimation is difficult in practice. For comparison purposes, glottal parameters were also estimated with the Iterative Adaptive Inverse Filtering (IAIF) method [8], and the mean relative error E_{IAIF} was calculated.

In Tab. 1, we present mean and standard deviations of E_{SGS} and E_{IAIF}, in percent, for different values of SNR, GNR, and f_0. The combination of SGS and SSM showed a good performance (small relative estimation errors) for all the condition here considered. Moreover, these results were comparable to those obtained by IAIF method. Even though the error values were similar, the proposed method presents the advantage of suitable combining SGS model with existing state-space voice signal decomposition strategies [1].

In Fig. 1, a typical result achieved with the proposed method for an artificial voice is shown. On the top, vowel /a/ waveform is presented. In the middle, glottal source estimation calculated with SGS model, g_{SGS}, is plotted. For comparison purposes, the actual LF glottal pulse, g_{LF}, and the glottal source estimated

Table 1: Error in the calculation of LF parameters $\{\alpha,\omega_g,\varepsilon\}$ using the proposed method and IAIF for different values of SNR, GNR, and f_0. It is presented the mean relative error followed by its standard deviation in parentheses.

	SNR (dB)			GNR (dB)			f_0 (Hz)	
	0-20	25-40	45-60	0-20	25-40	45-60	88-118	188-218
E_{SGS} (%)	-7.18 (11.43)	-11.35 (3.15)	-1.30 (5.81)	-2.67 (7.49)	3.05 (3.62)	2.26 (2.23)	3.52 (6.06)	-10.84 (6.90)
E_{IAIF} (%)	-6.86 (11.30)	-8.98 (4.06)	-1.81 (0.90)	-4.64 (2.91)	-1.47 (1.43)	-1.55 (1.03)	-0.72 (4.78)	-18.12 (4.02)

with IAIF, g_{IAIF}, are also shown. Whereas both g_{SGS} and g_{LF} are accurate estimations, the first estimate is smoother than the second one because state-space methods enable non-causal stochastic estimates [6, 7]. On the bottom, vocal tract power spectra computed from both SSM and IAIF are plotted, together with the actual vocal tract power spectrum used in the synthesis. It can be appreciated that both power spectra are very accurate estimations of vocal tract spectral information.

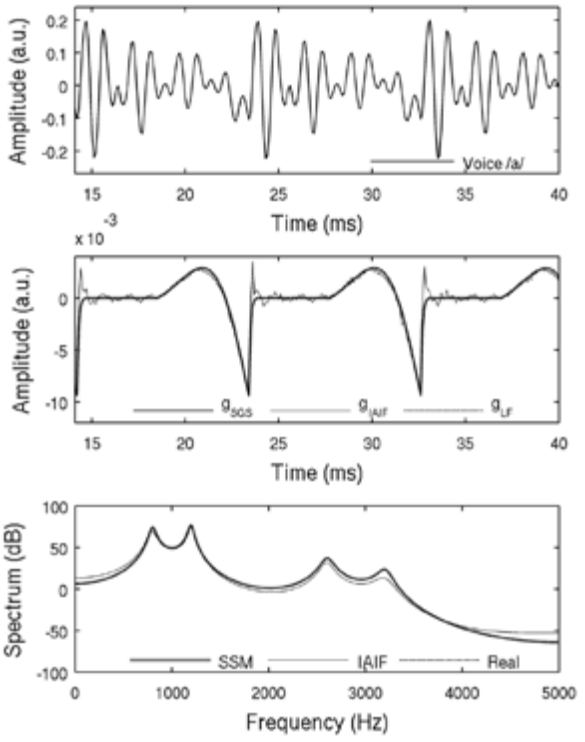


Fig. 1: Artificial voice decomposition into glottal source and vocal tract components (f_0=108 Hz, SNR=60 dB, and GNR=30 dB). *Top:* waveform of a vowel /a/. *Middle:* glottal source estimation applying SGS (g_{SGS}) and IAIF (g_{IAIF}) methods, in comparison to the actual LF glottal source (g_{LF}). *Bottom:* vocal tract power spectra obtained with state-space voice model (SSM) and IAIF method, in comparison to the real spectrum.

B. Real voices

The proposed state-space approach, considering the SGS model, was applied for pitch-synchronous decomposition of real voices. Glottal closure and opening instants were calculated from the electroglottographic signals [9]. Vocal tract filter and glottal source were estimated using Kalman filtering and smoothing methods.

Real voice decomposition is presented in Fig. 2. For comparison purposes, results obtained with IAIF method are also shown. On top, vowel /a/ waveform is presented. In the middle, glottal source estimates g_{SGS} and g_{IAIF} are plotted. Once again, even though both estimates are similar, it can be appreciated that the first estimate is smother than the second one. The 95 % confidence interval of g_{SGS} is also drawn in light gray filled region, demonstrating how accurate this estimate is. On the bottom, vocal tract power spectra computed from both SSM and IAIF are plotted. These power spectra differ slightly each other, especially in the first formant region.

Preliminary results with real voices suggest that SGS model suitable represents the real glottal information. Moreover, glottal parameters - A, B and C in Eq. (4)- were successfully calculated through E-M method, resulting in an alternative glottal source codification approach.

IV. Discussion

State-space models have proven to be useful for voice signal decomposition and vocal tract modeling [1, 10, 11]. Therefore, the preliminary results presented above support the hypothesis that the incorporation of the proposed SGS model provides a better representation and more accurate estimation of the glottal source information. However, more thorough studies are required in order to confirm it.

An important issue needs to be pointed out. Glottal closure and opening instants calculation is a very difficult task in the practice, and this information is essential for SGS model formulation in Eq. (4). As a

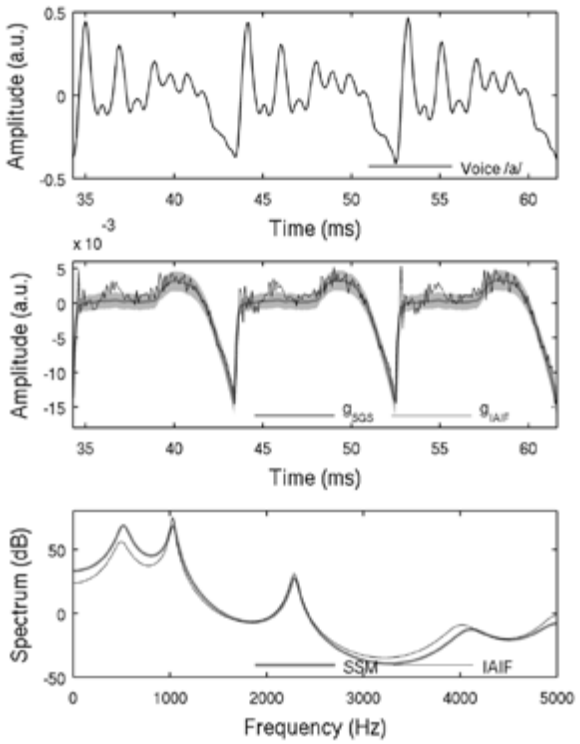


Fig. 2: Real voice decomposition into glottal source and vocal tract components. *Top:* waveform of a vowel /a/. *Middle:* glottal source estimates using SGS (g_{SGS}) and IAIF (g_{IAIF}) methods. Light gray filled region represents the 95 % confidence interval of g_{SGS} estimate. *Bottom:* vocal tract power spectra obtained with state-space voice model (SSM) and IAIF method.

consequence, SGS model is highly dependent on the quality of these values. Fortunately, several methods have been proposed in the last years to deal with this issue [9, 12, 13].

V. Conclusion

In this work, a stochastic glottal source model, based on the popular Liljencrants-Fant glottal model and in accordance with the state-space theory, was formulated. The combination of the proposed model and state-space methods was applied for voice signal decomposition, resulting in an accurate calculation of glottal source information. Preliminary results suggest that the proposed model could also successfully represent glottal information in real situations.

References

[1] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization", *IEEE Trans. Audio Speech Lang. Process.*, vol. 14(2), pp. 492-501, 2006.

[2] P. K. Ghosh and S. S. Narayanan, "Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter", *Speech Commun.*, vol. 53(1), pp. 98-109, 2011.

[3] O. Schleusing, T. Kinnunen, B. Story, and J. M. Vesin, "Joint Source-Filter Optimization for Accurate Vocal Tract Estimation Using Differential Evolution", *IEEE Trans. Audio Speech Lang. Process.*, vol. 21(8), pp. 1560-1572, 2013.

[4] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow", *STL-QPSR*, vol. 4, pp. 1-13, 1985.

[5] C. Drioli and A. Calanca, "Speaker adaptive voice source modeling with applications to speech coding and processing", *Comput. Speech Lang.*, vol. 28(5), pp. 1195-1208, 2014.

[6] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods.* 1st Edition. New York: Oxford University Press, 2001.

[7] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition", *IEEE Trans. Speech Audio Process.*, vol. 1(4), pp. 431-442, 1993.

[8] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: evaluation of an inverse filtering technique using physical modelling of voice production", *Folia Phoniatr. Logop.*, vol. 58(2), pp. 102-113, 2006.

[9] M. R. P. Thomas and P. A. Naylor, "The SIGMA Algorithm: A Glottal Activity Detector for Electroglottographic Signals", *IEEE Trans. Audio Speech Lang. Process.*, vol. 17(8), pp. 1557-1566, 2009.

[10] M. A. Berezina, D. Rudoy, and P. J. Wolfe, "Autoregressive modeling of voiced speech", in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5042-5045.

[11] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking", *J. Acoust. Soc. Am.*, vol. 132(3), pp. 1732-1746, 2012.

[12] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals", in *Proceedings of Interspeech Conference*, 2009, pp. 2891-2894.

[13] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of Glottal Closing and Opening Instants in Voiced Speech Using the YAGA Algorithm", *IEEE Trans. Audio Speech Lang. Process.*, vol. 20(1), pp. 82-91, 2012.

FP – Models 2

MEASUREMENT OF FUNDAMENTAL FREQUENCIES IN DIPLOPHONIC VOICES

P. Aichinger[1], M. Hagmüller[2], I. Roesner[1], W. Bigenzahn[1], B. Schneider-Stickler[1], J. Schoentgen[3], F. Pernkopf[2]

[1] Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Austria
[2] Signal Processing and Speech Communication Lab, Graz University of Technology, Austria
[3] Department of Signals, Images and Acoustics, Faculty of Applied Sciences, Université Libre de Bruxelles, Belgium
philipp.aichinger@meduniwien.ac.at, hagmueller@tugraz.at, imme.roesner@meduniwien.ac.at, wolfgang.bigenzahn@meduniwien.ac.at, berit.schneider-stickler@meduniwien.ac.at, jschoent@ulb.ac.be, pernkopf@tugraz.at

Abstract: **Fundamental frequency (F0) extraction in disordered voice is a prerequisite for many types of clinical analyses. In this paper, we make an approach to F0 extraction based on audio waveform modeling and evaluate it with regard to reference F0s obtained from laryngeal high-speed videos. We analyze 65 euphonic and 112 dysphonic (28 diplophonic, 84 non-diplophonic) intervals of sustained phonations during rigid telescopic high-speed video laryngoscopy. Waveform modeling has a total error rate E_{Total} of 46.28 % on diplophonic voices, which is a significant improvement compared to a benchmark method (95.74 %). The results illustrate that F0 extraction in diplophonic voice is challenging. Multiple F0s need to be considered in the analysis of disordered voice, which may lead towards more valid clinical voice assessment in the future.**

Keywords: Laryngeal high-speed videos, diplophonia, fundamental frequency measurement, voice disorders

I. INTRODUCTION

Voice assessment is important for evidence based clinical decision making. Many analysis procedures for clinical voice assessment assume that at most one F0 is observed at a time, which is not always true for all types of dysphonia. These procedures include cycle-based audio feature extractors, e.g. cycle length jitter and amplitude shimmer, but also stroboscopic image analyzers. Thus, the validity of these procedures is questionable for dysphonic voices, which makes F0 extraction crucial. Special attention must be paid if multiple oscillators with different F0s exist.

Reference F0s for the evaluation of audio-based extractors are usually obtained from electroglottographic (EGG) signals [1], [2]. They need sufficient glottal closure, which is not always observed in disordered voices. Thus, we evaluate the waveform modeling approach with regard to a reference obtained from laryngeal high-speed videos that does not rely on glottal closure [3]–[5].

The waveform modelling approach has recently been investigated in the context of disordered voice analysis. Unit-pulse FIR filtering has been used to determine doubled cycle marks in glottal area waveforms [5]. There it was assumed that the number of oscillators and their F0s were known a priori from spectral video analysis. Waveform modelling for joint F0 estimation and source separation with an unknown number of oscillators has been used for automatically distinguishing diplophonia from other types of dysphonia [7]. There F0 candidates are identified by spectral peak picking and waveform candidates are obtained by cross-correlation. The optimal waveform model is obtained by minimizing the time-domain model error with regard to all possible additive one- and two-oscillator waveform combinations.

The paper is structured as follows. First, details on the data collection are reported. Second, the audio- and video based F0 extractors are described and the error measures are introduced. We present examples of spectrograms together with F0 estimates, as well as summarized results of the analysis corpus. Finally, the results are discussed with regard to validity and future work is suggested.

II. SUBJECTS AND MATERIAL

Between September 2012 and August 2014 we recorded 40 euphonic and 80 dysphonic subjects (40 diplophonic, 40 non-diplophonic). From each subject up to six laryngeal high-speed videos with simultaneous high-quality audio recordings of sustained phonations were obtained. The used camera was a HRES Endocam 5562 (Richard Wolf GmbH) with a frame rate of 4 kHz. The microphone was a headworn AKG HC 577 L with windscreen AKG W77 MP. The original cap without presence boost was used. The microphone was connected to a portable digital audio recorder TASCAM DR-100 via a

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

phantom power adapter AKG MPA V L (linear response setting). The audio sampling rate is 48 kHz and the quantization resolution is 24 bits. The uncompressed PCM/WAV file format is used.

177 temporal intervals of homogeneous voice quality have been selected for analysis. The selection was based on video quality (visibility of the glottal gap, presence of vocal fold vibration, absence of artefacts), audio quality (video synchronization available, presence of phonation, absence of artefacts) and voice quality (euphonic, diplophonic, dysphonic/non-diplophonic) [8]. In total, 65 of the intervals are euphonic and 112 are dysphonic (28 diplophonic, 84 non-diplophonic). The intervals are between 128.7 and 2047.4 ms seconds long.

III. Methods

A. Video-based reference F0s

The reference F0s were obtained by spectral video analysis (SVA) [3], which enables tracking of spatially distinct glottal oscillators with different F0s. Up to two simultaneous F0s are so obtained from the laryngeal high-speed videos. The pixel intensity time series are normalized and windowed with a Kaiser window (β = 0.5, length = 128 video frames or 32 ms, no overlap). Discrete Fourier transformation (DFT) yields one spectrum per pixel position from which maximal spectral magnitudes are identified, i.e. the dominant frequency at each pixel position is obtained. All frequencies below 70 Hz or with spectral magnitudes below 2.45 are discarded. The remaining frequencies are summarized in a frequency histogram, from which peaks are picked and considered as F0s. If more than two F0s are found, the lowest two are chosen. Methodical illustrations are found in [3], and in [8], Section 3.2.1.

B. Extraction of F0s via modelling of audio waveforms

We propose waveform modelling (WM) for audio-based extraction of multiple F0s. A pragmatic waveform model that allows for automatic parameter estimation is structured as follows. The harmonic oscillators d_m are the sum of P partials and their DC component. The oscillators' angular frequencies ω, the partial index p and the time index n drive the trigonometric functions (1). The final signal model is obtained by summing up M harmonic oscillators (one for euphonia, two for diplophonia) and random noise η (2).

$$d_m(n) = \frac{1}{2} \cdot a_{m,0} + \sum_{p=1}^{P} \begin{bmatrix} a_{m,p} \cdot \cos(\omega_m \cdot p \cdot n) + \\ b_{m,p} \cdot \sin(\omega_m \cdot p \cdot n) \end{bmatrix} \quad (1)$$

$$d'(n) = \sum_{m=1}^{M} d_m(n) + \eta(n) \quad (2)$$

The model parameters M, a, b and ω are estimated from the observed signal d' as follows. The audio signal is resampled to 50 kHz and windowed with a rectangular window (length N = 3200 audio samples or 64 ms, overlap 50 %). It is assumed that the model parameters are constant within one analysis window. Zero padded DFT yields spectra with a resolution of 0.1 Hz. F0 candidates are identified by spectral peak picking, where it is further assumed that F0s lie between 70 and 600 Hz, and the spectral energy at F0s exceeds -15 dB respective the maximal spectral magnitude. If more than 12 candidates are identified, the lowest 12 are chosen to save computation time. For each F0 candidate $\hat{k}_\gamma$ with candidate index γ, a unit pulse train u_γ with period length N_γ is created and cross-correlated with the observed signal d' to obtain the time-domain pulse shapes r_γ (3 - 6).

$$N_\gamma = 2 \cdot \left\lfloor \frac{f_s}{2 \cdot \hat{k}_\gamma} \right\rfloor, \qquad \gamma = 1,2, \dots \Gamma \quad (3)$$

$$u_\gamma(n) = \sum_{\gamma} \delta(n - \mu \cdot N_\gamma), \ \ \mu \in \mathbb{Z}, \quad \mathrm{n} = 0,1,2, \dots \mathrm{N} - 1 \quad (4)$$

$$r_\gamma(l_\gamma) = \frac{1}{\sum_n u_\gamma(n)} \cdot \sum_n u_\gamma(n) \cdot d'(n - l_\gamma) \quad (5)$$

$$l_\gamma = 1 - \frac{N_\gamma}{2}, 2 - \frac{N_\gamma}{2}, \dots -1, 0, +1, \dots \frac{N_\gamma}{2} - 2, \frac{N_\gamma}{2} - 1 \quad (6)$$

The pulse shapes r_γ are Fourier transformed, which yields the oscillator candidates' Fourier coefficient estimates $\hat{a}$ and $\hat{b}$. In the resynthesis step, the coefficient vector is truncated to $P = 10$ elements for bandwidth limitation. The oscillator candidates' parameter estimates $\hat{\omega}$, $\hat{a}$ and $\hat{b}$ are inserted into waveform model equation (1), which yields the oscillator candidates $\hat{d}_\gamma$.

To estimate the number of oscillators M and obtain the optimal noiseless waveform model $\hat{d}$, a heuristic oscillator selector is used. For each additive oscillator combination S, the time domain model error e_S is obtained (7). Only zero-, one- and two-oscillator combinations are considered. The combination with the minimal relative root mean squared error $relRMSE$ (8) is considered as optimal.

$$e_S(n) = d'(n) - \hat{d}_S(n) \tag{7}$$

$$relRMSE(S) = 20 \cdot log\left(\frac{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} e_S(n)^2}}{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} d'(n)^2}}\right) \tag{8}$$

C. Wu's audio-based benchmark method (WWB)

The state-of-the-art F0 extractor for two simultaneously speaking individuals in noisy environments proposed by Wu et al. serves as a benchmark (WWB) [6]. The method is based on autocorrelation. It uses a 128-band auditory gammatone filterbank. Normalized autocorrelation functions of the sub-band signals are analyzed for periodicity. Above the center frequency of 800 Hz energy envelopes are analyzed, which accounts for F0 sub-band beating of higher harmonics and suppresses tracking of partials. Noisy channels are discarded and F0 information is summarized across channels, giving frame-wise F0 likelihoods. F0s are tracked with a hidden Markov model, assuming typical state transition probabilities for mixtures of two speakers.

IV. Results and Discussion

Figs 1 and 2 show the audio spectrograms of a voice sample with its video and audio F0 estimates. The video F0s are marked by circles (o). Lines indicate the audio-based estimates of our WM analysis (Fig. 1) and Wu's analysis (Fig. 2). At onset the voice is non-diplophonic, indicated by a normal harmonic structure. F0 is here correctly estimated by all three methods. At approximately 0.35 s, the voice becomes diplophonic. WM and SVA agree well after 0.6 seconds, while WWB is divergent. WWB obtains either an F0 in between the reference F0s or an artificially low one, which is due to undesired effects of WWB's autocorrelation.

Table 1 summarizes the error rates for the WM and WWB method for all 177 selected audio intervals. E_{01}, E_{02}, E_{10}, E_{12}, E_{20} and E_{21} are error rates that report if the correct number of F0s is estimated. E.g. E_{12} counts

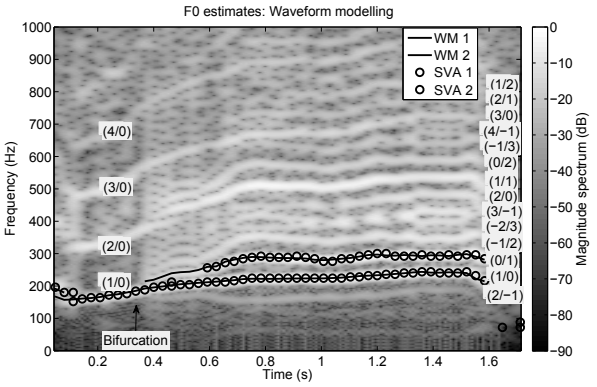


Figure 1: Audio spectrogram and F0 estimates for an example of diplophonic voice. The used F0 extractors are waveform modelling (WM) and spectral video analysis (SVA). On the right side, the frequencies of the sinusoidal components of the spectrogram are annotated in the format (a/b). Their frequencies are linear combinations of the oscillators fundamental frequencies f_1 and f_2, and given by equation $f = a \cdot f_1 + b \cdot f_2$.

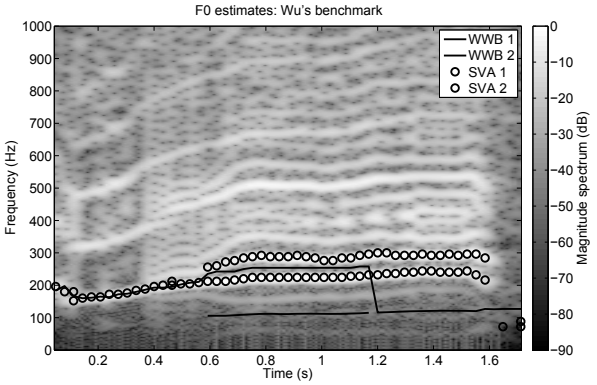


Figure 2: Audio spectrogram and F0 estimates for an example of diplophonic voice. The used F0 extractors are Wu's method (WWB) and spectral video analysis (SVA).

an error if one F0 exists but two are extracted. E_{Gross} is the rate of relative F0 errors greater than 20 %, given that the number of F0s is estimated correctly. E_{Total} is the sum of all above mentioned rates. E_{Fine} is the relative frequency error in percent, given that no total error exists in the analysis block [6]. Some of the error rates are high, i.e. E_{12}, E_{21} and E_{Gross} for WM and E_{21} and E_{Gross} for WWB, which results in high total error rates for both. The rate of total errors is higher in WWB than in WM, i.e. 47.18 versus 57.53 % respectively.

Table 2 illustrates E_{Total} with regard to voice quality and reveals that WWB is mostly incorrect for diplophonic voices, i.e. in 95.74 % of the analysis windows, while WM obtains 46.28 %. For other voice qualities E_{Total} is comparable between the approaches.

V. Discussion and Conclusion

Two audio-based F0 extractors have been evaluated with regard to reference F0s obtained from laryngeal high-speed videos. It has been shown that F0 extraction from diplophonic voice is a challenge. The results illustrate that voice analysis based on F0 measurement must be questioned for diplophonia and some types of dysphonia.

One may hypothesize that WWB fails most likely because it had been developed for two simultaneously talking speakers and not for coupled glottal oscillators. WWB extracts F0s via autocorrelation, which favors the extraction of metacycle frequencies. Perceptually, this frequency is closely related to residual pitch. However, spectral pitches are more salient in diplophonic voices, which are accounted for by our WM approach [9], [10]. Thus, WM obtains more valid F0 estimates than WWB for diplophonic voice, although its error rate is high.

Suggestions for future work are to improve the ground truth by visual annotation of laryngeal high-speed videos and to improve our WM approach by incorporating a F0 tracker that accounts for the system dynamics.

Table 1: Median error rates of multiple F0 extraction. Comparison of waveform modelling (WM) with Wu's method (WWB).

	WM (%)	WWB (%)	p-value
E_{01}	0.12	0.87	= 0.013
E_{02}	0.68	0.05	< 0.001
E_{10}	1.29	3.63	< 0.001
E_{12}	12.12	1.23	< 0.001
E_{20}	1.5	3.34	< 0.001
E_{21}	22.22	38.11	< 0.001
E_{Gross}	9.25	10.29	n.s.
E_{Total}	47.18	57.53	= 0.009
E_{Fine}	2.06	1.25	n.s.

Table 2: Total error rates of fundamental frequency extraction with respect to voice quality.

E_{Total} (Medians)	WM (%)	WWB (%)	p-value
Euphonic	37.5	39.37	n.s.
Diplophonic	46.28	95.74	< 0.001
Dysphonic	54.97	58.83	n.s.

Acknowledgements

The authors would like to thank Richard Wolf GmbH for providing the HRES ENDOCAM 5562.

References

[1] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 1509–1512.

[2] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

[3] P. Aichinger, I. Roesner, B. Schneider-Stickler, W. Bigenzahn, F. Feichter, A. K. Fuchs, M. Hagmüller, and G. Kubin, "Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic phonation," in *Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2013, pp. 81–84.

[4] P. Aichinger, I. Roesner, M. Leonhard, B. Schneider-Stickler, D. M. Denk-Linnert, W. Bigenzahn, A. K. Fuchs, M. Hagmüller, and G. Kubin, "Comparison of an audio-based and a video-based approach for detecting diplophonia," *Biomedical Signal Processing and Control (in press)*.

[5] P. Aichinger, B. Schneider-Stickler, W. Bigenzahn, A. K. Fuchs, B. Geiger, M. Hagmüller, and G. Kubin, "Double pitch marks in diplophonic voice," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7437–7441.

[6] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[7] P. Aichinger, I. Roesner, M. Leonhard, B. Schneider-Stickler, D. M. Denk-Linnert, W. Bigenzahn, A. K. Fuchs, M. Hagmüller, and G. Kubin, "Towards objective voice assessment: the diplophonia diagram," *Journal of Voice (accepted)*.

[8] P. Aichinger, "Diplophonic Voice - Definitions, models, and detection," Ph.D. dissertation, Graz University of Technology, 2015.

[9] E. Terhardt, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.

[10] R. Meddis and L. O'Mard, "A unitary model of pitch perception.," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997.

TUNING OF MODULATION SPECTRUM DISPERSION PARAMETERS FOR VOICE PATHOLOGY DETECTION

L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente

Universidad Politécnica de Madrid, Madrid, Spain
laureano.moro@upm.es, jorge.gomez.garcia@upm.es, igodino@ics.upm.es

Abstract: **Acoustic parameters are frequently used to assess the presence of pathologies in human voice. Many of them have demonstrated to be useful but in some cases its results could be optimized by selecting appropriate working margins. In this study two indices, CIL and RALA, obtained from Modulation Spectra are described and tuned using different frame lengths and frequency ranges to maximize AUC in normal to pathological voice detection. After the tuning process, AUC reaches 0.96 and 0.95 values for CIL and RALA respectively representing an improvement of 16 % and 12 % at each case respect to the typical tuning based only on frame length selection.**

Keywords: **Modulation Spectrum, voice pathology, CIL, RALA, AUC.**

I. INTRODUCTION

The acoustic analysis of voice [1] is a widely used method to assess the presence of a voice pathology due to it is a non-invasive, cost-efficient and easy-to-use technique. Although there are many indices helping clinicians to evaluate the voice perturbations, new parameters are needed to be employed whether in acoustic analysis or as the basis of automatic detectors being used as diagnostic support tools. Moreover, these parameters can be valuable in perceptual assessments to help specialist to increase reliability [2].

Modulation Spectrum (MS) [3] of acoustic signals contains information about the energy relative to modulation frequencies and it can be used as a source of features destined to measure perturbations in voice signal. Many works have used it in the automatic assessing and detection of voice pathologies such as [4] but there is still room for improvement. In this study two measures coming from the histogram of MS are presented. The main objectives of this work are to introduce these new parameters and to determine the optimal operational points for which these can be of use to distinguish between normal and pathological voices, following a simplified version of the methodology used in [5]. In the present case, the tuning is accomplished considering different frame lengths, acoustic and modulation frequency boundaries in order to obtain optimal Area Under the Curve (AUC) from the Relative Operating Characteristic (ROC) curve and its Standard Error (SE) as suggested in [6] in normal-pathological voice detection.

II. METHODS

A. Modulation Spectrum Dispersion Parameters.

MS provides information about the energy at modulation frequencies that can be found in the carriers of a signal. It is a three-dimensional representation where abscissa usually represents modulation frequency, ordinate axis depicts acoustic frequency and applicate, acoustic energy. To obtain MS, signal passes through a short-Time Fourier Transform (sTFT) filter bank whose output is used to detect amplitude and envelope. This output is finally analyzed using FFT producing a MxN complex matrix, being M the number of acoustic bands and N the number of modulation bands. Hence, a large amount of data is obtained depending on the size of M and N but in most of the cases MS matrix must be compressed to more specific parameters.

MS allows observing different voice features simultaneously such as fundamental frequency and harmonics and its corresponding modulations. For instance, the presence of tremor, understood as low frequency perturbations of the fundamental frequency, can be easily noticeable since it implies a modulation of pitch as a usual effect of laryngeal muscles improper activity. Fig. 1 shows the MS of a sinusoid without and with amplitude modulation. In the first case (a), only one point stands out from the overall matrix which is located at the central modulation band, corresponding with 0 Hz in the modulation frequency axe. On the other hand, when any type of modulation exists, new emerging points or areas appear in the modulation regions as it is illustrated on Fig. 1 (b). The study of statistics related with these standing out areas can provide new parameters destined to measure voice perturbations.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

The two proposed in this study are *Cumulative Intersection Level* (CIL) and *Ratio of points Above Linear Average* (RALA) which are intended to measure dispersion of energy across the modulation frequency axe respect to the acoustic axe. When modulation appears in human voice, due to voluntary or involuntary causes, the energy present in acoustic frequency spreads, going from the acoustic axe to the modulation bands. In these cases, dispersion arises.

Throughout this work, the MS has been calculated using the Modulation Toolbox library ver 2.1 [7].

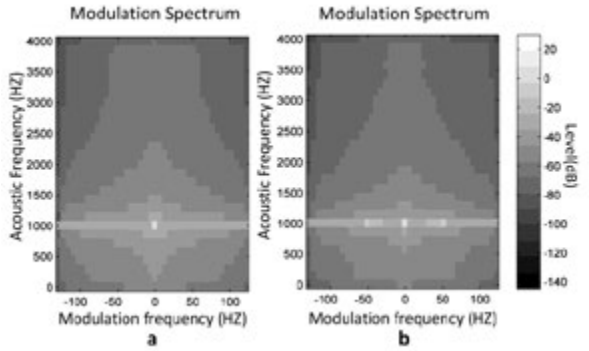


Fig. 1. *Modulation Spectra of 1kHz sinusoid without modulation (a) and with 50 Hz amplitude modulation (b).*

The first parameter, CIL, represents the intersection between the MS histogram increasing and decreasing cumulative curves. Histogram is obtained from MS modulus in logarithmic units (dB) using 29 bins. The higher the number of points in a bin, the nearer CIL index will be to this bin. As it is shown in Fig. 2, CIL tends to be higher in pathological than in healthy voices. On the other hand, RALA is the ratio between points in MS which are over the average of the modulus and the number of points which are above this average. Fig. 3 represents these points in a healthy and a pathological voice. It is noticeable that, as expected, the MS of dysphonic voices present more points above the modulus average.

After describing these new parameters, it is possible to perceive that the MS in Fig. 1 (a) will likely have fewer points over a certain threshold, which can be linear average, than the second one (b). Likewise, the high level bins of the histogram of the second MS (b) will have more cases than those of the first one (a) what will produce a higher CIL.

B. Database

The MEEI voice database is used on this study [8]. From the original 710 recordings, a corpus of 226 including the sustained vowel /ah:/ is selected according to the criteria found in [9]. All the files are sampled at 25 kHz and 16 bits. Before parameterization, all the recordings are normalized. Voice recordings of normal voices (53 files) have an average duration of 3 s while pathological voices recordings (173 files) have an average duration of 1 s.

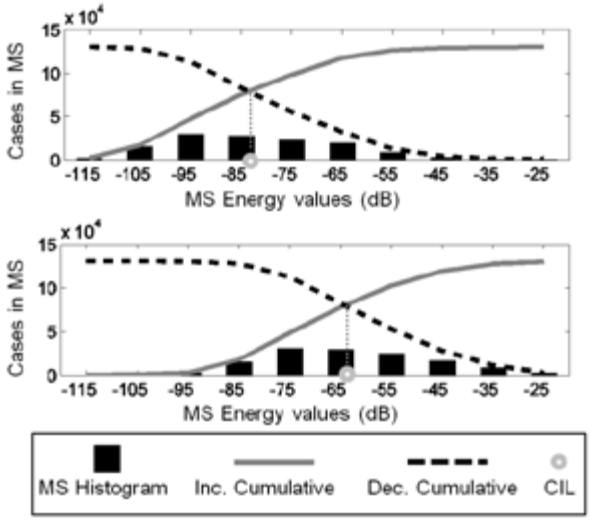


Fig. 2. *CIL calculation in a normal voice (top) and a pathological voice (bottom) diagnosed of bilateral laryngeal tuberculosis. In the second case, histogram presents more points with high levels.*

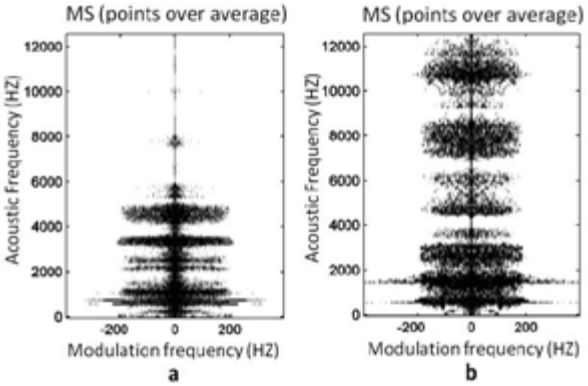


Fig. 3. *Points above (black) and below (white) modulus average in MS for a normal voice (a) RALA = 0.12, and a pathological voice due to bilateral laryngeal tuberculosis (b) RALA = 0.27.*

C. Tuning

To tune CIL and RALA in the voice pathology detection task, three degrees of freedom are selected: frame length, acoustic frequency range and modulation frequency range. Therefore, the corpus is parameterized three times, in which only one degree of freedom is modified at a time while the other two remain fixed. The purpose is to identify which frame

length and frequency ranges lead to the best AUC when using the proposed MS parameters to detect the existence of voice pathology.

Accordingly, in a first stage the corpus is parameterized with CIL and RALA varying only frame lengths in the range of 20 and 200 ms, 50 % overlapping, with fixed acoustic frequency band [0 - 12 kHz] and fixed modulation frequency band [0 - 220 Hz]. This is a basic tuning, widely used to optimize automatic detection systems. With these parameters AUC is calculated. Frame lengths providing the best AUC results are used to re-calculate both parameters separately in a second stage with a new modulation frequency margin, being minimum frequency fixed to 0 Hz and maximum ranging from 20 to 220 Hz in 20 Hz steps. The best AUC results obtained after these two initial stages serve to select optimum modulation frequency range. Using the best frame length and modulation frequency range, a last round of parameterizations and AUC computations are performed by modifying the lower boundary of acoustic frequency between 0 and 1000 Hz in 100 Hz steps and the maximum between 1.2 and 12 kHz in finer steps at low frequencies (300 Hz) and larger steps at high frequencies (from 1 to 3 kHz). After AUC calculation, optimal acoustic frequency range is obtained.

Lastly, two GMM classification systems, one for each parameter separately, are trained to test the ability of CIL and RALA to detect pathological voices. Validation is carried out using a k-Folds technique (8-folds).

III. Results

Regarding the first stage, best results are obtained in frames of 200 ms as it is observed in Fig. 4 (a). Using this frame length, new results are achieved when the maximum modulation frequency is varied, resulting 140 Hz as the optimum operating value as it can be deduced from Fig. 4 (b). Using these settings, a last round of parameterizations is performed varying acoustic frequency ranges, obtaining maximum AUC values of 0.96 for CIL and 0.95 for RALA respectively with SE under 0.01 in both cases.

As it is shown in Fig. 5, the optimum acoustic frequency margin is [0.0 - 1.8 kHz] for CIL and [0.9 – 3.0 kHz] for RALA.

Using these configurations, two GMM systems have been trained following an 8-fold validation scheme in EER. These systems are trained and tested using the tuned parameters, employing the frame length and frequency margins generating the highest AUC. The obtained efficiencies are 92.04 (± 3.53) % for CIL and 88.50 (± 4.16) % for RALA. DET curves are depicted on Fig. 6.

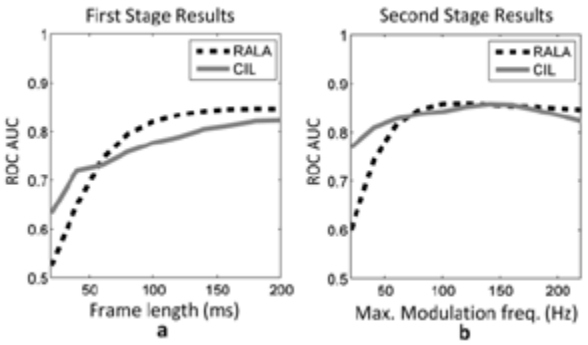


Fig. 4. *AUC variation in respect to frame length with fixed acoustic and modulation frequency (a) and to maximum modulation frequency when frame length is 200 ms (b).*

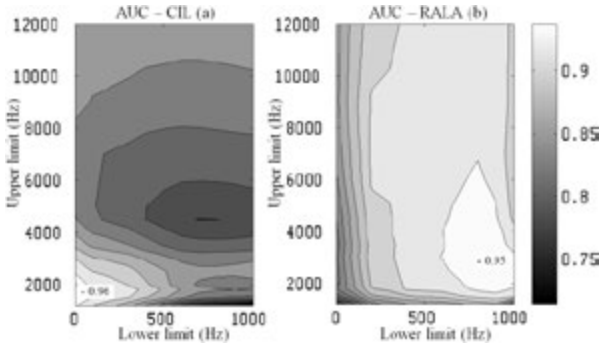


Fig. 5. *Influence of acoustic frequency boundaries on ROC AUC for MS parameters CIL (a) and RALA (b).*

IV. Discussion

In view of the results it is possible to infer that, at first, there is no relevant information to CIL and RALA over 140 Hz in modulation frequency as the best results are obtained using the margin [0 - 140 Hz] and no improvements were obtained beyond this range. Although this is the optimal margin, Fig. 4. (b) suggests that from 80 Hz as maximum modulation frequency, little improvements are achieved. Hence, it is possible to claim that the most relevant information is contained in the margin [0 - 80 Hz].

Regarding to CIL and the acoustic margins, the most important information seems to be around the fundamental frequency while RALA is optimal above the first formant. Taking into account that CIL aims to indicate if MS has a high number of points or regions with high level respect to the rest of the points, the obtained results suggest that the presence of high-level regions in the acoustic [0.0 – 1.8 kHz] and modulation

[0 – 140 Hz] frequency margins is indicative of the presence of a pathology or a dysfunction in the voice. Comparing the resulting AUC after the third stage with that obtained in the first stage, an improvement of 16 % is achieved. Regarding RALA, the amount of points above average seems to be representative of the presence of a perturbation especially in the range of [0.9 – 3.0 kHz]. In this case, the frequency tuning causes an improvement of 12% in AUC respect to only the frame length tuning.

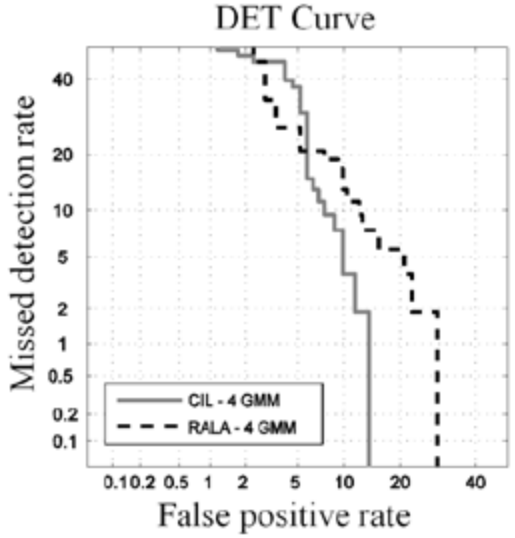


Fig. 6. *DET curves for CIL and RALA.*

CIL and RALA appear to be suitable for clinical assessment of voice but should be tested using other databases to verify the extent of these results. Moreover, the tuned parameters should be checked in other scenarios such as perceptual assessment of voice quality simulations. Mutual information and correlation can be studied respect to GRBAS subjective assessments.

V. Conclusion

In this study two new parameters, CIL and RALA, extracted from MS of voice are presented. These new indices are tuned choosing different frame lengths, acoustic and modulation frequency ranges to obtain maximum AUC values in normal/pathological detection. Results of up to 0.96 of AUC are obtained what suggest that these indices are useful for clinical applications. Further studies must be performed to evaluate the convenience of these indices and their usefulness in voice quality assessment.

VI. Acknowledgements

This research was carried out under grants: TEC201238630-C04-01 from the Spanish Ministry of Education and ayudas para la realización del doctorado (RR01/2011) from Universidad Politécnica de Madrid.

References

[1] C. Sapienza and B. Hoffman Ruddy, *Voice Disorders*. Plural Publishing, 2009.

[2] I. V. Bele, “Reliability in Perceptual Analysis of Voice Quality.” *Journal of Voice*, vol, 19 no. 1, 555-573, 2005

[3] L. Atlas and S. Shamma, “Joint acoustic and modulation frequency,” *EURASIP J. Appl. Signal Processing*, 2003.

[4] J. I. Markaki, M., Stylianou, Y., Arias-Londono, J. D., & Godino-Llorente, “Dysphonia detection based on modulation spectral features and cepstral coefficients,” *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5162–5165, 2010.

[5] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, “The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders.,” *Journal of Voice*, vol. 24, no. 1, pp. 47–56, Jan. 2010.

[6] J. Hanley and B. McNeil, “A method of comparing the areas under receiver operating characteristic curves derived from the same cases.,” *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[7] L. Atlas, P. Clark, and S. Schimmel, “Modulation Toolbox Version 2.1 for MATLAB.” University of Washington, 2010.

[8] “Voice Disorders Database.” Massachusetts Eye and Ear Infirmary, Kay Elemetrics Corp., Lincoln Park, NJ., 1994.

[9] V. Parsa and D. G. Jamieson, “Identification of Pathological Voices Using Glottal Noise Measures,” *J Speech Lang Hear Res*, vol. 43, no. 2, pp. 469–485, 2000.

EXTERNAL VALIDATION OF THE ACOUSTIC VOICE QUALITY INDEX VERSION 03.01 WITH EXTENDED REPRESENTATIVITY

B. Barsties[1,2], Y. Maryn[1,3,4]
[1] Faculty of Medicine and Health Sciences, University of Antwerp, Belgium
[2] Medical School, Hochschule Fresenius University of Applied Sciences, Hamburg, Germany 3 European Institute for ORL, Sint-Augustinus Hospital, Antwerp, Belgium
[4] Faculty of Education, Health & Social Work, University College Ghent, Belgium

ben.barsties@t-online.de
yourimaryn@vvl.be

Abstract: **The Acoustic Voice Quality Index (AVQI) is a six-factor acoustic model based on linear regression analysis to measure objectively overall voice quality in concatenated continuous speech and sustained phonation segments. This acoustic index correlates reasonably with the auditory-perceptual judgment of overall voice quality. The current version of AVQI 02.02 was investigated in several studies (i.e., diagnostic precision, concurrent validity, inter-language phonetic differences, and sensitivity in voice changes through voice therapy). The new further evaluated AVQI model (AVQI version 03.01) was found to be more representative and ecologically valid because the internal consistency of AVQI was balanced out through equal proportion of the two speech types. The present investigation aimed to explore its external validation and diagnostic precision in a large dataset. Therefore, 1058 voice samples were evaluated. Finally, 8 from 12 raters were chosen to reach acceptable reliability of the rater panel. Then the concurrent validity and diagnostic precision was investigated for every voice sample based on the mean results of perceptual overall voice quality level of the 8 raters and the acoustic results of the AVQI version 03.01. All results confirmed that the AVQI version 03.01 is a robust and ecologically valid measurement to objectify overall voice quality.**

Keywords: **acoustic voice quality index, overall voice quality, acoustic measurement, voice assessment, external validation**

I. INTRODUCTION

Voice quality assessment is standard in clinical practice and research. There are several methods to determine voice quality: perceptually, acoustically or aerodynamically [1]. Although auditory-perceptual judgement is gold standard in the evaluation of voice quality, it has several limitations that clearly affected clinical utility. Therefore, it is essential to explore reliability and validity of objective tools. Recently, the Acoustic Voice Quality Index, proposed by Maryn et al. [2], revealed sufficient diagnostic accuracy and reliability in the evaluation of continuous speech together with sustained phonation segments. The current version of AVQI (i.e., version 02.02) showed acceptable diagnostic precision [2-6], and high concurrent validity across studies [2-6] and languages [4-6]. Furthermore, it proved to be sensitive to voice quality changes through voice therapy [3]. In all these investigations the programs Speech Tool (Hillenbrand, 2008) and Praat (Boersma & Weenink, 2013) were used to analyze AVQI. Recently, the smoothed cepstral peak prominence (i.e., the main factor in the multivariate AVQI model and analyzed with Speech Tool) has been implemented in Praat, and thus, the use of Speech Tool might be expendable. A current investigation revealed that the outcomes of the original AVQI version with the two programs and the second AVQI version only in Praat are highly comparable in AVQI results [7].

The next step in the AVQI development was to establish equal proportion of continuous speech and sustained vowel to reach higher ecological validity and a more balanced internal consistency [8]. Therefore, the duration from continuous speech has been expanded from 17-22 syllables [6] to 34 syllables [8] because the length of continuous speech is significantly lower for the analysis, after separating voice to voiceless segments, than the constant duration of sustained vowel. Although Barsties and Maryn [8] found this AVQI 03.01 to be valid in sixty subjects, it is essential to externally substantiate it in a larger set of voice recordings. Therefore, this investigation aims to explore external validity (i.e., the ability to reproduce results with alternative subjects and in settings outside the initial study) of the new weighted equation in the AVQI version 03.01 with a completely new and independent large set of normophonic and dysphonic voice samples and an associated group of auditory-perceptual judges.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

II. Methodos

An expert panel of 12 speech-language therapists (i.e., professional experience in perceptual judgment of voice abnormalities ranged from 5 to 40 years) rated individually the overall voice quality (i.e., G from the GRBAS-scale [9]) of 1058 concatenated voice samples. Every voice sample contained a read aloud Dutch phonetically balanced text "Papa en Marloes" [10,11]. The samples also included a sustained vowel of three seconds of the mid-vowel portion of the vowel [a:] using for both speech types comfortable pitch and loudness. They were recorded using an AKG C420 head-mounted condenser microphone (AKG Acoustics, Munich, Germany) digitized at 44,100 samples per second, that is a sampling rate of 44.1 kHz and 16 bits of resolution using the Computerized Speech Lab model 4500 (Kay Pentax, Lincoln Park, NJ, USA). Furthermore, all recordings were conducted under the same circumstances and according to standards for hardware and software settings [1], the signal-to-noise ratio by Deliyski et al. [12,13], and using a soundproof booth. The voice samples consisted of 970 participants with dysphonia and 88 healthy subjects without any reported voice complaints and voice disorders. The dysphonia group presented various organic and non-organic etiologies and various degrees in dysphonia severity.

To minimize as many as putatively affecting factors in the perceptual evaluation, we strived for similar professional background and familiarity with voice quality evaluation across a large set of raters. Listener bias was minimized through blinding regarding the identity, diagnosis and disposition of the voice samples, through the use of anchor voices, and by considering internal factors such as fatigue, attention, and concentration level.

The statistical analysis covered perceived rater reliability with kappa coefficient (i.e., Cohen's Kappa coefficient to assess intra-rater reliability and Fleiss kappa coefficient for the evaluation of inter-rater reliability). Both Cohen's kappa and Fleiss kappa were considered reasonably reliable from k= 0.41 [14]. Furthermore, significant changes (i.e., considered statistically significant at $p \leq 0.01$) in all kappa values were tested using bootstrapping with 10,000 replications based on a script by Van Belle [15]. To establish a group of raters with homogeneous and representative reliability, the next criteria were followed (next to their longstanding experience in clinical rating voice quality as a speech-language therapist): Firstly, no significant differences were found in intra-rater Cohen's kappa results between all pairs of raters. Secondly, each rater approached intra-rater reliability with a level of Cohen's kappa$\geq$ 0.41 [14]. Thirdly, all leftover raters with representative and comparably high intra-rater reliability were analyzed to find a homogenous rater group with inter-rater reliability of Fleiss kappa$\geq$ 0.41 [14]. Therefore, no significant changes were found between the Fleiss kappa for all tested raters and the Fleiss kappa for one excluded rater of the group with a significantly better result as the Fleiss kappa for all tested raters. Thus, in each round we used a backward method to exclude the rater with the highest significant kappa value in comparison with the Fleiss kappa for all tested raters. This procedure were repeated until a minimum kappa value of k$\geq$ 0.41 [14] was achieved without significantly better Fleiss kappa results for one rater of the group who is excluded in comparison with the Fleiss kappa for all tested raters.

Concurrent validity was tested with Spearman rank-order correlation coefficient (r_s) and the coefficient of determination (r^2_s). Lastly, diagnostic accuracy was evaluated with several estimates: receiver operating characteristics (ROC), and likelihood ratio (LR+ and LR-).

III. Results

Intra-rater reliability showed no significant differences in Cohen's kappa values (t= 12.824, p= 0.306) between all twelve raters, but one rater did not reach the minimum of acceptable reliability level (Cohen's kappa= 0.32) and had to be excluded. The remaining eleven raters had a range of Cohen's kappa between 0.41 and 0.58. Inter-rater reliability was executed on the leftover eleven raters that reached an Fleiss kappa= 0.39 and five raters showed a significantly better Fleiss kappa result if they were excluded in comparison to the Fleiss kappa of all tested eleven raters (t= 18.985, p= 0.000 to t= 7.576, p= 0.006). After the fourth round an Fleiss kappa= 0.43 was found with a group of eight raters and simultaneously showed no significantly better Fleiss kappa results if one rater of this group was excluded (t= 7.25, p= 0.011 to t= 0.757, p= 0.384). Finally, 8 from the 12 experts were chosen because of acceptable reliability. Then the average G-scores of the 8 raters (i.e., G-mean) was taken as the perceptual dysphonia severity level for every voice sample. A strong correlation was identified between AVQI and G-mean (r_s = 0.815, p= 0.000). It indicated that 66.4 % of G-mean's variation was explained by AVQI (r^2_s = 0.664). Additionally, the ROC- and likelihood results showed again best diagnostic outcome at a cut-off score of AVQI=2.43 (sensitivity= 0.785, specificity= 0.932; LR+ =11.54, LR- =0.23).

IV. Discussion

The results indicated that the new version of AVQI with extended portion of continuous speech successfully corresponds to perceptual ratings of

overall voice quality. Furthermore, this AVQI 03.01 has high diagnostic accuracy. Although the present study has first, an extremely large dataset of more than 1000 analyzed voice samples, second, more exact selection of the rater panel based on the knowledge of several affecting factors [1] disturbing the perceived judgment, and third more critical statistical selection criteria in rater reliability, the results are comparable to the previous results about AVQI's version 03.01. Thus, the concurrent validity (i.e., r_s= 0.815 in the current study versus r_s= 0.929 in the study by Barsties & Maryn [8]) and diagnostic accuracy with the threshold of 2.43 (i.e., sensitivity = 0.785 and specificity = 0.932 in the present study versus sensitivity = 0.936 and specificity = 1 in the study by Barsties & Maryn [8]) are comparable. The previously analyzed data pool of 507 subjects across five studies [2-6] evaluated the initial AVQI model and auditory-perceptual judgement of overall voice quality. The results showed a homogeneous weighted mean correlation of r= 0.790 [7]. By comparison, under the same conditions of statistical analysis for the AVQI version 03.01 including 1118 subjects across two studies, the results showed not only a homogeneous weighted correlation but a slightly improved weighted mean r= 0.821. It can be concluded that the development of the AVQI model is a steady robust objective method in the evaluation of voice quality which has improved in ecological validity, concurrent validity and diagnostic accuracy.

V. Conclusion

The present results confirm AVQI as a robust and ecologically valid measurement to objectify overall voice quality. The AVQI version 03.01 demonstrates high validity and acceptable diagnostic accuracy in a large set of clinically representative voice recordings, reflecting different ages, genders, different types and degrees of voice quality, and including nonorganic as well as organic laryngeal pathologies. The independent external validation of the AVQI version 03.01 provided by this study accomplishes an important step in making practical, reliable, and reproducible objective voice assessments available to non-experts or professionals to support their clinical decision in practice or research in voice-disordered patients.

References

[1] B. Barsties , and M. De Bodt, "Assessment of voice quality: Current state-of-the-art," *Auris Nasus Larynx*, vol. 42, 183-188, 2015.

[2] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels," *J Voice*, vol. 24, pp. 540-555, 2010.

[3] Y. Maryn, M. De Bodt, and N. Roy, "The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders," *J Commun Disord*, vol.43, pp. 161-174, 2010.

[4] B. Barsties, and Y. Maryn, „[The Acoustic Voice Quality Index. Toward expanded measurement of dysphonia severity in German subjects]," *HNO,* vol. 60, pp. 715-720, 2012.

[5] V. Reynolds, A. Buckland, J. Bailey, et al., "Objective assessment of pediatric voice disorders with the acoustic voice quality index," *J Voice,* vol. 26, pp. 672.e1-7, 2012.

[6] Y. Maryn, M. De Bodt, B. Barsties, and N. Roy, "The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages," *Eur Arch Otorhinolaryngol,* vol. 271, pp. 1609-1619, 2014.

[7] Y. Maryn, and D. Weenink, "Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index," *J Voice,* vol. 29, pp. 35-43, 2015.

[8] B. Barsties, and Y. Maryn, "The improvement of internal consistency of the Acoustic Voice Quality Index," *Am J Otolaryngol,* In Press.

[9] M. Hirano, "Psycho-acoustic evaluation of voice," in *Disorders of Human Communication 5. Clinical Examination of Voice*, G.E. Arnold, F. Winckel, and B.D. Wyke Eds. Vienna, Austria: Springer-Verlag, 1981, pp. 81–84.

[10] J.C. Van de Weijer, and I.H. Slis, "Nasaliteitsmeting met de nasometer," *Tijdschrift voor Logopedie en Foniatrie*, vol. 63, pp. 97–101, 1991.

[11] K. Van Lierde, "Nasalance and nasality in clinical practice," Unpublished doctoral dissertation, University of Ghent, Ghent, Belgium, 2001.

[12] D.D. Deliyski, H.S. Shaw, and M.K. Evans MK, "Adverse effects of environmental noise on acoustic voice quality measurements," *J Voice,* vol. 19, pp. 15-28, 2005.

[13] D.D. Deliyski, H.S. Shaw, M.K. Evans, and R. Vesselinow, "Regression tree approach to studying factors influencing acoustic voice analysis," *Folia Phoniatr Logop,* vol. 58, pp. 274-288, 2006.

[14] J.R. Landis, and G.G. Koch, "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, pp. 159-174, 1977.

[15] S. Van Belle, "Agreement between raters and groups of raters" Unpublished doctoral dissertation, Univeristy of Liège, Liège, Belgium, 2009

September 3

FP - Singing-Infants

THE CORRECT FUNCTIONNING OF THE VOCAL CORDS IN THE PROFESSIONAL PRACTICE OF SINGING

Dominique Porebska-Quasnik
University of Siedlce, Poland
dporebska.quasnik@gmail.com

Abstract: **Contrary to the conclusion of the most of researchers about the functioning of the vocal cords in professional singing, we hypothesize that the vocal cords do (must) not vibrate, if it is only very slightly on the inner edges of the glottis. The emitter of sound (glottis) should not work that way. This error is the cause of many vocal pathologies (imperfect closure of the glottis, vocal cord nodules). The precision and the correctness of the vocal sound, its perfect height, is obtained (as for the wind instruments like flute) by the absolutely accurate opening of the glottis (the slot between the two vocal cords) combined with a corridor of air maintained by the infallible support of breath. We propose to verify this hypothesis by using scientific instruments (like the scanner and camera) to visualizing the functioning of the vocal cords during the emission of musical sounds into the time of professional performance . It means: on the complete scale required for each category of voice. By example in the case of a dramatic soprano: from A b (below the scope) to C # (above the scope)[see the role of Brunnhilde into the opera Sigurd from Ernest Reyer-1884]. The expected results are confirmation (or contestation) of the principles of vocal technique inherited from the Masters of the Past.**

I. INTRODUCTION

Each sound over the all musical scale has its place, unique, precise, infallible. Tones and semitones. The musical height is determined by the position of the vocal cords. That is to say, the slot between the vocal chords where passes the air. The support of the breath that keeps the volume of air is determined by the play of two muscles which go from sex (like inverted triangle) and are joined to those that allow the extending of the thoracic cage and to fill with air the inferior lungs. These muscles are like a second triangle (not reversed). The coordination of these two functions: exact slot of the glottis, corresponding to each musical sound, and stable volume of air sent through this slot, requires many years of work and appropriate exercises. Add to that all the technique of singing: the resonators, the modulators, the articulation. Then the interpretation: musicality and intelligence. We therefore propose to clear up three complementary aspects of vocal technique: the correct functioning of the two vocal cords in singing, breath support and resonators (thoracic cage, mask, veil of hard palate and of soft palate joined together).

II. DISCUSSION

A. The principles of vocal technique. The main errors

The vocal apparatus: glottis or both vocal cords, do not work as violin strings that vibrate. This is nonsense. The volume of air bounded by the gap formed between the two vocal cords determines the height of the musical sound. The airflow maintained by the support of breath, vibrates. Both vocal cords vibrate only very slightly on the interior edges. This is why the two basic functions of singing must be simultaneous, that is to say: the support of breath and the exact opening of the glottis. The fundamental problem is the correct functioning of the two vocal cords which depends of the command of the brain. Such as any other instrument, the voice has a very specific tuning that is independent of every other functions, like
support of breath, the resonators of the sound, the articulation. To mastering this tuning, one must possess a clear conception of the vocal instrument. Each musical sound has its unique place. That is to say an exact opening of the glottis. Which creates an precise airflow. For each musical sound exists an unique opening of the glottis. The challenge of the vocal technique consists into the mastering of all the notes on the whole vocal scale of each professional voice. For example, the dramatic soprano must give without difficulty the notes from the A b (below the scope) to C # (above the scope) . Figure 1 shows the vocal scale of dramatic soprano (and tenor).

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

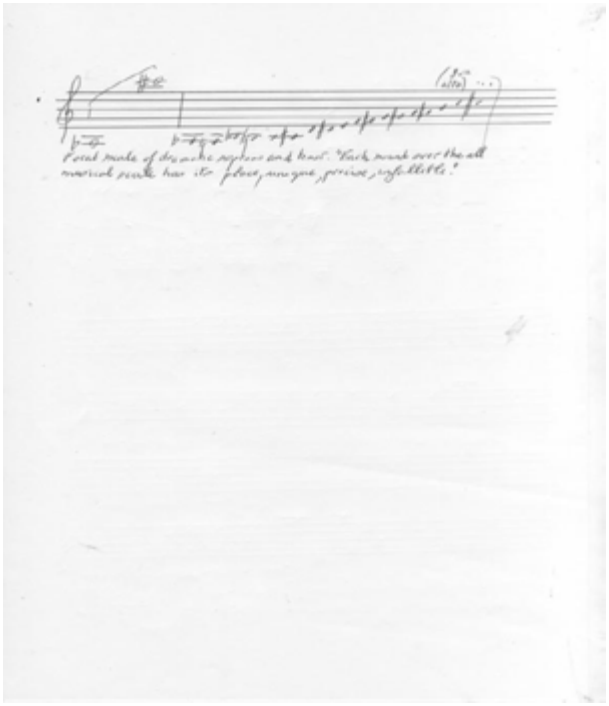


Fig 1: The vocal scale of dramatic tenor or soprano. Each sound over the all musical scale has its place, unique, precise, infallible. Tones and semitones.

B. About the formation of the vocal sound
A1.Sources external, and errors.
"The fact that the vocal cords are joined together and tense allows a resistance of the air expelled from the lungs and entails the rise of pressure upstream the glottis. It is called subglottic pressure. The air, trying to get out of the glottis makes then vibrate the vocal cords which produces the sound according to their position. When we want to make a high-pitched sound, the vocal cords are close and tense. In contrast, during the production of a bottom sound, they relax and are more distant One will also notice that more the sound is high more the vocal cords vibrate rapidly. [...] .The height of the sound depends on vibration frequency from the vocal cords. Its intensity depends in principle on the amplitude of vibration determined by the importance of the air stream which passes through the glottis, but also on the degree of closure of the glottis for each vibration. If the closure is incomplete, and if a part of the air stream passes without being converted into vibrations, the sound becomes more weak. The height and the intensity are also a function of the tension of the vocal cords (Sources-Internet from IRCAM, France).
We deny this assertion and launch the assumption that the vocal apparatus, issuer of the sound, works completely differently.

A2. The correct functioning of vocal apparatus.
Each tone and semitone have a specific place; that corresponds to an unique opening of the glottis. The appropriate tension of both vocal cords create this suitable opening. More the sound is acute, more the vocal cords must be close together, very precisely, to the nearest millimeter. The volume of air, expired from the lungs and maintained stable by the support of the breath, passes through the opening of the glottis, which determines the musical height. This is the alone function of the vocal cords (glottis). For a good functioning, they must be healthy: instantly and symmetrically respond to the command of the brain. For this, each voice must learn to emit each sound by controlling the adequate opening of the glottis. The height of the musical note is determined by the opening of the glottis, only. The other functions are independent: like volume and color of the sound. The amplification of the vocal sound is created by the resonators.
A3. The support of the breath and the vocal resonators
The sound emitted by the vocal cords alone is very small..The volume and the power come from the breath and the resonators. The principal and constant resonator is directly linked with the breath support. The true support of the breath comes from the two muscles that depart from sex, like a reversed pyramid, for joining the two muscles that allow to open the mobile part of the ribcage. The correct breathing in singing comes from the lungs inferiors of which volume can be increased by opening up extensively the ribcage. The breathing of the singing must coordinate two complementary and indispensable functions: the maximum opening of the ribcage for filling up the lower lungs with air and, simultaneously, the maintenance of the volume of air (by a stable support of breath). This volume is sent between the slot of the two vocal cords and forms the musical sound. By this support of the breath, the rib cage can remain open during expiration and thus serves as principal and constant resonator.
The other resonators of the vocal sound are: the mask (the bony part of the face above the nose) and the inside of the bucal cavity (hard palate and soft palate joined together).
The vibration of the vocal cords is a lure. Like the theory of the power of the sound produced by these vibrations. The metered air that passes through the slot of the glottis determines the musical height of the sound. Its volume and its power depend on the breath and on the resonators.

C. Professional vocal technique.

The vocal technique is divided into three distinct stages whose functioning must be coordinated during the vocal issuance (emission).

Mastery of the sound emitter: the glottis. That is to say: the right position of the both vocal cords for emission of each musical note.

Mastery of the breath support. That is to say: deep breathing from lower lungs and the maintaining the volume of air required for the emission of each sound.

Mastery of the resonators. That is to say: main resonance by all the bony part of the ribcage, resonance in the mask, resonance in the oral cavity.

In addition to these three faculties: the mastery of the articulation and of the vocal modulation.

D. Some secrets of vocal art.

The current deficiency of authentic tenors and dramatic soprano is due to a lack of depth technical in lyrical singing. The strength of these declamatory voice that must articulate and interpret the theatrical text requires a real technique: perfection of vocal emission, unfailing support of breath and mastery of all the resonators . That is to say, a minimum of ten years to study the instrument and all life to be master. The full possession of the main resonator of the thoraxrequires a large experience and a constant adaptation to physical and physiological possibilities that expected to grow with maturity. Therefore, it was once not recommended to sing the Wagnerian parts before the age of forty. With practice and age, rib cage increases in volume and it becomes increasingly difficult to sustain by the support of breath this growing magnitude without losing the accuracy of the sound emitting. For example, the main role of the opera Parsifal of Richard Wagner is extremely dangerous and allows no technical error. Because it plays both on the power and on the mastery of the most perilous notes for a tenor (see Figure 2/ a: Parsifal of Richard Wagner, act II, scene of the flower maidens: "Amfortas! die Wunde! ".

This vocal scale is a challenge for a dramatic tenor and the presence of little and very precise intervals (half-tones) don't allow the mediocrity (Figure 2/b: the perilous scale for dramatic tenor. Semitones ascending from the E natural to the A above musical stave).

III. CONCLUSION

The vocal art consists in the mastery of the diction and of the expression of the musical text. The main difficulty lies in the building of a voice, unique, and in maintaining of its quality during all the life.

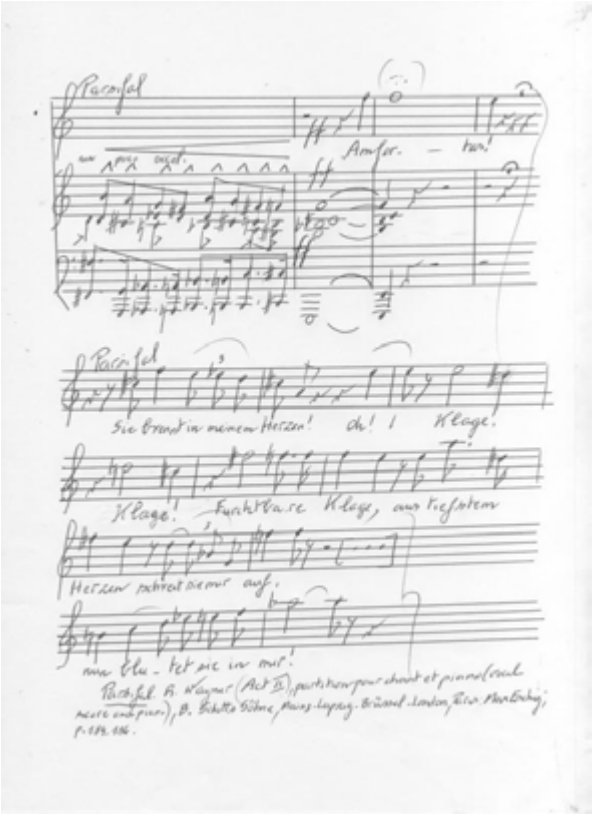


Fig. 2/a/b: Parsifal R. Wagner, Act II.

REFERENCES

Quasnik (Dominique), Marian Porebski, Wagnerian singer. His career and pedagogical work, Master Thesis of Musicology, Sorbonne, 1976.

Kwaśnik (Dominika), "Voice: vocal theory and technique", in Selected Elements of Therapy, Education and Arts, university of Zielona Góra, 2002.

Porebska-Quasnik, Marian Porebski, tenor of world reputation. Life and career, Kucharski, Toruń, 2012

SOME CONSIDERATIONS ABOUT THE RELATIONSHIP BETWEEN SINGING AND SCIENCE

Massimo Sardi
Singing teacher – Former President AICI-Associazione Insegnanti di Canto Italiana
massimo.sardi@teletu.it

I. INTRODUCTION

The aim of this speech is to point out the relationship between the scientific research about vocal phenomenology and the singing voice teaching methodology. We really hope that a more precise Knowledge of anatomy and physiology has to be achieved in order to get the teaching theory to organize its instruments more efficiently. But above all we wish the scientists' guiding principles could be better identified and to reach this achievement, in my opinion, they have to keep in touch with the singing teachers dealing with the most advanced phonatory phenomenon.

Stated the importance of this lecture we want also to stress the risks that we think are contained in it. For example very often it may happen to be conformed to a level of superficiality and approximation causing sometimes confused and misleading concepts.

Now what I have just stated will be analysed through its causes, which have their origin mainly in:
-diversity of specific language,
-difference of operating procedures,
-different point of view of the study .

In conclusion several issues to reflect about this theme will be proposed and we think they will have to be object in the future of more detailed responses by the scientific research in order to be acquired and used in singing teaching. Specifically:
- the attack of sound and its relationship with,
- breath support and larynx functions,
-neuronal functions in sound formation.
This lecture will be supported by slides pointing out more clearly the issues previously discussed.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

SINESWEEP-BASED METHOD TO MEASURE THE VOCAL TRACT RESONANCES

B. Delvaux[1], D. Howard[2]

[1] Audiolab, Department of Electronics, York University, University, United Kingdom
bertrand.delvaux@gmail.com, david.howard@york.ac.uk

Abstract: **In speech/singing, knowledge of the frequencies of the resonances of the vocal tract gives access to the vowel type (lower resonances) and the voice timbre (higher resonances). It is therefore crucial to be able to measure accurately these resonant frequencies directly. Several approaches have been developed such as glottal excitation, excitation at the lips, or medical imaging. More accurate measurements of the vocal tract resonances have emerged from physical modeling in recent years but a simple-to-use in vivo measurement tool would be much more valuable to give feedback to the singer/actor and their teacher(s), as well as voice therapists, language teachers, speech scientists, phoneticians and linguists. In this article, we suggest a method to measure simultaneously the frequencies of the Vocal Tract resonances with the voice spectrum.**
Keywords : **Vocal tract, Resonances, Sinesweep**

I. INTRODUCTION

The human voice production is usually described by the Source-Filter theory[1] in which the Glottis (source) emits a signal that is spectrally shaped by the natural resonances of Vocal Tract (filter). The lower resonances (*R1*, *R2* and to a lesser extent *R3*) define the vowel quality[2] whereas higher resonances (*R3*, *R4*, *R5*, …) determine the voice quality (or *timbre*[3] or *tone colour*[1,3]). The knowledge of the two first resonances is of particular interest in the context of speech and singing: to achieve distinction between vowels and use resonance tuning strategy[4]. In this article, we develop a new method to measure the voice spectrum and vocal tract resonances simultaneously by injecting a sinesweep directly into the mouth and we show how its portability can extend to smartphone applications.

II. METHODS

The method adapts the approach proposed by Epps et al.[5] with the methodology developed in Delvaux et al.[6]: a sinesweep signal is injected directly into the mouth of a singing subject and the output is processed to deliver the voice spectrum and vocal tract resonances simultaneously. The method consists of the following steps (see Fig 1 and Delvaux et al.[6] for further details):

1. **Input signal**: an input signal, in the form of an Exponential Sine Sweep (*ESS*) is injected into the mouth of the phonating subject.
2. **Inverse signal**: the recorded signal (input signal + phonation) is convolved with the inverse filter of the input signal, i.e. $(ESS)^{-1}$.
3. **Linearization**: this temporally separates the Linear Impulse Response (LIR) from the harmonic distortions (see Farina[7]).
4. **FFT**: an FFT is performed on the isolated LIR
5. **Subtraction**: processes 1 to 4 are reiterated in free space without the phonating subject, as a calibration. Both FFT's are subtracted in the frequency domain, on a log scale, to result in the spectrum of the voice + vocal tract impedance, with the loudspeaker-dependent impedance removed.

III. RESULTS

Fig 2 shows the results of the method applied on two professional singers: a male subject singing on three different pitches: A2=110Hz, A3=220Hz and A4=440Hz and a female subject singing in chest voice on f_0=218Hz and in head voice on f_0=345Hz. Both subjects were phonating on the vowel /ɑ:/ as in "hard". The solid lines represent the voice spectra, the arrows show the two first resonances *R1* and *R2*. The vocal tract transfer function (dashed line) is obtained by the suppression of the partials of the speech signal by a similar method to Epps et al.[5]

IV. DISCUSSION

In this paper, a method to measure the spectrum of a voice simultaneously with the resonances of the vocal tract is suggested. An acoustic current, under the form of a sinesweep, is directly injected into the mouth of a phonating subject and the output is recorded simultaneously. After some processing involving convolution and FFT's, the spectrum of the voice along with the vocal tract transfer function are obtained. The

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

shortcoming of the present method is to find a compromise between the relative spectral powers of the sinesweep and the voice; adjustment is crucial to see both the voice harmonics and the vocal tract resonances. In addition, measurements of closed vowels such as /i: / in "neap" are more difficult because the tongue impedes the acoustical current from owing through the airway passage. Due to the portability of the method, an application for iPhone has been developed and is being shown during the presentation of this paper.

REFERENCES

[1] G. Fant. *Acoustic Theory of Speech Production*, 2nd Edition, Mouton, The Hague, 1960.

[2] J Sundberg. *The science of the singing voice*., Northern Illinois University Press, Illinois, 1987.

[3] J. Wolfe, M. Garnier, and J. Smith, "Vocal tract resonances in speech, singing, and playing musical instruments" in *HFSP journal*, vol 3(1), pp. 6-23, January 2009.

[4] N. Henrich, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones" in *JASA*, vol 129(2), pp. 1024-1035, February 2011.

[5] J. Epps, J. R. Smith, and J Wolfe. "A novel instrument to measure acoustic resonances of the vocal tract during phonation" in *Measurement Science and Technology*, vol 8(10), pp. 1112-1121, October 1997.

[6] B. Delvaux and D. Howard, "A New Method to Explore the Spectral Impact of the Piriform Fossae on the Singing Voice: Benchmarking Using MRI-Based 3D-Printed Vocal Tracts", in *PLOS One*, vol 9(7), e102680, January 2014. ISSN 1932-6203. doi:10.1371/journal.pone.0102680.

[7] A. Farina. "Simultaneous measurement of impulse response and distortion with a swept-sine technique", in *Preprints Audio Engineering Society*, vol 108, pp. 1–24, 2000.

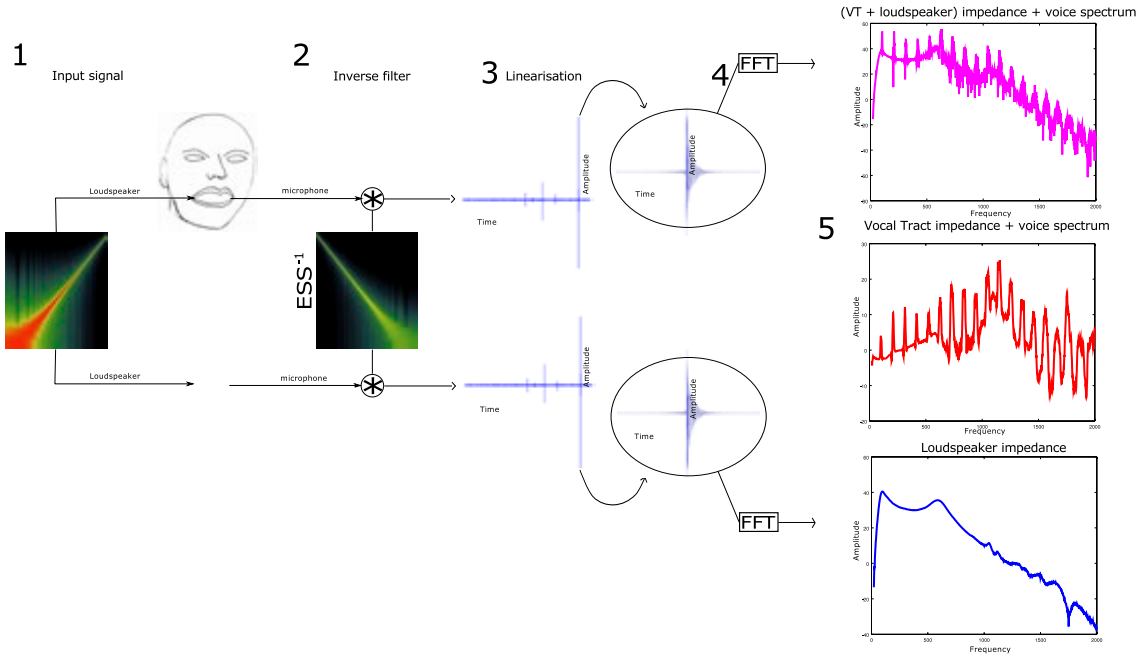


Figure 1

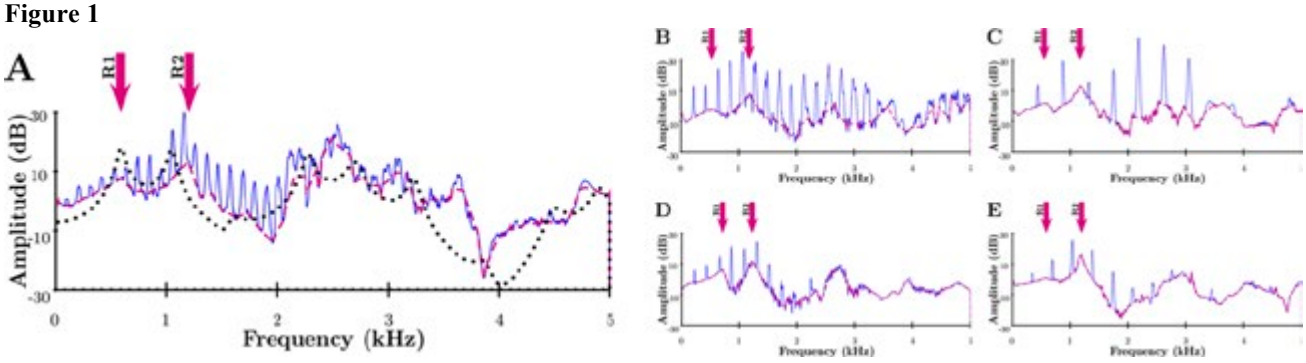


Figure 2

ELECTROGLOTTOGRAPHIC PARAMETERS IN EVALUATION OF VOICE QUALITY. ACOUSTIC ANALYSES FROM A SINGER AND AEROELASTIC MODELLING

T. Ikävalko[1], J. Horáček[2], D. Liu[3], A-M. Laukkanen[1]

[1] Speech and Voice Research laboratory, University of Tampere, Tampere, Finland
[2] Institute of Thermomechanics, The Academy of Sciences of the Czech Republic, Prague, Czech Republic
[3]Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
Ikavalko.Tero.O@student.uta.fi, Jaromirh@it.cas.cz, Dong.Liu@uef.fi, Anne-Maria.Laukkanen@uta.fi

Abstract: **Relations between electroglottographic (EGG) parameters and voice qualities were studied. Data were obtained from one male singer phonating on four pitches in three loudness levels (soft, medium, loud), and in three phonation types (breathy, normal and pressed), which represent varying degrees of adduction. Acoustic signal, EGG and oral pressure were recorded. For comparison, contact area was studied in relation to varying fundamental frequency, sound pressure level and pre-phonatory glottal half-width, using aeroelastic model of voice production. In the case of the singer the degree of firmness of phonation was perceptually analyzed from sound samples by three speech trainers. In modelling data, the pre-phonatory glottal half-width was used to simulate the degree of firmness. Pearson correlations between different parameters were calculated. Closed quotient (CQ) correlated best with the degree of perceived firmness in the human data. Modelling data, however, did not show significant correlation between CQ and pre-phonatory glottal half-width. Instead, for the model the maximum of the first time derivative of contact area increased together with pre-phonatory glottal half-width.**

Keywords: **Electroglottography, modelling, voice quality**

I. INTRODUCTION

Electroglottography (EGG) illustrates variation of vocal fold contact during phonation. Different parametrization methods have been proposed. Contact quotient (CQ, contact time/period time) has been found to correlate with loudness [1], phonation type [2, 3] and register [4], being larger in loud, hyperfunctional and modal voice than in soft, normal and falsetto voice. CQ may vary in opposite directions with fundamental frequency (F_0) increase [5]. In human phonation, pitch, loudness and quality are often interrelated. Normative data of CQ in different phonation types exists only for speech [2, 3]. CQ has been found to increase with impact stress (impact force / contact area of the vocal folds, collision pressure) posed on vocal fold tissue [6]. Previous findings also suggest that the maximum value of the first derivative of the amplitude-normalized EGG signal (MDEGG) correlates with phonation type [7]. Normalized amplitude quotient (NAQ = flow amplitude / (maximum of derivative * T0)), derived from inverse filtered signal, has been found to distinguish between voice qualities [8].

This study compared relations of CQ, MDEGG, NAQ and peak to peak (p-t-p) amplitude of EGG to sound pressure level (SPL), F_0 and phonation type in order to find the most effective parametrization method for evaluating voice quality with EGG.

II. METHODS

Human Data

One trained male pop-singer produced repetitions of [pa:] at four pitches (Bb2, Ab3, Db4, G4) a) by changing loudness (soft, medium, loud) and keeping phonation type the same, and b) by keeping loudness (and monitored sound pressure level, SPL) the same and changing phonation type (breathy, normal, pressed). Acoustic signal (B&K Mediator 2238, 40 cm mouth-to-microphone distance), EGG (Glottal Enterprises dual-channel; 20 Hz LP filtering) and oral pressure (P*oral*, MSIF, Glottal Enterprises) were recorded with Computerized Speech Laboratory (CSL 4150B, KayPentax) using 44.1 kHz sampling frequency and 16 bits quantization. During recording, acoustic signal was calibrated for further SPL measurements by recording a steady complex sound (BOSS TU-120), whose SPL was measured with B&K Mediator.

Praat software was used for further SPL measurements. Peak oral pressure P*oral* during [p] was measured for an estimate of subglottic pressure P*sub*. Analysis was made with CSL. EGG analysis was performed with custom-made multiparameter analysis scripts (Dong Liu) on Matlab. CQ (25% and 35% threshold levels), MDEGG, p-t-p amplitude of EGG signal, NAQ and F_0 were calculated.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

Three speech trainers analyzed the samples perceptually by marking the degree of 'firmness' on a visual-analogous scale 1-9 (1 = very breathy, 5 = normal, 9 = very pressed). Since interrater reliability was found to be good (Cronbach's alpha 0,88), mean of the perceived firmness reported by the listeners gave the value for perceived phonation type. Relations of the EGG parameters with F_0, SPL and both with aimed and perceived phonation type were studied with correlation analysis. Aimed phonation type was given arbitrary values as follows: -1 = breathy, 0 = normal, +1 = pressed. SPSS 21 was used for the statistical analyses.

Modelling data

For comparison with the human data, contact area was studied with an aeroelastic model of voice production [9, 10]. Subglottic pressure (P*sub)* and airflow rate (Q) were set within the values P*sub* < 3 kPa and Q < 0.6 *l*/s. Computations covered the range of P*sub* from phonation threshold pressures to phonation instability pressure of the model. Four F0 levels (100 Hz, 200 Hz, 300 Hz and 400 Hz) were considered. Different phonation types were simulated using four pre-phonatory glottal half-widths (0.2 mm, 0.3mm, 0.4 mm and 0.5 mm). The half-width 0.5 mm simulated the most breathy phonation. From contact area variation, the values for CQ (%), MDEGG (maximum derivative of contact area, mm^2/sec), p-t-p amplitude of EGG (i.e. maximum contact area, mm^2), NAQ and Impact Stress (*IS,* in *Pa*), were calculated.

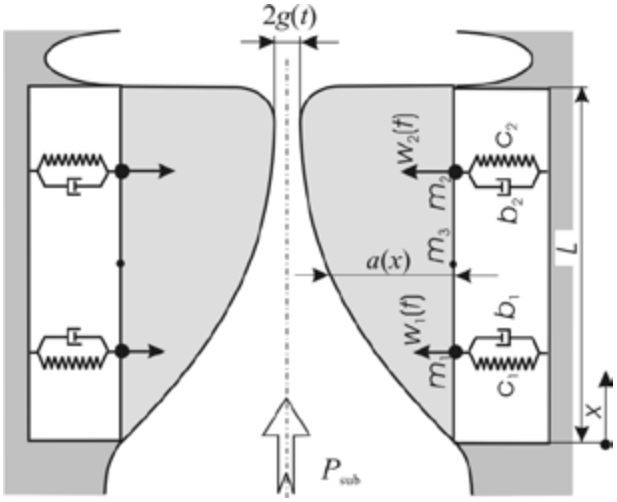


Fig. 1. Schema of the aeroelastic model [9].

Two-degrees-of-freedom dynamic system allowing rotation and translation of the vocal-fold-shaped element vibrating on two springs and dampers (see Fig. 1) modeled the vocal folds' self-oscillations, excited by nonlinear forces resulting from fluid-structure interaction. The impact Hertz force given as

$$F_H = k_H \, \delta^{3/2}, \tag{1}$$

where k_H = contact stiffness, δ = penetration of vocal fold through symmetry axis during collision, was calculated as maximum value during oscillation period:

$$IS = \frac{3}{2} \frac{F_{H,\max}}{\pi \, a^2}, \tag{2}$$

$$a = \sqrt[3]{\frac{3}{4} r \frac{\left(1-\nu^2\right)}{E} F_{H,\max}}, \tag{3}$$

$$F_{H,\max} = k_H \, \delta_{\max}^{3/2}, \tag{4}$$

$$k_H = \frac{4}{3} \sqrt{r} \frac{E}{1-\nu^2} \tag{5}$$

where E=8kPa is Young's modulus, and ν=0.4 is Poisson number. Parabolic vocal fold surface shape was considered, which gives radius r of vocal fold curvature at contact point. For on-line numerical simulations in time domain, the resulting system of ordinary differential equations describing the vocal fold vibrations was solved by 4th order Runge-Kutta method.

In correspondence to the data measured in human the numerically simulated outputs were the following quantities: CQ, MDEGG (simulated as maximum of the first time derivative of contact area in mm^2/s), p-t-p amplitude of EGG (simulated as maximum of contact area in mm^2), dimensionless NAQ (simulated as maximum of contact area/ maximum of the first time derivative of contact area times vibration period) and the impact stress (IS, in Pa).

III. Results

Table 1 summarizes the main results. For the singer, CQ correlated only with phonation type ('Firmness'). Results were similar for CQ25% and CQ35%. MDEGG correlated moderately with SPL and firmness, and NAQ with F_0. Singer's intended phonation type correlated both with the listeners' evaluation of firmness and with CQ 35 (Spearman´s rho > 0,75 for both).

Table 1. Correlation matrix (Pearson´s *r*) for relations between EGG parameters and F_0, SPL, perceptual evaluation of phonation type ('Firmness'), pre-phonatory glottal half-width (pre-gap), subglottal pressure (P*sub*) and Impact Stress (*IS*). NS = p >0,05.

	CQ35	MDEGG	p-t-p Ampl.	NAQ	SPL	F0
Singer						
Firmness	r 0,89, p 0,000	r 0,47, p 0,02	NS	NS	NS	NS
SPL	NS	r 0,48, p 0,02	NS	NS	1	r 0,80, p 0,000
F0	NS	NS	NS	r 0,65, p 0,001	r 0,80, p 0,000	1
Psub	NS	NS	NS	r 0,51, p 0,01	r 0,72, p0,000	r 0,83, p 0,000
Model						
Pre-gap	r -0,19, p 0,015	r 0,39, p 0,000	r 0,19, p 0,019	r -0,30, p 0,000	r -0,18, p 0,017	NS
SPL	r 0,80, p 0,000	r 0,69, p 0,000	r 0,56, p 0,000	r 0,73, p 0,000	1	r 0,64, p 0,000
F0	r 0,24, p 0,001	r 0,87, p 0,000	NS	NS	r 0,64, p 0,000	1
Psub	r 0,55, p 0,000	r 0,87, p 0,000	r 0,49, p 0,000	r 0,45, p 0,000	r 0,86, p 0,000	r 0,72, p 0,000
IS	r 0,82, p 0,000	r 0,31, p 0,000	r 0,96, p0,000	r 0,77, p0,000	r 0,57, p 0,000	NS

For the model, MDEGG correlated with pre-phonatory glottal half-width but the correlation was rather weak. CQ and NAQ correlated strongly with SPL and IS, MDEGG with F_0, and p-t-p amplitude with IS.

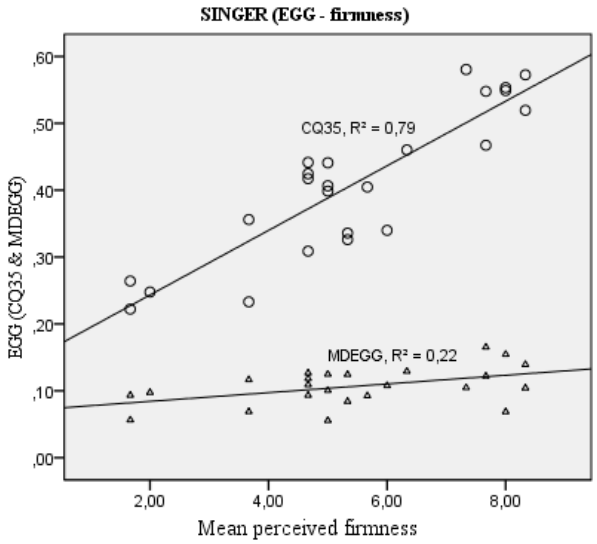


Fig. 2. Scatter plot for relations between EGG parameters (CQ & MDEGG) and perceptually evaluated firmness in human data. Circles represent CQ and triangles MDEGG.

IV. Discussion

MDEGG has been suggested to indicate voice quality better than CQ in speech, at least in sustained vowel samples at a unified SPL [7]. Results of this study suggest that CQ may better estimate phonation type in singers who supposedly have developed means of adjusting pitch, loudness and voice quality more or less independently from each other. Previously reported correlations between CQ and F_0 and SPL seem to reflect changes in adduction in relation to changes in F_0 and SPL.

For the model, MDEGG increased with increasing pre-phonatory glottal half-width, while in contrast the human data of the present study and results from previous studies concerning speech have shown higher MDEGG in more pressed phonation (suggesting a smaller glottal pre-gap). This discrepancy is due to the fact that only regular vibrations were studied with the model, and therefore a wider pre-phonatory glottal half-width required a higher P*sub*. In the singer, P*sub* and phonation type did not correlate with each other. However, for the model, CQ was in a low inverse correlation with pre-gap size, suggesting that simulation of phonation type was at least to some extent successful.

While NAQ from inverse filtered signal has been found to correlate well with phonation type, NAQ from EGG did not seem to reflect phonation type in the singer.

In the present study, the singer phonated on four relatively widely separated pitches (Bb2, Ab3, Db4 and G4). Therefore, it is likely that the laryngeal mechanism (register) was different between at least some of the pitches. According to Henrich et al. [10] in singers the relation between CQ and vocal intensity, and CQ and F0 depend on the laryngeal mechanism. A future study should investigate which EGG-parameters illustrate different voice qualities best within a given laryngeal mechanism, in which case more pitches should be recorded in each mechanism separately.

V. Conclusion

Out of the EGG parameters studied, CQ seems to be the best estimate for firmness of phonation, at least in singing. In the model, MDEGG correlated with pre-phonatory glottal half-width which meant a negative correlation between MDEGG and degree of firmness of phonation.

Acknowledgements

This study was supported by grants of The Academy of Finland (1128095 and 134868) and GACR P101/12/1306.

REFERENCES

[1] R.F. Orlikoff, "Assessment of the dynamics of vocal fold contact from the electroglottogram: Data from normal male subjects." *J Speech Hear Res*, vol. 34, pp. 1066-1072, 1991.

[2] K. Verdolini, D.G. Druker, P.M. Palmer, H. Samawi, "Laryngeal adduction in resonant voice," *J Voice*, vol. 12, pp. 315-327, 1998.

[3] E. Kankare, A-M. Laukkanen, I Ilomäki, A. Miettinen, T. Pylkkänen, "Electroglottographic contact quotient in different phonation types using different amplitude threshold levels," *Log Phon Vocol*, vol. 37, pp. 127-132, 2012.

[4] P. Kitzing, "Photo- and electroglottographical recording of the laryngeal vibratory pattern during different registers," *Folia Phon*, vol. 34, pp. 234-241, 1982.

[5] E. Yanagi, D.H. Slavit, T.V. McCaffrey, "Study of phonation in the excised canine larynx," *Otolaryngol Head Neck Surg*, vol. 105(4), pp. 586-595, 1991.

[6] K. Verdolini, R. Chan, I.R. Titze, M. Hess, W. Bierhals, "Correspondence of electroglottographic closed quotient to vocal fold impact stress in excised canine larynges," *J Voice*, vol. 12, pp. 415-423, 1998.

[7] E. Kankare, D. Liu, A-M. Laukkanen, A. Geneid, "EGG and acoustic analyses of different voice samples: Comparison between perceptual evaluation and voice activity and participation profile," *Folia Phon Log*, vol. 65, pp. 98-104, 2013.

[8] P. Alku, T. Bäckström, E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J Acoust Soc Am*, vol. 112(2), pp. 701-710, 2002.

[9] J. Horáček, P. Šidlof, J.G. Švec, "Numerical simulation of self-oscillations of human vocal folds with Hertz model of impact forces," *Journal of Fluids and Structures*, vol. 20, pp. 853-869, 2005.

[10] N. Henrich, C. D`Alessandro, B. Doval, M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *J Acoust Soc Am,* vol. 117(3), pp. 1417-1430, 2005.

WAVELET ANALYSIS OF NEWBORN INFANT CRY

S. Orlandi[1], A. Bandini[1,2], A. Perrella[1], J. Marjouee[1], G.P. Donzelli[3], 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
[3]Department of Neuroscience, Psychology, Medicines and Child Health, Università degli Studi di Firenze, Firenze
{silvia.orlandi, andrea.bandini, gianpaolo.donzelli, claudia.manfredi}@unifi.it

Abstract: **The acoustical analysis of the infant cry is a non-invasive approach to assist the clinical specialist in the detection of abnormalities in infants with possible neurological disorders. Along with the perceptual analysis, the automatic analysis of the cry is often carried out through commercial or free software tools. However, the neonatal cry is a signal extremely difficult to analyze with standard techniques due to its quasi-stationarity and the very high range of frequencies of interest. To address this issue, we present a new fully automatic method that exploits the wavelets high time-frequency resolution and low computing time properties for the estimation of the fundamental frequency F_0 and vocal tract resonance frequencies F_1-F_3. The method is tested on synthetic signals giving results comparable to existing tools. It is also applied to a set of 1669 newborn cry units (CU) coming from 10 very preterm babies and to a set of 3514 CUs of 20 full-term infants.**

Keywords : **acoustical analysis, wavelet transform, newborn infant cry, fundamental frequency, resonance frequencies**

I. INTRODUCTION

The acoustical analysis of the infant cry is a non-invasive approach to assist the clinical specialist in the detection of abnormalities in infants with possible neurological disorders. A brain dysfunction may lead to disorders in the vibration of the vocal folds and in the coordination of the larynx, pharynx and vocal tract.

The main parameters of the newborn cry are the fundamental frequency (F_0), the frequency of vibration of the vocal folds, and the first three resonance frequencies (RFs) of the vocal tract (F_1, F_2 and F_3) related to the varying shape of the vocal tract during the vocal emission. Indeed, in the newborn it is more appropriate to refer to resonance frequencies (RFs) rather than formants. In fact, the vocal tract is almost flat, the mobility of the oral cavity is reduced and the baby is unable to articulate vowel or consonant sounds, as the pharynx is too short and not wide enough for that purpose. For infants F_0 values are usually in the range 200 Hz - 800 Hz (in the case of hyperphonation they can reach and exceed 1000 Hz) [1-2]. Typical values for the first three RFs are approximately 1000 Hz, 3000 Hz and 5000 Hz [3]. Significant deviations from these ranges may be related to pathological conditions of the central nervous system.

The study of neonatal cry has its origins several decades ago, when the technology was limited and it was therefore mainly based on the perceptual analysis made by the clinician through listening to the cry signal and visually analyzing the recorded signal and its FFT-based spectrogram [4]. This approach is implemented in the MDVP^{TM}, the first and still used commercial tool, though developed for adult voices [5]. Currently, many researchers use PRAAT [6, 7] freely available on line. As MDVP, it was developed for the adult's voice and requires a careful manual setting of some parameters [7]. In the last years, a fully automatic adaptive parametric approach for the crying analysis was developed, named BioVoice [8, 9]. As for F_0, the difficulty in the estimation of the RFs is mainly linked to the quasi-stationarity and the very high range of frequencies of interest in the newborn cry, which requires sophisticated adaptive numerical techniques characterized by high time-frequency resolution.

To overcome such problems, this paper presents a new fully automatic method based on wavelet transforms specifically developed for the estimation of F_0 and the RFs of newborn cry that does not require any manual setting to be made by the user. The wavelet approach seems particularly suited to the study of neonatal cry thanks to its time-frequency high resolution characteristics and low computing time. In [10] a Continuous Wavelet Transform (CWT) with the Mexican hat was used on adult voice signals and in [11] the complex Morlet mother wavelet were applied.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

This paper presents the first attempt to apply wavelets to the analysis of newborn cry. The implemented approach, named InCA (Infant Cry Analizer) is currently implemented in MATLAB, but it is easily adaptable for any embedded processor. InCA is tested and compared on synthetic signals with BioVoice and PRAAT. Results are comparable as far as F_0, F_1 and F_2 are concerned while F_3 is slightly overestimated. InCA is also applied to a set of newborn cries coming from 10 preterm infants and 20 full-term infants for a total of 5183 Cry-Units (CU).

II. METHODS

A. Pre-processing

The analyzed signals, both simulated and real, were sampled at 44100 Hz and the time duration of the analysis window was chosen equal to 10 ms (441 samples). As compared to the use of longer windows, that might not take into account the variability of the signal, this leads to improved accuracy of the estimates.

The next step is the detection of the vocalic parts of the signal (the so-called "crying episodes" or Cry Units - CU) where F_0 and RFs are estimated. For the selection of CUs, the proposed approach takes advantage of the procedure developed in BioVoice whose higher robustness with respect to other software tools has been demonstrated [9].

An audio recording of crying usually includes several CUs. In the literature, different time lengths are considered for CUs, ranging from 60 to 500 ms [12]. However, CUs of very short duration do not allow the assessment of some relevant features such as their melodic shape. Moreover, inspiratory sounds that have duration less than 200 ms must be disregarded [12]. For these reasons, audio analysis is performed here on CUs longer than 260 ms.

B. Continuous wavelet transform

The wavelet transform filters a signal $f(t)$ with a shifted and scaled version of a prototype function $\psi(t)$, the so-called "mother wavelet", a continuous function in both the time domain and the frequency domain [13].

The scale parameter a of a Continuous Wavelet Transform (CWT) is related to the width of the analysis window: it either dilates or compresses the signal. The shift parameter b locates the wavelet in time. Varying a and b allows locating the wavelet at the desired frequency and time instant [13]. The relationship between a and the frequency is given by the so-called pseudo-frequency (F_a) in Hz, defined by the following equation:

$$F_a = \frac{F_c}{a\Delta} \quad (1)$$

where Δ is the sampling period, and F_c is the wavelet central frequency.

For F_0 estimation, a Mexican Hat CWT is used.

For each time window and in the frequency band of interest for F_0 [200-800], the highest coefficient of the CWT matrix is found. The autocorrelation (AC) is computed on the row of the matrix that contains this value, which corresponds to the optimal scale. F_0 is given by:

$$F_0 = F_s / \tau \quad (2)$$

Where τ refers to the position (lag) of the maximum of the AC.

The estimation of $F_1 - F_3$ is performed in a similar way, with different ranges for the band-pass filter as reported in Table 1 with a complex Morlet wavelet as prototype [13]. For this wavelet is defined:

$$\omega_c = 2\pi F_c \quad (3)$$

where ω_c as the center frequency of the wavelet; σ_t is the standard deviation (STD), that is the scale parameter which determines the amplitude of the wavelet. In fact $\omega_c\sigma_t$ sets the link between the bandwidth of the wavelet and its frequency F_c. For the Morlet wavelet, the latter must assume values such that [10, 11]:

$$\omega_c\sigma_t \geq 5 \quad (4)$$

Moreover, the following relationship is taken into account:

$$F_b = 2{\sigma_t}^2 \quad (5)$$

Where F_b is the bandwidth of the wavelet. Comparing the frequency ranges and on analogy to [10] the values of F_c and the corresponding values of F_b were set as in Table 1. Specifically, for each F_c relative to each frequency band, F_b was computed with $\omega_c\sigma_t = 5$ and according to Eq. (3) and (5).

Table 1. Frequency bands of interest in newborn cry, center frequency Fc and bandwidth Fb for the complex Morlet

Frequency band [Hz]	F_c [Hz]	F_b [Hz]
F_1 [800 - 2100]	0.8	1.98
F_2 [1500 - 3500]	0.75	2.25
F_3 [3400 - 5500]	1.5	0.56

C. Estimation of F_0

For F_0 estimation, the proposed method involves the following steps:

1. Band-pass filtering FIR with Kaiser window [200-800]Hz;
2. Mexican Hat CWT of the signal. A pxq matrix M of coefficients is obtained, where p = maximum value of the scale and q = number of frames of the signal;

3. Location of the scale (line) of M corresponding to the coefficients of maximum modulus and estimation of F_0 according to eq.(4).

On each time window the CWT scale parameter *a* was allowed to vary in the range 1÷55. This choice is related to a reasonable frequency range for F_0: 200 Hz-1050 Hz [1]. Therefore the Mexican Hat CWT was applied with a = 55, $\Delta = 1/F_s = 1/44.1$ s, Fc = 0.25 Hz. Consequently Fa = 200 Hz according to Eq. (1).

D. Estimation of RFs

The estimation of F_1–F_3 is carried out with a procedure similar to that used for F_0 but with different ranges for the band-pass filter, according to Table 1 and Complex Morlet as mother wavelet.

III. RESULTS

A. Synthetic signals

The method for the F_0 estimation was tested on a sine wave at 450Hz:

$$y(t) = \sin(450\,t) + e(t) \quad (8)$$

The RFs F_1-F_3 estimation method was tested on a sum of three sinusoids on analogy to [11]:

$$y(t)=5\sin(1000t)+10\sin(3000t)+15\sin(5000t)+e(t) \quad (9)$$

White noise e(t) set at 5% of the signal amplitude was superimposed through the Audacity® open source tool. Signals were sampled at Fs=44.1 kHz. Results were compared with those obtained with BioVoice and PRAAT.

BioVoice implements a robust method for the selection of the voiced parts of the signal (CUs) [9] and a variable window length for analysis: the higher the F_0 the shorter the analysis window. RFs F_1-F_3 are obtained by peak picking in a parametric PSD (AR models) whose variable order is estimated on the varying time windows previously found. Instead, PRAAT implements a method for the F_0 estimation based on the AC applied to a time window of fixed size while Linear Predictive Coding is applied for the RFs estimation. For proper use, and especially with newborn cry RFs, it requires the manual setting of some parameters. Therefore, its use must be made with caution [7]. Thus in this work the best parameters for PRAAT were preliminarily tested and set. Specifically, the range for F_0 was set at 200-800 Hz while for F_1-F_3 the maximum range was set up to 11025 (Fs/4) with the estimation of 5 formants instead of 3. The use of default values (5500 Hz and 3 formants) leads to wrong results.

To compare the three approaches for F_0 estimation a preliminary test was carried out on the sinusoid in Eq. (8). First results show that the CWT Mexican Hat allows to obtain better results than the other methods.

Table 2 shows the results obtained with the three approaches. The CWT has the best performance, as well as PRAAT (set with optimal parameters) though with a slightly higher standard deviation (STD), while BioVoice slightly underestimates F_0 (0.26%).

Table 2 – F_0 estimation. Comparison of BioVoice, PRAAT and CWT Mexican Hat on a synthetic signal (sinusoid at 450Hz with 5% white noise)

Method	**F_0 mean**	**STD**
Mexican Hat	450.00	0.00
BioVoice	448.81	2.08
Praat	450.00	0.88

On analogy to F_0, for F1-F3 a preliminary test was made with the synthetic signal in Eq.(9) with CWT Complex Morlet with 5% white noise. Table 3 shows that the CWT Complex Morlet provides good results especially for F_1 and F_2. All methods give comparable results although with significant differences on STD. BioVoice gives the best results, with the lowest STD for all RFs.

Table 3 – F1-F3 estimation. Comparison of BioVoice, PRAAT and CWT Complex Morlet

Method	**F_1 mean STD**	**F_2 mean STD**	**F_3 mean STD**
Morlet CWT	1024.09 91.00	2971.27 170.41	5163.62 411.32
BioVoice	985.47 5.12	2956.42 8.11	5050.56 11.20
Praat	1120.50 387.37	3068.69 346.26	5019.35 147.02

B. Real signals

Results concern spontaneous hunger cry of 20 full-term newborns (TN, 10 male and 10 female) and 10 preterm infants (PN, 5 male and 5 female). Gestational age (g.a.) of TN at birth was between 37 weeks and 2 days and 42 weeks; the weight was between 2400g and 4250g. Gestational age of PN at birth was between 23 weeks and 5 days and 34 weeks. The weight at birth was between 590g and 2700g. At the recording time (20-30 days after birth) the PN gestational age was between 35 weeks and 1 day and 43 weeks and 1 day; the weight ranged between 1380g and 2430g.

The TN infants were recorded within the first two days of life, while PN newborns could be recorded only about 20–30 days after birth, due to their long staying in the incubator. Specifically, the PN infants were recorded within the first 45 days after the normal end of pregnancy (37 weeks).

We collected an audio recording for each infant of at least 1 hour of duration consisting of at least 10% of crying. From the whole recording we manually selected 2 or 3 minutes of crying.
A total of 5183 CUs were extracted with BioVoice and the analysis performed with InCA. Table 4 summarizes the results.

Table 4 – Mean and STD values for F_0, F_1, F_2 and F_3 obtained with InCA.

	F_0 [Hz]	**F_1 [Hz]**	**F_2 [Hz]**	**F_3 [Hz]**
PN mean	481.9	1188.2	2743.6	4395.2
STD	65.2	204.0	535.5	736.0
TN mean	461.1	1090.0	3037.0	4324.2
STD	44,2	204.3	711.7	843.5

IV. Discussion

In this work an innovative method named InCA, based on the wavelet transform, is presented for the study of the acoustical features of the neonatal cry. Unlike most commonly used software tools, this method has been developed specifically for this kind of signals, characterized by high fundamental frequency F_0 and quasi-stationarity.
According to a careful selection of the wavelets, tested on synthetic signals, InCA implements the CWT Mexican Hat for F_0 estimation and the CWT complex Morlet for the RFs estimation.
The computing time is comparable to PRAAT: for 1 s of recording InCA requires 0.9 s for the estimation of F_0 and 2.8 s for the estimation of RFs, against less than 0.5 and approximately 2 s respectively with PRAAT. However, the CUs obtained with PRAAT are less reliable [9] and a careful manual setting of ranges and thresholds is required to avoid meaningless results especially for RFs [7].
InCA is applied to a quite large real data set coming from preterm and full-term newborns. Results are promising. The estimated values of F_0 and RFs are in the ranges reported in the literature.
The crying of newborns and infants is a functional expression of basic biological needs, emotional or psychological conditions such as hunger, cold, pain, cramps and even joy. It requires a coordinated effort of several brain regions, mainly brainstem and limbic system and is linked to the breath system. Its characteristics reflect the development and the integrity of the central nervous system. Thus, infant cry analysis is a suitable non-invasive complementary tool to assess the physical state of infants particularly important in the case of preterm neonates. Specifically, the distinction between a regular wailing and one with anomalies is of clinical interest. Preterm infants and infants with neurological conditions may have different cry characteristics when compared to healthy full-term infant.

For this reason is important to set up an efficient method for automatic cry analysis.
An automatic method for the estimation of crying acoustical characteristics provides a support to the perceptive analysis made by the clinician reducing the required amount of time often prohibitive in daily clinical practice.

REFERENCES

[1] H. Rothganger, “Analysis of the sounds of the child in the first year of age and a comparison to the language,” *Early Hum Dev*, vol. 75, pp. 55-69, 2003.
[2] P.S. Zeskind, “Infant crying and the synchrony of arousal,” in *The Evolution of Emotional Communication: From Sounds in Nonhuman Mammals to Speech and Music in Man*. Oxford University Press, Oxford, 2013; pp. 155-174.
[3] A. Fort, A. Ismaelli, C. Manfredi, P. Bruscaglioni, “Parametric and non-parametric estimation of speech formants: application to infant cry,” *Med Eng Phys,* vol. 18(8), pp. 677-691, 1996.
[4] P. Sirviö, K. Michelsson, “Sound-spectrographic cry analysis of normal and abnormal newborn infants,” *Folia Phoniatrica et Logopaedica*, vol. 28(3), pp. 161-173, 1976.
[5] http://www.kayelemetrics.com
[6] http://www.fon.hum.uva.nl/praat
[7] P. Boersma, *Acoustic analysis, in Research methods in linguistics*, R. Podesva and D. Sharma (eds.), Cambridge University Press, 2014.
[8] C. Manfredi, L. Bocchi, S. Orlandi, L. Spaccaterra, G.P. Donzelli, “High-resolution cry analysis in preterm newborn infants,” *Med Eng Phys*, vol. 31(5), pp. 528-532, 2009.
[9] 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,” *Biomed Signal Proces*, vol. 8(6), pp. 799-810, 2013.
[10] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebvre, F. Grenez, “Low-frequency vocal modulations in vowels produced by Parkinsonian subjects,” *Speech Commun,* vol. 50(4), pp. 288-300, 2003.
[11] L. Falek, A. Amrouche, L. Fergani, H. Teffahi, A. Djeradi, “Formantic Analysis of Speech Signal by Wavelet Transform,” Conf. Proc- of the World Congress on Engineering, vol. 2, pp. 1572-1576, 2011.
[12] M. A. Ruíz Díaz, C.A. Reyes García, L.C. Altamirano Robles, J.E. Xalteno Altamirano, A. Verduzco Mendoza, “Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis,” *Biomed Signal Proces*, vol. 7, pp. 43-49, 2012.
[13] Chui C. *An introduction to wavelets*, Academic Press, 1995.

FP – EEG-Imaging

DAMPING OF VOCAL FOLD OSCILLATION AT VOICE OFFSET

P. H. DeJonckere[1], J. Lebacq[2]

[1] Dept. Neurosciences, University of Leuven; Federal Institute of Occupational Diseases, Brussels, Belgium
[2] Neurosciences Institute, University of Louvain B-1200 Brussels, Belgium
philippe.dejonckere@med.kuleuven.be; j.lebacq@uclouvain.be

Abstract: **Vocal folds show a damped oscillatory movement while abducting at the end of a vocal emission. The phenomenon can be observed with high speed video and different glottographic methods. The damping reflects important mechanical properties of the vocal oscillator, and cannot be voluntarily controlled. It could become a valuable clinical parameter, particularly in a medicolegal context, but its large variability limits its use. The two main physiological factors accounting for the variability are analyzed, as well as possibilities and limitations of each recording method. For clinical / medicolegal applications, electroglottography with flowglottography seems the most interesting combination, when high speed video is not available. Additional research is required, particularly for defining an adequate protocol, taking in account the reported limitations.**

Keywords : Damping, High Speed Video, EGG, Photoglottography, Flow Glottography

I. INTRODUCTION

At the end of a vocal emission, when the voicing is not interrupted by a laryngeal closure and the airway remains open, the vocal folds (VF) are abducting from the median, phonatory position to the respiratory position, and the transglottic pressure suddenly drops [1], [2]. With adequate instrumentation, it is possible to observe a damped oscillatory movement on each VF after the last contact phase of the two fold edges on the midline. This phenomenon results from combined internal and external friction forces. It is very brief and obviously occurs at a level beyond the scrutiny of traditional videolaryngostroboscopic examination.

Nevertheless, the damping characteristics reflect important mechanical properties of the vocal oscillator, particularly pertaining to the efficiency of the voice production. In concrete terms, the amplitude decrement from cycle to cycle reflects the energy input requested to maintain a steady state oscillation. This means a direct link with the concept of vocal fatigue. It is expected that, in several cases of organic vocal fold pathology, the mechanical properties of the vocal oscillator become altered. Hence, damping characteristics could reflect the changes, e.g. a reduced vocal efficiency.

Furthermore, the damping of the vocal oscillator is an objective phenomenon that cannot be voluntarily controlled by the subject. This makes it particularly interesting in a medicolegal context for people claiming compensation.

However, standardizing the recording methodology as well as avoiding biases appear to be a major issue. For current clinical and medicolegal applications, non-invasiveness is mandatory. The method should also interfere as little as possible with spontaneous phonation.

The present study compares different techniques and approaches suitable for investigating and quantifying the damping phenomenon, discusses advantages and disadvantages, and points out pitfalls and limitations. Potential clinical applications are considered.

II. METHODS

The different available adequate techniques are :

(1) Microphone
(2) Intraoral pressure transducer
(3) Flow glottograph (Rothenberg mask; FLOG)
(4) Pneumotachograph
(5) Electroglottograph (EGG)
(6) Photoglottograph (PGG) Mode 1 (tracheal transillumination)
(7) Photoglottograph (PGG) Mode 2 (pharyngeal illumination)
(8) Ultrasonic glottograph
(9) Videokymograph (VKG; single line scanner)
(10) High speed film

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

III. RESULTS

Imaging techniques (high speed video and VKG) now make continuous digital recording possible. For VKG, the position of the scanned line needs to be controlled. Actually, only true high speed video with possibility to extract and display a posteriori the kymograms of several selected lines is suited. (Figs. 1 & 2, 3 & 4) [3]. Software is available for automatic computing of VKG-parameters [4].

The method requires laryngoscopy with a rigid endoscope, and full vocal fold visualization, but is not invasive. Only sustained sounds can be investigated. However, the required material currently limits this approach to research situations.

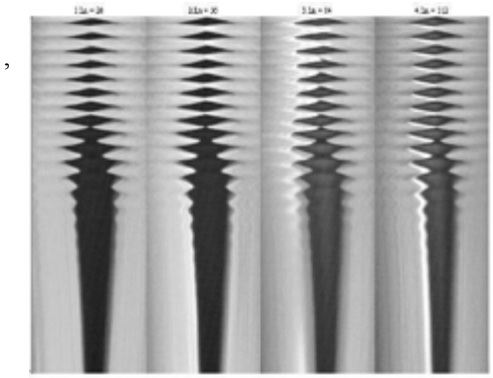


Fig.1. VKG at 4 levels of the glottis obtained from high speed video. Healthy male subject. End of a /a:/ at comfortable pitch and loudness.

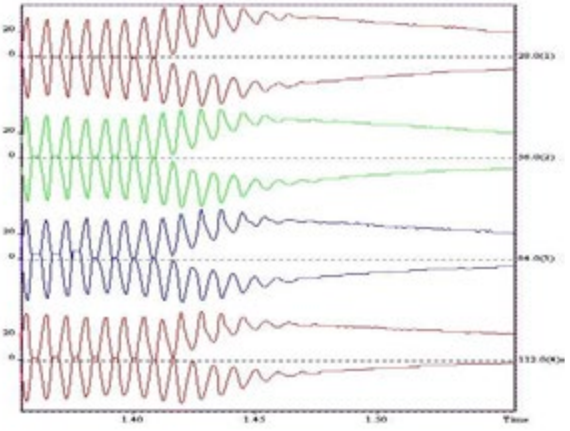


Fig.. 2. Movements of the vocal fold edges, computed from the videokymograms of Fig. 1. The damping phase lasts about 8 cycles.

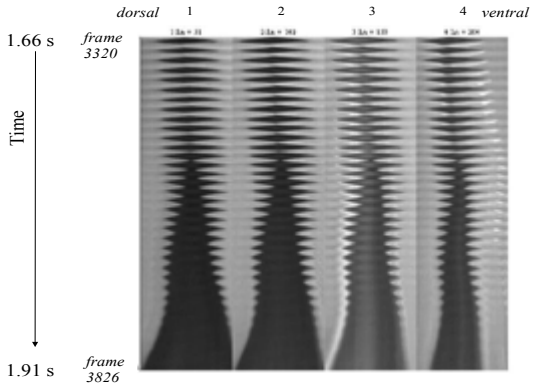


Fig.. 3. VKG at 4 levels of the glottis obtained from high speed video. Healthy male subject. End of a somewhat breathy /a:/ at comfortable pitch and loudness. Due to persistence of some airflow and slow vocal fold abduction, the damping phase lasts at least 20 cycles, starting with a progressive shortening of the closed phase

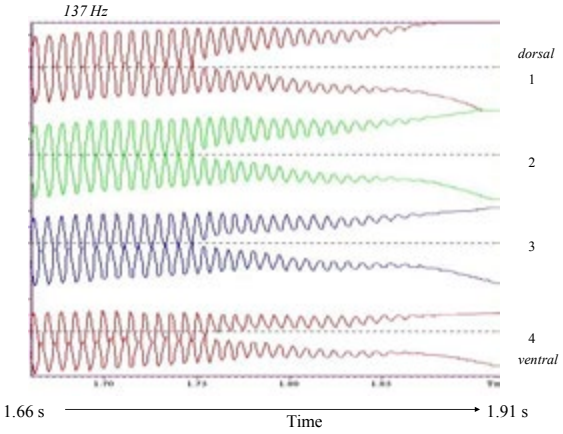


Fig..4 Movements of the vocal fold edges, computed from the videokymograms of Fig. 3.

Methods measuring variations of acoustic pressure (microphone, intraoral pressure, flow glottograph pneumotachograph) are patient-friendly and totally not invasive, but actually they measure the VF movements through an air buffer, which has inertia and resonance characteristics, the latter depending on the vocal tract configuration. To properly measure the damping, the signal needs to be corrected (inverse filtering) as does the flow glottography device (Rothenberg mask). Furthermore, a relevant information also provided by the flowglottogram, is the end of vocal fold contact

(closed plateau). The mask limits to some extent speech movements.

Photoglottography actually measures the true movements of the vocal folds, both in mode 1 (photodiode in pharynx, light source on pretracheal skin) and in mode 2 (light source in pharynx, photodiode on pretracheal skin) [5]. It is less invasive than imaging techniques, but more so than electroglottography. When combined with flowglottography, it can validate the signal : Fig. 5 shows a quasi-perfect temporal correspondence between the damping on photoglottographic and flowglottographic signals.

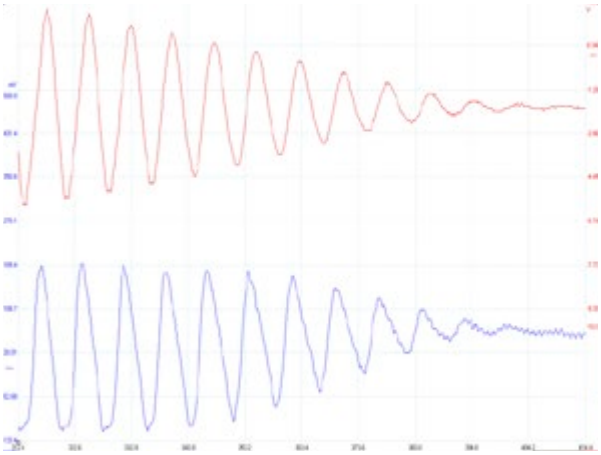


Fig. 5 : Simultaneous recording of flowglottogram (upper trace) and photoglottogram in mode 1 (lower trace). Healthy male subject. 130 Hz, 65 dBA. The damping is comparable.

Electroglottography is patient-friendly and does not interfere with vocalization. It allows precise phonetic tasks, with acoustic control. However, the sensitivity for detecting very small transglottic impedance variations (essential in this context) depends on the electronic design. Recent devices are characterized by a higher carrier-wave frequency (e.g. 1 MHz), and a more efficient feed-back control of the oscillator. Furthermore, the output uses a multipole filter, with sharper cut-off and flat bandwidth response. As a result, a better signal-to-noise ratio and a higher sensitivity are achieved with a larger bandwidth and better linearity [6]. As shown in Fig. 6, the EGG-signal can be as sensitive as the flowglottogram for detecting the smallest vocal fold oscillations, but – contrary to the flow signal - fails to actually show the final phase of the damping. The last ten sinusoidal EGG-cycles probably correspond to small (reduced amplitude) impedance variations at the level of the ventral commissure.

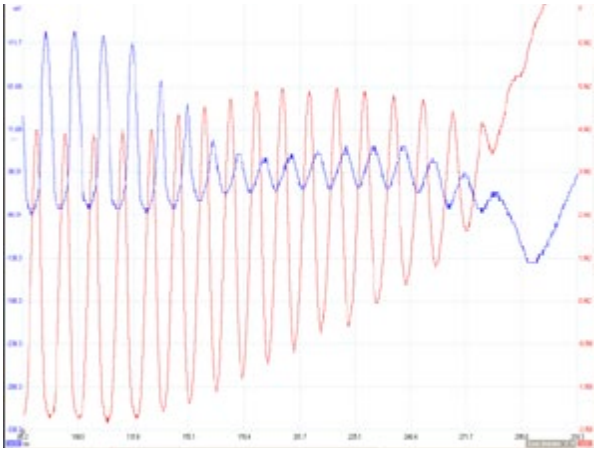


Fig. 6 : Simultaneous recording of the EGG signal (upper trace) and the flowglottographic signal (lower trace). Healthy male subject, 125 Hz, 65 dB. the EGG-signal is as sensitive as the flowglottogram for detecting the smallest vocal fold oscillations, but fails to actually show the final phase of the damping.

Imaging techniques, EGG, flowglottography and photoglottography adequately identify the first part of damping, when the vocal folds still have contact on the midline. In this part, the damping phenomenon is characterized by a progressive reduction of the closed phase, possibly concomitant with a slight increase in amplitude of oscillation.

Ultrasonic glottography is still experimental and does not allow sufficient control of position/orientation of the probe.

IV. Discussion

For phonation at comfortable pitch and loudness, two main physiological parameters seem to account for the major observed variability in damping of the vocal fold oscillation at voice offset. Both are actually related to some persistence of the driving force : (1) the timing dynamics of the expiratory pressure (muscular / elastic, depending on lung volume) with respect to the opening of the glottis; (2) the speed with which the vocal fold edges are abducted and the glottal resistance drops. The combined effect of (1) and (2) actually determines the persisting transglottal flow. Furthermore, from a certain degree of abduction, the morphology of the oscillator changes.

Creating an abrupt airflow interruption at subglottic level is practically ruled out in vivo. Airflow can be interrupted downstream, either artificially by an inflatable balloon within the pneumotachograph, or physiologically by linguo-palatal occlusion (e.g. in

/uk/). In the case of an artificial abrupt interruption during a /a:/, transglottal airflow (and vocal fold vibration) can persist by limited inflation of the upper vocal tract, upstream of the occlusion. In case of a /uk/ (or another denasal vowel followed by a voiceless occlusive), the vocal fold vibration usually stops (in a variable way) before the articulatory movement of the /k/. When the occlusive is voiced (/Ug/) the vocal fold vibration persists after the linguo-palatal occlusion, but again, as for artificial airflow interruption, in a very variable way.

A persisting transglottal flow after the last VF contact clearly slows the damping, as can be seen by comparing figs. 1-2 with figs. 3-4, obtained from the same (healthy) subject at a few seconds interval. In fact, this is the main problem and limitation with all imaging methods, as well as with photoglottography. The possibility of searching for a critical repetition rate of a denasal vowel followed by a voiceless occlusive was tried out : this means the rate (e.g. 6,2 s^{-1}) at which the oscillation is actually interrupted (Fig. 7). In such a task, repetition contributes to standardizing several parameters. This protocol could be of some interest, but it requires a trained vocalist, and seems unsuitable suited for clinical or medicolegal application.

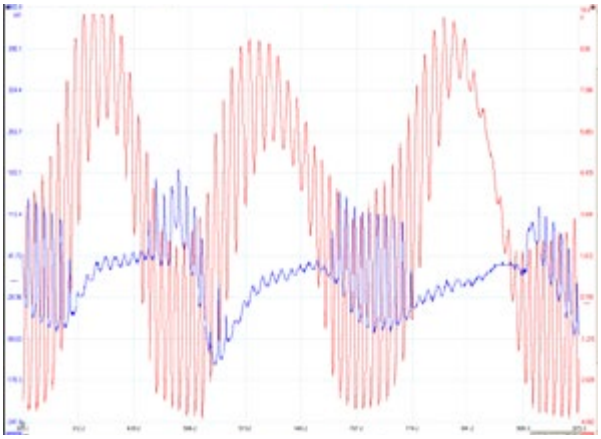


Fig. 7 : repetitions of /uk/ while slowering the rate. Between the first two repetitions, no interruption of vocal fold oscillation occurs, while this just occurs between the second and the third repetition.

All observations so far deal with phonation at comfortable pitch and loudness. Changes in intensity and fundamental frequency also influence the damping characteristics, but this effect is by far smaller than that of the other considered factors.

Our limited experience with organic pathology mainly shows irregularity during the damping phase, with chaotic patterns. This can probably be attributed to the fact that the absence of midline collision and the progressively reduced vibrating mass act as sensitizing factors for asynchronisms due to superficial lesions. This particularity could be of some use in objectifying micro-organic vocal fold pathology.

For clinical and medicolegal applications, additional work is necessary, mainly in the direction of defining an adequate protocol, taking in account the reported limitations. Apart high speed video, the combination of EGG and flowglottography appears the best formula.

V. Conclusion

The damping phenomenon is probably a major issue in voice pathophysiology. It could become a valuable clinical parameter, particularly in a medicolegal context, but its large variability, for physiological reasons, limits its application. Additional research is required, particularly for defining an adequate protocol, taking in account the reported limitations.

References

[1] P.H. Dejonckere, J. Lebacq, “Mechanical study of the damping of the phonatory oscillator.” *Arch. Internat. Physiol. Bioch.* vol. 88 : pp. 31-32, 1980.

[2] P.H. Dejonckere PH. “Damping biomechanics of vocal fold oscillation” , in *Vocal fold physiology : Acoustic, perceptual and physiological aspects of voice mechanisms*, J. Gauffin and B. Hammarberg Eds. San Diego : Singular Publishing Group, Inc. San Diego. 1991 , pp. 105 – 111.

[3] P.H. Dejonckere, H. Versnel, “High-speed imaging of vocal fold vibration : analysis by four synchronous single-line scans of onset, offset and register break.” , in *Proceedings of the XVIII I.F.O.S. (International Federation of Oto-rhino-laryngological Societies) World Congress* 25-30 June, D. Passali, Ed. Rome 2005, pp. 1–8.

[4] P.H. Dejonckere, J. Lebacq , L. Bocchi L, S. Orlandi, C. Manfredi, Automated tracking of quantitative parameters from single line scanning of vocal folds: a case study of the 'messa di voce' exercise. *Logoped Phoniatr Vocol.* Vol 40 : pp. 44-54, 2015.

[5] P.H. Dejonckere,“Comparison of Two Methods of Photoglottography in Relation to Electroglottography”. *Folia Phoniat.* vol 33, pp. 338–347. 1981.

[6] J. N. Sarvaiya, P.C. Pandey, V.K. Pandey, An impedance detector for glottography. *IETE Journal of Research* vol 55 : pp. 100-105, 2011.

EVALUATION OF GLOTTAL WAVES VARIABILITY BASED ON COMBINED AMPLITUDE-VELOCITY ANALYSIS

A. Nacci[1], A. Macerata[2], J. Matteucci[1], M. Manti[3], M. Cianchetti[3], S. O. Romeo[1], B. Fattori[1], S. Berrettini[1], C. Laschi[3], F. Ursino[4]

[1] ENT Audiology Phoniatrics Unit, Department of Neuroscience, University of Pisa, Pisa, Italy
[2] Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy
[3] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
[4] National Institute for Research in Phoniatrics, University of Pisa, Pisa, Italy
a.nacci@med.unipi.it; almacerata@gmail.com; jacopo.matteucci@hotmail.it; m.manti@sssup.it; matteo.cianchetti@sssup.it; s.romeo@gmail.com; bruno.fattori@med.unipi.it; s.berrettini@med.unipi.it; cecilia.laschi@sssup.it; profursino@gmail.com

Abstract: **A method for analyzing the EGG signals and for extracting features able to quantitatively characterize phonation is introduced. The EGG signal is processed in order to obtain the first derivative (velocity of vocal folds changes). The average fundamental frequency is computed and its corresponding period *T0* is taken as typical duration of the EGG cycle. The EGG signal and its derivative are processed for extracting each single glottal cycle which is normalized in time to the fixed length *T0*. For each glottal cycle, the amplitude and relative velocity signals are plotted in a X-Y graph so forming a multi-layer display where each EGG cycle appears as a circular loop; by using all the EGG loops the mean curve and related std values are computed and displayed. The mean std value is taken as Variability Index (*VI*) of the phonation process. The process can be characterized in more details by computing the same index as obtained by dividing the graph in 4 quadrants, roughly associated to the different phases of glottal cycle. In this preliminary study, the EGG analysis has been carried out for 30 cases (15 normal and 15 pathological voice), considering the variability based on the combined amplitude-velocity analysis. In normal subjects, the variability indices showed a definitely lower value than those obtained in pathological subjects. This difference was statistically significant ($p < 0.03$).**

Keywords: **EGG, first derivative, amplitude, variability index, glottal cycle.**

I. INTRODUCTION

The EGG is sensitive to changes in vocal fold contact area during phonation: it proves to be a valuable tool for both voice researchers and clinicians. Clinical observation and the application of various physical and mathematical models have been used to identify important EGG signal landmarks, as well as to correlate changes in signal morphology with specific aspects of laryngeal physiology and physiopathology. From a clinical point of view, the advantages of EGG are the following: the cycle is repeated at each contact, and its frequency is considered the most accurate indicator of the Fundamental Frequency (F0) [1-3]; still nowadays, the EGG tracking demonstrates to be the best representation of the oscillation of the glottis as a whole and particularly during its closing phase [2-4]. When used with high-speed imaging and acoustic analysis, EGG is able to analyze irregular vibratory patterns [5].

Actually, the EGG is a one-dimensional signal as obtained from the complex three-dimensional motion of the vocal folds. The speed of such motion is strictly related to the closing and opening phases of the vocal folds activity. The speed, i.e. the first mathematical derivative of the EGG waveform (DEGG), reflects the rate of change of the EGG with time [1]. The EGG and its first derivative, are rich of useful information on the vocal folds activity. Quantitative analysis of EGG could offer a valuable tool for evaluating the real behaviour of vocal folds in normal and pathological subjects. Quantitative analysis of the EGG waveform has been achieved by measuring the relative proportion of glottal closure within a glottal vibratory period [6], known as the "larynx closed quotient" [7] or "contact quotient" [8] (CQEGG). This quotient has been found useful in clinical as well as in basic voice research [9]. Research has shown, however, that the CQEGG is dependent on the choice of the algorithm used to determine the contacting and de-contacting events, and must therefore be used with caution [10,11].

In the present study, a new approach for data reduction of the electroglottographic signal will be presented. The method is based on the analysis of the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

EGG signal and its first derivative; it allows to extract quantitative indices about the EGG activity during opening-closing phases of the vocal folds process during steady-state vocal tests. Moreover, in this preliminary study, the EGG analysis of 15 cases of normal voice and 15 cases of pathological voice has been carried out, in order to demonstrate any quantitative differences between pathological and normal subjects.

II. METHODS

Thirty subjects participated in this study. The study group consisted of 15 patients (10 females, 5 men; mean age: 34.6 ± 3.4 SD) with voice quality disorders (disfunctional and/or organic dysphonia). The control group consisted of 15 subjects (11 females, 4 men; mean age: 35.2 ± 2.8 SD) with euphonic voice. Laryngeal electroglottography (EGG) (KAY Model 6103) was performed on all subjects while phonating the vowel [a] at a comfortable pitch and loudness. The recording of the EGG signal was performed four times for each subject during the same day (at 8 AM, 11 AM, 2 PM, 6 PM), on two different days. In this way, 8 recordings of EGG signal for each subject have been performed, with a total of 120 recordings in the pathological subjects and 120 recordings in the control group, in order to calculate and subsequently examine intra- and inter-individual variability, as well as the difference between different groups.

EGG Acquisition and Preprocessing: The EGG signal could be affected by changes in position of the plates, by muscle activities, by internal body parts movements. This provokes "noise" in the EGG signal which appears as low frequency baseline drift, high frequency noise and artefacts. In laboratory activities, an accurate protocol for signal acquisition has to be adopted. Despite these precautions, some noise could be still present and pre-processing has to take care for reducing residual artefacts and enhancing the real EGG component.

The original EGG signal is obtained from commercial instrumentation in form of standard WAV file with sampling rate of 44KHz. All the following signal processing is done by using the software package MatLab Release R2012a-Win64 (Mathworks Inc.).

The EGG signal was first filtered through a FIR band-pass filter (80-10000Hz) with window length=5000 samples, thus obtaining a series G_i for $i=1,\ldots N$ where N is the number of samples within the interval of the analysis. This filtering cleans the signal from the baseline slow movements and from the high frequency noise. The first derivative (DEGG) of this filtered signal was then computed, obtaining the series D_i for $i=1,\ldots N$. The DEGG signal was submitted to the Fast Fourier Transform (FFT) for spectral analysis and the frequency corresponding to the maximum amplitude of the spectrum is taken as Fundamental Frequency (***F0***).

Detection of glottal cycles: The EGG signal is a pseudo-periodical signal where the period is the time length of the wave representing the opening-closing movement of the vocal folds. In order to characterize each glottal cycle the beginning of each cycle has to be detected first. A proprietary software was applied to EGG and DEGG data obtaining an accurate estimate of the beginning of each glottal cycle, defined as the instant where the vocal folds are open and the closing phase starts.

Time normalization of glottal cycle: In this study the attention was focused on the variability of the EGG cycle shapes. In the hypothesis that the vocal folds behaviour (that is, the progress and sequence of body part movements during the phonation process) will be independent by the cycle duration, each EGG wave has been forced to the same time length. So each EGG cycle has been re-sampled by interpolation to obtain a new cycle of fixed length $L=1001$ samples. The same was done for the corresponding DEGG cycles; note that this procedure does not change the derivative values.

Glottal cycles as X-Y plot: The two distinct series EGG and DEGG have been then transformed into a sequence of blocks of consecutive glottal cycles, each one containing the same-length couple of EGG and DEGG samples. In order to show the whole information contained both in EGG and DEGG, an X-Y plot has been used, where the X axis represents the EGG and Y axis represents the DEGG. Overlapping all the glottal cycles on the same graph, a synthetic view of the entire process of phonation at vocal folds level is obtained in a single picture (Fig. 1).

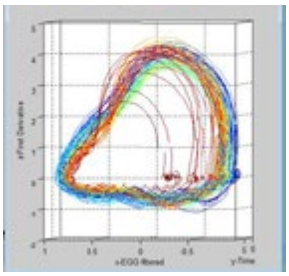


Fig. 1: EGG-DEGG plot.

Extraction of glottal features: Despite the potentiality of the above mentioned graph in the analysis of the vocal folds activity, the researcher is still required for its visual inspection and interpretation. To overcome this limit, this study proposes to extract

from the overall graph a set of features capable to characterize the entire vocal folds process in an objective and quantitative way. The graph is made by a defined number of points $U=M*L$, where M is the number of cycles and L is the fixed number of samples per cycle. Recalling what described in the previous section, the position of each point of the graph has a physiological meaning associated to the movement of the vocal folds (*VF*). The area of the graph can be roughly divided in 4 panels: 1) start of closure of *VF* up to the their maximal speed in closure, 2) from maximal speed in closure down to 0 and complete closure of *VF*, 3) start of *VF* opening up to the maximum of negative speed, 4) from maximum negative speed up to complete *VF* opening. The centre C_0 of this area is taken at X_c=mean value of the EGG samples and Y_c=0, i.e. first derivative (or speed) equal to zero. Reminding that the generic jth loop on the graph is actually the temporal sequence of points (x_{ji}, y_{ji}) where $i=1,\ldots,L$, at each ith instant it is possible to compute the mean value $P_i(X_{Mi}, Y_{Mi})$ and the standard deviation $SD_i(X_{SDi}, Y_{SDi})$ from the set of M cycles. A new variability index for the entire glottal process is defined as: $VI = (\Sigma\, SD_i)/T0$.

The mean and standard deviation of the M loops are represented in the Fig. 2, where the solid black line corresponds to the mean glottal cycle P_i, and the gray band is made by segments of rays proportional to the standard deviation SD_i, where $i=1,\ldots,L$. The mean loop surrounds the centre C_0 of the graph and passes through the 4 quadrants. As each quadrant is associated with a specific phase of the behaviour of the vocal folds, four variability indices were extracted, one for each of the quadrants:

VI_1: during the initial closing activity;
VI_2: during the last phase of *VF* closure;
VI_3: during the first phase of *VF* opening;
VI_4: during last phase up to complete opening of *VF*.

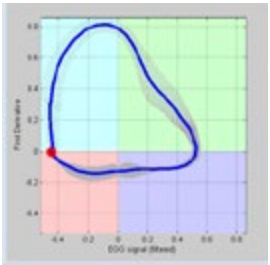


Fig. 2: Glottal cycle represented in a 4 panels graph.

Statistical analysis: A Student's t-test was used to compare the mean values of total VI, VI_1, VI_2, VI_3 and VI_4 obtained in the pathological group and in the control group. The results were considered statistically significant for p values < 0.05. Statistical analyses were carried out with Stat View 2.0 software.

III. Results

Intra-individual variability: The 8 EGG recordings carried out (four times during the same day at 8 AM, 11 AM, 2 PM, 6 PM on two different days) for each normal and pathological subject, showed no statistically differences for each subject (intra-individual variability has not been demonstrated).

Inter-individual variability: By averaging the VI (total VI, VI1, VI2, VI3 and VI4) of the 8 recordings of normal subjects and comparing them to each other in the same group, no statistically differences were found. Similarly, within the study group no statistically differences were found for the same parameters of VI.

Study group vs. Control group: Comparing the graphs obtained from EGG, there is a greater variability of the signal in pathological subjects compared to normal subjects. Comparing the mean VI of the study group with the mean of the control group, a statistically significant difference was found. Specifically for total VI, which represents the total variability in the four quadrants, the mean value was worked out at 0.283±0.13 for the study group, while for the group of normal subjects the mean value was worked out at 0.131±0.04 ($p < 0.001$). For VI1, in the study group the mean value was worked out at 0.109 by ± 0.15 while in the control group the mean value was worked out at 0.017±0.01 ($p < 0.007$); for VI2, the mean values were respectively worked out at 0.067±0.03 and 0.037±0.02 ($p < 0.001$); for VI3 the mean values were respectively worked out at 0.040±0.02 and 0.026±0.02 ($p = 0.03$). Finally, for VI4 the mean values were respectively worked out at 0.068±0.03 and 0.048±0.02 ($p = 0.01$).

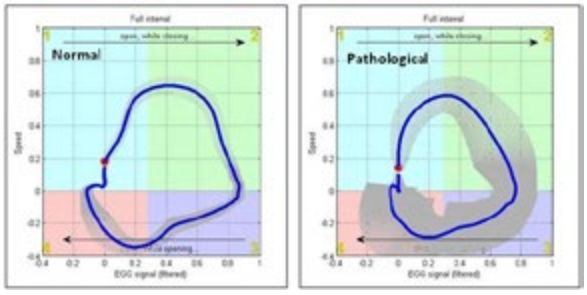


Fig. 3: Graph of normal vs pathological subjects.

IV. Discussion

The EGG tracking proves to be the best currently available representation of the oscillation of the glottis as a whole, and particularly during its closing phase [2-4]. Although EGG waveforms are unaffected by activity of vocal tract or environmental noise, this test shows some limits. In particular, variations between individuals impede the definition of pathological and normal voice, as well as the type of pathology.

Moreover, the EGG waveforms are easily affected by normal variations, such as mucous string across the glottis [1,12]. Clinical observation and the application of various physical and mathematical models have been used to identify important EGG signal landmarks and to correlate changes in signal morphology with specific aspects of laryngeal physiology and physiopathology. Many of these techniques are capable to offer a detailed view of the vocal folds process over the time, however the operator is always required to perform the analysis by visual inspecting the graphics [13-15].

In the present study, a new approach for data reduction of the EGG signal has been presented. The method is based on the analysis of the EGG signal and its first derivative; it allows to extract quantitative indices regarding the EGG activity during opening-closing phases of the vocal folds process in steady-state vocal tests. In particular, the Variability Indices (VI), that is the expression of Amplitude and Velocity variation, have been extracted.

The results of this preliminary study, performed on dysphonic and euphonic subjects, show that this type of EGG analysis allows to distinguish between pathological and normal voice. One of the limitations of the EGG examination with commercial devices is just the incapability to quantitatively characterize the normal and pathological voice. Instead, with this novel approach, and in particular through the calculation of the variability index (total VI), it is possible to demonstrate a highly significant difference between normal and pathological subjects ($p < 0.001$). The statistically significant difference was also confirmed for the variability index of each quadrant corresponding to the different phases of the glottic cycle, confirming that in the subjects with pathological voice, when compared to the normal, all phases of the vibratory cycle of the vocal folds are altered. Analyzing the data in detail, however, it is observed that the statistical difference becomes highly significant especially for the VI_2 parameter, which corresponds to the variability of the signal during the last phase of the vocal folds' closure ($p < 0.001$). In fact, in this phase the presence of an organic lesion of the vocal fold (and in particular of the vocal fold edge), determines a higher variability of the Amplitude and Velocity, in particular during the last phase of closure (VI_2), compared to the other stages.

V. Conclusion

The data exposed are preliminary but promising, since the quantitative evaluation of EGG has proved to be useful to distinguish pathological from normal voice, through the calculation of the Variability Index (VI). Further studies on larger groups of subjects will be required, in order to confirm these preliminary results, to assess any quantitative differences in the different phases of the glottal cycle and to ultimately show any differences in EGG signal between different pathologies of vocal fold.

REFERENCES

[1] D.G. Childers, and A.K. Krishnamurty, "A critical review of electroglottography", *Crit Rev Biomed Eng,* vol. 12, pp. 131-161, 1985.

[2] R.J. Baken, *Clinical measurement of speech and voice*. San Diego: Singular Publ Group Inc, 1997.

[3] A. Fourcin, "Precision stroboscopy, voice quality and electroglottography", in *Voice Quality Measurement,* R.D. Kent and M.J. Ball Eds. San Diego: Singular Publishing Group, 2000.

[4] M. Hirano, *Clinical examination of voice*. New York: Springer Verlag, 1981.

[5] M. Saito, H. Imagawa, K. Sakakibara, N. Tayama, K. Nibu, and M. Amatsu, "High-speed digital imaging and electroglottography of tracheoesophageal phonation by Amatsu's method", *Acta Otolaryngol*, vol. 126, pp. 521-525, 2006.

[6] M. Rothenberg, and J.J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area", *J Speech Hear Res,* vol. 31(3), pp. 338-351, 1988.

[7] D.M. Howard, "Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers", *J Voice*, vol. 9, pp. 163-172, 1995.

[8] R.F. Orlikoff, "Assessment of the dynamics of vocal fold contact from the electroglottogram: data from normal male subjects", *J Speech Hear Res,* vol. 34(5), pp. 1066-1072, 1991.

[9] D.G. Miller, H.K. Schutte, and J.Doing, "Soft phonation in the male singing voice: a preliminary study", *J Voice,* vol. 15(4), pp. 483-491, 2001.

[10] C. Herbst, and S. Ternström, "A comparison of different methods to measure the EGG contact quotient", *Logoped Phoniatr Vocol*, vol. 31(3), pp. 126-138, 2006.

[11] R.E. Kania, S. Hans, D.M. Hartl, P. Clement, L. Crevier-Buchman, and D.F. Brasnu, "Variability of electroglottographic glottal closed quotients: necessity of standardization to obtain normative values", *Arch Otolaryngol Head Neck Surg*, vol. 130(3), pp. 349-352, 2004.

[12] D.G. Childers, D.M. Hicks, G.P. Moore, L. Eskenazi, and A.L. Lalwani, "Electroglottography and vocal fold physiology", *J Speech Hear Res*, vol. 33, pp. 245-254, 1990.

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

[14] J.G. Svec, F. Sram, and H.K. Schutte, "Videokymography in voice disorders: what to look for?", *Ann Otol Rhinol Laryngol,* vol. 116(3), pp. 172-180, 2007.

[15] J. Lohscheller, U. Eysholdt, H. Toy, and M. Dollinger, "Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics", *IEEE Trans Med Imaging,* vol. 27(3), pp. 300-309, 2008.

EEG SONIFICATION FOR CLASSIFYING UNSPOKEN WORDS

Erick Fernando González-Castañeda, Alejandro Antonio Torres-García, Alejandro Rosales-Pérez
Carlos Alberto Reyes-García, Luis Villaseñor-Pineda

Biosignals Processing and Medical Computing Laboratory, Computer Science Department
National Institute of Astrophysics, Optics and Electronics (INAOE), Tonantzintla, Puebla, Mexico,
{erick.gonzalezc,alejandro.torres, arosales, kargaxxi,villasen}@inaoep.mx

Abstract: **Brain-computer interfaces (BCI) based on electroencephalograms (EEG) are an alternative technique that aims to integrate people with severe motor disabilities to their environment. However, they still are not used in everyday life because controlling electrophysiological sources is nowadays an unintuitive task. To address this problem, work has been carried out with the objective of classifying EEG signals recorded while imagining speech.**
In this paper sonication technique on EEG signals was used, which allows us to characterize EEG signal as an audio signal. The aim is to analyze whether the application of the sonification process to EEG signals can help to do a better discrimination or to highlight patterns in order to improve classification results of unspoken words. For proving this we processed signals with and without sonication. The results of the signals coming from the four nearest channels to the language areas of Broca and Wernicke were obtained. The average accuracy rates for signals without applying sonification and applying sonification are 48.1% and 55.88% respectively, for which it could be observed that the method of sonification of EEG improves slightly classification rates.

Keywords: **Electroencephalogram (EEG), Brain-Computer Interfaces (BCI), Sonification, Imagined / Unspoken Speech, Random Forest.**

I. INTRODUCTION

In the search for a way to integrate into society people with severe motor disabilities, the use of brain activity captures by electroencephalography (EEG) to control devices and interfaces has been explored. Generally a BCI can be seen as a recognition system where the EEG pattern is used as the primary source of information, a learning algorithm is used to learn an inference function from the EEG, and finally, according to the output predicted by the algorithm the desired output is executed in the device to be used. The BCIs requires neurological mechanisms to generate the control signals. The most used are the slow cortical potentials (SCP), potential P300, motor imagery (sensory motor rhythms mu and beta) and visual evoked potentials (VEP) [1]. These mechanisms have a long training period required for a user to employ a BCI. This is because these sources are generated by the user in an unconscious way. Another inconvenient are the low communication rates (one word or less, processed per minute) which are insufficient to allow natural interaction. This latter problem consists in that each of these sources requires a "mapping" or translation to the speech domain. To address these problems some works explored the use of the imagined speech also referred to as inner speech or unspoken speech. This term refers to the internal pronunciation of words without making sounds or articulating gestures to do it [2]. With the unspoken speech is expected that control patterns will be generated more consciously and training time to control a BCI will be minimal.

Sonification: The concept of sonification or 'Auditory display' concerns to the use of non-spoken sound to transmit information. The sonification of EEG has been used for exploratory analysis of EEG signals [3], to make musical compositions [4], to analyze patients with epileptic events [5] or to make early diagnosis of neurological diseases such as Alzheimer's disease by audible feedback [6]. There are various techniques for sonification of EEG signals as: audification, allocation by mapping parameters, sonification based in models [7], sonification based in modeling bumps [8], among others.
Our research aims to apply an EEG sonification method to obtain audible information of the cerebral signal which was previously registered during the imagined pronunciation of words and which also allows to highlight patterns to help an automatic classifier to improve the accuracy percentages reported in [9].

II. METHODS

This work consists of the following stages: acquisition of brain activity, preprocessing, sonification, feature extraction and classification. The methodology stages are shown in Figure 1. It is noteworthy that the EEG signals will also be processed without the sonification step, so as to have a comparison framework to the work described in [9].

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

Acquisition of brain activity: At this stage we used the set of EEG data recorded while imagining speech used in [9]. This data set consists of the EEG signals from 27 individuals whose native language is Spanish. EEG signals were recorded with the Emotiv EPOC kit. This is a wireless kit and consists of 14 high resolution channels (see Figure 2) whose frequency sampling is 128 Hz.

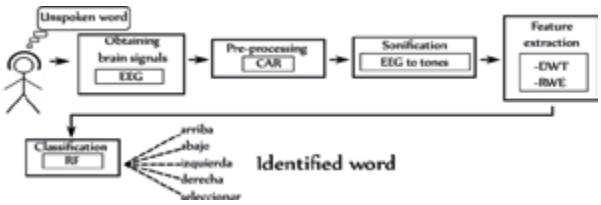


Figure 1: Diagram of the methods used to classify EEG signals captured while imagining speech.

The data set consist in the imagined pronunciation of five words in Spanish: “arriba” (up), “abajo” (down), “izquierda” (left), “derecha” (right), y “seleccionar” (select). Pronunciation of each word was repeated 33 times in succession during the EEG recording. Before each block the subject was told which of the words should be pronounced internally. All epochs from the same individual were recorded in a single session and all sessions were recorded on a laboratory isolated from external audible noise and visual noise.

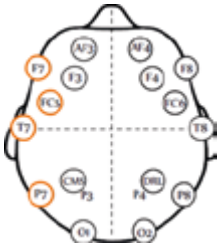


Figure 2: Location of the electrodes in the Emotiv EPOC kit. The channels corresponding to Geschwind-Wernicke model are marked with orange.

Pre-processing: The collected EEG signals are pre-processed using the Common Average Reference (CAR) method. This method is applied in order to improve the signal to noise ratio of the EEG. Basically, we seek to get rid of everything common to all simultaneous electrode readings. CAR can be calculated by subtraction of the potential among each electrode and the reference (the average potential of all the channels). This procedure is repeated for every time instant in the sampling frequency.

EEG Sonification: The sonification technique called 'EEG to tones' [10] was used in this work. In this technique the EEG frequencies are scaled to audible frequencies. This technique is based on the spectrogram of the input EEG signal, which is calculated using the fast Fourier transform (FFT). The dominant frequencies of the EEG signal are calculated from each window of the spectrogram. The dominant frequencies are scaled to tones, representing a frequency in the audible range. At the end of each set of dominant tones per window is joint to form the output audio. The pseudocode in Algorithm 1 explains in more detail the method used.

Algorithm 1 Sonification of EEG

Require: *EEG* (EEG signal), *NT* (Number of Tones), *LowF* (Minimum Frequency of. EEG signal), *HiF* (Maximum Frequency of EEG signal), *W* (Spectrogram Window Size), *Shf* (Overlap between Spectrogram Windows), *Dur* (Tones Duration), *LowFAu* (Audio Minimal Frequency), *HiFAu* (Audio Maximal Frequency) *Fs* (Audio Sampling Frequency)
Ensure: *Audio* (sonification of the EEG signal)
Spec ← Build Spectrogram of the EEG signal, using *EEG, LowF, HiF, W, Shf*)
Scale amplitudes of *Spec* frequency dividing by the maximum amplitude.
for $i \leftarrow 1$ to *Spec* horizontal size **do**
    Sort descending $Spec.column_i$ according to amplitudes.
    $Winners.column_i \leftarrow$ take from $Spec.column_i$ the first NT frequencies and their amplitudes.
    TonesF ← Scale the unique frequencies from *Winners* to the audible range using *LowFAu* and *HiFAu*.
    Tones ← Create the sine wave for each frequency in *TonesF* using *Fs* and *Dur*.
end for
for $j \leftarrow 1$ to *Winners* horizontal size **do**
    Audio ← Attach the previous value of *Audio* to the Sum of the signals in tones corresponding to the winning frequencies in $Winners.column_j$ according to *Tones* and *TonesF*.
end for
return *Audio*

Feature extraction using Discrete Wavelet Transform (DWT): The DWT is a technique that allows modeling variations in the time scale domain is the discrete wavelet transform. DWT analysis can be performed using a fast pyramidal algorithm described in terms of multi-rate filter banks, ie, those having more than a sampling rate performing conversions by decimation and interpolation operations. In the DWT each sub-band contains half of the samples of the highest neighboring frequency sub-band. In the pyramidal algorithm the signal is analyzed in different frequency bands with different resolutions by decomposing the signal into a rough approximation (approximation coefficients) and detailed information (detail coefficients). The rough approximation is then further decomposed using the same wavelet decomposition

step. This is achieved by a successive low-pass and high-pass time signal filtering, and a sub-sampling. Details of the process above are described can be found in [11].
In this paper the discrete wavelet transform is applied to sonified audio files. 6 decomposition levels were calculated using Daubechies mother wavelet of order 20 (db20). Likewise, for the case of EEG signals that were not sonified, the DWT is calculated with 5 decomposition levels using second order Daubechies (db2) as described in [9]. As is evident, the number of wavelet coefficients in each of the levels will vary depending on the size of the EEG signal delimited by the markers. This is because, similar to conventional speech, the duration of windows of the imagined pronunciation of a word is variable in windows of a single individual as in windows of different individuals. To deal with this problem, the wavelet coefficients are normalized by the relative wavelet energy described below.

Relative wavelet energy: After applying the DWT on the signal approximation and detail coefficients are obtained, from which is possible to calculate the relative wavelet energy. The relative wavelet energy represents the energy that some level of decomposition provides to the total of the wavelet energy of the signal. The relative energy provides information to characterize energy distribution of the signal in different frequency bands, where independence of the size of the EEG or audio signal window is obtained, as appropriate.
From the above description, in the EEG sonified signals it was determined to use 10 values representing wavelet energy at all decomposition levels and the last of approximation (D1-D9 and A9). While each imagined speech window of the non sonified EEG signals is represented by a set of 5 wavelet energy values, 4 of the decomposition levels and one of approximation (D2-D5 and A5) with respect to the total wavelet energy. As it was performed in [9], the value associated with D1 is discarded.

Classification: The aim of the classification is to infer a relationship between a data vector and a possible class (or category), for doing that we need create a model that automatically finds those relationships. The model is created based on a partition of the training data vectors. The learned models of the training data are then evaluated with a different testing set to determine if the models can be generalized to new cases. In this paper we train and test the Random Forest (RF) classifier with a 10 fold cross-validation approach.

Random Forest (RF): RF is a combination of predictor trees where each tree depends on the values of an independently sampled random vector and with the same distribution for all trees in the forest. Each tree casts a single vote for the most popular class for a given input *x*, and at the end the RF output is obtained using majority vote. In [12] the RF classification algorithm is described in detail including the process to build individual trees.
In this paper, we implemented the classifier using the Weka API with the following hyper-parameters: number of trees is 50 and the number of attributes considered at each node is log2 *(number Of Characteristics)+1*.

III. EXPERIMENTS AND RESULTS

Although the Emotiv EPOC provides the ability to register 14 channels will only be of interest for the experiments channels F7, FC5, T7 and P7. These channels, according to the Geschwind-Wernicke model, are the most related to the production of speech in the left hemisphere of the brain (except for some lefthanders) [13].

Selection of the sonification parameters: Since there are several parameters in the sonification process, we had to choose values that favor the percentages of correct classification, for which an iterative empirical parameter selection process was performed for each parameter. This process consists in the variation of one parameter at a time to evaluate its behavior according to the classification accuracy from which the best value elected. Later, using that value the following parameter was varied, choosing again the best value, this action is repeated until the best values of all parameters are obtained. The selected parameter values were: number of tones (14), minimum frequency of the EEG signal (1 Hz), maximum frequency of the EEG signal (60 Hz), spectrogram window size (26 samples), overlapping between the spectrogram windows (1 sample), tone length (0.6 sec), minimum audio frequency (50 Hz), maximum audio frequency (5000 Hz), audio sampling frequency (8000 Hz). Figure 4 shows examples of spectrograms obtained using the above parameters.

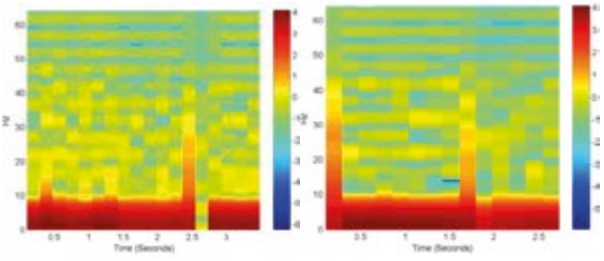


Figure 3: Examples of spectrograms of subject 11 during the imagined pronunciation of two different instances of the word *arriba* in channel F7.

Wavelet Selection: In the feature extraction process the same approach as in the selection of sonification parameters was followed. Experiments with diverse levels of different Daubechies Wavelets (db2, db6 and db20) testing a variation at a time were executed. The wavelet Daubechis 20 with 6 levels was the one with the best results.

Comparative Results: The average classification accuracy percentages using Random Forest for the 27 subjects in the two approaches are shown in Figure 5. When analyzing the table we can see that the method to sonify the EEG signal improves the average accuracy rates in 24 of the 27 analyzed subjects. Remarkably, differences above 15% in some cases can be noticed. In general, we can emphasize that the method of EEG sonification using the *EEG to tones* algorithm improves an average of 7.72% for whole set of 27 subjects.

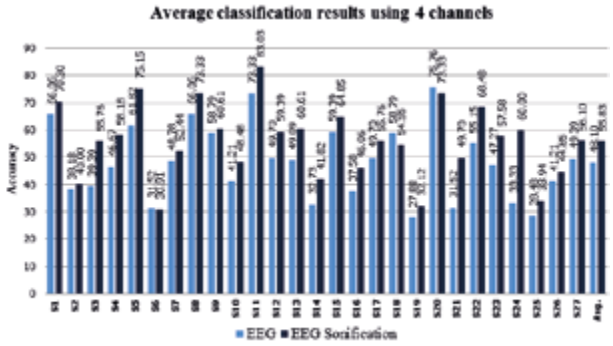


Figure 5: Comparison of EEG and EEG sonification methods. Classification accuracy average percentages with Random Forest for 27 subjects using 4 channels are shown. The last column shows the average percentages: EEG 48.10% and EEG using sonification 55.83%.

IV. CONCLUSION

This work presents the classification process of the EEG signals performed by transforming the original EEG signal to an EEG audio signal, with which average classification accuracy was improved at around 7.72%. Based on the results we can say that by choosing the dominant frequencies of the spectrogram of the EEG signal and mapping these EEG frequencies to audio frequencies, we achieved to highlight signal patterns that helped to improve the EEG classification accuracy. This happened even when we were using the same feature extraction methodology and the same classification method used in the work with which ours is compared. The results obtained motivate to experiment with other sonification, feature extraction and classification methods, looking to improve the accuracy rates achieved. One thing to consider, in future work is the parameter selection, which could be done by applying automatic search algorithms, such as evolutionary algorithms. We also need to compare the behavior of using the 14 channels available or by selecting the channels with the best information, regardless of their brain area location.

REFERENCES

[1] A. Bashashati, M. Fatourechi, R.K. Ward, and G.E. Birch. "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals". *Journal of Neural engineering*, 4:R32–R57, 2007.

[2] M. Wester and T. Schultz. "Unspoken Speech - Speech Recognition Based On Electroencephalography". Master's thesis, *Institut für Theoretische Informatik Universität Karlsruhe (TH), Karlsruhe*, Germany, 2006.

[3] T. Hermann, "Sonification for exploratory data analysis," Phd. Thesis, *Bielefeld University*, Germany, 2002.

[4] J. Eaton and E. Miranda, "Real-time notation using brainwave control," *Sound and Music Computing Conference*, 2013.

[5] G. Baier, T. Hermann, and U. Stephani, "Event-based sonification of EEG rhythms in real time," *Clinical Neurophysiology*, vol. 118, no. 6, pp. 1377–1386, 2007.

[6] M. Elgendi, B. Rebsamen, A. Cichocki, F. Vialatte, and J. Dauwels. Real-time wireless sonification of brain signals. *Advances in Cognitive Neurodynamics (III)*, pages 175–181. Springer, 2013.

[7] T. Hermann, P. Meinicke, H. Bekel, H. Ritter et al. "Sonification for EEG data analysis". *In Proceedings of the 2002 International Conference on Auditory Display*, 2002.

[8] F. Vialatte and A. Cichocki, "Sparse bump sonification: a new tool for multichannel EEG diagnosis of mental disorders; application to the detection of the early stage of Alzheimer's disease," *Neural Information Processing*, pp. 92–101, 2006.

[9] A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda. "Análisis de Señales Electroencefalográficas para la clasificación de habla imaginada". *Revista Mexicana de Ingeniería Biomédica*, 34(1):23–39, 2013.

[10] C. Anderson, "EEG to tones", *Department of Computer Science, Colorado State University*, 2005.

[11] M. A. Pinsky. "Introduction to Fourier analysis and wavelets". *Amer Mathematical Society*, vol.102, 2002.

[12] L. Rokach. "Pattern Classification Using Ensemble Methods". *World Scientific*, 2009.

[13] N. Geschwind. "Language and the brain". *Scientific American*, 1972.

NUMERICAL SIMULATION OF VIBRATION OF THE HUMAN VOCAL FOLDS – RECONSTRUCTION OF VIDEOKYMOGRAPHY RECORDS

T. Vampola[1], J. Horáček[2]

[1] Dept. of Mechanics, Biomechanics and Mechatronics, CTU in Prague ,Czech Republic
[2] Institute of Thermomechanics, Academy of Sciences of the Czech Republic
Tomas.Vampola@fs.cvut.cz, JaromirH@it.cas.cz

Abstract: **Three-dimensional (3D) finite element (FE) fully parametric model of the human larynx was developed and used for numerical simulation of stresses during vibrating vocal folds with collisions. The complex model consists of the vocal folds, arytenoids, thyroid and cricoid cartilages. The vocal fold tissue was modeled as a three layered transversal isotropic material. The modification of the layers is introduced for improved modelling of modes of vibration known from experimental investigations of the vocal fold self-oscillation. The results of numerical simulation of the vocal folds oscillations excited by a prescribed intraglottal aerodynamic pressure are presented. The videokymography records are reconstructed from the numerical simulation of the vocal fold vibrations and are used for tuning of the material characteristics of the separate layers.**

Keywords : **Biomechanics of human voice, 3D FE model of human vocal fold, tuning the model according to video-kymography records**

I. INTRODUCTION

Voice problems are common especially in professional voice users like teachers, actors and singers. The main reason may be a fatigue because of mechanical loading of the vocal fold tissue during voice production [1]. Design of a model of the human vocal folds, which would enable to model some pathological situations and voice disorders, is becoming an important part of the voice research. The loading of the vocal fold is caused by combination of the aerodynamic, inertial and impact forces. Excessive stresses during the impact may be responsible for tissues damage. With regard to the clinical practice, when basic investigative techniques include video-kymography, the question arises of whether the character of vibration recorded by videographic or the high-speed camera can be used for prediction of the damage to the vocal cords. In this contribution the reconstructed video-kymography records [2] from the numerical simulation of the vocal fold vibration are used for evaluation of the character of vibration of the damage vocal fold. Therefore, the computer model of the human vocal folds was designed enabling to model some pathological situations and voice disorders. Because of a complex mechanical loading of the vocal fold self-oscillations with collisions, and the complicated three-dimensional (3D) structure and material properties of the living tissue, it is necessary to assemble more sophisticated models based on Finite Element (FE) modelling. Such models enable us to estimate all main normal and shear stresses in the different vocal fold tissue layers in all three directions. However, the computational demands on computers and computer time needed are still very high and limited. These models can be used for the evaluation of the correlation between the deformation and stresses fields of the vocal fold tissues with the video-kymographic records of the vocal folds vibrations.

II. METHODS

The 3D complex dynamic FE model of the human larynx was developed by transferring the CT image data from the DICOM format to the FE mesh.

The developed fully parameterized 3D FE model enables to vary the thickness and material properties of the individual layers and to take into account longitudinal pretension and adduction of the vocal folds by positioning of the arytenoids and thyroid cartilages – see Fig.1.

The geometrical configuration of the cross-section of the vocal fold was taken according to [3] and the CT snaps. Three layers of vocal fold tissue are considered [1]: epithelium, vocal ligament and muscle with different physical and material properties - see Fig. 2. An additional superficial lamina propria layer formed by incompressible liquid was used and tested for improving character of numerically simulated vibration modes of the vocal fold - see Fig. 3.

The nonlinear elasticity theory for large-strain deformations with the linear transversal isotropic material model was used for modelling of the vocal

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

fold tissue, where the matrix of the elastic constants in strain-stress relations is defined according [4] as

$$\begin{bmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{bmatrix} = \begin{bmatrix} E_p^{-1} & -\mu_p E_p^{-1} & -\mu_{pl} E_p^{-1} & 0 & 0 & 0 \\ -\mu_p E_p^{-1} & E_p^{-1} & -\mu_{pl} E_p^{-1} & 0 & 0 & 0 \\ -\mu_{lp} E_l^{-1} & -\mu_{lp} E_l^{-1} & E_l^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & G_p^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & G_l^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 & G_l^{-1} \end{bmatrix} \begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{xy} \\ \sigma_{xz} \\ \sigma_{yz} \end{bmatrix} \quad (1)$$

where $E_p = 2\ G_p\ (1+\mu_p)$ is the Young modulus, μ_p is the Poisson number and G_p is the shear modulus in perpendicular plane xy to the ligament fibers, and analogical constants are denoted by the index l for the longitudinal direction z. The tissues material constants considered are summarized in Table 1.

Tab. 1 Nominal values of material constants of individual tissue layers according to [4] - E=Epithelium, L=Ligament, M=Muscle.

	E	L	M
G_p [kPa]	0.530	0.870	1.050
G_l [kPa]	10	40	12
μ_p	0.3	0.3	0.3
$E_l(\varepsilon)$ [kPa]	26	104	31
ρ [kgm^{-3}]	1020	1020	1020
μ_{lp}	0.3	0.3	0.3

For the incompressible liquid the bulk modulus k=2.1e6 kPa, density ρ=1020 kgm^{-3} and Poisson's number μ= from 0.4999 to 0.4999999 were used. The complete FE model of the larynx, developed in the FE system ANSYS, consists of about 500 000 linear and quadratic elements.

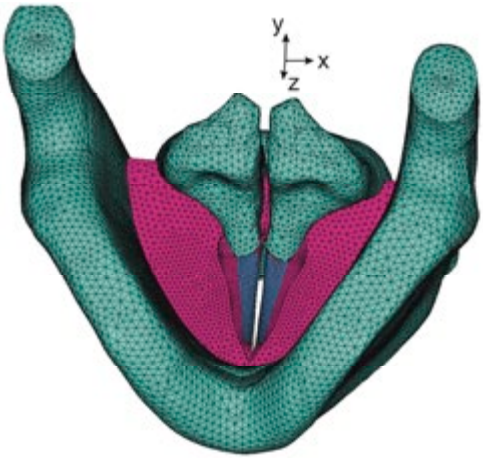


Fig.1 FE model of the human larynx with the vocal folds between the arytenoids and thyroid cartilages.

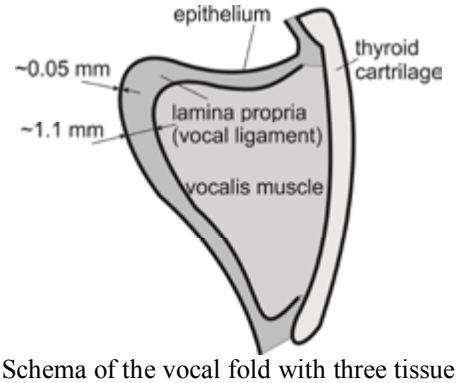


Fig. 2 Schema of the vocal fold with three tissue layers according to [3].

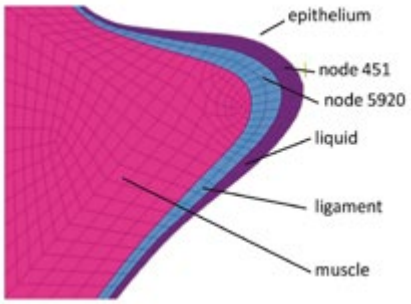


Fig. 3 Modification of the vocal fold with an additional liquid layer.

The motion of the vocal folds was numerically simulated for a prescribed intraglottal pressure $p(y,t)$ dependent on the vertical coordinate and given by a periodic function in the time domain. The intraglottal pressure signal loading the vocal fold surface was generated by the 2D aero elastic model [5] of the vocal folds during the vocal folds self-sustained vibrations for the airflow rate Q=0.179 l/s the prephonatory glottal half-gap g_0=0.2 mm and the fundamental frequency F_0=100.766 Hz, that corresponded to the subglottal pressure P_{sub}=378.4 Pa.

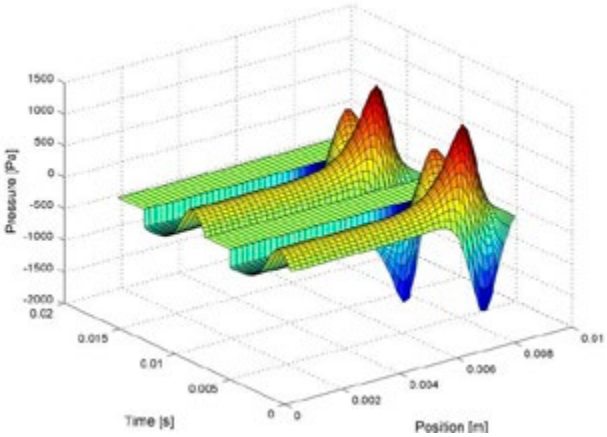


Fig. 4 Aerodynamic pressure $p(y,t)$ in time and space domain loading the vocal folds surface along vertical axis during two oscillation cycles.

III. Results

The computed trajectories given by the displacements $u_x(t)$, $u_y(t)$ in the selected node of the vocal fold tissue at the middle cross-section during the stabilized vocal folds oscillations are shown in Fig. 5 for the three-layer model with 5% prolongation of the human vocal folds. A very complicated motion is evident.

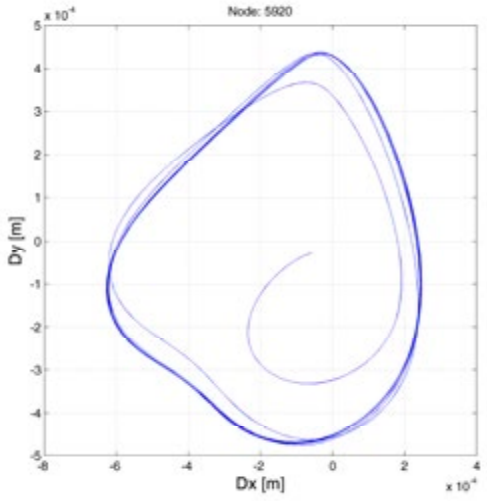


Fig. 5 Trajectory of node 5920 in the X-Y plane for three layers model.

For prescribed pressure field the three layers model gives nonrealistic character of vibration. Therefore the model with additional liquid layer was optimized in order to obtain the desired elliptical mode of vibration – see Figs. 6, 7 and [6]. The maximum thickness of the ligament plus liquid layers was approximately 1.5 mm.

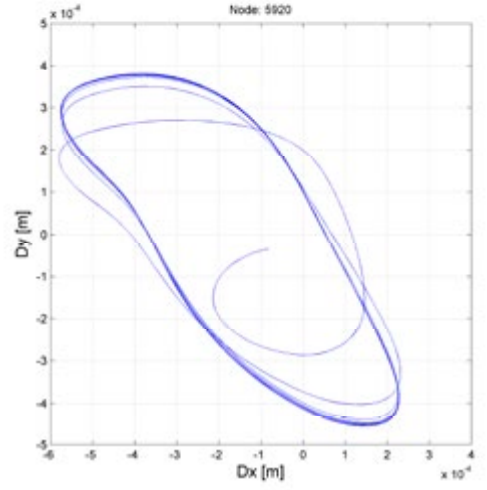


Fig. 6 Trajectory of node 5920 in the X-Y plane for model with an additional liquid layer, μ=0.4999997

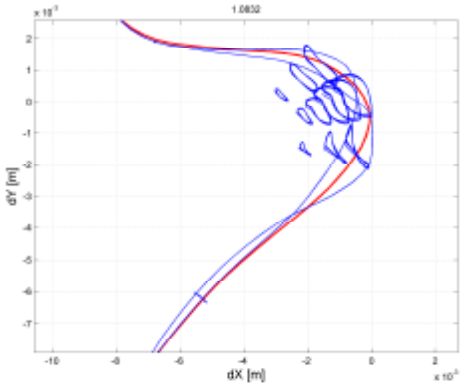


Fig. 7 Character of vibration of the human vocal fold in the X-Y plane for model with an additional liquid layer, μ=0.4999997

The mucous Rayleigh type waves are propagating near the vocal fold surface, especially in the upper part of the vocal fold. The maximum value of the peak to peak displacement in the medial (x) direction is about 1.2 mm and in inferior-superior (y) direction is about 1.1 mm. The medial displacement in x direction is limited by the vocal fold collisions. The anterior–posterior vibration amplitude in z direction is negligible. The maximum vibration amplitudes are on the vocal fold surface, and decreasing in the deeper tissue layers. - see Fig. 7.

Fig.8 shows the vocal folds deformation and the equivalent stress in medial mid/cross/section of the vocal folds at two time instant for closed and open glottis. The maxima of stress in the closed and open phases are comparable and are located in the ligament layer under the tissue surface.

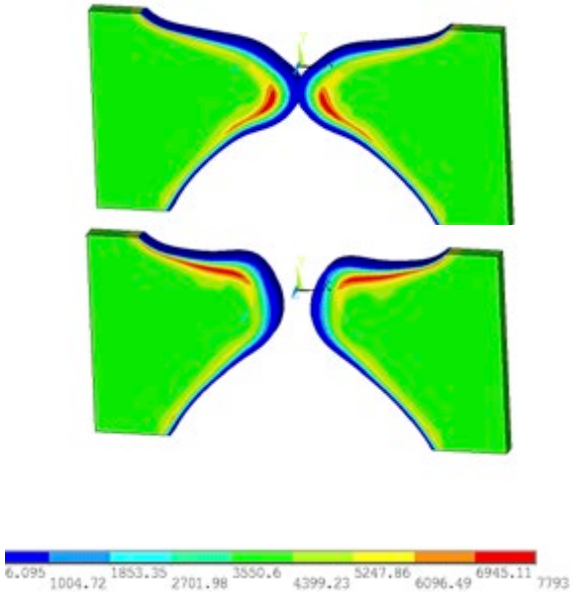


Fig. 8 Computed equivalent stresses σ_{eqv} [Pa] for the model with an additional liquid layer, μ=0.4999995.

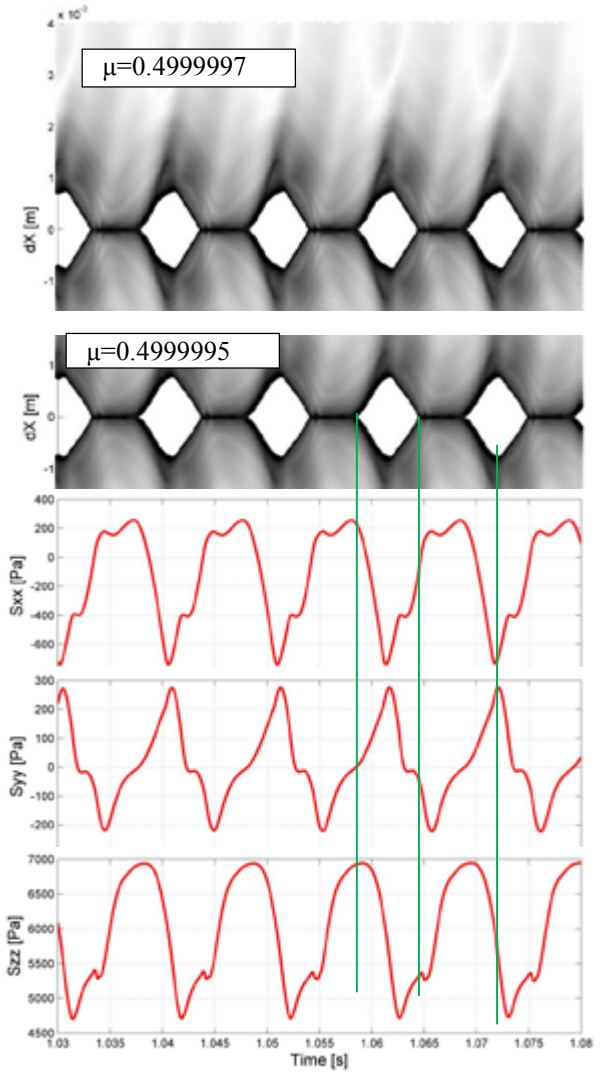


Fig. 9 Reconstruction of the video-kymographic records for the model with liquid layer (μ=0.4999997, μ=0.4999995) and the normal stresses of the vocal folds numerically simulated at the node 5920 in the ligament layer.

The computed deformations can be used for reconstruction of the video-kymographics records of the human vocal fold vibration. Fig. 10 demonstrates the sensitivity of vibration patterns in video-kimographs to a very small change of the material parameter of the liquid layer.

IV. Discussion

The geometry of the developed parametric FE model of the vocal folds as a part of the complex larynx model enables to modify the model easily and to apply tuning and optimisation procedures for finding proper model parameters related to the vocal folds vibration. The preliminary results are promising from the character of vibration and to predict approximate stresses in the vocal fold tissue due to the vibration of the vocal folds in normal phonation regimes with collisions in determination of the injury type of the vocal fold characteristics. The reconstructed videokymographic records are sensitive enough to the changes of the material parameters and geometric reconfigurations of the vocal fold and can be used for prediction of various vocal fold damages

V. Conclusion

The assembled model of the human larynx is sensitive to small changes of the geometric configuration and materials characteristics and can be used for multi-criterial global optimization for finding parameters of the human vocal folds model according prescribed character of vibration. This model can be used for prediction of the vocal fold injury from the video-kymographics records.

Acknowledgement
The research is supported by the Grant Agency of the Czech Republic by project No P101/12/1306 "Biomechanical modelling of human voice production - way to artificial vocal folds".

REFERENCES

[1] Titze, I.R., "Mechanical stress in phonation", *Jornal of Voice*, 8(2). pp. 99-105, 1994.
[2] J.G. Švec, H.K. Schutte, "Videokymography: high-speed line scanning of vocal fold vibration", *Journal of Voice*, vol. 10, pp. 201-205, 1996.
[3] Hirano, M., "Phonosurgery, basic and clinical investigations", *The 76th Annular Convention of the Oto-Rhino-Laryngological Society of Japan*, 1975.
[4] Luo, H. at al, "An immersed boundary method for flow-structure interaction in biology systems with application to phonation", *Journal of Computational Physics* **227** pp. 9303-9332, 2008
[5] Horáček, J., Laukkanen, A.M., Šidlof, P., "Estimation of impact stress using an aeroelastic model of voice production", *Logopedics Phoniatrics Vocology* **37** ,pp. 185-192, 2007.
[6] Doellinger, M., Berry, D.A., "Visualization and quantification of the medial surface dynamics of an excised human vocal fold during phonation", *Journal of Voice*, Vol.20, No. 3, pp. 401-413, 2006.

OPTICAL-FLOW KYMOGRAMS AND GLOTTOVIBROGRAMS: A NEW WAY TO PRESENT HIGH-SPEED DATA FOR LARYNGEAL ASSESSMENT

Gustavo Andrade-Miranda[1], Nathalie Henrich Bernardoni [2,3], Juan Ignacio Godino-Llorente[1]

[1] Center for Biomedical Technologies, Universidad Politécnica de Madrid
[2] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
[3] CNRS, GIPSA-Lab, F-38000 Grenoble, France
gxandrade@ics.upm.es, Nathalie.Henrich@gipsa-lab.fr, igodino@ics.upm.es

Abstract: **The use of high-speed videoendoscopy (HSV) in combination with image-processing tech- niques is the most promising approach to investigate vocal-folds vibration and laryngeal dynamics in speech and singing. The current challenge is to pro- vide facilitative and informative playbacks for clinical and research purposes. We present three new facilita- tive playbacks using an optical-flow framework (OF), which has the main advantage of requiring no glottis segmentation. The method has been tested on a data- base of 60 HSV sequences, which covers different voice qualities for spoken and sung vowels. The new data representations have been compared with com- monly used facilitative playbacks.**

Keywords: **Optical Flow, motion field, high-speed videoendoscopy, glottal dynamics, playbacks**

I. INTRODUCTION

Nowadays, high-speed videondoscopy has been increasingly used to assess glottal dynamics. It is the sole imaging technique capable to acquire the true intra-cycle vibratory behavior through a series of full-frame images of the vocal folds. It allows the study of cycle-to-cycle glottal variation. Due to the fast-growth of high-speed technology, it is possible to found cameras that can reach frame rates up to "twenty thousands", recording in color with high spatial resolution and excellent image quality for long durations. HSV allows to characterize many vocal-folds vibratory features that are not possible to visualize by means of videostroboscopic techniques. For instance, HSV helps to get insights into tissue vibratory characteristics, the influence of aerodynamical forces and muscular tension, vocal length and evaluation of normal laryngeal functioning in situation of rapid pitch change such as onset and offset of voicing or glides [1]. The use of HSV has been reported in the literature to evaluate variations in vocal-folds dynamics and extract important features such as vocal-fold vibratory amplitude, glottal open quotient, and glottal speed quotient. HSV provides a huge amount of images, whose analysis requires a great deal of human intervention and observation. Several playbacks have been proposed to reduce the spatio-temporal dimensionality while preserving the most relevant characteristics of glottal vibratory patterns. The most widespread and successful playbacks used either by clinicians or researchers are: Digital Kymograms [2], Mucosal Wave and Mucosal Wave Kymogram playbacks [3], Phonovibrogram [4], and Glottovibrogram [5]. Many common approaches make use of glottal segmentation algorithms, in which attention is focused on analyzing movements at vocal-fold edges. The widespread techniques are based on histogram equalization, region growing, watershed and active contours delineation methods (see [5] for a review). Nevertheless, motion analysis should not necessarily be focused only on the points belonging to the glottis contours but also in the regions where such movements originate. For that reason, the estimation of a global motion in which the different patterns relevant for voice production could be represented is desirable. Optical flow (OF) techniques estimate the motion of objects in consecutive frames by generating a motion field in which each pixel represents a vector displacement. In laryngeal HSV sequences, the vibrating vocal folds are most often the regions with greatest motion.

In this paper, OF image processing is investigated as a new approach for analyzing laryngeal images and for assessing glottal dynamics. Innovative playbacks are proposed within this framework. The paper is organized as follows. Section 2 details the principles of OF-based image processing, and describes the data- base. Section 3 presents the results for three new play- backs and provides a comparison with existing com- mon ones. Finally, Section 4 presents some conclu- sions and discussions.

II. MATERIALS AND METHODS

A. Principles of optical-flow estimation

The 3-D velocity vector of objects, projected onto the image plane, is known as the image flow field. This could be considered as the ideal and actual movement of objects that we expect to see. Unfortunately, image processing has to deal with the inverse problem: the movement of the objects has to be determined on the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

basis of a sequence of images. This leads to an approximation called optical flow field, which associate each pixel in the image with a motion vector.
There are many different ways to estimate the optical flow, which depend basically on the kind of chosen constraint. Most of the constraints are derived from the assumption that pixel intensities are translated from one frame to the next:

$$f(x+\delta x, y+\delta y, t+\delta t) \approx f(x,y,t) \quad (1)$$

Another type of constraint that has no obvious connection with the previous one is motion tensor (MT) [8]. The MT principle is that a video segment is a stack of images in which gray-value structures have certain orientations. The orientation in the *xy*-subspace is an indicator for the orientation of the structure in the space. In contrast, the orientation of the structure in the *xt*-subspace or *yt*-subspace relates to the image velocities. Thus, estimating the orientation of the structure in these two subspaces or a combination thereof allows estimating the OF. There are two main strategies for solving the OF problem: Sparse and Dense. The sparse optical flow finds the displacement only on a subset of features that have been specified beforehand; these features have certain desirable properties such as corners, dominant gradient orientation, or subpixel corner locations. In the other hand, dense OF finds out the vector displacement of all pixels in the image, requiring a more expensive computational burden, but providing more interesting information about the movements in the sequence.

B. Database

A database of 60 high-speed sequences was used to assess laryngeal dynamics in several phonatory tasks: spoken vowels with specific voice qualities (creaky, normal, breathy, pressed), pitch glides, sung vowels at different pitches, loudness and laryngeal mechanisms [6]. Two male subjects (one speaker, one singer) participated to the experiment. The recording took place at the University Medical Center Hamburg-Eppendorf (UKE) in Germany, in collaboration with Pr. Hess, Dr. Müller and Dr. Licht [5]. The high-speed sequences were acquired by means of Wolf high-speed cinematographic system (rigid endoscope Wolf 90 E 60491 and light source Wolf 5131, grayscale CCD camera). The laryngeal high-speed images were sampled at either 2000 or 4000 fps. They had a spatial resolution of 256x256 pixels. Audio and electroglottographic signals were recorded simultaneously to the high-speed sequences and synchronized in a post-processing step.

C. Image Processing Procedure

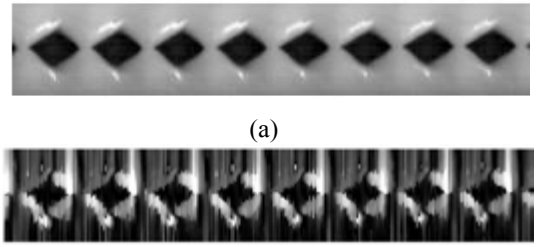
The algorithms were developed in C++ using the OpenCV library and integrated in Matlab for making the visualization of playbacks easier. Two different optical-flow methods were used. The first one, called TV-L1 OF, is based on the brightness constancy assumption [7]. This formulation adds a regularization term that allows discontinuities. Such feature is desirable when a complex motion is modeling. The brightness constancy term uses the robust L1 norm and is therefore less sensitive to intensity variations. The second OF method, called MT OF, is based on motion tensor computation [8]. It starts with computing 3D orientation tensors from the image sequence. These tensors are combined under the constraints of a parametric motion model to produce the velocity estimation. The formulation of this OF methodological approach does not use the common brightness constraint, and thus it is more sensible to the reflectance phenomena originated by the mucosa surface properties. Many additional techniques can be applied to mitigate these effects. The approach chosen here combines a nonlinear transformation with an anisotropic filter. Taking into account that the analysis of laryngeal HSV focuses attention on the dynamics of vocal-folds movements, a good strategy to reduce computational burden and mitigate the effect produced by noise regions is to calculate the OF field inside of a region of interest (ROI) that include the glottal gap and part of the vocal folds. The next step is to synthesize the motion-field information obtained between consecutive frames into visual playbacks, in which the information on the behavior of vocal-folds movement is readable. Three main representations have been elaborated. They will now be described and compared to the existing playbacks.

III. RESULTS

A. Local dynamics along one line: Optical-Flow Kymogram

The Optical-Flow Kymogram playback (OFKG) uses the same principle than Digital Kymogram (DKG) to compact high-speed information. However, the information used to condense the data is taken from the displacements originated in the x-axis. For rightwise OF movements, the direction angle of displacement ranges from $[-\pi/2, \pi/2]$ and is coded in white. On the other hand, the direction angle for leftwise displacements ranges from $[\pi/2, 3\pi/2]$ and is coded in gray tone of 128. The OFKG playback is illustrated in Fig.1 for a sequence of eight glottal cycles. Glottal-cycle shape and glottal dynamics present great similarity in both playbacks. The instants of change between opening and closing phases induce the presence of a discontinuity in the OFKG. This can be understood as the instants for which velocity comes close to zero. The

spread effect in the OFKG at given moments of the opening and closing phases may reflect the mucosal waves on vocal-folds surface. In DKG, mucosal waves are reflected as white flashing spots.



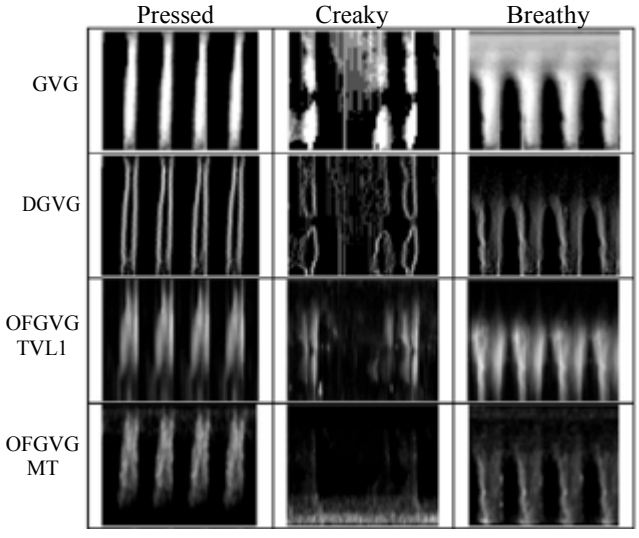
(a)

(b)

Fig.1: (a) DKG representation for a line located in the center of the main axis; (b) The new OFKG playback for the same line, in which gray scale distinguishes the direction of motion (rightwise: white gray; leftwise: pale gray).

B. Global dynamics along the whole vocal-folds length: Optical-Flow Glottovibrogram

The Optical-Flow Glottovibrogram (OFGVG) represents the velocity of glottal movement per cycle plotted along the vocal-folds length. It is obtained by averaging each row of the x component of the flow and representing it as a column vector. This procedure is repeated along time for each new frame. The aim of OFGVG playback is to complement the spatio-temporal information provided by the common techniques (glottovibrogram GVG, phonovibrogram PVG), by adding velocity information for each displacement of the vocal folds. For the purpose of visual comparison, four playbacks were performed in different phonation cases: GVG and its derivative DGVG were computed using [5]; two OFGVG playbacks were computed using TVL1 OF (OFGVG-TVL1) and MT OF (OFGVG-MT) respectively. Only OFGVG-MT includes the preprocessing step described in the section 2C. The corresponding plots for these playbacks are presented in Fig.2.



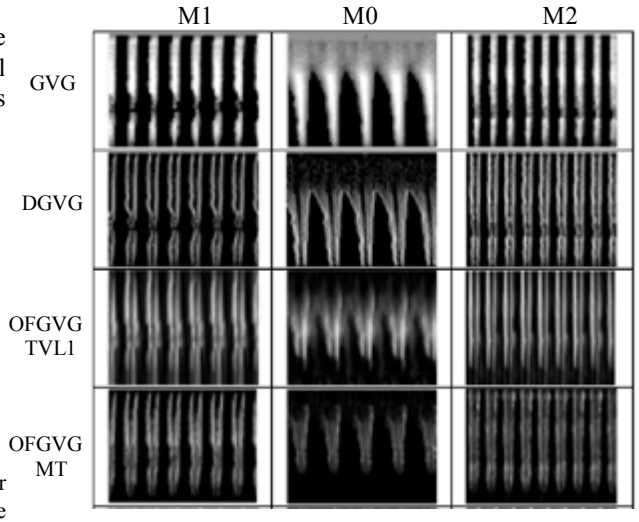


Fig.2: Representation of four playbacks (GVG, DGVG, OFGVG-TVL1, OFGVG-MT) for six different phonatory cases where either voice quality (pressed, creaky or breathy) or laryngeal mechanisms (M0, M1 or M2) are varied.

Similarities between DGVG and OFGVG playbacks are evidenced in Fig.2, especially in the shape appearance. However, OFGVG looks more blurred since movements taken into account by OF are not located only at the glottis edges (as in the DGVG case), but also in the vocal-folds surface. Breathy and M0 phonations present a posterior glottal chink that can be observed on GVG playback. Such regions are represented in DGVG and OFGVG playbacks with black color, as a result of the absence of movement. In creaky voice, OFGVG-TVL1 fails to provide accurate glottal cycles. The resulting playback may be improved by tuning the parameters of the preprocessing step.

Some features observed in DGVG playback and considered as artefacts due to segmentation problems do not appear in OFGVG-TVL1 playback. In M1 sequence for instance, the presence of mucus on vocal folds induces the appearance of a glottis splitted in two parts after the segmentation process. This is reflected by black spots in the median part of the glottis on GVG and DGVG playbacks. In both OFGVG playbacks, the glottis is not artificially splitted into two regions, as the motion field is robust to the presence of mucus.

C. Glottal velocity: Glottal Optical-Flow waveform

The Glottal Optical-Flow Waveform (GOFW) is a 1D representation of the velocity. GOFW is based on the same principle of the Glottal Area Waveform (GAW). The total magnitude of velocity is computed over the ROI for each instant of time. Graphically the GOFW represents the change of velocity as a function of time.

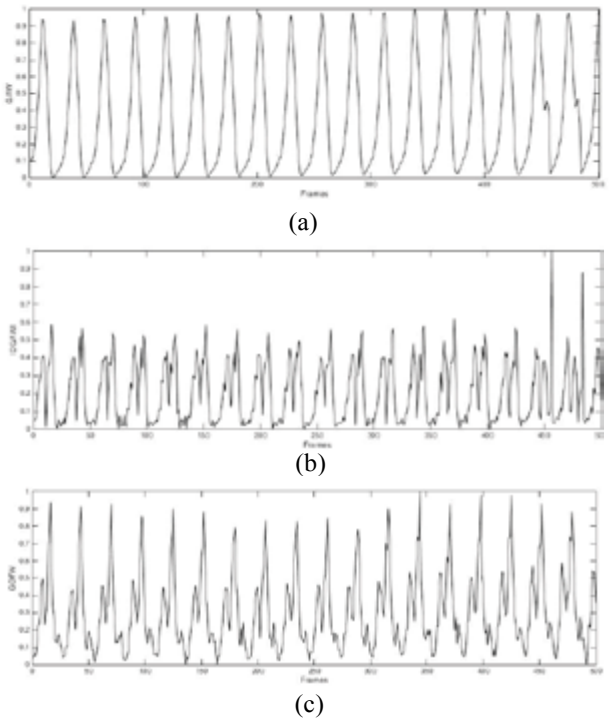


Fig.3: (a) GAW obtained by segmentation, (b) the absolute value of differentiated GAW (∠|DGAW∠|) and (c) GOFW.

The GOFW provide valuable information on the velocity instants over the HSV sequence. Additionally, if this information is overlapped with GAW one, it becomes feasible to analyze velocity variation with respect to glottal opening function. For instance, when maximal glottal opening is reached, GOFW shows a local minimum. Another interesting feature is that maximum speed is located during the closing phase. Since GOFW computes an absolute velocity, it is possible to obtain a similar representation by differentiating GAW and computing its absolute value (∠|DGAW∠|). As shown in Fig.3b, GOFW pulse shapes are similar to ∠|DGAW∠|, however with a stronger OF velocity during glottal closing.

IV. CONCLUSIONS AND DISCUSSION

High-speed videoendoscopy is probably the most promising technique for direct investigation of glottal dynamics in speech and singing. We have presented here a new approach to synthesize dynamical information from HSV recordings in a compact way, which does not depend on prior glottal segmentation. The glottis is treated as an unidentified object, and attention is focused on the motion field produced by vocal-folds vibration. Dense optical flow is computed among consecutive frames to extract dynamical information related to the pattern of glottal displacement. Three new playbacks are proposed to visualize the computed optical flow: OFKG, OFGVG and GOFW playbacks. There are some similarities in the information extracted from segmentation and OF, since both methods quantify the motion. However, the motion obtained from OF is raw information that include direction and magnitude of the pixels movements (displacement field map). Also using the displacement field is possible to segment the glottal gap, compute the contact time of the vocal folds and many other features. For the purpose of clinical diagnosis it seems a promising approach to complement, and eventually to replace, segmentation-based techniques.

ACKNOWLEDGEMENTS

This work has been funded by the Spanish Ministry of Economy and Competitivity under grant TEC2012 38630-C04-01.

REFERENCES

[1] K. Kendall and R. Leonard, *Laryngeal Evaluation: Indirect Laryngoscopy to High-speed Digital Imaging*. Thieme Publishers Series. Thieme, 2010.

[2] J. G. Svec and H. K. Schutte, "*Videokymography: high-speed line scanning of vocal fold vibration,*" J. Voice, vol. 10, no. 2, pp. 201–5, Jun. 1996.

[3] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris, and R. E. Hillman, "*Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution,*" Folia-Phoniatr. Logo, vol. 60, no. 1, pp. 33–44, Jan. 2008.

[4] J. Lohscheller and U. Eysholdt, "*Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics,*" IEEE Trans. Med. Imaging, vol. 27, no. 3, pp. 300–309, 2008.

[5] S-Z Karakozoglou, N. Henrich, C. d'Alessandro, and Y. Stylianou, "*Automatic glottal segmentation using local-based active contours and application to glottovibrography,*" Speech Communication, vol. 54, no. 5, pp. 641–654, 2011.

[6] B. Roubeau, N. Henrich, and M. Castellengo, "*Laryngeal vibratory mechanisms: The notion of vocal register revisited,*" Journal of Voice, vol. 23, no. 4, pp. 425 – 438, 2009.

[7] J. Sánchez Pérez, E. Meinhardt-Llopis, and G. Facciolo, *"TV-L1 Optical Flow Estimation"*, Image Processing On Line, 3 (2013), pp. 137–150

[8] G. Farneback, "*Two-frame motion estimation based on polynomial expansion,"* in Proceedings of the 13th Scandinavian Conference on Image Analysis, ser. SCIA'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 363–370.

SYNTHETIC KYMOGRAMS AND GLOTTAL AREA WAVEFORMS IN SIMULATED NON-NEUTRAL PHONATION

J. Schoentgen[1], P. Aichinger[2]

[1] F.N.R.S. & Université Libre de Bruxelles, Laboratories of Image, Signal Processing and Acoustics, Faculty of Applied Sciences, Brussels, Belgium

[2] Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Vienna, Austria

jschoent@ulb.ac.be, philipp.aichinger@meduniwien.ac.at

Abstract: **The presentation concerns the use of synthetic kymograms to facilitate the interpretation of natural kymograms as well as the study of glottal vibration patterns via simulation. Synthetic kymograms represent the evolving glottis via a static picture by stacking slices of modeled glottal shapes on top of each other. To illustrate the use of synthetic kymograms the presentation is devoted to the effects on glottal vibratory patterns of constant phase shifts between the left and right, the anterior and posterior as well as the entrance and exit. Possible causes of constant phase shifts are discussed by means of a generic model of phase-coupled oscillators.**

Keywords: **synthetic kymograms, phase-coupled oscillators, glottal vibratory patterns, voice quality**

I. INTRODUCTION

Video kymograms are obtained by recording video images at high speed, selecting one line per image and displaying the lines stacked on top of each other or, alternatively, by video equipment recording one line at a time in place of an image [1]. Kymograms are therefore a compressed representation of the evolving glottis that enables the static display of several glottal cycles in the same picture.

Video kymograms are increasingly popular in the framework of the documentation and assessment of laryngeal function, the diagnosis of laryngeal pathologies and research. Also, kymograms are frequently published, thus becoming available for inspection to readers who are not users of laryngeal high-speed imaging or dedicated kymography equipment. The interpretation of kymogram patterns of disordered voices may be difficult, however.

Here, we propose to simulate kymograms by means of a model of the glottis, the parameters of which enable mimicking a wide range of glottal patterns, left-right, entrance-exit and anterior-posterior asymmetries included. They offer the user the possibility to experiment with a range of glottal patterns the properties of which are known, thus aiding the interpretation of natural kymograms by means of simulations, assisted by separate entrance and exit kymograms as well as entrance, exit and effective width waveforms if so desired.

Similarly, kymograms may enable documenting by means of static pictures the behavior of numerical models of the vibrating vocal folds, behavior that can otherwise only be described either via scalar features or slowed-down video animations that are time-consuming to prepare and watch and that cannot be reported in static media.

To illustrate the use of synthetic kymograms we explore the effects of constant phase shifts between the left and right, the anterior and posterior and the entrance and exit on the vibratory patterns of the glottis. The causes of constant phase shifts are discussed by means of a generic model of phase-coupled oscillators.

II. METHODS

A. The PDOS model of the glottis

The *PDOS* model designates the *p*hase-*d*elayed *o*verlapping *s*inusoidal model of the glottis proposed by Titze [2]. It consists of glottal hemi-widths (1) that evolve sinusoidally. The sinusoids animating the glottal exit are phase delayed with regard to the glottal entry because of the out-of-phase motion of the edges of the cover of the vocal folds. The glottal slit is rectangular or elliptical. Glottal wall collision is simulated by means of a *max* operator and the effective width by means of a *min* operator.

$$\begin{aligned} w_{entr} &= \xi_{entr,0} + \xi_{entr}\sin(\varphi) \\ w_{exit} &= \xi_{exit,0} + \xi_{exit}\sin(\varphi - \phi) \\ w_g &= \min[\max(0, w_{entr}), \max(0, w_{exit})] \\ \frac{d\varphi}{dt} &= 2\pi f_0 \end{aligned} \quad (1)$$

Symbols w and ξ designate the glottal hemi-widths as well as the abduction and the amplitude of vibration of the glottal entrance and exit. Constant phase ϕ accounts for the delay between entrance and exit and f_0 for the instantaneous vocal frequency.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

The original mirror-symmetric *PDOS* model has 7 parameters (glottal length included) [1]. When allowing for left-right and anterior-posterior asymmetries, the total number of model parameters is four times as large, enabling the simulation of a wide range of evolving glottal shapes.

B. Constant phase shifts

Simulated kymograms are used in this presentation to illustrate vibratory patterns owing to fixed phase shifts between left and right, anterior and posterior as well as entrance and exit glottal vibrations.

The origin of constant phase shifts can be qualitatively explained by means of a generic model (2) of two phase-coupled oscillators [3]. Symbols φ designate the phases and ω the natural angular velocities. The generic model involves the first two terms of the Fourier series of a smooth coupling function F the inverse $(F)^{-1}$ of which exists in the vicinity of the fixed points of model (2). When the coupling constants are positive or negative, they favor anti-phase or in-phase motions respectively.

$$\begin{aligned} \dot{\varphi}_1 &= \omega_1 + K_{o,1}\sin(\varphi_1-\varphi_2) + K_{e,1}\cos(\varphi_1-\varphi_2) \\ \dot{\varphi}_2 &= \omega_2 + K_{o,2}\sin(\varphi_2-\varphi_1) + K_{e,2}\cos(\varphi_2-\varphi_1) \\ \frac{d(\varphi_1-\varphi_2)}{dt} &= \omega_1-\omega_2 + (K_{o,1}+K_{o,2})\sin(\varphi_1-\varphi_2) \\ &\quad +(K_{e,1}-K_{e,2})\cos(\varphi_1-\varphi_2) \end{aligned} \tag{2}$$

Constant phase shifts are found by zeroing the time derivative of the phase difference in model (2). In that model, three mechanisms shift the phases between oscillators by a constant amount. Two are due to coupling and one to natural phase shifts that are intrinsic to the uncoupled oscillators.

Unequal natural frequencies of the oscillators shift the phases proportionally to the frequency difference and inversely proportionally to the coupling strength (3). A condition is that the frequency difference is smaller than the total coupling strength. For simplicity's sake, coupling constants K are assumed to be identical and positive by default.

$$\begin{aligned} \dot{\varphi}_1 &= \omega_1 + K\sin(\varphi_1-\varphi_2) \\ \dot{\varphi}_2 &= \omega_2 + K\sin(\varphi_2-\varphi_1) \\ 0 &= \omega_1-\omega_2 + 2K\sin(\varphi_1-\varphi_2) \\ -1 \le \frac{\omega_2-\omega_1}{2K} \le +1, &\varphi_1-\varphi_2 = \arcsin(\frac{\omega_2-\omega_1}{2K}) \end{aligned} \tag{3}$$

Examples are minor differences in mass and tension of the left and right vocal folds.

Natural phase shifts: Phase shifts that are intrinsic to the uncoupled oscillators are unaffected by (phase) coupling when coupling constants K are the same and negative and the natural frequencies identical.

$$\begin{aligned} \dot{\varphi}_1 &= \omega + K\sin(\varphi_1+\phi-\varphi_2) \\ \dot{\varphi}_2 &= \omega + K\sin(\varphi_2-\varphi_1-\phi) \\ 0 &= +2K\sin(\varphi_1+\phi-\varphi_2) \\ \varphi_1-\varphi_2 &= -\phi \end{aligned} \tag{4}$$

An example is phase delay ϕ between the glottal entrance and exit owing to the out-of-phase motion of the fold cover edges.

Unequal coupling functions that are not odd move the oscillators out of phase, but continue synchronizing the frequencies when the frequency difference is not too large. When the natural frequencies are identical, the phase shift is proportional to the difference between the even coupling constants and inversely proportional to the sum of the odd coupling constants (5).

$$\begin{aligned} \dot{\varphi}_1 &= \omega + K_{o,1}\sin(\varphi_1-\varphi_2) + K_{e,1}\cos(\varphi_1-\varphi_2) \\ \dot{\varphi}_2 &= \omega + K_{o,2}\sin(\varphi_2-\varphi_1) + K_{e,2}\cos(\varphi_2-\varphi_1) \\ 0 &= (K_{o,1}+K_{o,2})\sin(\varphi_1-\varphi_2) + (K_{e,1}-K_{e,2})\cos(\varphi_1-\varphi_2) \\ \varphi_1-\varphi_2 &= \arctan\frac{K_{e,2}-K_{e,1}}{K_{o,2}+K_{o,1}} \end{aligned} \tag{5}$$

C. Dispersion

Natural phase shifts between glottal entrance and exit vibrations are explained by a wave propagating with a finite speed from the bottom to the top of the fold cover. This raises the question whether the time of propagation is frequency-dependent. A lack of dispersion would predict that time delay τ of the propagation is frequency-independent and that phase delay ϕ is proportional to the frequency of vibration.

$$\begin{aligned} \phi &= 2\pi f_0 \tau \\ \tau &= \frac{T_h}{c} \end{aligned} \tag{6}$$

One expects, however, that an increase of the vocal frequency is associated with a thinning as well as stiffening of the cover and therefore an increase of the speed of propagation. One therefore expects time delay τ to decrease with frequency, given its dependency on speed c and fold thickness T_h. The *PDOS* model, for instance, assumes by default that $\tau \sim 1/f_0$.

D. Synthetic kymograms

Selecting a glottal slice orthogonally to the anterior-posterior glottal axis and stacking subsequent slices vertically obtains synthetic kymograms. The edges of the evolving entrance and exit of the slice are tracked separately as follows. When the entrance or exit slice width is positive, the edge positions are reported exactly, when the slice width is negative (i.e. the glottal walls interpenetrate virtually), the tracking stops and the edge positions are reported via a linear interpolation between the edge position when the walls touch and the following edge position when the walls move apart. The final step consists in removing edges

of the glottal entrance that are hidden by edges of the glottal exit. The result is the outline of the visible glottal edges that report the evolving shape of a *3D* glottal model by means of a *2D* static picture. An optional final step consists in coloring the kymogram surfaces so that the glottal neighborhood (i.e. fold surface) seen from above is light grey, the glottal walls when visible are dark grey and the visible glottal opening is black.

The synthetic effective, entrance and exit kymograms may be presented together with time series that report the *PDOS* entrance and exit glottal widths as well as effective glottal width w_g, which is the minimum of the entrance and exit widths [2]. The *PDOS* effective width does not agree exactly with the numerical kymogram width because the *min* operator does not take into account accurately the visual masking of the glottal entrance by the exit.

III. Simulations

Fig. 1 shows an example of a simulated neutral voice. The vertical axis is the time axis and the horizontal axis the width in arbitrary units. Kymogram and glottal width time axes are shifted horizontally by 5 units to ease comparisons. The figure displays from left to right the effective kymogram, the exit and entrance kymograms followed by the effective, exit and entrance glottal widths. The phase shift between glottal entrance and exit is moderate (1 rad).

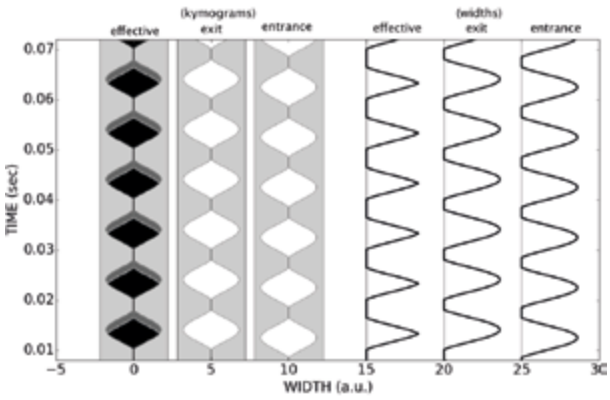


Fig. 1: Simulated neutral voice

A. Entrance-exit phase shifts

A likely cause of phase shifts between glottal entrance and exit is the out-of-phase motion of the vocal fold cover. Fig. 2 illustrates phase shifts for a frequency-independent propagation delay (no dispersion) and a propagation delay proportional to $1/f_0$ and $1/(f_0)^2$ with the vocal frequencies equal to 100 Hz (left) and 150 Hz (right).

The dispersion-free case predicts that the phase delay increases with frequency, whereas, the two

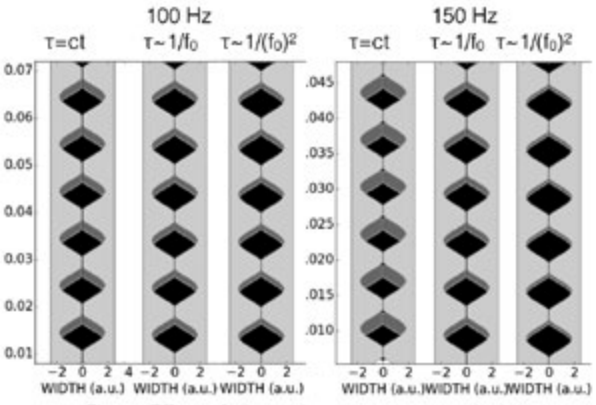


Fig. 2: Effect of dispersion on entrance-exit delay

dispersive cases predict that the phase delay stays the same or decreases with vocal frequency. A constant or decreasing entrance-exit phase delay is expected given the dependency of the temporal delay on cover thickness and stiffness.

In Fig. 3, the vocal frequency is equal to *80Hz* and temporal delay τ is assumed to be $\sim 1/f_0$. It shows double pulsing, which is the apparent twofold opening of the glottis within one cycle due to the shifting of the closed phase towards the middle of the cycle.

Double pulsing is the effect of large entrance-exit phase delays and moderate to large relative abductions. The relative abduction is the ratio in *%* of amplitudes ξ_0 and ξ.

A case of mild double pulsing can also be seen in Fig. 2 in the third kymogram from the right. In Fig. 2 double pulsing is the effect of a hypothetical frequency-independent temporal entrance-exit shift that if it existed would predict that increasing vocal frequencies favor double pulsing. Due to dispersion, however, double pulsing is the most likely to be observed at extra-low vocal frequencies when the vocal folds are thick and slack.

B. Left-right phase shifts

A possible cause of left-right phase shifts are mild differences in the natural frequencies of the left and right vocal folds, which vibrate at the same frequency. Model (3) predicts that the sign of the phase shift is identical to the sign of the frequency shift, when coupling constant $K < 0$. This suggests that observing the direction of the phase shift enables discovering the fold with the larger natural frequency.

Fig. 4 shows positive right-left phase shifts (left) and negative right-left phase shifts (right), i.e. the natural frequency of the right fold is larger on the left-hand side and smaller on the right-hand side. In Fig. 4, the vocal frequency equals 100 Hz and the entrance-exit phase shift (6) is proportional to the temporal entrance-exit delay.

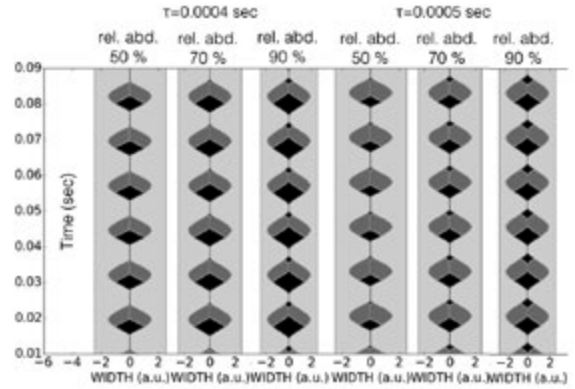


Fig.3: Simulated double pulsing

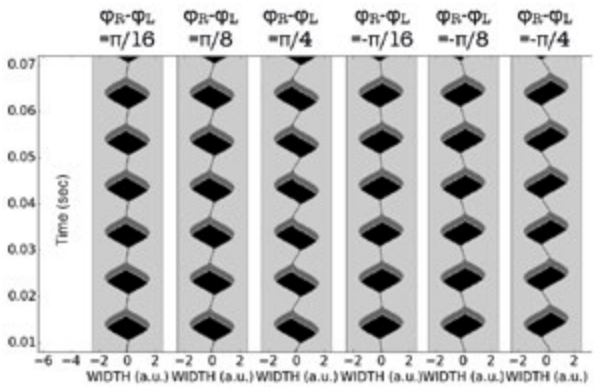


Fig. 4: Left-right phase asymmetry

C. Anterior-posterior phase shifts

Anterior-posterior asymmetric glottal vibrations may be observed when the folds are quasi-immobile at a position along the longitudinal axis and the elliptical glottal area is turned into a figure eight pattern. The unequal figure-eight top and bottom may vibrate at different frequencies and amplitudes. The glottal area is equal to the sum of the anterior and posterior sub-areas.

The anterior-posterior phase shift as well as the closed quotient of the effective glottal area can be assessed visually by slicing the kymograms longitudinally and gluing the anterior left to the posterior right and vice versa. The effective width of the glottis equals the sum of the anterior and posterior widths weighted by the relative lengths of the sub-glottises, lengths that cannot be inferred from kymograms alone. If the glottal shape is rectangular, the effective width can be inferred from the total glottal area and length.

Anti-phase vibrations at the same frequency of the figure-eight top and bottom are compatible with simple model (2) when the coupling constants are positive. Fig. 5 shows an anterior-posterior phase shift of $\pi/2$ and Fig. 6 shows anti-phase anterior and posterior vibrations. Figs 5 and 6 demonstrate that the effective closed quotient of the glottis is smaller than the closed quotients of the sub-glottises and that it is zero when the sub-glottises vibrate in anti-phase.

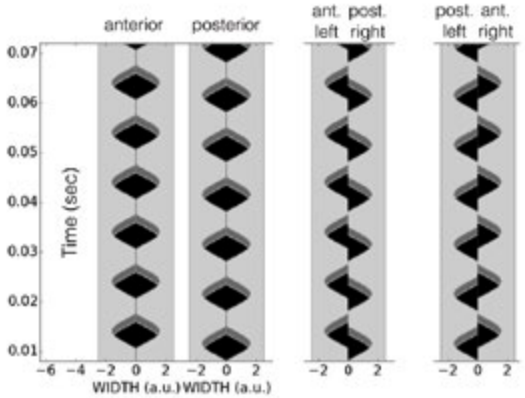


Fig. 5: Anterior-posterior phase shifts

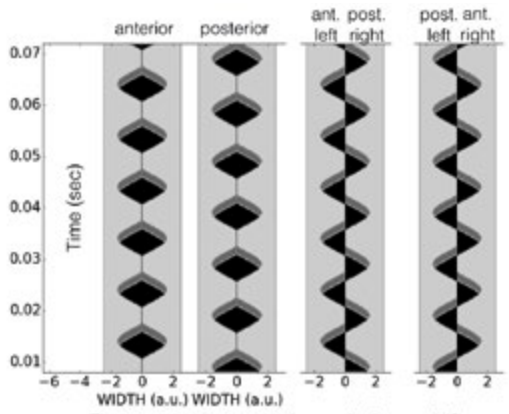


Fig. 6: Anterior-posterior anti-phase shift

IV. Summary and Conclusion

Synthetic kymograms enable simulating kymogram patterns the cause of which is known, facilitating the interpretation of natural kymograms. Conversely, synthetic kymograms provide a more detailed and user-friendly picture of the dynamic behavior of glottal models than the time series of scalar quantities, such as glottal area or glottal width.

IV. References

[1] J. Svec, On Vibration Properties of Human Vocal Folds, PhD Thesis, Rijksuniversiteit Groningen, 2000

[2] I. R. Titze, The Myoelastic Aerodynamic Theory of Phonation, The National Center for Voice and Speech, 2006, p. 259

[3] A. Pikovsky, M. Rosenblum, J. Kurths, Synchronization, Cambridge University Press, 2001, p. 225

Workshops

STUDYING THE PHYSICS OF VOICE PRODUCTION USING MECHANICAL REPLICAS

X. Pelorson[1], S. Becker[2]

[1] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France, CNRS, GIPSA-Lab, F-38000 Grenoble, France
[2] Fluid System Dynamics and Aeroacoustics, University Erlangen-Nuremberg Erlangen, Germany
xavier.pelorson@gipsa-lab.grenoble-inp.fr.
sb@ipat.uni-erlangen.de

Abstract: **Making measurements on the human voice organ *in vivo* is invasive and difficult. Reproducibility is usually hard to achieve with living or excised tissue. For studying fundamental phenomena of the physics of voice production, it is often more convenient to resort to mechanical replicas of the various parts of the voice organ. In comparison to computer simulations, the physicality of replicas can impart a more immediate and intuitive appreciation of the structural dimensions and the mechanical and/or acoustical aspects of voice production. In some cases, a highly simplified geometry for the vocal folds or the vocal tract is useful; it facilitates the specification of parameters, and the validation by measurement of numerical simulations of the corresponding geometries. In other cases, a more realistic 3-D geometry is needed to answer the research questions; such as for the study of the influence of particular pathologies or the cause of particular voice features. Modern 3-D printers have opened up the possibility to recreate complex shapes from MRI data, recently even in soft materials.**
In this workshop, we illustrate some applications of mechanical replicas in voice research.

Keywords : **Voice production, experiments, fluid mechanics, acoustics.**

I. INTRODUCTION

A crucial aspect of physical modeling deals with the experimental validation of the theoretical or numerical models. Experiments on excised larynxes or in-vivo observations are interesting for the purpose of analysis, for measuring the order of magnitude and the range of physiological parameters such as the subglottal or the oral pressure together with some relevant acoustical properties of the generated sound and for comparison with the models outputs. However these experiments are not suitable for a direct validation itself. Experimental quantitative validation of physical models requires indeed a highly controlled set-up, reproducible and on which accurate measurements of relevant quantities (flow velocity, pressure distribution, elastic response…) can be performed, which is not usually possible in-vivo on even on excised larynxes. Since van Den Berg et al. [1] pioneer work, many mechanical replicas of the vocal folds have been designed with an increasing complexity and thus realism. From very stylized static replicas under steady flow conditions ([2], [3], [4]), specific set-ups have been improved in order to add unsteadiness in the flow [5], moving vocal folds in forced motion [6], deformable self-sustained oscillating structures ([7], [8], [9]). In addition, novel techniques such as Particle Image Velocimetry or Laser Doppler Anemometry have become available and allow for flow measurements with details that were unbelievable twenty years ago. In this workshop, we present some basic general principles for the design of mechanical replicas and develop two examples, one for the vocal folds and one for the vocal tract.

II. METHODOS

A mechanical replica of the vocal tract including all the details and complexity of the human anatomy is obviously not possible to achieve at present time. The challenge of mechanical replicas involves therefore a necessary simplification while reproducing the major physical effects. These effects can be quantified *a priori* using a dimensionless analysis as follows.

For the sake of simplicity we consider here the equations of motion for a Newtonian fluid under the assumption of (local) incompressibility. These equations are derived from the principle of mass and momentum conservation, respectively:

$$\vec{\nabla}.\vec{v} = \vec{0} \qquad (1)$$

$$\rho_0 \frac{\partial \vec{v}}{\partial t} + \rho_0 \left(\vec{v}.\vec{\nabla}\right)\vec{v} = -\vec{\nabla}p + \mu_0 \vec{\nabla}^2 \vec{v} \qquad (2)$$

where p = p(t,x,y,z) and $\vec{v} = \vec{v}(t,x,y,z)$ are the flow pressure and velocity, respectively, ρ_0 is the constant air density and μ_0 the kinematic viscosity coefficient.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

Dimensionless analysis consists in a rewriting of equations (1) and (2) using dimensionless quantities :

$$\vec{v}^* = \frac{\vec{v}}{v_0}, p^* = \frac{p}{\rho_0 v_0^2}, t^* = \frac{t}{t_0}, x^* = \frac{x}{l_0}, ...$$

where v_0 is a caracteristic flow velocity, l_0 a dimension representative of the geometry and t_0 a caracteristic time scale. Using these quantities, the Navier-Stokes equation (2) can be rewritten as:

$$Sr\frac{\partial \vec{v}^*}{\partial t^*} + \left(\vec{v}^*.\vec{\nabla}^*\right)\vec{v}^* = -\vec{\nabla}^* p^* + \frac{1}{\mathrm{Re}}\vec{\nabla}^{*2}\vec{v}^* \qquad (3)$$

with :

$$Sr = \frac{l_0}{t_0 v_0} \text{ and } \mathrm{Re} = \frac{\rho_0 v_0 l_0}{\mu_0}$$

Sr and Re are respectively the Strouhal and the Reynolds numbers. In the form (3), thanks to the normalization, all terms become comparable with each other. The Strouhal number, for example, is a measure of the importance of inertial terms ($\rho_0 \frac{\partial \vec{v}}{\partial t}$) compared with the convective ones ($\rho_0\left(\vec{v}.\vec{\nabla}\right)\vec{v}$). The Reynolds number compares the relative importance of the convective term with respect to the effects of viscosity ($\mu_0 \vec{\nabla}^2 \vec{v}$).

A more detailed analysis reveals that the unsteadiness of the flow is mainly dominated by two factors: the time, or period, characteristic of the deformation of a constriction of the vocal tract (such as the glottis), t_0, and the time needed for the flow to pass the constriction. If d_0 is the length of the constriction, this time can be estimated by d_0/v_0. The relevant Strouhal number is therefore :

$$Sr = \frac{d_0}{t_0 v_0}$$

The analysis of flows through constrictions teaches that viscous losses are a function of the aperture of the constriction, h_0. The relevant Reynolds number is then:

$$\mathrm{Re} = \frac{\rho_0 v_0 h_0}{\mu_0}$$

Using typical values collected from the literature [10], [11] one obtains: $v_0 = 20$ m.s^{-1}, $t_0 = 10$ ms, $h_0 = 1$ mm, $d_0 = 10$ mm, leads to the following values for the Reynolds and the Strouhal numbers:

$\mathrm{Re} = O(10^3)$, $\mathrm{Sr} = O(10^{-2})$.

To be relevant, any mechanical replica of the vocal folds should therefore be able to operate with Reynolds and Strouhal numbers of this order of magnitude.

A collection of the physical parameters determining the process of human phonation is summarized in Table 1.

Table 1: Physical parameters of the human phonation process according to [9]

Vocal fold length (anterior-posterior)	$l_{vf} = 10 - 17$ mm
Vocal fold thickness (inferior-superior)	$t_{vf} = 9 - 10$ mm
Glottal gap diameter	$d_G = 0 - 5$ mm
Glottal duct length	$t_{Gd} = 2 - 5$ mm
Subglottal oscillation onset pressure	$P_{sub,on} = 200 - 1000$ Pa
Maximum intraglottal velocity	$U_{G,max} = 10 - 40$ m/s
Fundamental frequency	$f_0 = 80 - 300$ Hz
Reynolds number	$Re = O(10^3)$
Mach number	$Ma = O(10^{-1})$
Strouhal number	$Sr = O(10^{-2})$
Average vocal fold Young's modulus	$E = 5 - 20$ kPa

Note that in the case of an up-scaled replica such as in [2] or [8], this implies a reduction of the flow velocity and of the oscillation frequency. As an example, for a glottal replica up-scaled by a factor 3, the flow velocity will be a factor 3 lower than those expected within the (human) glottis while the oscillation frequency of the mechanical replicas will be a factor 9 lower.

III. A mechanical replica of the vocal folds

There are static and externally-driven vocal fold models in the literature. These can almost exclusively be applied to investigate aerodynamic properties of the phonation process. Thus, self-oscillating models that emulate real vocal folds very closely need to be taken into account to understand the fluid-structure-acoustic interaction process and its fully-coupled physics.

Different approaches were made to realize self-sustained, flow-driven oscillations in artificial vocal folds.

One attempt is to use membranous-type models, consisting of different material layers and thus making it possible to reproduce the different anatomical layers of vocal folds. Thereby, the membrane starts to oscillate by flow induction when the onset pressure in the subglottal region is high enough [14]. The oscillation stops when the pressure is below an offset pressure. Several studies showed that a hysteresis between onset and offset pressure can be observed, meaning that a higher pressure needs to be applied to start the oscillation than to keep it self-sustained (c.f. [14], [16] among others).

The membranous-type self-oscillating vocal folds described above lack of precise control over their geometry [14]. Thus, a better reproduction of original human shaped vocal folds was achieved by molding models into predetermined shapes defining the initial geometry.

Based on the M5 model as proposed by Scherer [17], self-oscillating vocal folds made of homogeneous isotropic polyurethane rubber were introduced and further used in several studies (c.f. [9], [14], [15]). These models are denoted as single-layer models.

Using the geometrical form of these models, an anatomically more realistic approach to reproduce the different layers of human vocal folds has been presented later [18]. However, these multi-layer models still exhibit some differences in behaviour when compared to human vocal folds.

The research of the newer, molded vocal fold models, both single-layer and multi-layer, has just recently started. Therefore, there is still a lot of research necessary to better reproduce the behavior of human vocal folds while still offering a high amount of reproducibility.

In Figure 1 and Figure 2, an exemplary model based on the M5 geometry is respectively shown as a single-layer and a multi-layer model [14].

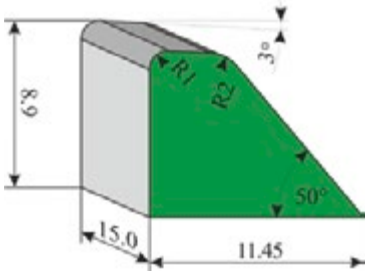


Fig. 1: Schematic of a single-layer vocal fold model

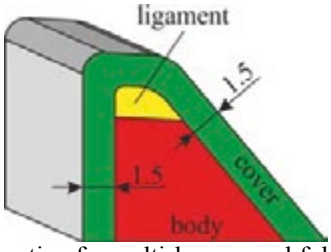


Fig. 2: Schematic of a multi-layer vocal fold model

V. A mechanical replica for the vocal tract

A. Static vocal tracts

The previously described experimental studies were based on simplified supraglottal channel geometries. When it comes to a more realistic approach concerning these geometries, little is known with respect to the phonation process in a synthetic human larynx model.

Thanks to imaging techniques such as MRI, one can now obtain detailed vocal tract shapes in 3-D and for a large variety of speech sounds. Using a 3-D printer it is thus possible to obtain a mechanical replica of the vocal tract with a high degree of geometrical accuracy. Such an accuracy is needed when considering the high frequency behavior of the vocal tract. As an example, we present if Fig. 3 three mechanical replicas obtained from different geometrical approximation of the vowel /a/ [12].

Fig. 3: Three different mechanical vocal tracts obtained using a 3-D printer. All share the same area function but with a different geometrical approximation: (from left to right: case 1: circular and centered shape, case 2: circular eccentered shape and case 3: elliptical centered shape).

The transfer function between different points was measured inside each replica using a microphone probe (B&K 4182) as detailed in [12]. Fig. 4 shows an example of comparison between the three different replicas.

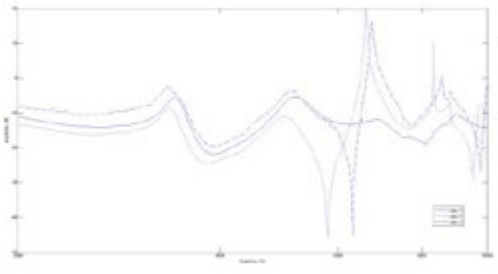


Fig. 4 : Example of measured transfer function for the three mechanical replicas of the vowel /a/.

As can be clearly seen from this example, above 5 kHz details of the geometry can have a considerable influence on the acoustical field.

B. Deformable vocal tracts

Allowing for the vocal tract wall to move, in order to mimic articulation, is a challenge especially when one wants to control precisely the time and spatial motion of the deformation. An attempt in this direction is presented in figure 5.

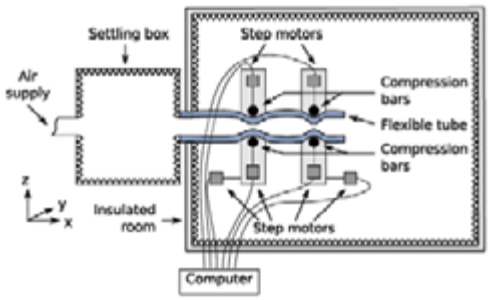


Fig. 5: deformable vocal tract set-up

The set-up consists of a flexible tube whose shape can be varied dynamically using bars which can laterally compress it in four locations along its length. The bars are controlled using step motors which allows, after calibration [13] to determine accurately the shape of the tube for each time step. Video and sound demonstrations of the capabilities of this set-up to mimic diphthongs and fricatives will be presented during the workshop.

ACKNOWLEDGMENTS

This work is supported by EU-FET grant EUNISON 308874.

REFERENCES

[1] Van Den Berg Jw., Zantema J.T., Doornenbal P. "On the air resistance and the Bernoulli effect of the human larynx". *J. Acoust. Soc. Am. 29 (5):* 625-31, 1957.

[2] R.C. Scherer and I.R. Titze, "Pressure-flow Relationship in a Model of the Laryngeal Airway with a Diverging", in *Vocal Fold Physiology*, College-Hill Press, San Diego, California, 179-193, 1983

[3] Gauffin J., Liljencrants J. "Modelling the air flow in the glottis". *Annual Bulletin RILP 22*: 41-52, 1988.

[4] Pelorson X., Hirschberg A., Wijnands A.P.J., Baillet H. "Description of the flow through in-vitro models of the glottis during phonation". *Acta acustica 3*: 191-202., 1995.

[5] Vilain C., Pelorson X., Fraysse C., Deverge M., Hirschberg A, Willems J. "Experimental validation of a quasi-steady theory for the flow through the glottis". *J. Sound and Vibration*, 276, 475-490, 2004.

[6] A. Barney, C. H. Shadle and P. O. A. L. Davis, "Fluid flow in a dynamic mechanical model of the vocal folds and tract. 1- Measurements and theory". *J. Acoust. Soc. Am.* 105, 444-455, 1999.

[7] Thomson, S. L., Mongeau, L. and Frankel, S. H.. "Aerodynamic transfer to the vocal folds". *J. Acoust. Soc. Am*., 118 :1689–1700, 2005.

[8] Ruty, N., Pelorson, X., Van Hirtum, A., Lopez, I., and Hirschberg, A. "An in-vitro setup to test the relevance and the accuracy of low-order vocal folds models". *J Acoust. Soc. Am.*, 121(1):479-490, 2007.

[9] S Becker, S Kniesburges, S Müller, A Delgado, G Link, and M Kaltenbacher. "Flow-structure-acoustic interaction in a human voice model". *J Acoust Soc Am,* 125(3):1351–1361, 2009.

[10] Hirano, M. "Morphological structure of the vocal cord as a vibrator and its variations". *Folia Phoniatrica*, 26(2) :89–94, 1974.

[11] Baken, R.J. *Clinical Measurements of Speech and Voice*, Allyn & Bacon eds, Boston, 1987.

[12] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. Van Hirtum and X. Laval "Effects of higher order propagation modes in vocal tract like geometries"*J Acoust Soc Am*, 137(2), pp. 832-843, 2015.

[13] Van Hirtum A. "Deformation of a circular elastic tube between two parallel bars: quasi-analytical geometrical ring models". *Mathematical Problems in Engineering*, 2015:1-15, 2015.

[14] Kniesburges, S., Thomson, S.L., Barney, A., Triep, M., Sidlof, P., Horáčcek, J., Brücker, C. and Becker, S."In vitro experimental investigation of voice production". *Current Bioinformatics*, 6(3):305-322, 2014.

[15] Kniesburges S, Hesselmann C, Becker S, Schlücker E, Döllinger M "Influence of vortical flow structures on the glottal jet location in the supraglottal region". *J Voice* 27(5):531-544, 2013.

[16] Lodermeyer A, Becker S, Döllinger M, Kniesburges S "Phase-locked flow field analysis in a synthetic human larynx model", *Exp Fluids* 56: 77-89, 2015.

[17] RC Scherer, KJD Witt, C Zhang, BR Kucinschi, and AA Afjeh. "Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees". *J Acoust Soc Am,* 109(4):1616 – 1630, 2001.

[18] JS Drechsel and SL Thomson. "Influence of supraglottal structures on the glottal jet exiting a two-layer synthetic, self-oscillating vocal fold model". *J Acoust Soc Am*, 123(6):4434–4445, 2008.

MODELLING VOICE PRODUCTION WITH LARGE-SCALE PHYSICS-BASED NUMERICAL SIMULATIONS

O. Guasch[1], J. Jansson [2]

[1] GTM Grup de recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, Barcelona 08022, Catalonia, Spain
[2] Computational Technology Laboratory, Basque Center for Applied Mathematics and KTH
Royal Institute of Technology, Bilbao, 48009, Basque Country, Spain
oguasch@salleurl.edu jjan@kth.se

Abstract: **The human voice organ fits in a small space having a characteristic length of ~20cm. Large amounts of complex physical phenomena combine in it so as to produce sounds. Despite of the reduced dimensions of the voice organ, however, a complete numerical simulation of its physics is still out of reach, even when using massively parallel supercomputers.**
This has led researchers to split the problem of voice generation into parts, independently focusing for instance, on simulating the self-oscillation of the vocal folds to generate the glottal pulse, the propagation of acoustic waves in moving vocal tracts to produce diphthongs, or the diffraction of the glottal jet pressure by the teeth, which results in fricative sounds. In this workshop, a review will be given of the type of equations and difficulties encountered when trying to solve these type of phenomena and show that, under some assumptions, the first unified simulations coupling the mechanics, aerodynamics and acoustics of the vocal folds and vocal tract, may not be as far as one might think. A workflow from 3D biomechanical models, to the generation of vocal fold self-oscillations, flow and acoustic waves to the radiated sound may be feasible in the short term.

Keywords: **Voice production, Finite element method, Vocal folds, Vocal tract, Articulatory synthesis**

I. INTRODUCTION

The voice organ consists of four main components, namely the lungs, the larynx that includes the vocal folds (VF), the vocal tract (VT) and the nasal tract. The air expelled from the lungs flows through the larynx generating a pressure drop at the glottis, which induces self-oscillations of the vocal folds. Because of such oscillations, acoustic waves are generated and propagate through the vocal tract, which modulates them up to the mouth exit where they become emitted outwards.

From a physical point of view, most interesting phenomena take place at the larynx and at the vocal tract. Despite of the reduced dimensions of the voice organ, these phenomena may be very intricate involving aperiodic and/or turbulent flows, fluid-structure interaction with elastic solids (the VF), which collide, distort and vibrate, propagation of acoustic waves in a moving VT of complex geometry, etc. Consequently, a complete physics-based simulation of the voice generation process implies solving the fully coupled equations describing the interaction between the mechanical, aerodynamic and acoustic fields. At present, this seems totally out of reach, even when resorting to massive supercomputer facilities.

The natural option when a problem is too hard to be addressed in all its complexity is that of split it in smaller, easier subproblems. This is what traditionally has been done in numerical voice production, some research teams aiming at simulating the behavior of the VF (see e.g., [1,2,3]), others at the acoustics of the VT either in the frequency domain (see e.g., [4]) or in the time domain (see e.g., [5,6,7]).

In this workshop we will first comment on some of the equations governing voice production. Due to space restrictions, only the case of diphthong generation will be described in detail in this short communication, together with some issues related to the simulation of the VF self-oscillations. The inherent numerical difficulties that have to be faced when addressing these type of problems using the finite element method (FEM) will be outlined. The case of vowels and fricatives will be left for the oral presentation.

Besides, a brief discussion on the status of unified simulations that involve both, the VF and the VT, will be presented. Though as said, a full coupling of the mechanical, aerodynamic and acoustic fields seems beyond the capability of current computers, some promising steps can be done in that direction.

II. NUMERICAL SIMULATIONS

A. Production of diphthongs

To have a glimpse at the numerical approach to voice production let us first consider the case of diphthong generation. It is to be noted, however, that

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

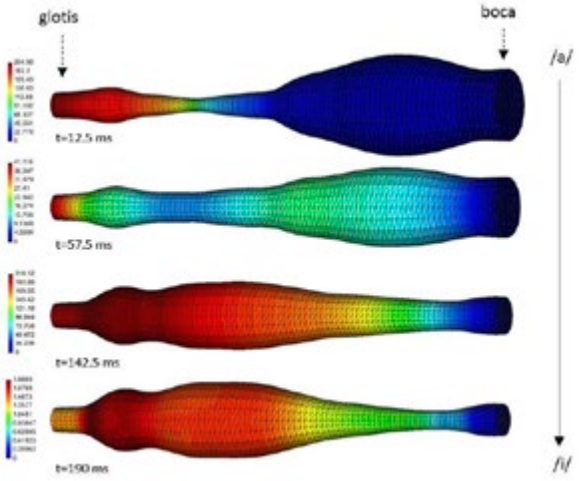


Fig.1 Snapshots of vocal tract evolution form /a/ to /i/. Surface pressure distribution.

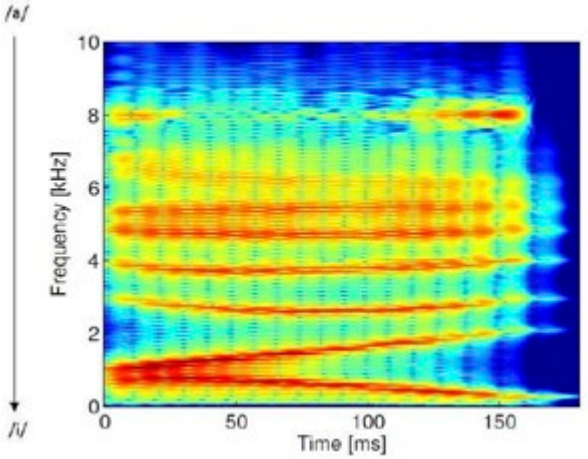


Fig.2 Spectrogram corresponding to the production of diphthong /ai/.

from a computational point of view, the easiest sound to be produced is a vowel. Once we get an appropriate VT geometry corresponding e.g., to the pronunciation of /a/, one has to simply solve the reduced wave equation $\partial_{tt}^2 p - c_0^2 \nabla^2 p$ for the acoustic pressure p in that domain. The equation has to be complemented with suitable boundary conditions, namely prescription of a glottal pulse at the glottis, admittance condition at the VT walls and free field radiation condition at the outer boundary of the computational domain.

From a numerical point of view, the wave equation does not present special difficulties if one does not aim at determining wave propagation at very far distances. The Laplacian operator is well-behaved and most difficulties lie on imposing a non-reflecting boundary condition to let acoustic waves propagate to infinity. This is usually achieved either by means of a perfectly matched layer (PML), by using infinite elements or by imposing high order non-reflecting Sommerfeld conditions. The situation can get more complex, though, as long as one wants to include finer details in the simulation, like considering the flexibility of the VT walls [8], or imposing a wall frequency dependent admittance in time domain simulations.

If we next focus on diphthongs [8], at first sight the main obvious difference with respect to vowels is that one has to deal with moving vocal tracts. For example, to generate /ai/ the VT transitions from the shape of an /a/ to that of an /i/ (see Fig. 1). The corresponding vowel formants will evolve accordingly (see Fig. 2). However, this apparently simple fact has notorious implications from a numerical point of view.

The first one is that the reduced wave equation used for vowels is no longer useful for diphthongs. This equation directly stems from the linearized versions of the mass and momentum conservation equations for a perfect fluid, and it is stated in an Eulerian frame of reference.

Withal, for moving vocal tracts this option is awkward and the physics are better described in an arbitrary Lagrangian-Eulerian (ALE) frame of reference. In the particular case of resorting to a quasi-Eulerian approach the mass and momentum conservations equation read

$$\begin{aligned} \tfrac{1}{\rho_0 c_0^2}\partial_t p - \tfrac{1}{\rho_0 c_0^2}\boldsymbol{u}_d\cdot\nabla p + \nabla\cdot\boldsymbol{u} = Q, \\ \rho_0\partial_t\boldsymbol{u} - \rho_0\boldsymbol{u}_d\cdot\nabla\boldsymbol{u} + \nabla p = \boldsymbol{f}, \end{aligned} \tag{1}$$

where p stands for the acoustic pressure, $\boldsymbol{u}$ for the acoustic particle velocity, $\boldsymbol{u}_d$ for the distorting computational mesh velocity, Q for a volume source distribution and $\boldsymbol{f}$ for an external force. As usual ρ_0 represents the air density and c_0 the speed of sound.

The equations in (1) are known as the wave equation in mixed form, in an ALE framework, and cannot be combined to get a unique reduced wave equation for the acoustic pressure. From a numerical point of view, solving (1) becomes more intricate than dealing with the reduced wave equation. For instance, the standard Galerkin FEM approach to solve the variational form of (1) has to satisfy an inf-sup compatibility condition which does not allow to use equal interpolations for the acoustic pressure and particle velocity fields, unless one resorts to stabilization strategies [8]. Moreover, the mesh node displacements due to distortion of the vocal tract have to be computed as time evolves. In the case of using simple geometries, as those in Fig. 1, this can be done by solving a Laplacian equation for the node displacements. Nonetheless, if detailed MRI vocal tract geometries are used, remeshing strategies become

necessary, which considerably increases the computational cost.

B. Self-oscillations of the vocal folds

Diphthongs can still be generated without resorting to supercomputer facilities if one makes some simplifying assumptions such as imposing zero pressure release conditions at the mouth exit and dealing with simplified VT with circular or elliptical cross sections.

The situation gets more intricate when moving to fricatives. For instance, in the case of an /s/ sound is mainly produced by the diffraction of the constricted glottal jet flow pressure at the upper and lower incisors. Typically, one has to follow a hybrid approach using an acoustic analogy or resorting to acoustic perturbation equations. First, a computational fluid dynamics (CFD) simulation is carried out to solve e.g., the incompressible Navier-Stokes equations, from which an acoustic source term is derived. In a second step, that term is input in an acoustic wave equation to compute the aerodynamically generated sound. Most of the numerical problems concern the CFD computation. Non-linearity has to be dealt with at each time step of the simulation. An inf-sup compatibility condition is again to be satisfied and demands using different polynomial fields for the aerodynamic pressure and velocity. Instabilities may also occur for convection dominated flows and may also arise for very small time steps, at the beginning of the evolutionary process. Moreover, turbulence has to be modelled somehow. Stabilization strategies become again a suitable framework to deal with all these mathematical and/or physical problems.

Even when considering simple geometries, three-dimensional simulations for fricatives require using supercomputer facilities. Another turn of the screw appears if one considers computing the fluid-structure interaction governing the VF self-oscillations. Modelling phonation demands high-fidelity mathematical models posing conservation laws for mass, momentum and energy in terms of differential equations for the fluid (air) and structure (vocal folds) materials, as well as for the interaction of these two phases. This Fluid-Structure Interaction (FSI) problem is a very challenging one involving multiphysics, multiscales and non-linearity.

Two strategies can be followed. The first one relies on a separated approach, where the fluid and structure sub-problems are formulated independently, and iteratively solved on different computational domains with coupling conditions on the fluid-structure interface [9]. Alternatively, one could follow a monolithic scheme that defines and solves the problem on a single computational domain. This results in more

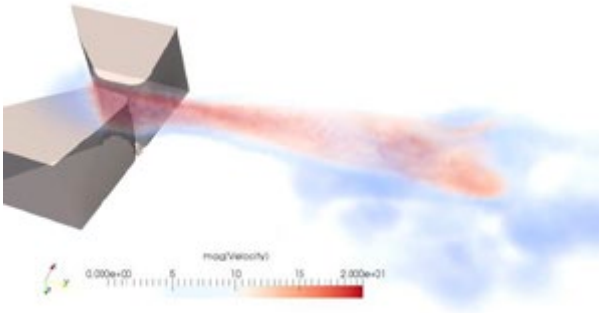


Fig.3 Snapshot of VF self-oscillation simulation using FEM.

robustness though at the price of larger systems of unknowns [10].

We have favored the second option and developed a simulation for the larynx, which takes the form of a monolithic Unified Continuum FSI model (UC-FSI) with implicit contact modeling. This has been implemented in the FEniCS-HPC automated PDE solving framework, optimized for massively parallel hardware architectures.

The problem to be dealt with is that of solving [10]

$$\begin{aligned} \rho_0(\partial_t \boldsymbol{u}_0 + \boldsymbol{u}_0 \cdot \nabla \boldsymbol{u}_0) + \nabla \cdot \boldsymbol{\sigma} &= \boldsymbol{f}_0, \\ \nabla \cdot \boldsymbol{u}_0 &= 0, \\ \partial_t \theta + \boldsymbol{u}_0 \cdot \nabla \theta &= 0 \end{aligned} \tag{2}$$

with $\boldsymbol{u}_0$ denoting the incompressible velocity, $\boldsymbol{f}_0$ an external force and θ a phase function that takes the value 0 at the solid and 1 at the fluid. The incompressible pressure p_0 is subtracted from the Cauchy stress tensor so that $\boldsymbol{\sigma} \equiv \overline{\boldsymbol{\sigma}} - p_0 \boldsymbol{I}$. The phase function θ allows one to define in a closed formulation the constitutive laws for both, the fluid and solid phases. A Newtonian law is used for the fluid stress $\overline{\boldsymbol{\sigma}}_f$ while a Neo-Hookean law is needed for the solid stress $\overline{\boldsymbol{\sigma}}_s$, if an Eulerian description of the problem is to be followed. We get

$$\begin{aligned} \overline{\boldsymbol{\sigma}} &= \theta \overline{\boldsymbol{\sigma}}_f + (1-\theta)\overline{\boldsymbol{\sigma}}_s, \quad \overline{\boldsymbol{\sigma}}_f = 2\mu_f \epsilon, \\ \mathrm{D}_t \overline{\boldsymbol{\sigma}}_s &= 2\mu_s \epsilon + \nabla \boldsymbol{u}_0 \overline{\boldsymbol{\sigma}}_s + \overline{\boldsymbol{\sigma}}_s \nabla \boldsymbol{u}_0{}^T, \end{aligned} \tag{3}$$

where μ_f stands for the fluid viscosity, μ_s for the solid shear modulus and ϵ for the strain rate tensor.

Equations (2) and (3) supplemented with appropriate initial and boundary conditions are solved again using a stabilized FEM approach, with a moving mesh tracking the phase interface. The proposed overall strategy allows one to set the problem in a unified PDE (partial differential equation) modeling, which includes contact, adaptive error control, interface robustness and an implicit model for turbulence.

The simulations of phonation presented herein (see Fig. 3) have involved meshes of ca. 200k-300k nodes and were performed on a supercomputer using 300-1000 cores. A typical computation time is ca. 5 hours (wall clock time) for an oscillation cycle with the contact model enabled.

III. TOWARDS UNIFIED SIMULATIONS OF VOICE

As mentioned in the Introduction, a unified simulation that accounts for the complete interaction with feedback between the three fields involved in the generation of voice seems out of the capabilities of current supercomputer facilities. Such an approach would imply performing an FSI simulation for the VF plus VT using a compressible model for the flow. The mesh should be fine enough and the time step size appropriate to generate acoustic waves covering completely the audible spectrum. In addition, the equations should be set in an ALE framework to account for diphthong and syllable generation. If the problem was to be tackled in its full complexity, a biomechanical model should be included to determine the VT shape according to muscle activation.

Aiming at a less ambitious goal, some intermediate steps between the split problems in the preceding section and a complete unified simulation can be foreseen to be feasible in the short term. The generation of a vowel simulating the VF self-oscillations and the VT acoustics using an acoustic analogy type approach is a realistic objective. Actually, it was almost achieved in [15] though the VF motion was still prescribed in the acoustic computation of that work. Extension to diphthongs could follow. The numerical production of fricative sounds is also viable and syllable generation, though more demanding, may be the natural subsequent step. Besides, coupling a biomechanical model with the VT acoustics to generate diphthongs seems also a realistic possibility.

IV. CONCLUSIONS

In this workshop we have reviewed some of the numerical challenges and difficulties encountered when trying to numerically simulate the physics of the voice organ. This includes the self-oscillations of the vocal folds and the generation of sounds like vowels, diphthongs, fricatives or syllables.

The inherent complications of those problems and the impossibility of a fully coupled simulation, justify the subproblem splitting strategy followed to date. However massively parallel supercomputers will soon allow one to deal with simulations involving the vocal folds and vocal tract biomechanics, aerodynamics and acoustics.

ACKNOWLEDGMENTS

This work is supported by EU-FET grant EUNISON 308874.

REFERENCES

[1] W. Zhao, C. Zhang, S. Frankel, and L. Mongeau, “Computational aeroacoustics of phonation, Part I: Computational methods and sound generation mechanisms,” *J. Acoust. Soc. Am*., vol 112(5), pp. 2134–2146, 2002.

[2] G. Link, M. Kaltenbacher, M. Breuer, and M. Döllinger, “A 2d finite-element scheme for fluid–solid–acoustic interactions and its application to human phonation,” *Comput. Methods Appl. Mech. Engrg*., vol 198(41), pp. 3321-3334, 2009.

[3] H. Luo, R. Mittal, X. Zheng, S. A. Bielamowicz, R. Walsh, J. K. Hahn, “An immersed boundary method for flow-structure interaction in biological systems with application to phonation,” *J. Comput. Phys*., vol. 227(22), pp. 9303 – 9332, 2008.

[4] K. Motoki, “Three-dimensional acoustic field in vocal-tract,” *Acoust. Sci. & Tech.*, vol. 23(4), pp. 207–212, 2002.

[5] T. Vampola, J. Horacek, and J.G. Svec, “FE modeling of human vocal tract acoustics. Part I: Production of czech vowels,” *Acta Acust*., vol. 94(5), pp. 433–447, 2008.

[6] M. Arnela and O. Guasch, “Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method,” *J. Acoust. Soc. Am.*, vol. 133(6), pp. 4197–4209, 2013.

[7] M. Arnela, O. Guasch, and F. Alías, “Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations,” *J. Acoust. Soc. Am.*, vol. 134(4), pp. 2946–2954, 2013.

[8] O. Guasch, M. Arnela, R. Codina and H. Espinoza, "Stabilized finite element formulation for the mixed convected wave equation in domains with driven flexible boundaries", *Noise and Vibration: Emerging Technologies (NOVEM2015)*, April 13-15, Dubrovnik (Croatia), 2015.

[9] J. Hoffman, J. Jansson and M. Stöckli, “Unified continuum modeling of fluid-structure interaction,” *Math. Mod. Meth. Appl. S.,* vol. 21(3), pp.491-513, 2011.

[10] J. Hoffman and C. Johnson, *Computational turbulent incompressible flow,* Applied Mathematics Body and Soul Vol. 4, Springer-Verlag Publishing, 2006.

[11] M. Kaltenbacher, S. Zörner, A. Hüppe and P. Sidlof, “3D simulations of human phonation”, *WCCM XI (11th World Congress on Computational Mechanics),* July 20-25, Barcelona, Catalonia (Spain), 2015.

FP – Mechanical-Markerless

THE FACE VIBRATION IN RESONANCE EXERCISES MEASURED BY THREE DIFFERENT METHODS - FIRST RESULTS

M. Frič

Musical acoustics research centre, Academy of performing arts in Prague, Prague, the Czech Republic
marekfric@centrum.cz

Abstract: **Results of three vibration measurement methods (by accelerometers, vibrometers, and optical measurement of the facial surface vibration by high speed camera) are presented. A subject (male, 37 y. o.) performed two vocal exercises (glissando and crescendo) in three different types of vocal tract settings and three types of resonance tubes extending the vocal tract. In the case of glissandos peaks were found on the frequencies corresponding to the maximal vibration amplification of monitored areas (larynx, ala of nose, forehead). Results confirmed the effect of the vocal tract shape and the resonance tubes as for the frequency position of this maxima so for the maximal amount of the vibration amplification. The optical measurement showed position of the facial regions with the maximal vibration amplitudes. With resonance tubes attached the area near the lips dominates, and in brumendos the maximal amplitudes were observed in the nose region.**
Keywords : **voice, resonance tubes, facial vibration**

I. INTRODUCTION

The facial region vibration is usually connected with the production of resonant voice, which is the phenomena associated to professional voice production. Several methods have been applied to measure the face vibrations, using accelerometers [1,2] and vibrometers [3,4].

Those methods showed an increase in vibration of the nose for nasal sounds and resonant voice in comparison to non-nasal and non-resonant [2]. Scanning by vibrometers showed distinct vibrant parts of the face. Their position depends on the measured vowels or voice register [4]. Semi-occlusion and extending of vocal tract by resonance tubes are used in voice education and therapy [5,6]. These methods increase a singer's (or patient's) sensation of face vibration and improve the voice quality. Modeling of this phenomenon revealed up to 3-times increase of pressure in the mouth, just behind the lips, when the resonance tube is used compared to the vowel "u" [7].

The goal of the study is to document by three parallel methods the effect of resonance tubes to the vibrations of the face and larynx.

II. METHODS

The same subject (male, 37 years old) participated in two experiments. In the first experiment the surface vibrations were measured using the piezoelectric accelerometers type PCB Electronics 352C23 (with sensitivity 1.02 mV/(m/s^2) placed on the **larynx**, and with sensitivity 0.5 mV/(m/s^2) placed on **the ala of nose**). Laser Doppler vibrometers Polytec Fiber interferometer OFV 518 (sensitivity 125 mm/(s.V) measured the **upper lip** and the middle part of the **forehead**. The subject performed upward and downward **glissandos** (two times) in 3 different types of vocal tract setting (opened and closed brumendo and vowel "u"), and with 3 types of attached resonance tubes (60 x 1 cm, 40 x 1,2 cm and 29 x 1 cm).

In the second experiment, in addition to the above mentioned methods, the **crescendos** from *mf* to *ff* of tone E3 (165 Hz) was recorded using a high-speed camera (HSV) (Vision Research Phantom v611, FPS: 8000, a macro lens Nikon AF MICRO NIKKOR 60 mm). All utterances were also recorded using Laryngograph D-200 (acoustic and EGG signal).

All signals (audio, electroglottographic, two vibrometers and two accelerometers) were recorded synchronously using AUDACITY (96 kHz, 24-bit). Their analysis was performed on the windows with the length of 50 ms and in 25 ms steps. Data from vibrometer and accelerometer were indefinite integrated to calculate the displacement amplitude. The resulting amplitudes of vibration for the windows were calculated as the mean values of peak-to-peak magnitudes of vibrations. Recordings of high-speed camera were analyzed by software VIC-2D 2009 Digital Image Correlation, version 2009.1.0. Image calibration shows that the displacement change of 1 pixel corresponds to the value 8,95e-05 m.

From the calculated values of vibration amplitudes the frequency position of the maxima and maximal values of amplitudes were detected. One-way analysis of variance (ANOVA) was carried on vibration

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

amplitudes with the type of vocal tract and the length of resonance tube as an independent variable.

III. Results

A. Glissandos

The results of glissandos measurement show that the dependence of the vibration amplitudes on the pitch has several distinct maxima. Significant effect of vocal tract and the length of the tube were found for pitch positions and values of maximum vibration amplitudes in all detected peaks.

For the larynx vibration (see Fig. 1A), two peaks were generally measured by accelerometer. The **first major peak** (see Figure 2AB left column) was typically near E3 (165 Hz = 52nd midi semitone) in modal register especially during phonation into the tubes. Effect for pitch and amplitude was found (*F*=15.13, df=5, *P*<1e-05 and *F*=128.52, df=5, *P*<1e-13, respectively). Pitch position for that peak gradually decreased with the length of the resonance tube, for brumendos and vowel "u" it was relatively low. The amplitude was significantly higher when resonance tubes were attached.

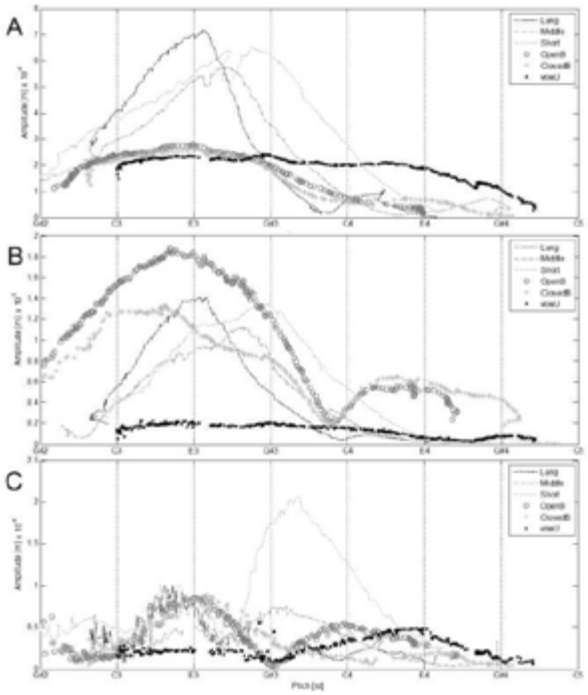


Fig. 1 Examples of measured vibration amplitudes in glissandos for different shapes of vocal tract for A) larynx vibration B) vibration of nasal ala measured by accelerometers and C) forehead vibration measured by laser vibrometer.

The **second peak** was typically placed in falsetto register near D#4 (63rd midi semitone), except the measurement of shortest tube and open brumendo. Effect for pitch and amplitude was found (F=56.83, df=3, P<1e-07 and F=33.99, df=3, P<1e-05 respectively).

The vibrations of the nasal ala (see Fig. 1B and Fig. 2 middle column) showed more complicated frequency-amplitude relations. When tubes were attached, the positions of the **first peak** were similar to those measured in larynx. The effect for pitch and amplitude was found (*F*=12.99, df=5, *P*<1e-04 and *F*=46.47, df=5, *P*<1e-09 respectively). The most significant differences from the larynx measurement were found for amplitudes of brumendos. Brumendos had noticeably higher maximum amplitudes of vibration and lower frequency positions of that peak. For the vowel "u" the results was reversely, the maximum amplitude was low and frequency position of the peak was high.

Effect for pitch and amplitude was also found for the **second peak** (*F*=12.99, df=5, *P*<1e-04 and *F*=46.47, df=5, *P*<1e-09 respectively). Similar to first peak both brumendos had the highest amplitudes and relatively low pitch positions.

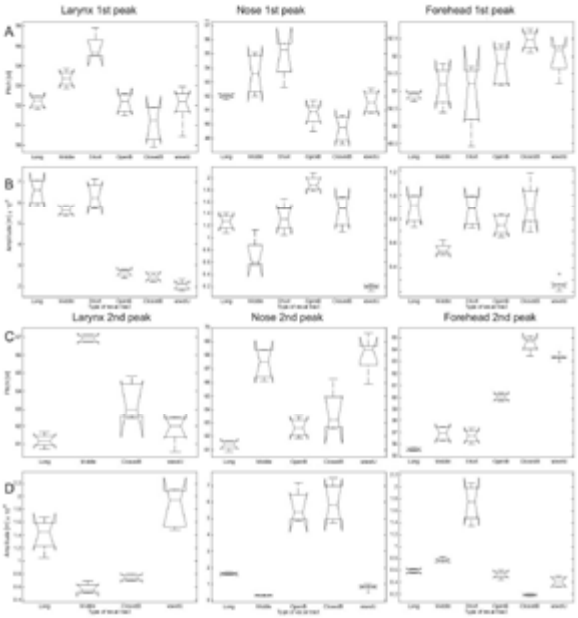


Fig. 2 Mean values and variations of positions in semitones (rows A and C) and maximal amplitudes in 10e-06 m (rows B and D) for vibration of larynx (left column), ala of nose (middle column) and forehead (right column), and for the first (rows A and B) and second (rows C and D) peaks respectively.

Unlike to previous measurements, the forehead vibrations (measured by vibrometer) revealed three distinct vibratory peaks (see Fig. 1C and Fig. 2 right column), especially for longest and middle tubes.

Effect for pitch and amplitude was also found for all of them (1st: F=5.8, df=5, P<1e-02 and F=26, df=5, P<1e-07; 2nd: F=299.91, df=5, P<1e-17 and F=67.1, df=5, P<1e-11; 3rd: F=399.69, df=1, i<1e-05 and i=72.84, df=1, <1e-04 respectively).

The **first maximum** was placed near the E3 with the smallest variation between types of vocal tract, brumendos and resonance tubes had higher amplitude than vowel "u". The position of **second peak** was considerably dependent on the type of vocal tract, the position of this maxima had the largest pitch range. Brumendos had typically higher pitch but the shortest tube had the largest amplitude. The **third peak** was detectable only for the long and middle tubes, pitch decreased and the amplitude increased with the length of the tube.

The measurement of the upper lip vibration by laser vibrometer failed due to relatively large movement of the mouth in glissandos so the second experiment was prepared especially to measurement vibration of the face by mean of high-speed camera (HSV).

B. Crescendos

ANOVA reveals that the significant effect of the resonance tube and shape of vocal tract was found for all measured parameters in crescendos (see Table 1).

Measured differences in vibration amplitudes of lip and nose by vibrometer or accelerometer respectively, and by HSV in x and y direction is depicted in Fig. 3. Similar to glissandos for lip vibration, long and middle resonance tubes had proportionally higher amplitudes of larynx than brumendos and vowel "u". Nose vibrations showed significantly higher amplitudes in brumendos.

Tab. 1 Results of ANOVA for measurements of the crescendos by different methods.

Param.	Method	df	*F*	*P*
SPL [dB]	10 cm	5	1385.17	6.39875e-144
CQ [%]	dEGG	5	645.27	1.38385e-114
Larynx	Acc.	5	9923.76	1.10745e-221
Forehead	Vib.	5	571.49	5.40086e-108
Lip	Vib.	5	4286.5	4.59646e-185
Lip X	HSV	5	670.09	1.0179e-060
Lip Y	HSV	5	477.91	2.31039e-055
Nose	Acc.	5	15940.18	1.46496e-240
Nose X	HSV	5	40.54	4.19291e-020
Nose Y	HSV	5	160.31	6.86016e-039

The measurement results for other parameters is depicted in Fig. 4. Similarly to glissandos, the amplitude of larynx grew with the length of the resonance tube; forehead amplitudes in opposite way decreased with the length of resonance tube, the highest amplitude reached in open brumendo; SPL was highest for vowel "u" and contact quotient was longest in the shortest resonance tube.

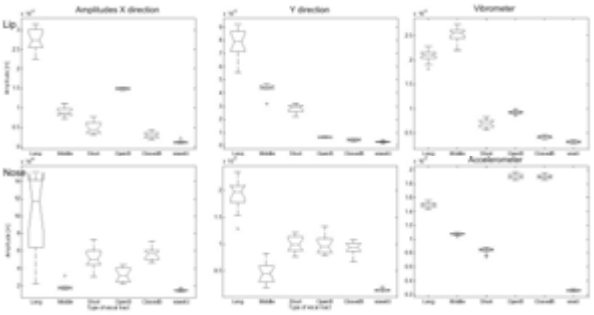


Fig. 3 Amplitudes of lip (upper row) and nose (bottom row) vibration measured by high-speed camera (left an middle column) and by vibrometer and accelerometer respectively.

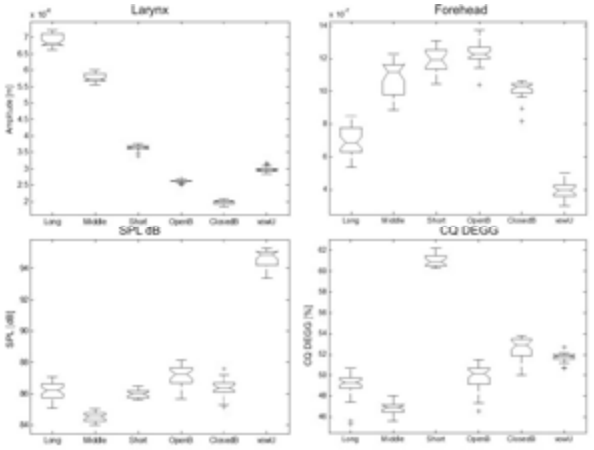


Fig. 4 Mean values and variation of measurement of larynx and forehead vibration amplitudes measured by accelerometer and vibrometer respectively, SPL at 10 cm from the opening, CQ DEGG -contact quotient measured from the first derivative of electroglottographic signal.

Fig. 5 shows the standard deviation of the displacement measured by HSV in x and y direction. These measurements are proportional to the maximal displacement. Comparison of different resonance tubes and vocal tracts reveals that the highest displacements in y direction are located near the mouth for the longest resonance tube. Amplitudes in y direction decreased for shorter tubes and for opened and closed brumendo. The smallest vibration was measured for vowel "u".

IV. Discussion and Conclusion

For all of the measured vibration parameters in glissandos, significant effects of the vocal tract setting (or the resonance tubes' length) were found for the

oscillation amplitudes maxima and for the peak frequency positions in all measured positions, on the larynx, nasal ala and the forehead.

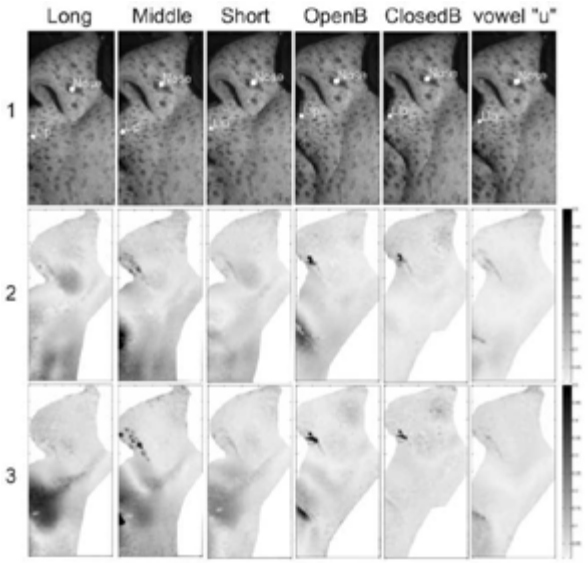


Fig. 5 Display of the face measurement by high-speed camera (first row), standard deviation of measured displacement in *x*-direction (second row), standard deviation of measured displacement in *y*-direction (third row). Measured vocal tracts and resonance tubes are depicted in columns.

Measurements of resonance tube provided the main effect on the larynx vibration and the lip vibration, which were modeled in [7]. Different positions of the peak frequency document significant effect of the voice pitch on the amplitudes of vibration. This observation implies a substantial influence of the length of the tube for the efficient use in voice therapy. The observed effect on the separation of voice registers can be used in the training of professional voice.

Larynx vibrations were in relation with the vibrations of the lip area in the face, which were very good recognized using the high speed camera. The highest amplifications of the lip and larynx vibration were reached with the longest resonance tube.

The nose vibration maximum was observed for open brumendo in modal register, which is in accordance with previous results [1–3], but the most significant effect on the peak's pitch position with resonance tubes were observed in subject's falsetto range.

Forehead vibration showed the most complicated shape, typically with three resonance peaks. The second peak was placed between modal and falsetto, which is the main difference from larynx and nose vibration character. This observation are assumed to be connected with the starting point of head resonance, because the forehead's vibration reached the maximum.

Optical measurement of vibration showed the highest amplitudes near the mouth especially with the resonance tubes, and the highest vibration of nose area for brumendos. There was a little discrepancy between measurement of vibration amplitudes by high speed camera (in *x* and *y* direction) and accelerometer or vibrometer respectively. It could be caused by measurement of vibration only in two dimensions. Using two cameras will allow to track the movement of the face in all three dimensions.

The results provide a basis for future searching of the relationships between the resonant frequencies of different cavities or resonance tubes and efficiency of vocal folds vibration.

V. Acknowledgment

This research has been supported by the Ministry of Education, Youth and Sports of the Czech Republic in the Long term Conceptual Development of Research Institutes grant of the Academy of Performing Arts in Prague: The "Sound quality" project.

REFERENCES

1. F.C. Chen, E.P. Ma, and E.M. Yiu, "Facial bone vibration in resonant voice production," J Voice 28 (5): 596-602 (2014).
2. E.M. Yiu, F.C. Chen, G. Lo, and G. Pang, "Vibratory and perceptual measurement of resonant voice," J Voice 26 (5): 675-679 (2012).
3. T.Kitamura. Measurement of vibration velocity pattern of facial surface during phonation using scanning vibrometer. Acoust.Sci.& Tech. 33[12], 126-128. 2012.
4. T.Kitamura, H.Hatano, T.Saitou, Y.Shimokura, E.Haneishi, and H.Kishimoto. A pilot study of vibration pattern measurement for facial surface during singing by using scanning vibrometer. Stockholm Music Acoustics Conference 2013. Proceedings of the Stockholm Music Acoustics Conference 2013 , 275-278. 2013.
5. I.V. Bele, "Artificially lengthened and constricted vocal tract in vocal training methods," Logoped.Phoniatr.Vocol 30 (1): 34-40 (2005).
6. S. Simberg and A. Laine, "The resonance tube method in voice therapy: description and practical implementations," Logoped Phoniatr Vocol 32 (4): 165-170 (2007).
7. I.R. Titze and A.M. Laukkanen, "Can vocal economy in phonation be increased with an artificially lengthened vocal tract? A computer modeling study," Logopedics Phoniatrics Vocology 32 (4): 147-156 (2007).

THE STUDY OF ACOUSTIC-ARTICULATORY RELATIONS IN PRODUCING SINGING VOWELS WITH THE USE OF EMA

Karina Evgrafova[1], Vera Evdokimova[1] , Pavel Skrelin[1] ,Tatiana Chukajeva[1]

[1]Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia
evgrafova@phonetis.pu.ru, postmaster@phonetics.pu,ru, skrelin@phonetics.pu.ru, chukaeva68@mail.ru

Abstract: **The given paper is aimed at investigating acoustic-articulatory relations in producing singing vowels. The study employs the method of electromagnetic articulography (EMA) to obtain exact data on articulatory characteristics in singing. Two types of recording experiments with the use of EMA are conducted involving four trained female singers. They were instructed to sing one of the Russian classical romances with the AG500 sensors attached to their main articulators in the first type of the experiment. The second type supposed reading aloud the text of the same romance. The obtained material (both singing and reading) is annotated and analyzed in terms of the kinematics of articulatory organs in singing as opposed to that in speech. The acoustic-articulatory relation from point of view of intelligibility in singing is considered.**

Keywords: **singing vowels, EMA, intelligibility**

I. INTRODUCTION

The problem of the intelligibility of singing vowels has been addressed in many studies [4-14, 17-18] which are reviewed well in [14]. They maintain that the intelligibility of isolated vowels sung by trained singers is relatively low for higher pitches, especially in case of isolated context. The distortion of the vowel quality is due to formant repositioning. Morozov [3] studied intelligibility of syllables sung by professional singers (males) as a function of fundamental frequency. Nelson and Tiffany [14], Scotto di Carlo [15] found that vowel intelligibility differed for different vowels. Their results showed that vowel identification is much better if the vowel is produced in the consonant context. The main articulatory reasoning of these modifications may be caused by the adjusting of articulatory movements required for the generation of the singer's formant and other acoustic characteristics which are specific for singing. As a result, some vowels assume formant frequencies typical of a different vowel in singing, which causes confusion in vowel identification. The given paper is aimed to investigate acoustic-articulatory relations in producing singing vowels to get insight on the issue of their intellibility. The study employs the method of electromagnetic articulography (EMA) to obtain exact data on articulatory characteristics in singing and reading.

II. METHODOS

Speech and singing are both dynamic processes involving the rapid and small movements of many different articulators. Besides, the tracking of speech kinematics is fairly hampered as the movements of the tongue and other structures inside the mouth are largely obstructed from direct observation.

The method of electromagnetic articulography (EMA) is quite popular technology for tracking articulatory movements. It is often used in the study of speech as its high spatial (0.3mm) and temporal (sampling rate of 200-400Hz) resolution allows for the measurement of small and rapid articulatory movements. It uses electromagnetic sensor coils which are attached directly to the articulators. These sensors are of a small size (2x3mm), which allows minimizing the physical obstruction to natural speech production [2]. A set of transmitters mounted around the head of a subject produce an oscillating complex magnetic field that induces a current in the sensors allowing their position in space to be calculated. The principles of this calculation are explained in more detail at http://www.articulograph.de/.

A. Experimental Procedure

Two types of recording experiments with the use of EMA were conducted. The goal of the experiments was to obtain the parallel samples of singing and reading from the same subject.

Four trained female singers were involved in the experiments. They were instructed to sing one of the Russian classical romances with the AG500 coil sensors attached to their main articulators in the first type of the experiments. The second type supposed reading aloud the text of the same romance. Thus, we obtained the samples of singing and read speech and registered the EMA data in both types of vocal activities. Audio was recorded by means of Sanken cs-1 condenser shotgun microphone.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

In our research AG500 system by CarstensMedizinelektronik GmbH (Carstens) was used. The six transmitters of the AG500 are mounted on specific locations of an open clearcube structure that is stationed around the head of the seated subject, as shown in Figure 1.

Fig 1. The EMA technician is attaching thecoil sensors to the subject.

The AG500 system can support up to 12 sensor channels for tracking positions in three Cartesian dimensions and two angular dimensions. In our experiments 10 coil sensors were employed.

1. Tongue tip
2. Tongue dorsum (5.5-6.0cm from tip of tongue)
3. Back of the tongue
4. Upper lip
5. Lower lip
6. Head [reference]
7. Nasal bridge [reference]
8. Left ear [reference]
9. Right ear [reference]
10. Lower jaw[stationary]

Figure 2 below shows the coil attachment locations.

Fig. 2. The subject with coil sensors attached to the main articulators.

III. RESULTS

The obtained in the EMA experiments' material (both singing and speech) was annotated and analyzed in terms of the kinematics of articulatory organs in singing as opposed to that in speech.

The movement of each of the 10 coils were plotted separately in three Cartesian dimensions (x, y, andz). In our study tongue tip, tongue dorsum, tongue back, upper and lower lips movements in singing and speech were of special interest. The plots of these articulators' movements in x-plane (forth and back) and z-plane (up and down) were obtained and analyzed in terms of the kinematics of articulatory organs in singing as opposed to that in speech.

In the paper some major tendencies are illustrated by the plots of back of the tongue movements in x-plane and z-plane.

Fig. 3 and fig.4 show the back of the tongue movements in x-plane (forth and back) and z-plane (up and down) in singing and reading respectively. The initial phrase of romance (sung and read by the same subject) is presented in the plots.

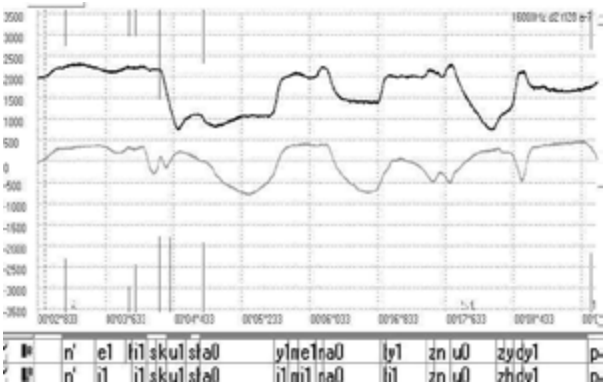


Fig. 3. Singing. The back of the tongue movements in x-plane (upper line) and z-plane (lower line).

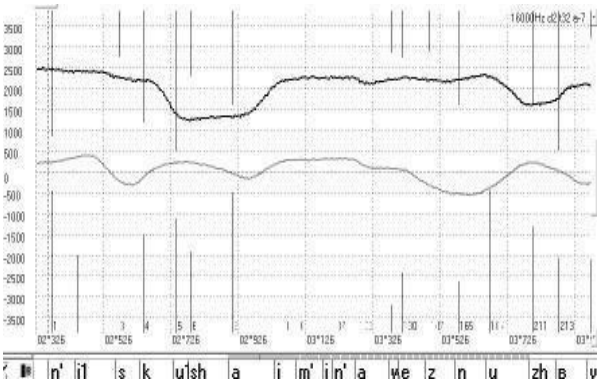


Fig. 4. Reading. The back of the tongue movements in x-plane (upper line) and z-plane (lower line).

The following set of figures (5-10) shows tongue (both dorsum and back) movements in the singing and reading of 3 cardinal vowels /a/, /i/. /u/.

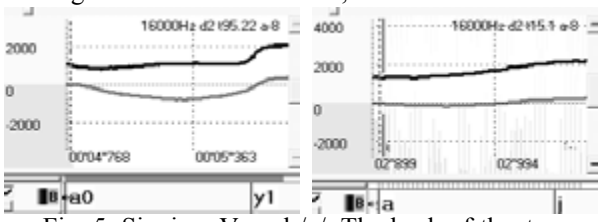


Fig. 5. Singing. Vowel /a/. The back of the tongue movements in x-plane (upper line) and z-plane (lower line).

Fig. 6. Reading. Vowel /a/. The back of the tongue movements in x-plane (upper line) and z-plane (lower line)

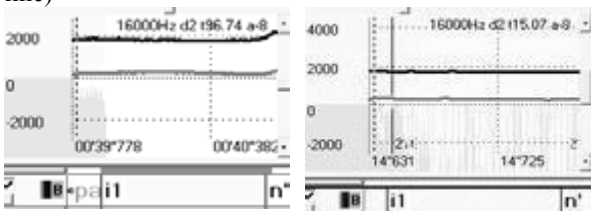


Fig. 7. Singing. Vowel /i/. The back of the tongue movements in x-plane (upper line) and z-plane (lower line).

Fig. 8. Reading. Vowel /i/. The back of the tongue movements in x-plane (upper line) and z-plane (lower line)

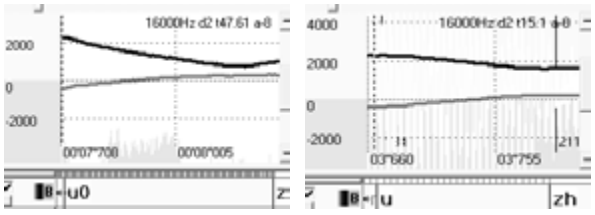


Fig. 9. Singing. Vowel /u/. The back of the tongue movements in x-plane (upper line) and z-plane (lower line)

Fig. 10. Reading. Vowel /u/. The back of the tongue movements in x-plane (upper line) and z-plane (lower line).

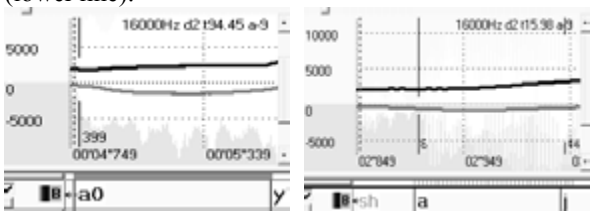


Fig. 11. Singing. Vowel /a/. The tongue dorsum movements in x-plane (upper line) and z-plane (lower line).

Fig. 12. Reading. Vowel /a/. The tongue dorsum movements in x-plane (upper line) and z-plane (lower line)

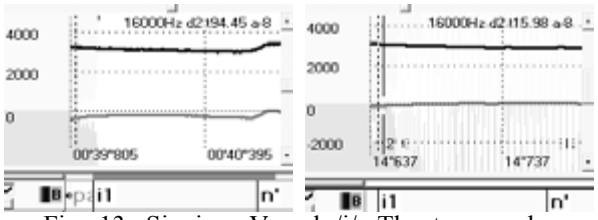


Fig. 13. Singing. Vowel /i/. The tongue dorsum movements in x-plane (upper line) and z-plane (lower line).

Fig. 14. Reading. Vowel /i/. The tongue dorsum movements in x-plane (upper line) and z-plane (lower line)

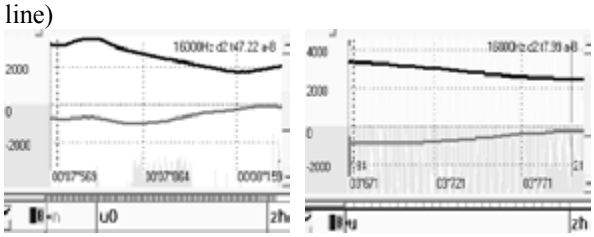


Fig. 15. Singing. Vowel /u/. The tongue dorsum movements in x-plane (upper line) and z-plane (lower line).

Fig. 16. Reading. Vowel /u/. The tongue dorsum movements in x-plane (upper line) and z-plane (lower line).

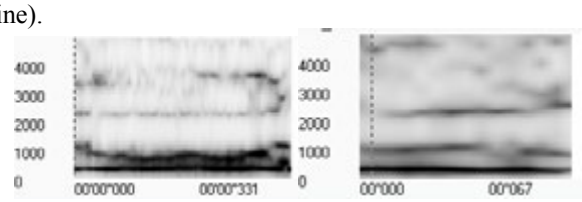


Fig. 17. Singing. Vowel /u/. Spectogram.

Fig. 18. Reading. Vowel /u/. Spectogram.

One can see that the trajectories of the forth and back movements of the tongue back for /a/ are quite similar in singing and reading (fig. 5-6). However, the trajectory of up and down movement shows that the tongue back is lower in singing.

For a singing /i/ the back of the tongue is slightly more backward and upper than in reading (fig. 7-8).

The positions of the tongue back contrast sharply for /u/. It has more backward and upper position in singing (fig. 9-10).

The tongue dorsum movements for /a/ and /i/ in singing and reading also differ (fig. 11-14). Yet, the most evident difference is again noticeable for /u/ (fig. 15-16). The tongue dorsum has a more backward and upper position in singing. These articulatory differences cause the change in the acoustic characteristics of the sound which one can observe in fig.17-18.

IV. Discussion

The comparing kinematics data in singing and reading shows that in general the amplitude and

patterns of articulatory movements in singing differs considerably from those in reading. The articulators perform more frequent and abrupt movements. The amplitude of the forth and up movements is larger than that in reading. The positions of the tongue (dorsum and back) tend to be more backward and upper in singing.

The comparison of the kinematic data for the /a/, /i/, /u/ vowels shows that there were no considerable differences in the articulatory movements for /a/, while the plots for /i/ and /u/ demonstrate different amplitude magnitudes and the trajectories of movements in singing as compared to reading. This is consistent with the results we obtained earlier using acoustic and perception analysis. It was shown that the sung vowels /i/ and /u/ tend to have lower intelligibility, while /a/ keeps its phonetic quality [1, 16].

V. Conclusion

The method of electromagnetic articulography employed in the study allows obtaining exact data on articulatory characteristics in singing as opposed to those in reading. The analysis of difference in kinematic characteristics provides reasons of acoustic distortion in singing vowel quality.

REFERENCES

[1] K. Evgrafova, V. Evdokimova, “Perception Of Russian Vowels In Singing”, Proceedings Of The Fifth International Conference Baltic Hlt2012, 2012. — Vol. 247, — № 42-49.

[2] G. Fant, “Acoustic Theory of Speech Production”, Mouton: The Hague, 1960 (second edition, 1970).

[3] V. P. Morozov, “Intelligibility in singing as a function of fundamental voice pitch”. Soviet Physics Acoustics, 1965. 10, pp 279-283.

[4] D.S. Lundy , S. Roy , R.R. Casiano , J.W. Xue , J. Evans, “Acoustic Analysis of the Singing and Speaking Voice in Singing Students”. Journal of Voice, 2000 Dec;14(4), pp 490-493.

[5] A.P. Mendes , H.B. Rothman, Sapienza Ch., W.S.Jr. Brown, “Effects of Vocal Training on the Acoustic Parameters of the Singing Voice”. Journal of Voice, 2003 Vol. 17, No. 4, pp 529–543.

[6] H. K. Schutte, D. G. Miller , J. G. Svec, “Measurement of formant frequencies and bandwidths in singing”. Journal of Voice, 1995, Vol. 9, No. 3, pp 290-296.

[7] J. Sundberg, “Articulatory differences between spoken and sung vowels in singers.” Speech Transmission Laboratory / Quarterly Progress Status Report, 1969. Vol. 1, pp.31-42.

[8] J. Sundberg, “Research on the singing voice in retrospective”. TMH-QPSR, 2003, Vol. 45 nr: 1, pp 11-22.

[9] J. Sundberg, “The acoustics of the singing voice”. Scientific American, March 1977, pp 82-91.

[10] J. Sundberg, “The science of the singing voice”. DeKalb, Illinois: Northern University Press, 1987.

[11] I. R. Titze, “Speaking vowels versus singing vowels” Journal of Singing 1995, Sept/Oct, p 487.

[12] Ch. Watts, K. Barnes-Burroughs, M. Andrianopoulos, M. Carr, “Potential Factors Related to Untrained Singing Talent: A Survey of Singing Pedagogues” Journal of Voice, 2003. 17(3), pp 298–307.

[13] H. Hollien, A. P. Mendes-Schwartz, K. Nielsen, “Perceptual confusions of high-pitched sung vowels”. Journal of Voice 2000; 14, pp 287–298.

[14] H. D Nelson, W. R. Tiffany, “The intelligibility of song”. The NATS Bulletin. 1968; 25, pp 22–28. [15] N. Scotto Di Carlo, A. Germain, “Perceptual study of the influence of pitch on the intelligibility of sung vowels”. Phonetica. 1985; 42, pp 188–192.

[16] O.N. Glotova, K.V. Evgrafova, V.V. Evdokimova. “The perception of sung vowels by the native speakers of Russian”. Homo loquens: the studies of XXI century. Ivanovo, 2012, pp 59-67 (in Russian).

[17] J. Howie, P. Delattre, “An experimental study of the effect of pitch on the intelligibility of vowels.” The National Association of Teachers of Singing Bulletin, 1962, 18:4, pp 6-9.

[18] J. G. Westerman, R. C Scherer, “Vowel Intelligibility in Classical Singing. Journal of Voice,” Vol. 20, No. 2, pp 198–210.

EVALUATING A MARKERLESS METHOD FOR STUDYING ARTICULATORY MOVEMENTS: APPLICATION TO A SYLLABLE REPETITION TASK

A. Bandini[1,2], S. Ouni[3], S. Orlandi[1], 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
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France

Abstract: **The analysis of the articulatory movements allows investigating the kinematic characteristics of some speech disorders. However, the methodologies most used until now, as electromagnetic articulography and optoelectronic systems, are expensive and intrusive which limit their use to specialized laboratories. In this work, we use a completely markerless and low-cost technique to study lip movements during a syllable repetition task. By means of a Kinect-like and an existing face tracking algorithm, we are able to track the movements of the lower lip, testing the performances against a reference method (marker-based optoelectronic system). Good results were obtained in terms of RMSE for the tracking of the lower lip during the repetitions. Some kinematic measures, as opening and closing velocities and accelerations, were also computed. Despite the limitations in terms of image resolution, these results are very promising in the optic of developing a new markerless system for studying speech articulation.**

Keywords : **speech articulation, markerless, Kinect, contactless, accuracy evaluation**

I. INTRODUCTION

Kinematic analysis of the articulatory movements (i.e., the movements of tongue, lips and jaw) allows investigating the characteristics of some speech disorders, like hypokinetic dysarthria. Walsh *et al.*, 2012 [1] studied jaw and lower lip movements in patients with Parkinson's disease (PD) using an optoelectronic system. They demonstrated that these patients exhibit reduced ranges of movements and velocities of jaw and lips during the pronunciation of plosive consonants. Yunusova *et al*., 2008 [2] studied the articulatory movements in patients with hypokinetic dysarthria due to PD and amyotrophic lateral sclerosis (ALS) by means of the X-ray microbeam technique, tracking the position of several markers located on tongue, lips and jaw. In that work, they showed that tongue movements in PD and ALS patients could be more discriminative in the comparison with healthy subjects, although there are also alterations in lips and jaw movements.

Wong *et al.*, 2011 [3] has also investigated articulatory movements in patients with speech disorders, where the tongue kinematics is studied by means of electromagnetic articulography (EMA), in dysarthric and non-dysarthric PD patients. They demonstrated that both categories exhibit different patterns of tongue movements with respect to healthy subjects.

These researches show clearly that techniques for studying movements are useful to describe the kinematic characteristic of the articulatory organs in dysarthric patients. However, the methodologies most used until now (EMA, optoelectronic systems, X-ray techniques, etc.), which are actually very accurate, have the big disadvantage of being expensive which limit their use to specialized laboratories [4]. Moreover, some of these techniques need long and tedious preparation protocols, resulting in a discomfort for patients. Thus, the use of these methodologies for studying speech articulation is limited to the research field. In order to broaden the kinematic studies of speech articulation, (e.g., for speech therapy purposes, or to track the disease progression), the use of a low-cost and fully contactless system would be desirable.

In the last five years the spreading of 3D video sensors (like Microsoft Kinect), has revolutionized the world of videogames and not only, providing new possibilities to study body movements without any sensor attached to the subject. These devices, unlike a normal camera, provide a 3D information about the observed scene. Even for speech therapy purposes, some applications with the Kinect sensor has been proposed, in order to study and automatically identify the therapeutic exercises that involve facial movements [5]. To our knowledge, no existing work has tested the accuracy of a fully markerless technique to study speech articulation. For these reasons, our aim is to test the performance of a system composed by a 3D depth sensor and a face tracking algorithm in order to track

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

lip movements during speech. In this study the accuracy is verified against an established optoelectronic method.

II. Methods

The markerless system proposed in this study is composed of a 3D structured light sensor (Primesense Carmine 1.09) and an existing face tracking algorithm [6], in order to study lips movements in the 3D space without any sensor attached to the subject's skin.

Two healthy volunteers (an Italian native speaker and a French one) were recruited for the experiment. The speech task consists in the repetition of the syllable /pa/ for at least 30 times with a single breath. The acquisitions were performed in a room with reduced environmental noise. Each subject had to repeat the syllables avoiding large head movements. The subjects' face was kept under constant and uniform illumination during the whole acquisition time.

Markerless system: during the experiments the subjects' faces were acquired by means of the depth sensor Primesense Carmine 1.09. This device was chosen for its ability to work at short distances (0.4-1.5 m), thus appropriate for face movements. As classical structured-light sensors, it provides two video streams: the color video (like a normal webcam) and the depth stream, where the pixels of each frame code the distance of a point in the scene from the camera plane. The image resolution of both streams was set at 320 x 240 pixels. Both videos were acquired synchronously at 30 frames per second, and stored as avi files by means of the OpenNI (ver. 2.2) and OpenCV (ver. 2.4.9) libraries.

The device was located in front of the subject's face (at the height of the mouth) at a distance around 0.7-0.8 m from the lips, according to the specifications provided by the manufacturer.

For the automatic identification of the facial features, the tracking algorithm *Intraface* was used. This algorithm fits to the video frames a face model composed of 49 points, on the basis of texture descriptors like SIFT (Scale-Invariant Feature Transform) [6, 7]. This algorithm was chosen for its robustness against illumination changes, for its ability to describe asymmetrical face movements (very important in the context of speech therapy applications) and for its efficiency [6]. In particular, lips are modeled as a set of 18 points: 12 on the outer border and 6 on the inner border. In the case of our study, only 7 points on the outer border were considered for the analysis (Fig. 1) to compare the performance of the system against the marker-based method.

Since the coordinates of the points that were computed with the *Intraface* tracker are on the image plane, a further step to extract the 3D locations of the points of interest is required, as this algorithm works only on the color image. Starting from the coordinates on the image plane and using the depth values (Z) retrieved from the depth image, it was possible to calculate the 3D coordinates in mm. In fact, before each acquisition the color and depth frames were aligned and synchronized, then, we just sampled the depth image in the same pixel coordinates of the model points provided by the tracker.

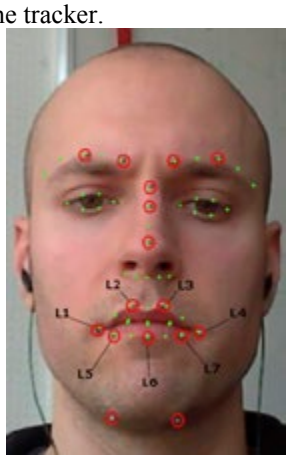


Fig. 1: *Intraface* tracker model points (green dots) and optical markers locations (red circles). The markers were located in the same position of some model points, in order to estimate the 3D rigid transformation to register the two sets of points.

According to the scheme in Fig. 2, we calculated the X and Y coordinates with the following formulas [8]:

$$X = Z\,\frac{(x-c_x)}{f} \qquad with\ f = \frac{W}{2}\left[\tan\left(\frac{FOV_h}{2}\right)\right]^{-1} \quad (1)$$

$$Y = Z\,\frac{(y-c_y)}{f} \qquad with\ f = \frac{H}{2}\left[\tan\left(\frac{FOV_v}{2}\right)\right]^{-1} \quad (2)$$

Where x and y are the coordinates on the image plane (in pixels) of the 3D point $[X\ Y\ Z]^{\mathrm{T}}$, (c_x, c_y) are the coordinates (in pixels) of the principal point (i.e. the point where the optical axis intersects the image plane) of the color camera, f is the focal length (in pixels) of the camera, W and H are the dimensions of the image in pixels (width and height, respectively), FOV_h and FOV_v are the horizontal and vertical field of view of the camera (58° and 45° respectively).

Marker-based system: To compare the performance of the aforementioned markerless method, we used an optoelectronic system (Vicon Motion Systems Ltd., UK) as a reference. This system was composed by four cameras (MX3+ model) with special optics for near range applications. Sixteen reflective markers of 3mm diameter were glued on the faces of the subjects. This size is suitable to study facial movements without interfering with the face tracker.

Before each acquisition, the markers were accurately located in some precise facial points defined by the

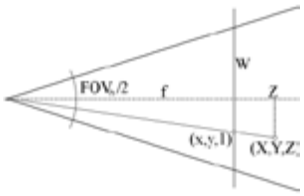


Fig. 2: Pinhole camera model. This model was used to retrieve the 3D coordinates of the face points (estimated with the markerless system), starting from the image coordinates plus the depth information Z (according to equations 1 and 2).

Intraface model: two for each eyebrow, three on the nose, seven on the outer border of the lips (one for each corner – L1 and L4, two on the upper lip – L2 and L3 – and three on the lower lip – L5-L7) and two on the chin (Fig. 1). The 3D trajectories of these markers were acquired synchronously using the markerless system at 100 Hz and reconstructed using the Vicon Nexus software.

Data Processing: To compare the trajectories of the points of interest extracted with the markerless system with those of reference, the two sets of points must be aligned in the space, since the two reference frames are different. To do this, since we paid a lot of attention to locate the markers in the same position as some *Intraface* points, the 3D rigid transformation that allows mapping the markerless points in the marker-based reference frame can be estimated. Using pairs of corresponding points provided by the two systems, the rotation matrix R and the translation vector T were estimated through a least squares solution and this transformation was applied to each point extracted from the markerless system.

Once the two sets of points were aligned, the trajectories extracted with the markerless system were resampled at 100 Hz using a spline interpolation technique. In this way, the comparison with the reference trajectories was possible by means of the calculation of the root-mean-square error (RMSE) in mm, according to the following formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (3)$$

Where N is the number of samples of the trajectories during a single repetition, y_i is the *i-th* sample of the marker-based trajectory and $\hat{y}_i$ is the corresponding sample extracted from the markerless trajectory.

Afterwards, for each syllable repetition the following kinematic parameters (for both systems) were computed: the maximum velocity (V_{open}) and acceleration (A_{open}) during the opening phase, the maximum velocity (V_{close}) and acceleration (A_{close}) during the closing phase. These parameters were calculated differentiating in time the trajectory on the vertical axis of the central point of the lower lip (point L6 in Fig. 1). V_{open} was calculated as the minimum speed value during the first half of the repetition, while V_{close} was identified as the maximum speed value from the time instant of V_{close} up to the end of the utterance (Fig. 3). The same criteria were adopted to extract A_{open} and A_{close} from the acceleration values of the same lip point (Fig. 3). Moreover, for each syllable repetition the Pearson's correlation coefficient between trajectories, velocities and accelerations extracted with both systems was computed. Correlation values close to 1 indicate that the trends of displacement, speed and acceleration calculated with the proposed method are very similar to the ground truth, as shown in Fig. 3.

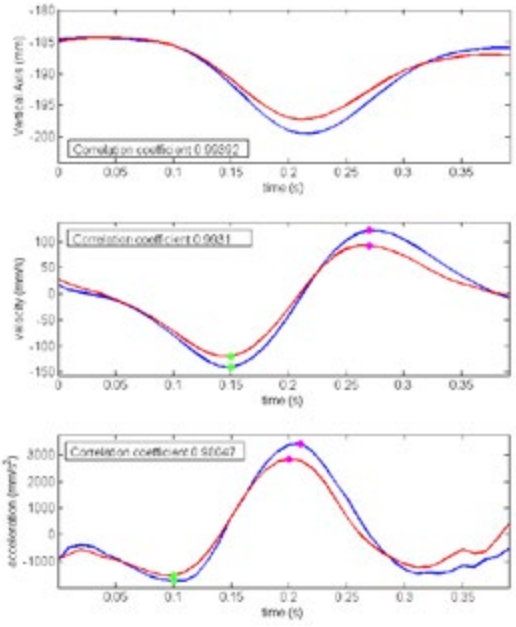


Fig. 3: vertical trajectory of the central point of the lower lip (upper plot) during the repetition of the syllable /pa/; speed (central plot) and acceleration (lower plot) on the vertical axis. The blue lines are relative to the reference method (marker-based), while the red lines are estimated with the markerless technique. The green points indicate the maximum velocities and accelerations during the opening phase, while the magenta points are the maximum velocities and accelerations during the closing phase.

III. Results

The analysis was conducted on a total of 80 utterances. RMSE values of the central point of the lower lip (point L6, Fig. 1) were around 2 mm on the three axes, respectively (1.89 ± 0.82) mm on the lateral axis, (1.61 ± 0.45) mm on the frontal axis and (2.16 ± 0.70) mm on the vertical axis. The kinematic parameters (mean values and standard deviations) for speed and acceleration during opening and closing phases relative to the same point were reported in Tab. 1.

The correlation coefficient for the trajectory on the vertical axis of the point L6 was (0.96 ± 0.03), while those for speed and acceleration were respectively (0.95 ± 0.05) and (0.88 ± 0.10).

Tab.1: Mean values and standard deviations of the kinematic parameters during the opening and closing phases of the syllable repetition

	Marker-based	Markerless
V_{open} (mm/s)	-114.37 ± 30.55	-96.39 ± 23.48
V_{close} (mm/s)	100.01 ± 45.82	79.77 ± 26.97
A_{open} (mm/s^2)	-1689.63 ± 559.86	-1759.10 ± 665.73
A_{close} (mm/s^2)	2619.02 ± 1068.82	2141.91 ± 910.86

IV. Discussion

Good results were obtained in term of accuracy with RMSE for the point L6 around 2 mm on the three axes. The kinematic parameters reported in Tab. 1 show a tendency to underestimate the module of the maximum and the minimum speed values (closing and opening phases) with differences around 20 mm/s. An underestimation is visible also for the closing acceleration, while during the opening phase the two estimates seem to be closer.

Although the results on kinematic parameters seem to be inconsistent, from the plot in Fig. 3 and from the correlation values between the two systems, it is possible to observe that the trajectories, the velocities and the accelerations extracted with the markerless technique were very similar when compared with the reference. This suggests that a bias is present in the estimation of the kinematic parameters.

This bias might be due to the distance from the face at which the device was located (about 0.8 m), or to the different framerate of the systems (30 Hz for the depth sensor, 100 Hz for the marker-based method). This distance was a trade-off between the need to move the sensor as close as possible to the subject's face and its characteristic (range of work: 0.4-1.5 m), without interfering with the field of view of the Vicon cameras. The distance, in conjunction with the low image resolution (320 x 240 pixels) probably explain these differences. However, further experiments with structured light sensors should consider an experimental design with higher frame resolutions (at least 640 x 480 pixels) and smaller distances from the subject's face (i.e., 0.5-0.6 m, according to the specification provided by the manufacturer).

V. Conclusion

In this work, we have introduced a fully contactless and low-cost method to track the articulatory movements, in particular those relative to the lower lip during a syllable repetition task. We demonstrated that good accuracies could be reached in terms of RMSE with respect to a marker-based reference methods. These results are very promising in the optic of developing new systems to study speech articulation, that could be implemented also in domestic environments. This would allow enlarging the number of patients who undergoes to speech therapy, in particular elderly people who suffer from hypokinetic dysarthria due to Parkinson's disease.

Further developments will be oriented to test the performance of this system with different configurations (image resolution, distance from the camera), as well as to use this contactless technique with PD patients (but not only), in order to check if it is possible to highlight kinematic differences due to the dysarthria with respect to healthy control subjects, as demonstrated with other more expensive techniques [1-3].

References

[1] B. Walsh, A. Smith, "Basic parameters of articulatory and acoustics in individuals with Parkinson's disease", *Movement Disorders*, vol. 27, no. 7, pp. 843-850, 2012.

[2] Y. Yunusova, G. Weismer, J.R. Westbury, M.J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls", *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, pp. 596-611, 2008.

[3] M.N. Wong, B.E. Murdoch, B. Whelan, "Lingual kinematics during rapid syllable repetition in Parkinson's disease", *International Journal of Language and Communication Disorders*, vol. 47, no. 5, pp. 578-588, 2012.

[4] M.M. Earnest, L. Max, "En route to the three.dimensional registration and analysis of speech movements: Instrumental techniques for the study of articulatory kinematics", *Contemporary Issues in Communication Science and Disorders*, vol. 30, pp. 5-25, 2003.

[5] C. Lanz, B.S. Olgay, J. Denzler, H-M. Gross, "Automated classification of therapeutic face exercises using the kinect", *Proceedings of the 8th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2013), Barcelona, Spain,* pp. 556-565, 2013.

[6] X. Xiong, F. De la Torre, "Supervised descent method and its applications to face alignment" in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, June 23-28, Portland-OR, USA,* pp. 532-539, 2013.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[8] R. Szeliski, "Computer vision: algorithms and applications", Springer London, 2010.

TEST BENCH FOR HUMAN EXCISED LARYNX STUDIES

T. Legou[1], A. Lagier [1,2,3], F. Silva[4], N. Henrich[5], P. Champsaur[3], A. Giovanni[1,2]

[1] Laboratoire Parole et Langage UMR 7309 CNRS-Université Aix-Marseille, Aix-en-Provence, France
[2] Service d'ORL, CHU Timone, AP-HM, Marseille, France
[3] Laboratoire d'Anatomie, Aix-Marseille Université, Marseille, France
4 Laboratoire d'Acoustique de Marseille, UPR 7051 CNRS, France
5 Gipsa Lab, UMR 5216 CNRS - Grenoble INP - Université Joseph Fourier - Université Stendhal, Grenoble, France
thierry.legou@lpl-aix.fr, aude.lagier@ap-hm.fr, silva@lma.cnrs-mrs.fr,
pierre.champsaur@ap-hm.fr, antoine.giovanni@ap-hm.fr

Abstract: **Many questions about the physiology of vocal folds vibration, mechanical properties of the vocal folds, impacts of pathologies on the vocal mechanism are investigated in different ways: physical modeling, experiments in vivo on animals or in human volunteers or patients, and also on excised larynges (animals or human/autopsic). Many parameters are involved in phonation and quality of the voice such as the source (flow, pressure), the action of laryngeal muscles and glottal configuration, temperature and humidity.**
Objectives: In order to independently control these parameters with a biologic model as similar as possible to *in vivo* human conditions, and also to reach a level of reproducibility that permits measurements comparisons, we have developed a dedicated test bench for human excised larynges. Several classic measurements are systematically recorded such as airflow, subglottal pressure, glottograph signal, sound (F0 and intensity), but also contact pressure between vocal folds. A high speed camera is also synchronized for glottis area and motion analysis. The glottal configuration depends on each study, and has been automated to achieve dynamical control, reproducing the action of laryngeal muscles. The humidified airflow is applied to the larynx via an intubation tube introduced in the three last upper rings of the trachea.
Keywords : **Larynx, Test Bench, control, measure.**

I. INTRODUCTION

Phonation is a complex activity which implies aerodynamic and muscular controls. Humans are able to produce a wide variability of sounds from whispering to shouted voice. The use of excised larynx and on a dedicated test bench permits to decouple these control parameters and therefore understand their individual role. It therefore permits to evaluate the effect of a parameter on phonation for a given configuration. The capacity of measuring and controlling parameters also offers the possibility to reach a sufficient level of reproducibility to make possible measurements comparisons for a given larynx but also to cope with the intrinsic differences observed between larynges.

II. LARYNX PREPARATION

Larynges are dissected, keeping them intact from the thyroid to the third tracheal ring. The vocal folds adduction was constant, using concomitant arytenoid adduction and membranous vocal folds medialization with Montgomery implants. Subglottal pressure is measured 1 cm under the glottis by a tracheal puncture.

III. TEST BENCH STRUCTURE

The test bench can be divided in six main principal functions (PF). The larynx holder & positioning system (PF1), the airflow apparatus (PF2), the pulling system (PF3), the vocal folds pressure contact module (PF4), the high speed camera (PF5) and the measurement/acquisition unit (PF6).

A. The positioning system. (PF1)

On the bench the larynx lays horizontally, the posterior edges of the thyroid cartilage are in contact with an adaptable holder. Adjusting the space between the two parts of the holder is mandatory in oder to identically set the larynx on the bench independently of its size. As the larynx is secured by the strap of the glottograph well tied on its older, the cricoid cartilage is kept free to permit the crico-thyroid tilt (see Fig. 1).

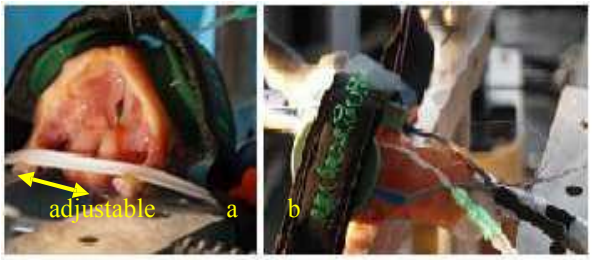


Fig. 1

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

B. The Airflow apparatus(PF2)

The airflow is generated by a turbine (Werie Rietschle) with a capacity above 10dm^3/s. Before being injected into the larynx, the air goes through a heater/ humidifier (Drager, Aquador). Then, the humidified airflow is applied to the larynx via an intubation tube introduced in the three last upper rings of the trachea. The flow is adjustable manually by a ¼ turn valve. To set the desired flow, the operator has a permanent real time monitoring of the airflow actually applied to the larynx.

C. The pulling system(PF3)

In order reproduce the cricothyroid tensor muscle, we have developed a pulling system (Fig. 2) based on a micro linear actuator with a 0.2mm accuracy (Firgelli L12-100) controlled by an Arduino board, an open hardware/open software microcontroller platform. The microcontroller can be controlled on the fly, and its motions are recorded by the measurements unit.

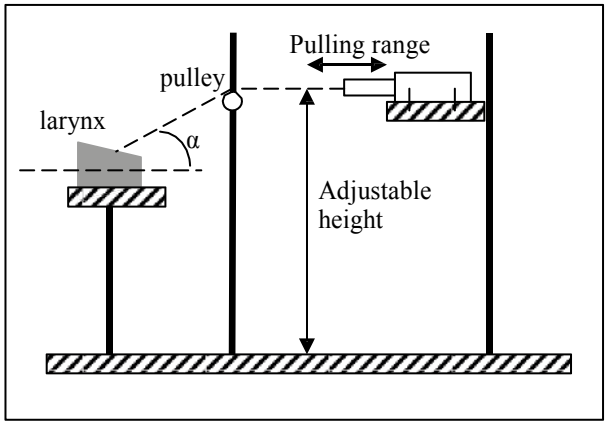


Fig. 2

As the thyroid is secured on the bench, the linear actuator pulls on the cricoids (free to move) via a nylon wire to modulate the cricothyroid tilt. The pulling range (100 mm) of the linear actuator permits to study the impact on phonation of the cricothyroid tilt over is full natural variation range. A dynamic control is possible up to 12 mm/s.

The distance between the larynx horizontal axis, the actuator axis and the height of its associated pulley is adjustable giving the possibility to change the pulling angle (α).

D. Pressure/contact unit(PF4)

Vocal folds synchronization is a complex process. Many simulations have taken into account aerodynamic effects and tissues properties that could explain vocal folds motions. The pressure of contact between folds has also been studied to understand its role in the folds synchronization and also to evaluate its effect on phonotrauma in vocal abuse. To study the pressure of contact, several studies used tiny piezoelectric transducers type 060S from Precision Measurement Company (Michigan, USA). Jack J. Jiang and al used it in exvivo canine larynges experiment [1], and other studies used it in human invivo measurements [2] and [3].

We selected the available type 105S from the same manufacturer, with stainless-steel diaphragms (visible between vocal folds in Fig. 1.a). The pressure range covered by the sensor is 0-60 psi. The sensor is a part of wheastone bridge, from which the output voltage is amplified by a differential amplifier. The high gain (x5000) of two cascaded amplifiers permit the readout of very weak vocal folds pressure signals.

E. High Speed Camera(PF5)

To study vocal folds motions, and to measure the glottis area, we use a high speed camera (Fastcam SA-3, Photron) equipped with a micro-lens (AF Micro-Nikkor 105 mm f/2.8D, Nikon). Films are recorded at 5000fps and the high sensitivity of the camera sensor gives very high contrasted images processed by a homemade Matlab script to measure the open glottis area. To synchronize films and signals recorded by the measurements/acquisition unit, the camera sends a trigger signal at the beginning of each video.

F. The measurements/acquisition unit(PF6)

On the present version of the test bench, all data are acquired using a specialized speech aerodynamic workstation (EVA2) [4]. This unit usually records simultaneously audio, airflow, air pressure, glottograph signal. Table 1 lists the possible measurements variation range for the recorded parameters.

Table 1

Parameter	Range
Airflow	0-2 dm^3/s
	0-10 dm^3/s
Subglottal pressure	0-20 hPa
	0-40 hPa
	0-100 hPa
	0-200 hPa

In addition to these classic parameters, three exogenous signals are also recorded simultaneously. One analog signal which is proportional to the folds pressure. And two digital signals for post synchronization. The first one is sent by the linear actuator that pulls the cricoid and the second one sent by the high speed camera at the beginning of each film sequence. Fig.3 is the sketch of the acquisition system in its environment.

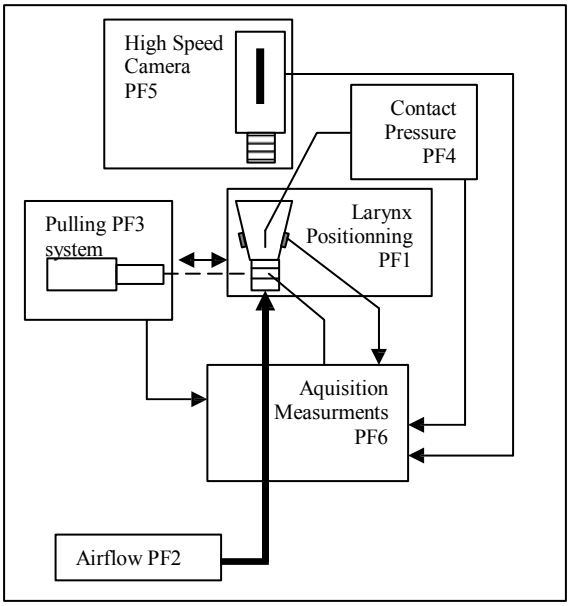


Fig. 3

The sound is recorded by a C420 AKG microphone digitized with a 25000Hz sampling frequency over 16bits. The microphone is set at 15 cm from the larynx, with an angle of 30° from the larynx sagittal axis. Apart from the audio signal, all analog signals (glottograph, sublglottal pressure, airflow, pressure of vocal folds contacts) as well as digital ones (puller system and camera trigger signals) are digitized with a sampling frequency of 25kHz over 16 bits.

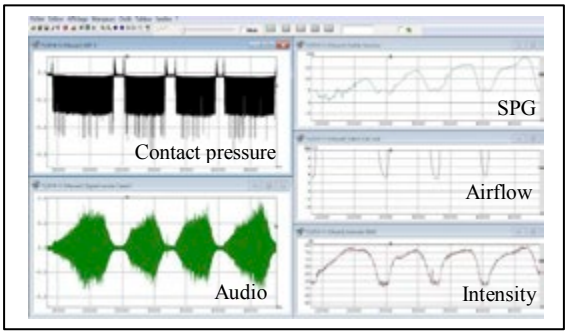


Fig.4

All recorded signals can be displayed simultaneously with a common moving time marker using Phonedit software (see Fig. 4). For quantitative analysis dedicated Matlab scripts are used.

IV. PERSPECTIVES

To remove any human action during tests, the airflow control will be soonly automated.

REFERENCES

[1] Jack J. Jiang, Anand G. Shah, Markus M. Hess, Katherine Verdolini, Franklin M. Banzali, Jr and David G. Hanson, "Vocal fold Impact Stress Analysis" *Journal of Voice,* vol. 15,No. 1, pp. 4-14, 2001.

[2] M. Hess, Katherine Verdolini, Wolfgang Bierhals, Ulrich Mansmann, and Manfred Gross, "Endolaryngeal Contact Pressures", *Journal of Voice,* vol. 12,No. 1, pp. 50-67, 1998.

[3] Katherine Verdolini, Markus M. Hess, Ingo R. Titze, Wolfgang Bierhals and Manfred Gross, "Investigations of Vocal Fold Impact Stress in Human Subjects**",** *Journal of Voice,* vol. 13,No. 2, pp. 184-202, 1999.

[4]A.Ghio, G.Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C. De Looze, D. Robert, F. Viallet, A. Giovanni, "How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?", Speech Communication, 54(5), 664-679, 2012.

TOWARDS SIBILANT PHYSICAL SPEECH SCREENING USING ORAL TRACT VOLUME RECONSTRUCTION

A. Van Hirtum[1], K. Nozaki[2], Y. Fujiso[1]
[1]Gipsa-lab, UMR CNRS 5216, Grenoble University, France
[2]Osaka University Dental Hospital, Japan
annemie.vanhirtum@grenoble-inp.fr

***Abstract*: The current paper considers physical speech screening by supplying air to a rigid replica based on oral tract volume reconstruction of a speaker uttering phoneme /s/ (Reynolds number about 5300). Radiated acoustic pressure spectra were measured for different flow conditions upstream from the recon- structed portion: Reynolds number (2800, 5300, 8900) and turbulence intensity (<4% and >4% by varying the method of air supply). Acoustic spectra obtained with the replica were compared to spectra of phoneme /s/ uttered by the same speaker at different loudness levels. It is found that noise emitted by the replica reproduces the spectral shape (in particular for frequencies up to 8 kHz) and the order of magnitude of spectral features (spectral slopes and dynamic amplitude) of phoneme /s/. Nevertheless, spectral differences (energy discrepancy, peak frequency, negative spectral slope) between phoneme /s/ and sound generated with the replica are observed as a function of Reynolds number as well as a function of upstream turbulence intensity. Therefore, current data suggest that in order to perform sibilant physi- cal speech screening Reynolds number as well as upstream flow conditions need to be taken into account.**

***Keywords* : sibilant fricative, oral tract reconstruction, speech/noise screening**

I. INTRODUCTION

The human oral tract is due to its multiple functions and its crucial role in daily life of interest to many disciplines ranging from physiology to entertainment. A common way to evaluate some aspects related to the oral tract either for clinical, oral health care or research purposes is physical speech screening. Consequently, many research efforts focus on quantifying speech features as a function of oral tract features (e.g. geometrical) [1]. Nevertheless, some difficulties are inherently related to speech screening on human subjects such as intra- and inter speaker variability or inherent correlation of flow and geometrical quantities [2]. To avoid these problems, some researchers and developers working in relation to oral health care, make use of the fast development of volume reconstruction and rapid prototyping to study noise production using a reconstructed oral tract portion instead of a human speaker. Sibilant physical speech screening is a concrete example of direct use for speech researchers as well as for clinical researchers e.g. dentistry. Therefore, in the following, we consider sibilant fricative physical speech screening based on a partial reconstruction of the oral tract during phoneme /s/ production. The underlying mechanism of sibilant fricative sound production is generally described as noise produced due to the interaction of a turbulent jet, issued from a constriction between the tongue and the hard palate – i.e. the sibilant groove – somewhere in the vocal tract, with a downstream wall or obstacle [1]. Consequently, acoustic features of sibilant noise are influenced by both flow and geometrical parameters. Sibilant physical speech screening using a reconstructed geometry allows to study the potential impact of flow parameters on noise production independently from geometrical parameters. Recently, evidence was provided suggesting that different upstream flow conditions affect spectral features used to describe sibilant fricatives such as phoneme /s/, but comparison with spectra of phoneme /s/ uttered by a human speaker was lacking [3]. Such a comparison is needed when aiming sibilant physical speech screening using a replica. Concretely, in the current study spectral features of noise generated using a replica containing a reconstructed oral tract portion are gathered as a function of two flow parameters. Firstly, different Reynolds numbers Re are assessed to account for the effect of mean velocity on the sound outcome. Secondly, the turbulence intensity upstream from the reconstructed portion is altered by varying the method of air supply to the replica in order to change the contribution of aerodynamic noise production expressed by the Reynolds stress tensor. Spectral features ob- tained with the replica are then compared to spectral features of sibilant /s/ uttered by the subject for which the oral tract reconstruction was made. In addition, flow downstream from the replica is visualised in order to qualitatively assess the flow field.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

It is discussed to which extent a replica based on a reconstructed oral tract can be used for sibilant physical speech screening.

II. METHODS

Noise was generated at the entrance of the quasi-anechoic chamber in three different ways labelled (I), (II) and (III):

1. Phoneme /s/ (I). Phoneme /s/ was uttered following three different loudness instructions 'soft', 'medium' and 'loud'. These instructions are commonly used in sibilant fricative speech screening to evaluate the impact of volume flow rate Qoral [2]. The turbulence intensity (Tu, root mean square of streamwise velocity) upstream from the constricted vocal tract portion is unknown. Note that the oral tract geometry is not controlled.

2. Flow facility (with settling chamber) + replica (II). Air was supplied to the replica (Fig. 1) so that volume flow rate $Qcomp$ is controlled [3]. The replica (containing a reconstructed portion corresponding to sibilant /s/ utterance [3] as depicted in Fig. 1) is mounted to an upstream settling chamber (0.4m×0.4m×0.5m) tapered with acoustic foam to avoid acoustic resonances due to the experimental setup upstream from the replica. The turbulence intensity upstream from the reconstructed vocal tract portion (Tu) is smaller than 4% [3] for all assessed volume flow rates Qcomp.

3. Human blowing + replica (III). Air was supplied to the replica (Fig. 1) by human blowing following three different effort instructions: 'soft', 'medium' and 'loud'. The turbulence intensity upstream from the reconstructed vocal tract portion (Tu) is greater than 4% for all assessed volume flow rates Qblown [3].

The qualitative estimation of the volume flow rate Qoral (I) and Qblown (III) was obtained using the volume flow meter (TSI 4000 series) in combination with the instruction. The upstream turbulence intensity was measured at the centre of the exit of the upstream circular duct using hot-film anemometry (TSI1201-20 and IFA 300). Reynolds number Re upstream from the reconstructed portion (II) is estimated using the exit diameter of the upstream circular duct of the vocal tract replica. A qualitative estimation for the Reynolds numbers associated with the different loudness conditions ('soft', 'medium' and 'loud') of noise generation systems (I) and (III) was obtained using again diameter 8 mm in combination with the measured volume flow rate Qoral (I) and Qblown (III). It is seen that the resulting Reynolds numbers (2800, 5300 and 8900) are the same for all noise generation systems so that they can be used as quantitative labels. The same labels are used to indicate the imposed condition of either loudness (I) – flow (II) or blowing effort (III) in the figure legends further on.

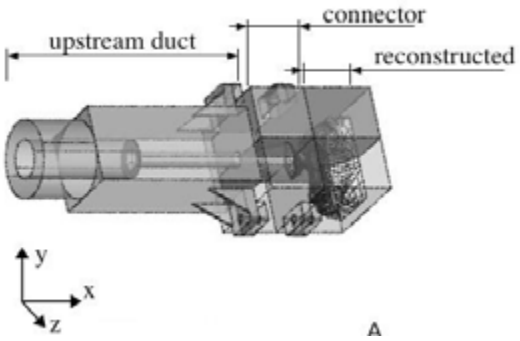


Figure 1: Replica with reconstructed portion.

Acoustic measurements were done using pressure-field microphone located in a quasi-anechoic chamber. The sound pressure spectra were parameterized by considering dynamic amplitude Ad , spectral peak frequency fm and linear regression slopes S1 ($fmin < f < fm$, with fmin the frequency associated with minimum spectral amplitude, positive slope) and S2 ($fm < f < 20$ kHz, negative slope) in accordance with fricative spectral characterisation [2]. Dynamic amplitude Ad is commonly reported to depend on volume flow rate (loudness level), peak frequency fm relates to the geometry (front-cavity resonance) and spectral slopes depend on volume flow rate (loudness level) as well as geometry (front-cavity resonance and hence peak frequency). In addition, recently [3] evidence is provided that varying upstream flow conditions might affect negative spectral slope S2 and dynamic amplitude Ad as well, so that the relationship between spectral features and Reynolds number might be altered by varying upstream flow conditions such as turbulence intensity. The difference of normalised acoustic energy is quantified (as a function of the octave band with centre frequency fc) with respect to noise generation system II (flow facility + replica) since this is the most reproducible generation system (volume flow rate and geometry are controlled/known).

In addition the flow field in the vicinity of the replica exit was visualized (300fps).

III. ESULTS

In general, generated sounds are perceived by a human listener as noisy as expected for a natural human phoneme /s/ utterance. A more quantitative analysis is obtained by analysing acoustic spectra (Fig. 2). All acoustic spectra, i.e. for all noise generation systems (I, II and III) and Reynolds numbers (2800, 5300 and 8900), are dominated by frequencies greater than 3.5 kHz, which corresponds to the order of magnitude of the first cut-on frequency of higher order propagation modes. Spectra observed for Re~5300 match well for all three noise generation systems (I, II and III) up to 8 kHz. This is reassuring since Re~5300 corresponds to the condition for which the oral cavity reconstruction was obtained. Beyond 8 kHz spectral differences occur, which are more important when the geometry is not controlled (I versus II/III: different shape and amplitude) than when the method of air supply is varied for a same geometry (II versus III: overall similar shape and different amplitude). Increased amplitude in the high frequency noise portion might be due to the increase in upstream turbulence intensity Tu causing an increase in aerodynamic noise (Reynolds tensor) when comparing air supplied using a flow facility (Tu<4%) and air supplied by human blowing (Tu>4%) (II versus III). Comparison of spectra for noise generation systems II and III for Re~8900 reveals the same tendencies as observed for Re~5300 (overall similar shape and different amplitude) whereas spectra for phoneme /s/ (I) differ in shape as well as amplitude from those obtained with the replica (II and III). Moreover, amplitude and shape of phoneme /s/ spectra (I) for Re~8900 and Re~5300 differ in general whereas the impact of Reynolds number on spectra obtained with the replica is limited to amplitude and the spectral shape is less affected.

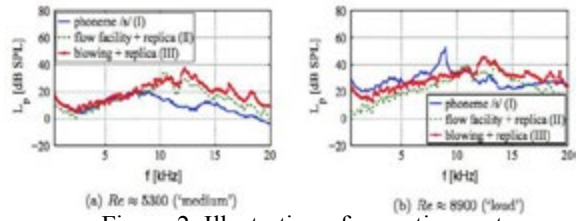


Figure 2: Illustration of acoustic spectra

The given qualitative spectral characterisation is confirmed when considering the normalised acoustic energy discrepancy (Fig. 3). Spectral features are quantified (Fig. 4): dynamic amplitude Ad , spectral peak frequency f m and linear regression slopes S1 and S2. For all noise generation systems (I, II and III) and Reynolds numbers (2800, 5300 and 8900) spectral parameters yield values within the range reported for sustained /s/ phonemes. Dynamic amplitude Ad increases with Reynolds number Re for all noise generation systems (I, II and III). Nevertheless, the increase is more important for phoneme /s/ (<110% for I) than when the replica with constant geometry is used (<10% for II and III). This is due to the sharp spectral peak observed for phoneme /s/.

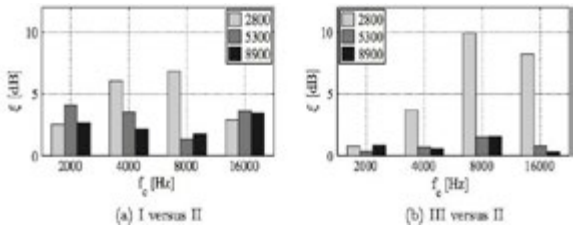


Figure 3: Difference in acoustic energy as a function of octave band with centre frequency fc and Reynolds number Re – 2800 ('soft'), 5300 ('medium') and 8900 ('loud') – normalised by the total energy with noise generation system II taken as a reference.

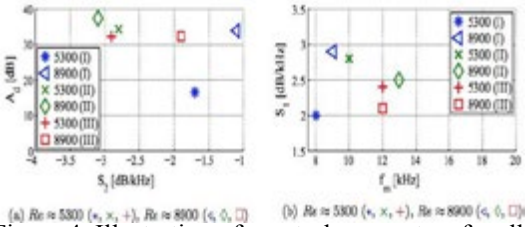


Figure 4: Illustration of spectral parameters for all noise generation systems.

For the highest assessed Reynolds number Re~8900, the order of magnitude of the dynamic amplitude (Ad +/-2 dB) is the same for all noise generation systems (I, II and III). For phoneme /s/ (I), spectral slopes increase with Reynolds number (>30% for S1 and > 55% for S2) as observed for sustained phonemes /s/. This increase of spectral slopes with Reynolds numbers is not observed when air is supplied to a fixed geometry (II and III). Spectral peaks fm for the replica (II and III) occur in the range 10 kHz up to 13 kHz, which is of the same order of magnitude as the constriction and front-cavity resonances of the reconstructed portion of the replica. In the case of phoneme /s/, the peak frequency (fm < 10 kHz) is lower than observed with the replica indicating a change in geometry compared to the reconstructed oral tract geometry. Moreover, the sharp spectral peak (about 9 kHz) observed for phoneme /s/ at Re~8900 suggests that whistling (Helmholtz resonance) might interfere with fricative sibilant production. For phoneme /s/ (I), spectral slopes increase with Reynolds number (>30% for S1 and > 55% for S2) as observed for sustained phonemes /s/. This increase of spectral slopes with Reynolds number is not always observed when air is supplied to a fixed geometry (II and III). The current comparison of spectral features for phoneme /s/ and for noise

generated with a replica suggests that the relationship between loudness level and vocal tract geometry is at least a triple relationship between Reynolds number, turbulence intensity and geometry needs to be considered. Since turbulence intensity and more in general upstream flow conditions is currently not measured for human speakers, physical sibilant speech screening using replicas with reconstructed portions seems more appropriate for studies aiming understanding. At the same time, current data suggest that the order of magnitude of spectral features of phoneme /s/ can be reproduced using such a replica.

IV. Discussion and Conclusion

*For all assessed Reynolds numbers and for all noise generation systems (either with the replica or from a human speaker) the acoustic energy is situated at frequencies above the cut-on frequency of the first non-plane mode (>3.5 kHz). Consequently, higher order acoustic propagation modes potentially affect the radiated sound field and need to be taken into account when aiming sibilant sound screening either using replicas or using human speakers.

*Spectral similarity between phoneme /s/ and noise generated with the replica is most obvious for loudness instruction Re~5300) for which the oral volume reconstruction is realised. The spectral match for frequencies up to 8 kHz illustrates that such a replica can at least partially reproduce acoustic spectra of the same speaker and hence might be useful for speech screening applications, modelling or experimental studies in the frequency range below 8 kHz. Additional replicas based on volume data obtained for a 'soft' and 'loud' loudness instruction of the same speaker are of interest in order to understand how acoustic radiated sound spectra are affected by a potential change in oral tract shape.

*Noise generated with the replica is characterised by spectral features (dynamic amplitude, spectral slopes and peak frequency) of the same order of magnitude as observed for phoneme /s/. This supports and encourages the use of volume oral tract reconstruction for speech screening applications. Nevertheless, the peak frequency of phoneme /s/ is lower than the one observed with the replica for all assessed Reynolds numbers. The decrease is probably due to intra-speaker variability of the oral tract geometry between different phoneme /s/ utterances. Moreover, current data show that spectra of phoneme /s/ can be influenced by additional noise sources such as whistling observed for the highest assessed Reynolds number (Re~8900). Moreover, the tendency of spectral features as a function of Reynolds number is different for noise produced with the replica as for phoneme /s/. This illustrates the interest to study flow parameters in combination with a constant replica geometry.

*Spectral features depend on flow conditions upstream (Reynolds number as well as turbulence intensity) from the reconstructed portion of the replica. The upstream turbulence intensity (method of air supply) influences the energy discrepancy observed between the replica and phoneme /s/. The influence of upstream turbulence intensity is more pronounced for the lowest assessed Reynolds number (Re~2800). The Reynolds dependence of spectral features (peak frequency and negative spectral slope) is affected by the turbulence intensity (method of air supply) as well. Therefore, current data provide evidence that in order to perform physical speech screening it is not sufficient to define a constant geometry based on volume reconstruction and vary Reynolds number. In addition, a detailed study of upstream flow conditions in relation to sibilant sound production and radiated spectra using a replica with constant geometry is needed. Such a study can also contribute to criteria characterising upstream flow conditions relevant to human sibilant fricative production since such a direct measurement on a human subject is extremely difficult to achieve.

*The current study focused on the impact of turbulence intensity on the spectral discrepancy between noise generated with a replica and phoneme /s/. Flow visualisation immediately down- stream from the replica suggests that future studies need to account for different exit conditions (geometry, rigidity) as well in order to consider the impact of the visualised highly three-di- mensional flow field on noise production and radiation as well as its relevance for phoneme /s/ uttered by human subjects.

REFERENCES

[1] G. Fant. The acoustic theory of speech production. Mouton, The Hague, 1960.

[2] L.M.T. Jesus and C.H. Shadle. A parametric study of the spectral characteristics of European Portuguese fricatives. J. Phonetics, 30(3):437–464, 2002.

[3] A. Van Hirtum, Y. Fujiso, and K. Nozaki. The role of initial flow conditions for sibilant fricative production. J. Acoust. Soc. Am., 136:2922–2925, 2014.

ACKNOWLEGDMENTS

This work was partly supported by EU-FET grant (EUNISON 308874).

INFLUENCE OF A SOFT TISSUE OF VOCAL TRACT ACOUSTIC CAVITIES PROLONGED BY A TUBE ON FORMANT FREQUENCIES

V. Radolf[1], J. Horáček[1]
[1] Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Prague, Czech Republic
radolf@it.cas.cz, jaromirh@it.cas.cz

Abstract: **A mathematical model, which can help to clarify physical background of an influence of the soft tissues of vocal cavities on the formant frequencies, is introduced. Strong acoustic-structural interaction is demonstrated on the vocal tract cavity prolonged by a tube that is used for voice training and therapy purposes. The lower an oscillating mass in the glottis is, the higher the formant frequencies are. Especially the first formant frequency is very sensitive to the mass even if it was very small, order of 1 gram or less.**
Keywords: **Biomechanics of voice, phonation into tubes, effect of low frequency mechanical resonance**

I. INTRODUCTION

Phonation into tubes of various dimensions is used for voice training and therapy purposes, see e.g. [1,2]. Phonation into a resonance tube in air has been used for voice training of normal voiced subjects to improve loudness and voice quality, see [3]. According to the results of Story et al. [4], first acoustic resonance (formant frequency F_1) lowers with a prolongation or a semi-occlusion of the vocal tract (VT). The authors calculated that F_1 was approximately in the range of 200-300 Hz for phonation into hard-walled tubes of 10 cm and 30 cm in length and inner diameter 8 mm. The vocal tract wall was considered to be yielding, having a mechanical resonance frequency of its own. The parameters of such a VT model were set to the values suggested by Sondhi and Schroeter [5]. However in [5], the lowest resonant frequency 200 Hz of the closed VT is used as a parameter, although this value can differ significantly considering the VT prolonged by a tube. This is the motivation to introduce another mathematical model, which can help to clarify physical background of an influence of the soft tissues of vocal cavities on the formant frequencies.

II. METHODS

Let us consider a coupled mechanical-acoustical system shown in "Fig. 1", consisting of a vocal tract cavity 1 and a tube 2 with cross-sectional area S_2 and length L_2. The glottis is closed by a yielding wall having a mass m_W and vibrating with a displacement $w(t)$ on a spring of stiffness k_W and a damper b_W.

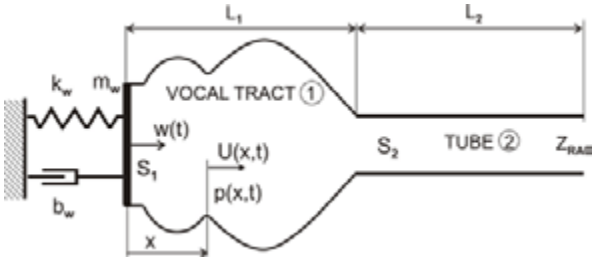


Figure 1. Schema of the vocal tract model prolonged by a tube and closed with a yielding wall.

The equation of motion for the wall is

$$m_W \ddot{w}(t) + b_W \dot{w}(t) + k_W w(t) + F(t) = 0\,, \tag{1}$$

where $\ddot{w}, \dot{w}$ denotes second and first derivative of w with respect to time, respectively, and F is the force loading the mass by the pressure in the vocal tract

$$F(t) = S_1 \cdot p(x=0,t)\,. \tag{2}$$

The wave equation for air in the acoustic cavities can be written as

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{1}{S}\cdot\frac{\partial S}{\partial x}\cdot\frac{\partial\varphi}{\partial x} - \frac{1}{c_0^2}\cdot\left(\frac{\partial^2\varphi}{\partial t^2} + c_0\cdot r_N\cdot\frac{\partial\varphi}{\partial t}\right) = 0, \tag{3}$$

where ϕ is the flow velocity potential related to the acoustic pressure p and the acoustic volume velocity U as

$$p = -\rho_0\,\partial\phi/\partial t\,,\; U = S\,\partial\phi/\partial x\,, \tag{4}$$

x is the longitudinal coordinate along the VT, t is time, c_0 is speed of sound and ρ_0 is fluid density. Using boundary and continuity conditions for the pressure and acoustic volume velocity we get the frequency equation

$$\omega^2 + j\omega\left(\frac{B - D\cdot Z_{RAD}}{A - C\cdot Z_{RAD}}\cdot\frac{S_1^2}{m_W} - 2\zeta\omega_0\right) - \omega_0^2 = 0\,, \tag{5}$$

where $\omega_0^2 = k_W/m_W$ is the squared angular frequency of mechanical resonance, $\zeta = b_W/(2m_W\omega_0)$ is the damping ratio of the yielding wall, Z_{RAD} is radiation impedance at the open tube end and A, B, C, D are components of the transfer matrix of the complete acoustical system (VT + tube):

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

$$\begin{bmatrix} p_{OUT} \\ U_{OUT} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} . \begin{bmatrix} p_{IN} \\ U_{IN} \end{bmatrix} . \quad (6)$$

We assumed the output loaded by the acoustic radiation impedance of a vibrating circular plate placed in an infinite wall, see e.g. [6].

III. Results

The numerical solution of equation (5) was performed for the following parameters: $\rho_0 = 1.2$ kgm^{-3}, $c_0 = 353$ ms^{-1}, $L_2 = 26.4$ cm, $S_2 = 0.36$ cm^2, and VT geometrical configuration for vowel /u:/ (see [6]). The mechanical resonance frequency was kept constant at 15 Hz as in [5], $\omega_0 = 2\pi \cdot 15$ rad/s, the damping ratio was set to $\zeta = 0.2$, whereas the mass varied from 0.0001 grams to 200 grams. The resulting first four natural frequencies are shown in "Fig. 2" for the mass m_W varying from 0 to 4 grams.

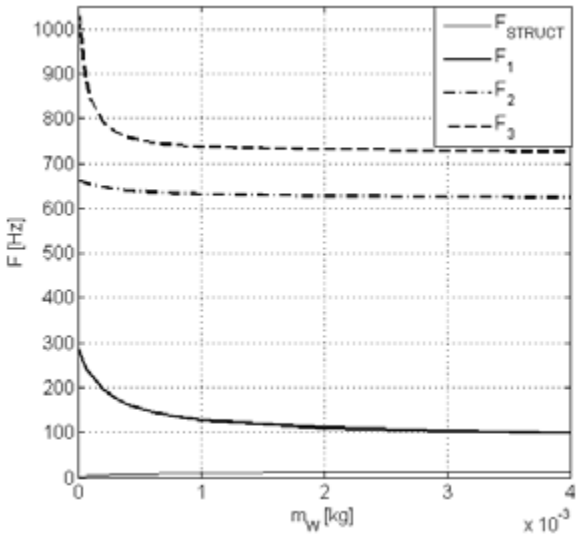


Figure 2. Natural frequencies of VT model prolonged by the tube in dependence on the yielding wall mass m_W.

First natural frequency F_{STRUCT} corresponds to the mechanical resonance ω_0 and varies from zero for $m_W \to 0$ to $F_{STRUCT} = 14.4$ Hz for $m_W > 32$ grams. Second natural frequency of the coupled system corresponds to the first acoustic resonance F_1. The solid line in Fig. 2 reveals that F_1 is strongly influenced by coupling with the vibrating wall, when the mass m_W decreases below about 1 gram. The higher acoustic resonances F_2 and F_3 are influenced by the vibrating wall in smaller range of the mass m_W, only up to about 0.5 g.

IV. Discussion and Conclusion

A brief analysis of the frequency equation (5) can be done for two extreme cases. When the mass m_W goes to zero (and so does the stiffness k_W, because we assume ω_0 to be constant), then the numerator in brackets must be zero and thus we get solution $\omega = 0$ and a frequency equation for acoustic resonances of the open-open system of the acoustic cavities:

$$B - D \cdot Z_{RAD} = 0 . \quad (7)$$

If the mass goes to infinity (and so does k_W), then $\omega = \omega_0$ or the denominator must be zero, which yields the frequency equation for the closed-open system:

$$A - C \cdot Z_{RAD} = 0 . \quad (8)$$

The acoustic resonances of the system computed for $m_W = 0.0001$ g are equal to frequency values corresponding to the acoustic system with both ends opened: $F_1 = 285$ Hz, $F_2 = 662$ Hz and $F_3 = 1039$ Hz, see eq. (7). The acoustic resonances for $m_W > 100$ g are equal to the acoustic resonances when the vocal tract is closed by a rigid wall at the glottis: $F_1 = 85$ Hz, $F_2 = 623$ Hz and $F_3 = 724$ Hz, see eq. (8).

The lower an oscillating mass in the glottal region is, the higher the formant frequencies are. Especially, the first formant frequency is very sensitive to this mass when it is in range below 1 gram.

References

[1] I.R. Titze, E.M. Finnegan, A.M. Laukkanen and 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, pp. 151-161, 1992.

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

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

[5] M.M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, pp. 955-967, 1987.

[6] 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* 94 (2008) 433-447.

The study was supported by the grant No P101/12/1306 of the Czech Science Foundation.

September 4

FP – Models 3

ON THE HARMONIC-TO-NOISE RATIO AS A CUE FOR AUTOMATIC CLASSIFICATION OF PARKINSON'S DISEASE

A. Kacha[1], F. Grenez[2], J. Schoentgen[2,3], S. Skodda[4]

[1] Laboratoire de Physique de Rayonnement et Applications, Université de Jijel, Jijel, Algeria
[2] LISA Department, Universite´ Libre de Bruxelles, Brussels, Belgium
[3] National Fund for Scientific Research, Belgium
[4] Department of Neurology, Knappschaftskrankenhaus, Ruhr-University of Bochum, Germany
akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be, sabine.skodda@kk-bochum.de

Abstract: **In this paper, the effectiveness of the HNR as a cue for the classification of Parkinson disease is investigated. The HNR is known to be very effective in the evaluation of the overall quality of the disordered voices produced by dysphonic speakers. However, the question whether the HNR provides a good feature for PD classification is still unanswered. The empirical mode decomposition-based method is used to estimate the HNR by decomposing the log of the magnitude spectrum of the speech signal into its harmonic, envelope and noise components. Experimental results carried out on stimuli of sustained vowels [a] produced by 265 Parkinson speakers and 78 control speakers show that the HNR values corresponding to Parkinson speakers are in the same range as the HNR values corresponding to control speakers. However, it is observed that the absolute difference between local HNR values corresponding to the first analysis window and a sliding analysis windows tend to increase for Parkinson speakers compared to control speakers.**

Keywords: **harmonic-to-noise ratio, vocal dysperiodicities, empirical mode decomposition, Parkinson disease.**

I. INTRODUCTION

Parkinson's disease (PD) is a chronic neurological disorder basically caused by a loss of the nerve cells in the part of the brain called the substantia nigra. The core motor symptoms of PD include slow physical movements (hypokinesia), tremor, muscle stiffness (rigidity) and postural instability. Secondary symptoms include - among others - cognitive impairments, a variety of autonomic and neuropsychiatric features, sleep problems and speech and voice impairments. In a study based on a large sample of patients with PD, it has been reported that between 70% and 90% of patients with PD have problems related to speech and voice impairments [1]. Recently, there has been a considerable interest to acoustic analysis of speech for automatic detection and classification of PD [2-4]. A reliable and accurate method for PD classification is of great importance to fulfill the clinicians´ need for objective and reproducible measures to classify the severity and progression of the disease and to monitor therapeutic interventions.

A number of acoustic cues have been proposed in the literature to characterize the voice and speech of PD speakers. In many studies, it has been found that the average fundamental frequency of Parkinson speakers is higher than that of normal speakers [5]. Several acoustic measures grouped into different feature sets that share common attributes have been proposed in [2] for classification of PD from speech. Most of these acoustic measures are well known in the general framework of analysis of disordered voices. It has been found that only ten acoustic measures enable to achieve almost 99% classification accuracy. Among all the sets of tested measures, MFCCs and measures that quantify noise appear to be consistently selected. These measures include the harmonic-to-noise ratio (HNR), glottal-to-noise excitation (GNE) and vocal fold excitation ratio (VFE). In [4], several acoustic measures derived from the fundamental frequency and noise level (HNR and adaptive normalized noise energy) extracted from sentences using different conventional software programs have been used to characterize PD. Experimental results showed that there were no significant differences between the values of the noise-based acoustic measures of PD and control speakers. Taking into account that the corpora used in [2] and [4] are different, the findings of both studies still appear to be-contradictory.

The HNR is known to be very effective in the evaluation of the overall quality of the disordered voices produced by dysphonic speakers. However, the question whether the HNR provides a good feature for PD classification is still unanswered. In the present study, the effectiveness of the HNR as a cue for the classification of PD is investigated. The empirical mode decomposition-based method is used to estimate the HNR by decomposing the log of the magnitude spectrum of the speech signal into its harmonic, envelope and noise components.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

The remainder of the paper is organized as follows. The EMD-based method for HNR estimation is briefly reviewed in Section II. The corpus used to test the analysis method is described in Section III. Results are presented in Section IV. Finally, conclusions are given in Section V.

II. METHODS

A. Estimation of the harmonic-to-noise ratio

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, $v(t)$ [6]:

$$x(t)=e(t)*v(t) \quad (1)$$

where * denotes the convolution.
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:

$$\log|X_w(f)|=\log|E_w(f)|+\log|V(f)| \quad (3)$$

It appears 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 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 [7], so that the decomposition of the log magnitude spectrum via the EMD algorithm results into several oscillating components (IMFs, intrinsic mode functions) that can be clustered into three classes and each class of components is assigned to some part of the log magnitude spectrum

The method used to estimate the HNR is presented in [8]. For a given utterance, the empirical mode decomposition (EMD) algorithm is used to decompose the log of the magnitude spectrum of the speech signal into its harmonic, envelope and noise components. The analysis interval is divided into L frames and the HNR is computed as the average of the HNR_i (i=1,…, L) of the L frames. The frame length for HNR estimation is set to 200 ms.

In order to investigate the time-evolution of the HNR, a local HNR is computed on non-overlapping analysis windows. For each window, the local HNR is computed by averaging the HNRs of the corresponding 200 ms frames. The entire analysis interval has been divided into five windows.

B. Corpus

The corpus comprises sustained vowels [a] produced by 265 Parkinson speakers and 78 control speakers. The stimuli have been sampled at 44.1 kHz and recorded in wav format at the Department of Neurology of the Knappschaftskrankenhaus/Ruhr-University Bochum /Germany. The stimuli have different lengths. The length of the stimuli produced by control speakers ranges between 2 s and 32 s while the length of the stimuli produced by Parkinson speakers range between 0.8 s and 26 s. For each stimulus, the analysis has been carried out on the whole interval.

III. RESULTS AND DISCUSSION

The EMD-based method has been used to compute the HNR values of the stimuli produced by control speakers and Parkinson speakers. As an illustration, Fig. 1 shows a frame of 200 ms extracted from a vowel [a] and its corresponding estimated harmonic component for a control speaker and a Parkinson speaker. The global HNR values corresponding to the control speaker and Parkinson speaker are 13 dB and 21.4 dB, respectively.

The quartiles of the HNR values (in dB) computed via the EMD-based method are depicted as a boxplot in Fig. 2. As observed, the HNR values corresponding to Parkinson speakers are in the same range as the HNR values corresponding to control speakers. A two-tailed t-test shows that the difference between HNR values of control speakers and Parkinson speakers is not statistically significant. The null hypothesis is retained at the 5% significance level. We note that the voice/speech samples of the PD group derive from patients in all the different stages of the disease (from mildly to very severely affected) and with very different disease durations. To investigate the time-evolution of the HNR over the analysis interval, local HNRs corresponding to the different sliding analysis windows have been computed. To measure the variation of the HNR, the first analysis window is taken as a reference window and the absolute difference between the local HNR corresponding to this reference window and the local HNR associated to the j-th window (j=2, . . ., 5) have been used.

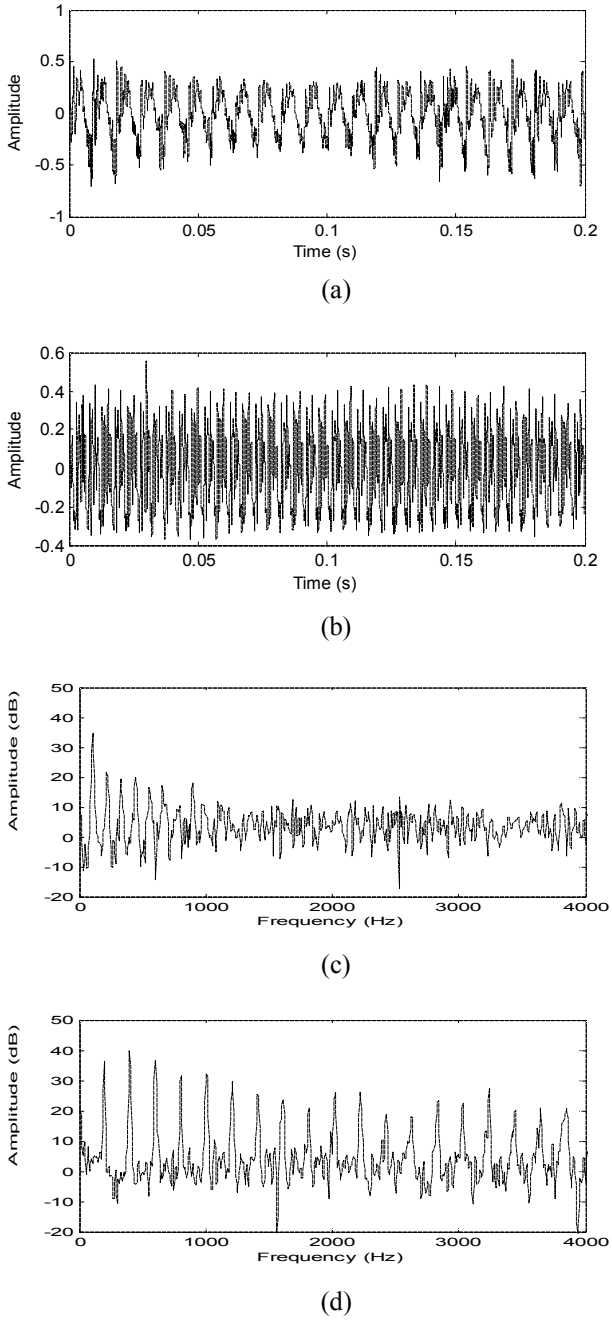


Fig.1 : Frame of 200 ms extracted from a sustained [a] produced by a control speaker (a) and a Parkinson speaker (b) and the harmonic component estimate corresponding to the control speaker (c) and Parkinson speaker (d).

The absolute differences between the local HNR values associated to the reference window and those corresponding to the second and fourth sliding windows are depicted as a boxplot in Fig. 3. The number j between parentheses indicates that de difference is computed between the HNR values corresponding to the reference (first) window and those associated to the window j. One observes that the absolute difference between HNR values corresponding to the first and second windows are slightly in the same ranges for both Parkinson speakers and control speakers while the absolute difference between HNR values corresponding to the first and fourth windows tend to increase for Parkinson speakers compared to control speakers. A two-tailed *t*-test shows that the difference in HNR values corresponding to the first and fourth windows between control speakers and Parkinson speakers is statistically significant. The null hypothesis is rejected at the 5% significance level.

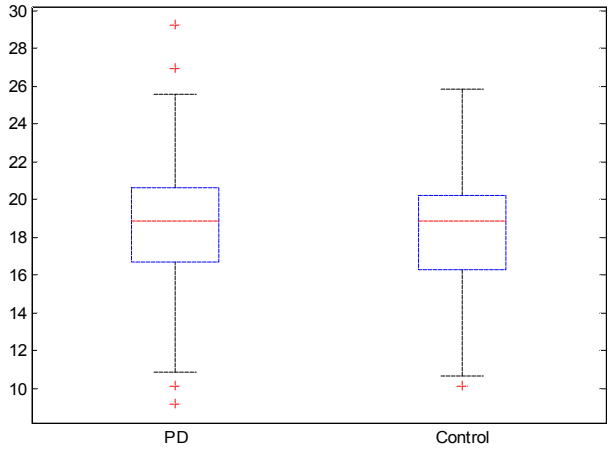


Figure 2: Boxplot of HNR values (in dB) estimated via EMD algorithm for control and Parkinson speakers.

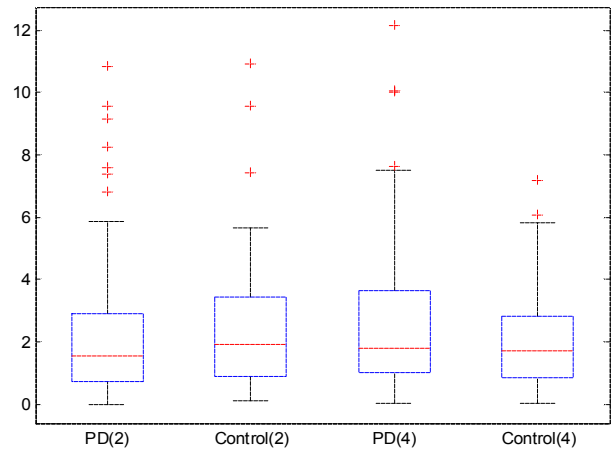


Figure 3: Boxplot of the absolute difference (in dB) between local HNR values corresponding to the reference window and local HNR values associated to the second and fourth window for control and Parkinson speakers.

Experimental results carried out on sustained vowels show that the HNR per se computed on the whole analysis interval does not provide a useful feature that enables to discriminate between Parkinson speakers and normal speakers. Although the corpora are different, our conclusions concerning the effectiveness of the HNR as a cue for PD classification are in agreement with those drawn in [4].

Fig. 4 shows the variation of the first quartile, the median and the third quartile of the relative difference in percent between local HNR values corresponding to the reference window and local HNR values associated to the sliding window for control speakers and Parkinson speakers. As can be seen the percentile values corresponding to Parkinson speakers increase as the analysis window moves from the second position (j=2) to the fourth position (j=4) with respect to the reference window. However, one observes that the percentile values decrease as the analysis window moves from the fourth position to the final position (j=5). A possible explanation of the reason for the decrease of the percentile is that at the end of the analysis interval, the speakers are unable to produce correctly the stimuli. The decrease of the percentile values is likely due to the vocal fatigue. The time-evolution of the HNR quantified as the absolute difference between the HNR values corresponding to a reference window and those corresponding to successive sliding analysis windows appear to be more effective.

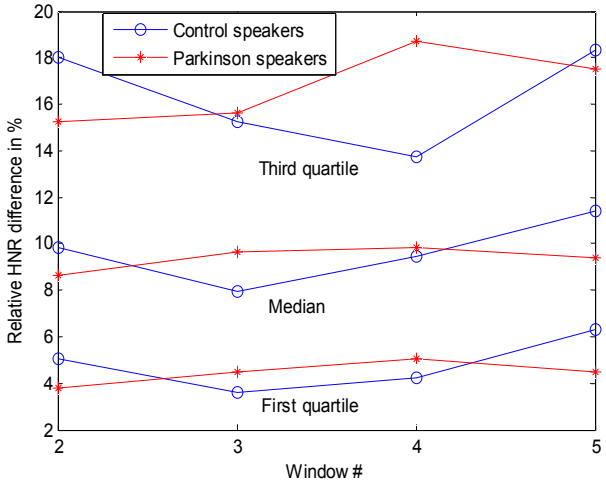


Figure 4: variation of the first quartile, the median and the third quartile of the relative difference in percent between local HNR values corresponding to the reference window and local HNR values associated to the sliding window for control speakers and Parkinson speakers.

IV. Conclusion

In this paper, the effectiveness of the HNR as a cue for the classification of PD is investigated. The empirical mode decomposition-based method has been used to estimate the HNR by decomposing the log of the magnitude spectrum of speech signal into its harmonic, envelope and noise components. Experimental results carried out on stimuli of sustained vowels [a] produced by Parkinson speakers and control speakers show that the HNR values corresponding to Parkinson speakers are in the same range as the HNR values corresponding to control speakers so that the HNR per se computed on the whole analysis interval does not provide a useful feature that enables to discriminate between Parkinson speakers and normal speaker. However, it has been observed that the absolute difference between HNR values corresponding to the first and fourth windows tend to increase for Parkinson speakers compared to control speakers.

References

[1] J. A. Logemann, H. B. Fishe, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[2] A. Tsanas, M. -A. Little, P. -E. McSharry, J. Spielman, and L. -O. Ramig, "A novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Trans. Biomed. Eng., vol. 59, no. 5, pp. 1264–1271, 2012.

[3] T. Bocklet, S. Steidl, E. Nöth, S. Skodda, "Automatic Evaluation of Parkinson's Speech - Acoustic, Prosodic and Voice Related Cues," in: Proc. Interspeech 2013, Lyon (France), Aug. 2013, pp. 1149-1153.

[4] A. Brandini et al., "Automatic identification of dysprosody in idiopathic Parkinson's disease," Biomed. Signal Proc. Control, vol. 17, pp. 47-54, 2015.

[5] L. –K. Bowen, G. –L. Hands, S. Pradhan, C. -E. Stepp, "Effects of Parkinson's Disease on Fundamental Frequency Variability in Running Speech," J. Med. Speech. Lang. Pathology, vol. 21, no 3, pp. 235–244, 2014.

[6] 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.

[7] P. Flandrin, G. Rilling, and P. Conçalvès, "Empirical Mode Decomposition as a Filter Bank", *IEEE Signal Proc. Letters*, vol. 11, pp. 112-114, 2004.

[8] A. Kacha, F. Grenez, and J. Schoentgen, "Multiband vocal dysperiodicities analysis using empirical mode decomposition in the log-spectral domain," Biomed. Signal Proc. Control, vol. 17, pp. 11-20, 2015.

MODELING OF GRBAS PERCEPTUAL EVALUATION USING SPECTRAL FEATURES OBTAINED FROM AN AUDITORY-BASED FILTERBANK

R. Fraile[1], K. Neumann[2], J.M. Gutiérrez-Arriola[1], N. Sáenz-Lechón[1], V.J. Osma-Ruiz[1]

[1]Signal Theory & Communications Department, Universidad Politécnica de Madrid, Madrid, Spain
[2]Department of Phoniatrics & Paedaudiology, St. Elisabeth Hospital, Ruhr-University Bochum, Bochum, Germany
rfraile@ics.upm.es, Katrin.Neumann@ruhr-uni-bochum.de, jmga@ics.upm.es, nslechon@ics.upm.es, vosma@ics.upm.es

Abstract: **Perceptual voice evaluation according to the GRBAS scale is modelled using a linear combination of acoustic parameters calculated after a filter-bank analysis of the recorded voice signals. Modelling results indicate that for breathiness and asthenia more than 55% of the variance of perceptual rates can be explained by such a model, with only 4 latent variables. Moreover, the greatest part of the explained variance can be attributed to only one or two latent variables similarly weighted by all 5 listeners involved in the experiment. Correlation factors between actual rates and model predictions around 0.6 are obtained.**

Keywords: **Perceptual evaluation, Linear modelling, Auditory models.**

I. INTRODUCTION

Since the primary function of human voice is interpersonal communication, voicing is closely related to hearing. For this reason, many protocols for voice quality assessment currently in use include perceptual evaluation of voice quality [1]. The widespread use of standardized scales such as GRBAS [1] or CAPE-V [2] has contributed to increasing the value of perceptual rating as a clinical tool, in spite of the reliability issues identified by some researchers [3].

Specifically, GRBAS has been recommended as a minimum standard for perceptual evaluation in the voice clinic [4]. It includes the evaluation of five aspects of voice: overall grade (G), roughness (R), breathiness (B), asthenia (A), and strain (S). For each one, the rater has to assign a mark ranging from 0 (best quality) to 3 (worst). In general, G tends to be easier to evaluate that R, B, A or S [5], "easier" meaning that a lower degree of variability is to be expected, both inter-rater and intra-rater.

Regarding reliability of GRBAS, values around 0.6 have been reported for inter-rater Pearson's correlation coefficient in scales G, R and B [1]. Reliability may also be measured in terms of the Cohen's kappa statistic [5] but, since inter-rater agreement is greater when continuous scales are used [6], correlation coefficients are to be preferred for evaluating such agreement in authors' view.

Due to the limited reliability of perceptual evaluation of voice, the search for objective acoustic descriptors of voice quality has received a great deal of attention in the scientific community for years. The relation between these descriptors and perceptual ratings of voice have also been investigated. For instance, correlations of G, R, B, and A with noise parameters have been found [7]. Similarly, correlations of R and B with both noise measures and pitch/amplitude perturbation measures have been reported [8]. Some spectral measures have also been proposed and they have shown to provide relevant correlations mainly with B [9]. A low-dimensional coding of the overall spectral shape in cepstral domain has shown to provide fair correlations with R and B [10] and the cepstral peak prominence (CPP) also exhibits significant correlations with G, R, and B [11].

This paper presents an analysis of the perceptual rates assigned by five raters to the voice of 47 individuals according to the GRBAS scheme. Inter-rater correlations of rates corresponding to the same scale are studied. The relationship between this set of rates and voice measures obtained after a filter-bank analysis similar to that presented in [12] is also studied. Such processing scheme models the front end of the auditory system and, consequently, it is expected to provide acoustic measures that are relevant to perception. The conclusions in [13] prevent against the use of measures describing the spectral shape and other researchers also point out that the temporal dynamics of the outputs of filters in the filter-bank are more relevant to perception than the overall spectral shape, even when calculated in short-time frames [14]. Consequently, the acoustic measures used here intend to provide simple, low-level descriptions of such dynamics.

The obtained results indicate that for B and A up to 55% of the variance of the perceptual rates can be explained by a few factors combining these low-level measures, and that most of such variance can be explained by factors that are common and similarly weighted for all raters.

II. MATERIALS

Voice recordings corresponding to 20 patients (14 females, 6 males) and 27 healthy speakers (15 females, 12 males) were available. Average age for female patients

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

was 45.3 while for female healthy speakers it was 40.5. Similarly, for male patients the average age was 57.2 and for healthy speakers it was 36.9.

All voices were recorded in the Phoniatrics & Paedaudiology Department of the St. Elisabeth Hospital (Ruhr-University Bochum) in a quiet room within a normal clinical environment. No special attempt was made to prevent the appearance of background noises. Recordings were collected at a sampling rate equal to 22,050 Hz and with 32 quantization bits using a system from XION-Medical (XION GmbH). All recordings were normalised to have a root mean square value equal to 1.

All patients and healthy speakers were asked to pronounce at comfortable pitch and intensity. A head-mounted microphone was used in order to keep distance between lips and microphone constant (approximately 20 cm). The recording used in this investigation corresponds to the reading of a German translation of Aesop's fable "The northwind and the sun".

III. Methods

A. Perceptual evaluation

Recordings corresponding to the reading of Aesop's fable were presented to five listeners with diverse levels of experience in voice evaluation. Specifically, the listeners were a phoniatrician with 17 years of experience (the author KN) and four advanced bachelor students of logopedics (in their 6th semester). All of them were asked to assign labels according to the GRBAS scales to each recording. Label values for all five scales were allowed to vary between 0 and 3 with a resolution equal to 0.25, though the students kept resolution of their labels coarser (0.5). All five listeners held a joint meeting for training and discussion before performing the GRBAS evaluation.

B. Processing of voice recordings

From the recordings, the first sentence of the fable was selected: "*Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges kam*". As a pre-processing stage, intervals corresponding to voiced sounds were selected, according to the algorithm described in [15] with prior μ-law compression so as to attenuate the highest peaks.

Recording segments corresponding to voiced sounds were processed by a filter-bank consisting of 22 filters with pass bands corresponding to the 22 first auditory critical bands, as detailed in [12]. However, instead of Hamming-based filters, gammatone responses were preferred since they provide a better model for the front end of the auditory system. Specifically, the implementation proposed by Slaney was used [16].

The dynamics of the filter-bank output signals were described in terms of two parameters defined in [12]: the average energy for each band and the band energy decorrelation time, which is a measure of the stability of the signal energy (longer decorrelation times correspond to more stable signal energies). Consequently, for each recording 44 parameters were obtained, corresponding to the average energy and the energy decorrelation time at the output of each one of the 22 filters.

C. Statistical analysis

Correlations between GRBAS rates and between rates and acoustic parameters were measured in terms of the Spearman correlation coefficient. This measure was preferred instead of the more common Pearson correlation coefficient because of its capability for measuring non-linear relations between variables. The correction for ties proposed in [17, chap. 5] was implemented in the computation of the correlation coefficients.

Modelling of the GRBAS rates as linear combinations of the values of the acoustic parameters was done by means of partial least squares (PLS) regression [18]. If $\mathbf{Y}$ is a 47×5 matrix containing the rates assigned by each one of the five listeners to each recording and corresponding to either G, R, B, A or S, and $\mathbf{X}$ is a $47\times N_p$ matrix containing the values of a selected set of N_p out of the available 44 acoustic parameters associated to each voice recording, then the PLS model approximates the rates as:

$$\mathbf{Y} \approx \mathbf{X} \cdot \mathbf{F} \cdot \mathbf{W} \qquad (1)$$

where $\mathbf{F}$ is a $N_p\times N_f$ matrix that reduces the dimensionality of the space of input variables from N_p down to N_f. N_f is the number of latent variables or factors of the model. $\mathbf{W}$ is a $N_f\times5$ matrix that models the weight that each listener assigns to each one of the factors in order to generate the corresponding rates.

Inputs and outputs of the PLS model were linearised and normalised following standard procedures [18]. Namely, acoustic parameters were transformed according to a fourth-root law (a substitute for the logarithmic law when null or almost null values occur) and mean subtraction and variance normalisation were applied to both transformed acoustic parameters and GRBAS rates.

IV. Results

A. Inter-rater correlations

Tab. I shows the values of the Spearman coefficients measuring correlation between rates corresponding to the same scale and assigned by different listeners.

As expected, the values of the correlation coefficients for scale G are larger than for the rest of scales. This is consistent with results from other researchers [5]. The values for the Spearman correlation coefficients are also greater than the Pearson coefficients reported in [1], but this is also as expected because Pearson coefficients are not sensitive to non-linear relations.

Table I. Minimum, maximum and average inter-rater correlation coefficients for all five scales in GRBAS.

Scale	Minimum correlation	Maximum correlation	Average
G	0.73	0.88	0.80
R	0.66	0.80	0.75
B	0.55	0.84	0.73
A	0.65	0.87	0.78
S	0.39	0.84	0.64

Table II. Positive and negative relevant correlations between GRBAS rates and acoustic parameters. *AE* stands for *average energy* and *DT* stands for *decorrelation time*. f_c is the central frequency of the corresponding critical band in Hz.

f_c	G		R		B		A		S	
	AE	*DT*	*AE*	*DT*	*AE*	*DT*	*AE*	*DT*	*AE*	*DT*
60								+		
150										
250				–						
350										
455	–		–		–		–			
570		–			–	–	–			–
700		–		–						
845										
1000					–					
1175	–	–	–	–	–	–	–	–	–	–
1375	–				–		–		–	
1600					–		–			
1860					–	–	–	–		
2160		–		–	–	–	–	–		–
2510										
2925										
3425	–									
4050										
4850					+		+		+	
5850		–		–	+					
7050					+					
8600					+		+		+	

B. Identification of the relevant acoustic parameters

In order to identify which acoustic parameters among the 44 available ones were the most relevant for modelling GRBAS rating, the following procedure was implemented. Firstly, the correlation coefficients between each parameter and the rates corresponding to each scale and each listener were calculated. This resulted in a set of 44×5×5=1100 values. Secondly, the 90-percentile of the absolute values of these correlation coefficients was set as a threshold for selection. Last, all parameters with correlation coefficients having absolute values greater than the threshold for at least one listener were selected as relevant for the corresponding scale. The number N_p of parameters selected as relevant by this procedure was 9 for G, 7 for R, 16 for B, 13 for A, and 7 for S.

Tab. II summarises the signs of the relevant correlation coefficients identified by the afore-mentioned procedure. For the average band energy, a negative correlation implies higher rates for lower energies and a positive correlation means higher rates for higher energies. As for decorrelation time, all relevant correlations are negative, which means higher rates for shorter decorrelation times (more instability in energy).

C. PLS modelling

Fig. 1 shows how the fraction of variance in GRBAS rates explained by the PLS model in (1) evolves as the number of latent variables N_f varies. The PLS model has been built using the acoustic features in Tab. II as inputs.

Due to the limited number of input variables, models with more than 7 latent variables did not converge. The graphs in Fig. 1 indicate that a quasi log-linear dependence of the fraction of explained variance from the number of latent variables happens for up to 4 latent variables. Beyond that number, the fraction of explained variance in G, R and S only experiences minor changes and although it has a more relevant growth for B and A, this is still lower than 10% of its value for $N_f = 4$.

Considering the previous observations, a PLS model with 4 latent variables has been selected. For all GRBAS scales, the variable which corresponds to the greatest fraction of variance explained by the model is an average of all relevant acoustic parameters with similar weights for all of them and with weight signs as indicated in Tab. II. For the scales with the highest fraction of variance explained by the model (A and B, as shown in Fig. 1) the weights assigned to the first two variables for the five listeners are similar, while the most relevant differences happen for the third and fourth variables. In contrast, for R and S similarity only happens in the first variable.

Tab. III shows the correlation coefficients between rates assigned by listeners and rates predicted by the PLS models. Not surprisingly, the highest correlations occur for the scales with the highest fractions of variance explained by the model. The mean values of such correlation coefficients, averaged for all scales, are similar for all listeners. These mean values are around 0.6, below the average values of inter-rater correlations in Tab. I.

V. Conclusions

The spectral analysis reported in [12] indicated that the presence of dysphonia was closely related to low energy in frequencies from 1080 to 2700 Hz and high energy in bands over 5300 Hz for running text recordings. For the same type of recordings, there also was a looser relationship between dysphonia and shorter energy decorrelation times in frequencies from 630 to 2700 Hz.

Results reported here confirm these relations and they indicate that, among the five dimensions included in the GRBAS scheme, the spectral distribution of energy is more closely related with B and A than with the rest. This may be a cue that B and A mainly depend on low-level auditory features such as the ones used here, while rating of G, R and S requires more complex processing.

The characteristics of the linear model built to relate GRBAS rates and acoustic features also revealed that the greatest part of the rate variations explained by the model can be attributed to a few factors common for all listeners.

ACKNOWLEDGEMENTS

This work has been partially financed by the Spanish Government, through project grant number TEC2012-38630-C04-01. Voice recording was carried out in the context of project AIB2010DE-00304, jointly financed by the Spanish Government and the Deutscher Akademischer Austauschdienst (DAAD).

REFERENCES

[1] P.H. Dejonckere, C. Obbens, G.M. DeMoor and G.H. Wieneke, "Perceptual evaluation of dysphonia: Reliability and relevance", *Folia Phoniatr Logop*, vol.45, n.2, pp.76–83, 1993.

[2] G.B. Kempster, B.R. Gerratt, K. Verdolini, J. Barkmeier-Kraemer and R.E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol", *Am J Speech-Lang Pat*, vol.18, n.2, pp.124–132, 2009.

[3] R. Buekers, "Perceptual evaluation of vocal behaviour", *Logoped Phoniatr Vocol*, vol.23 (PEVOC-II Suplem.), pp.23–27, 1998.

[4] P. Carding, E. Carlson, R. Epstein, L. Mathieson, C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom", *Logoped Phoniatr Vocol*, vol. 25, n.3, pp.133–138, 2000.

[5] M.S. De Bodt, F.L.Wuyts, P.H. Van de Heyning, C. Croux, "Test-retest study of the GRBAS scale: Influence of experience and profesional background on perceptual rating of voice quality", *J Voice*, vol.11, n.1, pp.74–80, 1997.

[6] J. Kreiman, B.R. Gerratt, M. Ito, "When and why listeners disagree in voice quality assessment tasks", *J Acoust Soc Am*, vol.122, n.4, pp.2354–2364, 2007.

[7] T. Bhuta, L. Patrick, J.D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements", *J Voice*, vol.18, n.3, pp. 299–304, 2004.

[8] A. McAllister, J. Sundberg, S.R. Hibi, "Acoustic measurements and perceptual evaluation of hoarseness in children's voices", *Logoped Phoniatr Vocol*, vol.23, n.1, pp.27–38, 1998.

[9] R. Shrivastav, C.M. Sapienza, "Objective measures of breathy voice quality obtained using an auditory model", *J Acoust Soc Am*, vol.114, n.4, pp.2217–2224, 2003.

[10] N. Sáenz-Lechón, R. Fraile, J.I. Godino-Llorente, R. Fernández-Baíllo, V. Osma-Ruiz, J.M. Gutiérrez-Arriola, J.D. Arias-Londoño, "Towards objective evaluation of perceived roughness and breathiness: An approach based on mel-frequency cepstral analysis", *Logoped Phoniatr Vocol*, vol.36, n.2, pp.52–59, 2011.

Table III. Values of the Spearman coefficients of correlations between rates assigned by each listener and those predicted by its corresponding PLS model.

	R1	**R2**	**R3**	**R4**	**R5**	**Aver.**
G	*0.62*	*0.57*	*0.48*	*0.63*	*0.55*	0.57
R	*0.57*	*0.52*	*0.45*	*0.63*	*0.50*	0.53
B	*0.64*	*0.53*	*0.80*	*0.65*	*0.68*	0.66
A	*0.68*	*0.75*	*0.61*	*0.71*	*0.67*	0.68
S	*0.69*	*0.56*	*0.53*	*0.57*	*0.41*	0.55
Aver.	0.64	0.59	0.57	0.64	0.56	0.60

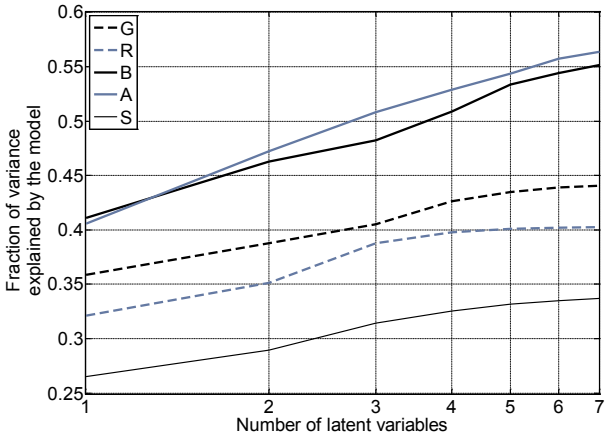


Figure 1. Fraction of variance in GRBAS rates explained by the PLS model vs the number of latent variables N_f.

[11] Y.D. Heman-Ackah, D.D. Michael, G.S. Goding, "The relationship between cepstral peak prominence and selected parameters of dysphonia", *J Voice*, vol.16, n.1, pp.20–27, 2002.

[12] R. Fraile, J.I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, J.M. Gutiérrez-Arriola, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech", *J Voice*, vol.27, n.1, pp.11–23, 2013.

[13] D.M. Howard, E. Abberton, A. Fourcin, "Disordered voice measurement and auditory analysis", *Speech Commun*, vol.54, n.5, pp.611–621, 2012.

[14] H. Hermansky, "Speech representations based on spectral dynamics", *MAVEBA 2013*, pp.191–194, 2013.

[15] 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", *Biomed Signal Process Control*, vol.8, n.6, pp.799–810, 2013.

[16] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank", *Apple Computer, Perception Group*, Tech rep, 1993.

[17] J.J. Higgins, *An Introduction to Modern Nonparametric Statistics*, Brooks/Cole, 2004.

[18] S. Wold, M. Sjöström, L. Eriksson, "PLS-regression: A basic tool of chemometrics", *Chemometr Intell Lab Syst*, vol.58, n.2, pp.109–130, 2001.

COMPARISON OF DEVELOPMENTAL AND NEUROGENIC STUTTERING

T. Tykalová[1], R. Čmejla[1], E. Růžička[2], J. Rusz[1,2]

[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
tykalova.tereza@gmail.com

Abstract: **Although the full etiological nature of developmental stuttering is still unknown, the key role of disturbed basal ganglia function along with the role of dopamine system have been thoroughly discussed. Therefore, the aim of the current study was to survey the characteristics of neurogenic stuttering based on patients with Parkinson's disease (PD) and compared them to the characteristics of developmental stuttering. The database consists of 14 persons with developmental stuttering (pDS) and 14 patients with idiopathic PD. In addition, 14 sex-matched healthy controls (HC) were recruited. Each participant was instructed to perform two-minutes long monolog on given topic. Analysis of dysfluency was conducted according to the Lidcombe behavioral taxonomy of stuttering. Our results showed that pDS subjects manifested significantly more vocal blocs ($p < 0.01$) and filling words ($p < 0.05$) compared to PD patients. On the other hand, the neurogenic stuttering of PD was mainly characterized by incomplete syllable repetitions ($p < 0.05$) and prolongations ($p < 0.01$) as compared to HC. In conclusion, our study demonstrates that there seems to be more differences rather than similarities between neurogenic and developmental stuttering.**

Keywords : **Developmental stuttering, Neurogenic stuttering, Parkinson's disease, Acoustic analyses.**

I. INTRODUCTION

Stuttering is a chronic speech disorder characterized by involuntary repetition of speech movements, prolongation of sounds, and vocal blocks. Developmental stuttering has an early onset, typically between the age of 2 and 7, whereas neurogenic stuttering occurs later in live as a consequence of neurodegenerative diseases, stroke or traumatic brain injury. Although the full etiological nature of developmental stuttering is still unknown, the key role of disturbed basal ganglia function along with the role of dopamine system have been thoroughly discussed [1-3].

Parkinson's disease (PD) is a neurological disorder characterized by the progressive loss of dopaminergic neurons in the part of the basal ganglia called substantia nigra pars compacta. The cardinal motor symptoms include tremor, rigidity, slowness of movement, and postural instability. The treatment management of motor symptoms consists mainly in dopamine replacement therapy through the usage of levodopa and dopamine agonists. In addition to the main motor symptoms, up to 90 % of patients with PD develop speech impairment in the course of their illness [4,5]. Although the most prominent signs of hypokinetic dysarthria are related to monotonous speech, reduced stress and vocal intensity, articulatory imprecision, and harsh voice quality [4], disruption of speech fluency has also been documented in some PD patients [5-8]. For instance, within the group of 53 PD patients, the acquired speech dysfluencies were pronounced in 15 patients [6]. Further, Shahed & Jankovic [9] reported 12 PD patients with a former history of developmental stuttering who remitted and subsequently recurred stuttering after the onset of PD symptoms.

Therefore, the aim of the current study was to survey the characteristics of neurogenic and developmental stuttering.
We assume that neurogenic stuttering will show some typical features, on the basis of which this form of stuttering will be distinguished from developmental stuttering. On the other hand, we also expected to find out some similarities between both types of stuttering.

II. METHODOS

A. Participants and recording procedure

A total of 14 consecutive patients with the diagnosis of idiopathic PD (3 women, 11 men) were recruited for the present study. Their mean age was mean age 61.8 ± 10.4 (41-80) years. At the time of the examination all PD subjects were on stable medication for at least 4 weeks consisting of levodopa alone or in combination with different dopamine antagonist (772 ± 348, 320-1500 mg/day). As a part of the examination, PD

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

patients were also scored according to Unified Parkinson's Disease Rating Scale (21.4 ± 11.3, 4-30). In addition, 14 persons (2 women, 12 men) with the diagnosis of developmental stuttering (pDS) and 14 healthy controls (HC; 3 women, 11 men) were included. The mean age in pDS group was mean age 31.8 ± 12.0 (18-52) years while in HC group was equal to 59.2 ± 14.2 (40-80) years. None of the participants in pDS or HC group have a history of neurological disorders. Among the PD or HC participants no history of developmental stuttering on any other communication disorders was registered. The participants were a part of previous studies [3,10].

Speech recordings were performed in a quiet room with a low ambient noise using a head-mounted condenser microphone. Speech signals were sampled at 48kHz with 16 bit resolution. Each participant was instructed to perform monolog on given topic for at least two minutes. The final length of the elicited monologs range between 1.5 to 3.5 minutes. The median number of words was equal to 200 words (235 ± 150, 58-708).

B. Acoustic analyses

Analysis of dysfluency was conducted according to the behavioral taxonomy of stuttering, where stuttering events are categorized into three primary types: repetitions, superfluous verbal behaviour and fixed posture with or without audible air flow [11]. This taxonomy was further elaborated and particular dysfluent events were divided into 6 primary types including incomplete syllable repetitions, syllable and multisyllable unit repetitions, vocal blocks, prolongations, filling words, and reformulations and unfinished words (see Table 1 for examples of specific events). Furthermore, dysfluent events were also classified in terms of within-word dysfluencies (events occurring on only part of the word or in the middle of the word) and between-word dysfluencies (events that embrace a whole word or multiple words). Each phenomenon was counted as a single entity, irrespectively of the number of its repeated speech elements (phonemes, syllables, words) based on the audiotape samples. To avoid confounding effects due to different speakers' speaking rates, the overall amount of particular dysfluent event was normalized as if each participant had been producing 200 words long monolog.

C. Statistical analyses

To assess group differences, each acoustic parameter was compared across all three groups (pDS, PD, HC) using Kruskal-Wallis test with post-hoc Bonferroni adjustment. Effect sizes were measured with Cohen's *d,* where $d > 0.5$ indicates a medium effect and $d > 0.8$ a large effect. The level of significance was set to $p < 0.05$.

Table 1. Definitions of particular dysfluent events

Category	Examples of corresponding behaviour
Incomplete syllable repetitions	"I went to S…S…S…Sydney." "Nice car…r…r…r."
Syllable and multisyllable unit repetitions	"Where…where…where's the ball?" "It's my…it's my …it's my daughter." "That's a beauti…beautiful place"
Prolongations	"mmmmmy sister is 40." "ffffishy gone!"
Vocal blocks	"I …(no sound) bought …" "It's a great oppor…(no sound) tunity."
Filling words	"I went—oh well—ah—oh well—I—well I went over…" Grunting
Reformulations and unfinished words	"It was his… her idea." "I can cook some sandy… sandwitch." "Prague is a city in Czech Repub…"

III. Results

The results of the statistical analysis revealed that pDS subjects had a significantly higher within-word as well as overall level of dysfluency ($p < 0.001$) as compared to HC. In particular, the pDS performances was mainly characterized by incomplete syllable repetitions ($p < 0.01$), syllable and multisyllable unit repetitions ($p < 0.01$), vocal blocs ($p < 0.001$), and filling words ($p < 0.001$) while PD patients performances by incomplete syllable repetitions ($p < 0.05$) and prolongations ($p < 0.01$). The direct comparison between PD and pDS showed that pDS subjects manifested significantly more vocal blocs ($p < 0.01$) and filling words ($p < 0.05$). However, it should be mentioned that overall severity of dysfluency was also greater in pDS group compared to PD ($p < 0.01$). Table 2 provides detailed numerical data and comparisons between pDS, PD, and HC groups across all acoustic measurements.

IV. Discussion

According to the current data, the developmental stuttering compared to neurogenic was mainly characterized by the higher occurrence of vocal blocks and filling words. These findings are in accordance with previous report [12] where a panel of professionals was presented at random speech samples from four developmental and four neurogenic stutterers in order to classify them accordingly. In this study, blocks and accessory stuttering behavior were mentioned as the distinguishing features for developmental stutterers while word-finding

Table 2. Results of statistical analyses

	HC	pDS	PD	Group difference	Effect size		
	mean ± SD	mean ± SD	mean ± SD				
Category	(range)	(range)	(range)	*p*	HC vs. pDS	HC vs. PD	pDS vs. PD
Incomplete syllable repetitions	0.46 ± 0.96 (0-3.6)	7.1 ± 9.6 (0-29.0)	1.7 ± 1.5 (0-5.5)	0.005	0.97**	0.96*	-0.79
Syllable and multisyllable unit repetitions	1.3 ± 1.4 (0-4.4)	10.1 ± 10.9 (0-38.2)	2.5 ± 2.7 (0.35-8.8)	0.005	1.13**	0.58	-0.95
Prolongations	0.3 ± 0.7 (0-2.2)	2.5 ± 3.9 (0-11.6)	1.3 ± 0.8 (0-3.0)	0.01	0.76	1.32**	-0.40
Vocal blocks	1.5 ± 2.4 (0-8.0)	20.2 ± 20.4 (0-78.2)	2.2 ± 1.8 (0-5.0)	<0.001	1.29***	0.32	-1.25**
Filling words	0.5 ± 0.8 (0-2.4)	10.3 ± 10.2 (0-36)	2.0 ± 2.1 (0-6.7)	<0.001	1.35***	0.94	-1.12*
Reformulations and unfinished words	1.5 ± 2.0 (0-5.9)	2.3 ± 1.7 (0-6.7)	1.9 ± 1.5 (0-5.0)	0.37	0.41	0.22	-0.23
Within-word dysfluent events	1.4 ± 1.5 (0-4.8)	27.5 ± 27.2 (0.9-95.7)	5.3 ± 3.9 (0-13.7)	<0.001	1.35***	1.31*	-1.14
Overall number of dysfluent events	5.5 ± 3.7 (1.1-12.0)	52.4 ± 33.7 (6.6 - 133.3)	11.5 ± 7.4 (1.5-25.8)	<0.001	1.96***	1.02	-1.68**

difficulties were most frequently mentioned as the clue leading to the neurogenic stuttering diagnosis [12]. However, we do not observed this word-finding difficulties in our patients. The main reason behind this discrepancy might be the differences in subjects, as we used only patients with the diagnosis of PD, no more than 6 years after diagnosis, while in the study by Borsel et al. [12] patients with different etiologies were used.

The neurogenic stuttering of PD was mainly characterized by incomplete syllable repetitions and prolongations. In general, these findings are in accordance with previous research [7]. However, our results are not in agreement with a study by Benke et al. [6], as we did not observe a hyperfluent, formally resembling palilalia, type of dysfluency. This discrepancy may be due to differences in clinical data, as patients in the previous study were generally in the advanced stages of PD with frequent on-off fluctuations, while the present study was focused on patients in the early to middle stages of PD with no orofacial dyskinesia. Notably, incomplete syllable repetitions and prolongations of sounds are typical manifestations related to freezing of speech, which has been hypothesized to have a shared pathophysiology with freezing of gait [13].

Admittedly, this study is accompanied by several limitations. First, only half of our PD patients exhibited a greater level of dysfluency that HC subjects and, in general, the severity of dysfluency was greater in pDS compared to PD group. Furthermore, our PD patients were treated by levodopa medication which was reported to influence stuttering [2,3]. Therefore the effect of medication and different severity levels cannot be excluded. Last but not least, our PD subjects were significantly older than pDS subjects, thus the effect of normal aging process cannot be ruled out.

V. Conclusion

Our study demonstrates that there seems to be more differences rather than similarities between neurogenic and developmental stuttering. Further studies are warranted in large, representative samples of patients with similar dysfluency severity to verify if the proposed method has potential to become a helpful tool for the differentiation of various stuttering subtypes.

ACKNOWLEDGEMENT

This study was supported by the Czech Ministry of Health (AZV CR 15-28038A) and Czech Science Foundation (GACR 102/12/2230).

REFERENCES

[1] P.A. Alm, “Stuttering and the basal ganglia circuits: a critical review of possible relations,” *J. Commun. Disord.*, vol. 37, pp. 325–369, 2004.
[2] J.C. Wu, G. Maguire, G. Riley, A. Lee, D. Keator, et al., “Increased dopamine activity associated with stuttering,” *Neuroreport*, vol. 8, pp. 767–770, 1997.
[3] T. Tykalova, J. Rusz, R. Cmejla, et al., “Effect of dopaminergic medication on speech dysfluency in Parkinson's disease: a longitudinal study,” *J. Neural Transm.*, 2015; in press.
[4] F.L. Darley, A.E. Aronson, J.R. Brown, “Differential diagnostic patterns of dysarthria. *J. Speech Hear. Res.*,” vol. 12, pp. 246–269, 1969.
[5] J.A. Logemann, H.B. Fisher, B. Boshes, E.R. Blonsky, “Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients,” *J. Speech Hear. Res.,* vol. 43, pp. 47–57, 1978.
[6] T.H. Benke, C. Hohenstein, W. Poewe, B. Butterworth, “Repetitive speech phenomena in Parkinson's disease,” *J. Neurol. Neurosur. Ps.,* vol. 69, pp. 319–325, 2000.
[7] A.M. Goberman, M. Blomgren, E. Metzger, “Characteristics of speech disfluency in Parkinson disease,” *J. Neurolinguist*, vol. 23, pp. 470–478, 2010.
[8] L. Hartelius, “Incidence of developmental speech dysfluencies in individuals with Parkinson's disease,” *Folia Phoniatr. Logo.*, vol. 66, pp. 132–137, 2014.
[9] J. Shahed, J. Jankovic, “Re-emergence of childhood stuttering in Parkinson's disease: A hypothesis,” *Mov. Disord.*, vol. 16, pp. 114–118, 2001.
[10] T. Lustyk, P. Bergl, R. Cmejla, “Evaluation of disfluent speech by means of automatic acoustic measurements,” *J. Acous. Soc. Am.*, vol. 135, pp. 1457–1468, 2014.
[11] K. Teesson, A. Packman, M. Onslow, “The Lidcombe Behavioral Data Language of Stuttering,” *J. Speech Lang. Hear. R.*, vol. 46, pp. 1009–1015, 2003.
[12] J. Borsel, C. Taillieu, “Neurogenic stuttering versus developmental stuttering: An observer judgement study,” *J. Commun. Disord.*, vol. 34, pp. 385–395, 2001.
[13] H.K. Park, J.Y. Yoo, M. Kwon, J.H. Lee, S.J. Lee, et al., “Gait freezing and speech disturbance in Parkinson's disease,” *Neurol. Sci.*, vol. 35, pp. 357–63, 2014.

FP – Emotions-Therapy

A SPECTRAL ANALYSIS OF F0-CONTOURS IN BIPOLAR PATIENTS

A. Guidi[1, 2], J. Schoentgen[3], G. Bertschy[4], C. Gentili[5], E. P. Scilingo[1, 2], N. Vanello[1, 2]

[1]Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy
[2]Research Center "E. Piaggio", University of Pisa, Pisa, Italy
[3]L.I.S.A. – Signals, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium
[4]Dept. of Psychiatry, University Hospital and University of Strasbourg, INSERM u1114, Strasburg, France
[5]Dept. of Surgical, Medical, Molecular Pathology and Critical Care, Univ. of Pisa, Pisa, Italy
andrea.guidi@for.unipi.it, jschoent@ulb.ac.be, claudio.gentili@med.unipi.it, e.scilingo@centropiaggio.unipi.it, nicola.vanello@iet.unipi.it

Abstract: **Mental diseases are increasingly common. Among these, bipolar disorders heavily affect patients' lives given the mood swings ranging from mania to depression. Voice has been shown to be an important cue to be investigated in this kind of diseases. In fact, several speech-related parameters were used to characterize voice in depressed people. The goal is to build a decision support system (DSS) improving diagnosis and possibly predicting mood changes. In the literature several works have been conducted regarding depression. Lately, some efforts were devoted to studies concerning bipolar patients. Here a spectral analysis of F0-contours extracted from audio recordings of text reading will be performed. The algorithm is completely automatic so that it can be easily integrated into a DSS. The proposed features are related to both speech rhythm and intonation. Analyses on both bipolar and healthy subjects are reported. The former ones were recorded while subjects were experiencing different mood states, while the latter were recorded at different days. Some coherent features trends are detected in bipolar patients across different mood states, while no significant differences are highlighted in healthy subjects. Preliminary results indicate that the proposed features could be fruitfully explored to characterize mood states in bipolar patients.**

Keywords: **bipolar disorder, mood state, voice analysis, fundamental frequency, spectral analysis**

I. Introduction

Mental illnesses have an increasing impact in contemporary society. In particular, the lives of persons suffering from bipolar disorder may be impaired, due to periodic, and sometimes extreme, mood swings. Patients may experience oscillations between depression, mania, euthymic condition in which symptoms are mostly absent, or mixed condition, which associates depressive and manic symptoms. The development of decision support systems could be very useful in helping physicians in formulating diagnosis. With this aim some work has been conducted on biomedical signal processing to detect physiological correlates of mood changes [1, 2]. Moreover, several studies have focused on possible relations between voice and mental disease, especially in persons suffering from depression. Several speech-derived features have been shown to differ in such patients from healthy subjects [3, 4]. Specially, speaking rate has been found to negatively correlate with the Hamilton Depression Rating Scale score [3]. With regard to bipolar patients, an intra-subject study [5] has reported significant differences in speech fundamental frequency (F0) variability and average among different mood states. Moreover, in [6] the speech intonation contour has been found to be a reliable indicator of mood changes from a euthymic to an either depressed or manic state. Despite the relevance of the results, several limitations have been observed. Particularly, the direction of the features changes was not coherent across subjects. Better consistency may be achieved both by improving subject status characterization, e.g. by evaluating anxiety level [6], and by investigating other features. In this work a spectral analysis of the F0 contour is proposed to investigate differences in mood states in patients suffering from bipolar disorder. Patients have been recorded reading a neutral text at several days. In addition, a study on healthy control subjects is presented. Preliminary results are reported and discussed.

II. Methods

A. Experimental protocol and data

Eight patients (6 males and 2 females, 42.00 ± 9.95 years) suffering from bipolar syndrome were recruited for this study. The experimental protocol, approved by the clinical ethical committee, consisted in the reading of a neutral text during each session. Recording sessions were performed at two or three different days. Six patients out of eight were recorded twice each day. A physician labeled the patient's mood status before each recording using clinician administered rating

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

scales. Four different mood states were identified in this study, namely depressed, euthymic, hypomanic and mixed. A high quality directional microphone was used to record signals at a sampling frequency of 48 KHz and with a resolution of 32 bits.

Ten healthy control subjects (6 males and 4 females, 30.00 ± 5.00 years) were recorded twice at two different days to test for inter-day variability. Typically, the second session was recorded 7 days after the first one.

B. Algorithm

In a first step, voice activity detection (VAD) is carried out by means of autocorrelation coefficients and speech energy as described in [7]. Later, the F0 contour is estimated within voiced segments by means of Camacho's Swipe' algorithm [8]. A cubic spline interpolation is used to obtain F0 in unvoiced segments, while F0 within silent pauses is set to 0Hz. Finally a set of 7 features is extracted from the spectrum of each mean-subtracted F0-contour. Power spectral density is estimated from each recording using the periodogram. The features are: median frequency (F_{median}), power amplitude at the median frequency (A_{median}), maximum peak power amplitude (A_{peak}), and the corresponding frequency (F_{peak}), the ratios between amplitudes and corresponding frequencies, and Slope according to (eq. 1-3) (Fig. 1).

$$Ratio_{peak} = A_{peak}/F_{peak} \quad (1)$$

$$Ratio_{median} = A_{median}/F_{median} \quad (2)$$

$$\text{Slope} = \frac{A_{peak}\text{-}A_{median}}{F_{peak}\text{-}F_{median}} \quad (3)$$

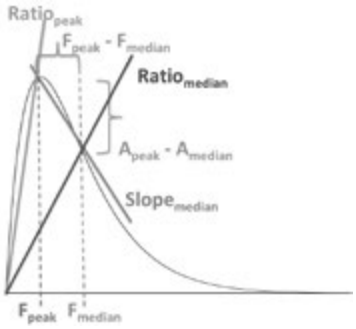


Figure 1: Scheme of extracted features.

C. Statistical analysis

When the same mood state had been recorded twice for the same speaker, the features extracted from the two audio recordings, were averaged. Thus, for each subject and for each mood state one sample for each feature is estimated. Friedman's test was used to check for statistical differences in paired data corresponding to different mood states (i.e., same patients, different mood states), while Mann-Whitney U-test was used to investigate such differences for independent samples (i.e., different patients and different mood states). The latter test was performed with and without normalization with respect to the same patient's features estimated from euthymic state data. In every test, a p-value lower than or equal to 0.05 was considered significant.

III. RESULTS

The proposed features showed a similar behavior across all subjects. Specifically, F_{peak} was always lower than F_{median} thus resulting in a negative Slope parameter and in a $Ratio_{peak}$ that was consistently higher than $Ratio_{median}$.

Analysis of the data recorded for healthy control subjects did not return statistically significant differences between features obtained from audio samples acquired at two different days. In table 1 the relative p-values are reported, while in Fig. 2 F_{median} trends in healthy control subjects are displayed.

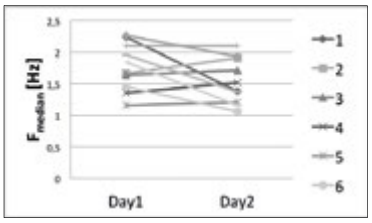


Figure 2: F_{median} trends in healthy control subjects.

Table 1: Healthy control subjects: p-values.

F_{median}	A_{median}	F_{peak}	A_{peak}	Slope	$Ratio_{peak}$	$Ratio_{median}$
7,39E-01	1,00E+00	2,06E-01	5,27E-01	5,27E-01	1,00E+00	5,27E-01

Table 2: Patients' mood label.

	Day 1	Day 2	Day 3
1	Hypomania	Euthymia	
2	Hypomania	Euthymia	Depression
3	Hypomania	Euthymia	
4	Depression	Euthymia	
5	Depression	Euthymia	
6	Hypomania	Euthymia	
7	Depression	Euthymia	
8	Mixed	Depression	Euthymia

Each bipolar patient reported a euthymic state in one of the three recording days (table 2). By exploring such table it is possible to select the subjects that can be used for paired and independent data tests. On average the reading recordings of bipolar patients lasted about 4 minutes. Analysis of paired data (table 3) showed statistically significant differences between hypomania and euthymia states (patients 1, 2, 3 and 6) for A_{peak}, F_{peak}, $Ratio_{peak}$ and their Slopes.

In all subjects but one (patient 3), F_{peak} (Fig. 3) showed a lower value in the hypomanic state, while for all the subjects A_{peak} (Fig. 3) and $Ratio_{peak}$ (Fig. 4)

showed a higher value in the hypomanic state compared to the euthymic one. Opposite trends were observed for the Slope features (Fig. 5).

Table 3: Bipolar patients: p-values. Significant p-values, lower than 0.05, are highlighted in bold.

	F_{median}	A_{median}	F_{peak}	A_{peak}	Slope	$Ratio_{peak}$	$Ratio_{median}$
Hyp Vs. Eut	3,17E-01	3,17E-01	**4,55E-02**	**4,55E-02**	**4,55E-02**	**4,55E-02**	3,17E-01
Dep Vs. Eut	**2,53E-02**	6,55E-01	6,55E-01	**2,53E-02**	**2,53E-02**	6,55E-01	1,80E-01

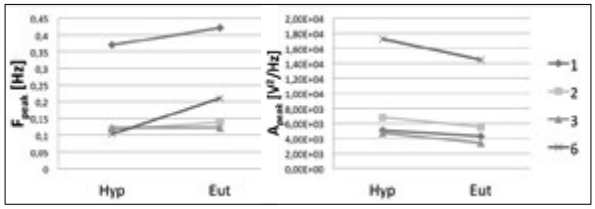


Figure 3: F_{peak} (left) and A_{peak} (right) trends in patients passing from hypomania to euthymia.

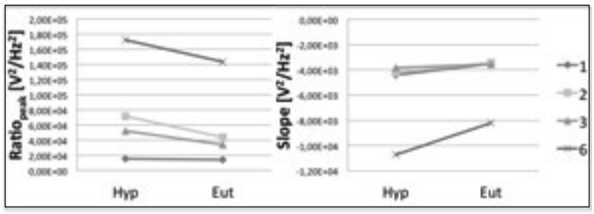


Figure 4: $Ratio_{peak}$ (left) and Slope (right) trends in patients passing from hypomania to euthymia.

Moreover, further analysis on paired data (patients 2, 4 5, 7 and 8) showed that differences between depression and euthymia states were statistically significant for F_{median}, A_{peak} and Slope. For all subjects, F_{median} (Fig. 5) and Slope were lower for the depressed state, while A_{peak} (Fig. 6) was higher.

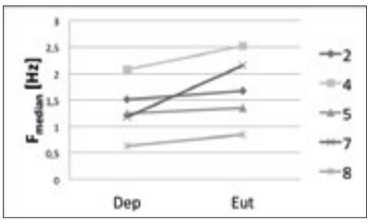


Figure 5: F_{median} trends in patients passing from depression to euthymia.

Comparisons carried out via the Mann-Whitney U-test on unpaired normalized data between depression and hypomania and depression and euthymia, showed statistically significant differences for F_{median} (Fig. 7) and Slope. In addition A_{peak} reported significant results just for the depression-hypomania comparison. F_{median} and Slope for the depressed state were lower with respect to the other mood states, while A_{peak} was higher. Without normalization, the proposed features did not show any statistically significant differences.

In Mann-Whitney U-test different data groups were formed by the features extracted from different patients' audio in different mood states.

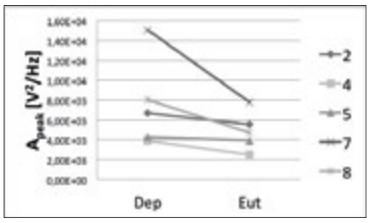


Figure 6: A_{peak} trends in patients passing from depression to euthymia.

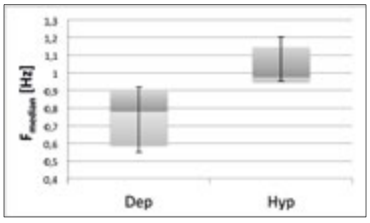


Figure 7: Boxplot of F_{median} in patients passing from depression to hypomania. F_{median} values are normalized with respect the corresponding values in euthymic state.

IV. Discussion

In this work an automatic spectral analysis of the F0-contours is carried out. Conventionally, the F0-contour is studied in the time domain. We believe that an analysis in the frequency domain can provide a compact description of the F0-related prosodic information. Specifically, the proposed features are related to the shape of F0 spectrum profile. Since F0 in silent segments was set to zero, the proposed features summarize the contribution of rhythm as well as of intonation. Specifically, results depend not only on syllabic rhythm (4Hz typically), but also on pauses between words and sentences.

The statistical analysis was performed on bipolar patients experiencing different mood states, and on control subjects. Statistically significant differences were found between features across different mood states. Interestingly, the proposed features showed a good specificity, whereas they are similar for control subjects.

Notwithstanding the small number of patients who have been analyzed, we think that the results may be relevant because coherent feature trends have been detected in patients across mood states. Due to the sample size, it was not possible to perform any statistical test on paired data with regard the comparison between depression and hypomania.

As regards the paired data analysis, the comparison of euthymia and depression, showed that A_{peak}

increases and F_{median} decreases in depressed state with respect to euthymic stated. Since F_{peak} is lower than F_{median} this behavior indicates a higher contribution at lower frequency in depressed state. The paired analysis of hypomanic and euthymic states revealed a significant decrease of F_{peak} and an increase of A_{peak} in the former state, while no coherent change of F_{median} was observed. Moreover, a decrease of Slope parameter was found. These results show a behavior of the proposed features that could possibly differentiate hypomania from depression. Preliminary results on independent samples seem to confirm this hypothesis. In fact, a decrease of F_{median} and Slope was reported in depressed with respect to hypomanic patients, thus indicating a higher contribution at lower frequencies of the F0 profile spectra in the former subjects. The significant results, obtained from the statistical tests on independent samples, were reached after normalizing the feature values by the corresponding value in the euthymic state. This normalization was performed under the hypothesis that euthymia represents the emotional point of reference since it is characterized by the absence of relevant symptoms.

It is important to highlight that the choice of the text can play a crucial role both as regards content and structure. In fact, a specific content might elicit an emotional response. Moreover, since the reading rhythm is one of the parameters under study, it is important to use text with similar or equal lexical structure. In this work, a neutral text was adopted, i.e. "*The universal declaration of human rights*" for the different recording sessions.

V. Conclusion

This work has proposed a spectral analysis of F0-contours to characterize mood changes in bipolar patients. The limited number of enrolled patients does not allow generalizing the result, but anyway we believe that interesting indications can be drawn from this study. In fact, the obtained results showed that some significant differences could be found analyzing reading of neutral text by bipolar patients recorded while they were experiencing different mood states. Interestingly, coherent feature trends were reported in enrolled patients across mood states. On the contrary, no statistically significant differences were found investigating neutral readings in healthy control subjects. Such results highlight that the spectral analysis of rhythm of speaking and intonation can characterize different mood states.

Integration of such approach with other ones could allow reaching new important results in the mood recognition research field. Especially, a jointed investigation of such proposed approach with energy-related or micro-prosodic features could be very informative.

Finally, the proposed approach is completely automatic and could be easily integrated in a decision support system to help clinicians in the difficult task of making diagnosis and tailoring treatments.

Acknowledgements

This research is partially supported by the EU Commission under contract ICT-247777 Psyche.

References

[1] G. Valenza, C. Gentili, A. Lanatà, and E.P. Scilingo. "Mood recognition in bipolar patients through the PSYCHE platform: preliminary evaluations and perspectives." Artificial intelligence in medicine 57, 1, 2013, pp. 49-58.

[2] O. Mayora, B. Arnrich, J. Bardram, C. Drager, A. Finke, M. Frost, S. Giordano et al. "Personal health systems for bipolar disorder Anecdotes, challenges and lessons learnt from MONARCA project." In Pervasive computing technologies for healthcare (PervasiveHealth), 2013 7th international conference on, pp. 424-429. IEEE, 2013.

[3] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder. Voice acoustical measurement of the severity of major depression. Brain and cognition, 56, 1, 2004, pp. 30-35.

[4] K.E.B. Ooi, M. Lech, and N.B. Allen. Multichannel weighted speech classification system for prediction of major depression in adolescents. Biomedical Engineering, IEEE Transactions on, 60, 2, 2013, pp. 497-506.

[5] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanatà, and E.P. Scilingo. Speech analysis for mood state characterization in bipolar patients. In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pp. 2104-2107, IEEE.

[6] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E.P. Scilingo. "Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients." Biomedical Signal Processing and Control (2014).

[7] J.L Blanco, J. Schoentgen, and C. Manfredi. Vocal tract settings in speakers with obstructive sleep apnea syndrome. In Proc. 8th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications, 2013, pp. 211-214, Firenze University Press.

[8] A. Camacho, and J.G. Harris. "A sawtooth waveform inspired pitch estimator for speech and music." The Journal of the Acoustical Society of America 124, 3, 2008, pp. 1638-1652.

ASSESSMENT AND PSYCHOACOUSTIC MODELLING OF AUDITORY STREAMS IN DIPLOPHONIC VOICE

P. Aichinger[1], B. Schneider-Stickler[1], W. Bigenzahn[1], M. Hagmüller[2], A. Sontacchi[3], J. Schoentgen[4]

[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
[3] Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria
[4] Department of Signals, Images and Acoustics, Faculty of Applied Sciences, Université Libre de Bruxelles, Belgium
philipp.aichinger@meduniwien.ac.at, berit.schneider-stickler@meduniwien.ac.at, wolfgang.bigenzahn@meduniwien.ac.at, hagmueller@tugraz.at, sontacchi@iem.at, jschoent@ulb.ac.be

Abstract: **Auditory diplophonia is the simultaneous presence of two pitches in disordered voice. It is used as treatment indicator and treatment effect descriptor in clinical practice and thus needs objectification. A procedure for computational auditory scene analysis is proposed. The procedure is organized into signal segregation and mutual masked loudness modelling. The functionality of the procedure is demonstrated on a representative example of diplophonic voice. The model's accuracy is evaluated as compared to humanly drawn loudness curves on a corpus of ten diplophonic sustained phonations. The predictor model performance is better than guessing but limited, reasons for which are discussed.**

Keywords: Auditory diplophonia, auditory streams, psychoacoustic modelling, loudness, pitch salience

I. INTRODUCTION

Auditory diplophonia is an often misunderstood sign of severe voice disorders. Its presence triggers clinical actions and is used to assess the effectiveness of treatment techniques. Its perception is underresearched and hardly understood. Auditory diplophonia is the presence of two simultaneous pitches, which is ambiguous according to well-known psychoacoustic principles of pitch perception [1], [2]. Thus, more robust descriptors of perceptual aspects of diplophonic voice are needed.

Diplophonic sound signals are mixtures of harmonic complex tones, its sinusoidal intermodulation effects and noise. Each sinusoidal component of the sound may provoke a pitch sensation, and also noise may be perceived tonal, depending on its spectral envelope. Pitches corresponding to spectral maxima are called "spectral pitches", whereas "virtual pitches" arise at greatest common divisors of the spectral pitches' frequencies. In diplophonic voice detection, spectral pitches are more often decisive than virtual pitches. Pitch salience can be computed from signals' spectra, and more than one pitch exists for most signal types [3]. A definition of diplophonia that is compatible with the concept of pitch salience is the following: "Diplophonia is the presence of exactly two pitches with above-threshold salience". Although the threshold is unknown and may depend on some additional factors, this definition enables scientific descriptions of the diplophonia phenomenon.

"Psychophysics provide tools to insure that perceptual phenomena are real and general and codified in a form that can be shared" [2], which is of utmost importance in voice science because pure introspection for probing the reality of perceptual phenomena cannot be communicated reliably. The use of external references, such as reference stimuli or perceptual models, would help calibrate what one hears and make comparisons possible. Generality, reproducibility and credibility of perceptual observations may be increased to foster the quality of scientific and clinical communication. Thus, subjectivity and disagreements may be reduced [2].

The circumstances under which humans hear out pitches from mixtures of harmonic complex tones and noise are not fully understood, because studies on the psychoacoustics of competing pitches are sparse [2]. Factors that may play a role include but are not limited to mutual energetic and informational masking of the distinct perceptual streams, peripheral harmonic resolvability as well as listener related covariates (experience, musical training, attention).

In this paper the loudness curves of humanly segregated auditory streams are predicted in a model-based automatic way. The model's functionality is demonstrated on a representative phonation example, and time-average prediction errors are evaluated for a corpus of ten diplophonic phonations.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

II. Material and Methods

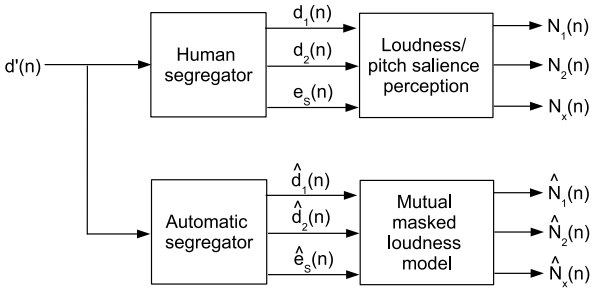


Figure 1: Block diagram of auditory stream segregation and consequent perceptual weighting. The natural case and the model prediction are shown.

The analyzed recordings stem from a database of 120 subjects (40 euphonic, 40 dysphonic/non-diplophonic, 40 dysphonic/diplophonic) [4]. The audio recordings of sustained phonation have been obtained during rigid telescopic high-speed laryngoscopy. A headworn microphone AKG HC 577 L with windscreen AKG W77 MP and its original cap (no presence boost) was used. It was connected via a phantom power adapter AKG MPA V L (linear response setting) to a digital portable recorder TASCAM DR-100. The sampling frequency and the quantization resolution were set to 48 kHz and 24 bits.

Ten audio fragments of diplophonic voice were selected. The momentary presence of diplophonia was here defined as the presence of two pitches, or a distinct impression of beating. If the perceptual decision was doubtful, waveforms and spectrograms were visually inspected for the presence of metacycles or subharmonics [4]. Additionally, only fragments for which fundamental frequency extraction is correct were used, because the automatic segregation procedure relies on it. The analyzed audio fragments were between 0.18 and 1.24 s long.

The audio recordings are put into an automatic segregator based on waveform modelling [5]. The segregator finds the optimal waveform model of the recorded signal comprised of at most two harmonic oscillators $\hat{d}_1$ and $\hat{d}_2$. The algebraic difference of the natural signal and the harmonic oscillators is the residual signal $\hat{e}_S$ that contains the natural signal's inharmonic/noise components.

The automatically segregated signals are put into a psychoacoustic model for mutual masked loudness [6]–[9], which predicts their loudness curves. $\widehat{N}_1$ is the loudness curve of $\hat{d}_1$, $\widehat{N}_2$ of $\hat{d}_2$ and $\widehat{N}_x$ of $\hat{e}_S$. Each of the signals is masked by the supplementary sum of the others. The model considers outer ear filtering, middle ear filtering, cochlear filtering, masking threshold calculation, calculation of loudness over frequency, as well as binaural and temporal integration.

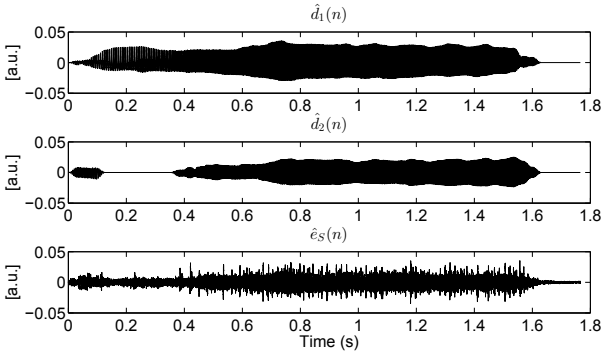


Figure 2: Waveforms of automatically segregated auditory streams of a representative example of diplophonic phonation.

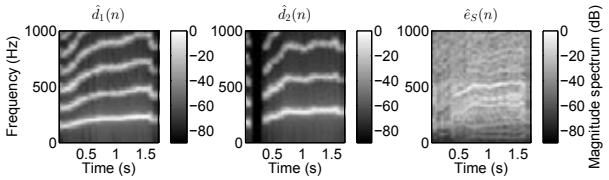


Figure 3: Spectrograms of automatically segregated auditory streams of a representative example of diplophonic phonation.

Fig. 1 shows the block diagram of the segregation and perceptual weighting. The humanly segregated signals d_1, d_2 and e_S are hidden variables. Only the percepts of their internal representations are observable via human reporting. To evaluate the segregation model and the loudness model, the predicted loudness curves are compared to loudness curves drawn by a listener. The recorded audio signals were presented via AKG K 271 MK II headphones. A 1 kHz reference sine tone preceded the recorded audio. It was adjusted to a predicted loudness of 10 sone, i.e. a sound pressure level of 73.2 dB [10]. Its duration was 1 s, with 0.1 s linear fades at the onset and offset. The recorded audio signals were adjusted to an average sound pressure level of 70 dB. The listener was allowed to listen to the reference tone, the recorded audio and the automatically segregated signals arbitrarily often. The listener drew the perceived loudness curves N_1, N_2 and N_x obtained from the recorded signal onto scale paper by hand. Time was reported on the x-axis (5 cm/s) and loudness on the y-axis (0.5 cm/sone). The durations of the stimuli were known to the listener and the model predictions were unknown.

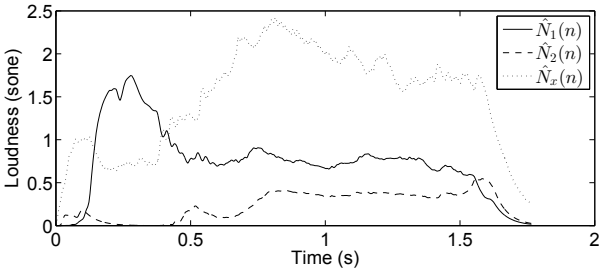


Figure 4: Example of predicted loudness curves.

The drawn loudness curves were subtracted from the model predictions in sampling steps of 0.1 s, which is a typical time constant of auditory integration. The root mean squared (RMS) prediction error has been obtained for each audio fragment.

To test statistical significance of the results, the RMS prediction error is compared to a random numbers control, i.e. the model is tested for being better than guessing. The random numbers are uniformly distributed between 0 and 10 sone.

III. Results

Fig. 2 shows the harmonic oscillator waveforms $\hat{d}_1$ and $\hat{d}_2$, as well as the residual signal $\hat{e}_S$ of a representative audio fragment. The fragment organizes into four distinct time intervals. At onset and between 0.35 and 1.65 s the voice is diplophonic, thus all three streams are active. Between approximately 0.15 and 0.35 s, only one harmonic oscillator is active, i.e. $\hat{d}_2$ is inactive. The offset is unvoiced, i.e. $\hat{d}_1$ and $\hat{d}_2$ are inactive after 1.65 s.

Fig. 3 shows the spectrograms. All four distinct time intervals can be recognized. One can identify the harmonic structures of $\hat{d}_1$ and $\hat{d}_2$, i.e. partials exist at the fundamental frequencies and their integer multiples. $\hat{e}_S$ is a mixture of sinusoidal components and noise. Sinusoidal components arise from coupling and modulation of glottal oscillators. They lie at linear combinations of the oscillators' partials.

To investigate the validity of the automatic segregator, the waveforms are auditorily compared to humanly segregated waveforms of which only the percepts of their internal representations are accessible via human reporting. The sound files d1+d2+e.wav, d1.wav, d2.wav, e.wav, d1+d2.wav, d1+e.wav and d2+e.wav can be downloaded [11]. The file d1+d2+e.wav contains the original signal. d1.wav and d2.wav contain the segregated waveforms of the low and the high pitched oscillators. Their musical pitch interval is approximately a major third. d1+d2.wav contains the sum of the segregated harmonic oscillators. d1+e.wav contains the original signal from which the high-pitched diplophonic component has

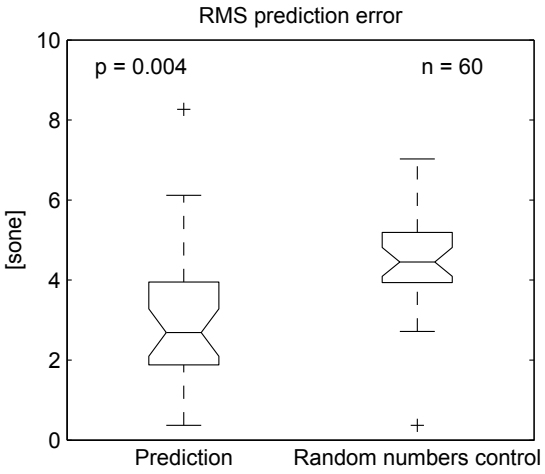


Figure 5: RMS prediction error $\sqrt{\overline{(\hat{N}-N)^2}}$.

been removed. The perceived degree of diplophonia is decreased. d2+e.wav contains the original signal from which the low-pitch oscillator has been removed. Compared to the original signal, the high pitch auditory stream is perceived as louder, because it is not masked by the low-pitch stream. e.wav contains the residual signal that sounds noisy, hoarse and harsh. In the residual signal, one therefore hears a pitch that is not clearly audible in the original sound, because it is suppressed by $\hat{d}_1$ and $\hat{d}_2$.

The loudness curves of the auditory streams of the original signal have been predicted from the automatically segregated signals and compared to the perceived loudness of the original. Fig. 4 shows examples of predicted loudness curves. Again, four temporal intervals are distinguished: At onset the residual signal is the loudest (dotted line), the secondary high-pitch oscillator is less loud at 0.15 sone (dashed line) and the primary low-pitch oscillator is suppressed (solid line). In the interval in which $\hat{d}_2$ is inactive (0.15 – 0.4 s) the primary low-pitch oscillator is the loudest of the three signals. The loudness of the residual signal is approximately 0.8 sone, and the high-pitch oscillator is absent. After 0.4 s the residual signal is dominant. The loudness of the primary high-pitch oscillator is at approximately 0.8 sone and the loudness of the secondary low-pitch oscillator at approximately 0.4 sone. At phonation offset all curves decay. The secondary high-pitch oscillator briefly exceeds the loudness of the primary low-pitch oscillator. The predicted loudness curves appear to be plausible compared to perceived loudness of humanely segregated auditory streams.

Loudness curve predictions of the whole corpus are evaluated by comparing them to loudness curves drawn by a human listener. Fig. 5 shows the boxplot of the RMS prediction error as compared to the random

number control. The prediction error is smaller than the hypothetical error that would have occurred if the predictor would have put out random numbers that were uniformly distributed between 0 and 10 sone. The two distributions differ significantly, as tested by a two-sided paired Wilcoxon rank sum test. This result suggests that the numbers that are put out by the proposed predictor are better than guessing. The agreement between curve drawings and predictions suggests that a real perceptual effect exists. However, the model prediction errors are large, reasons for which are discussed hereafter.

IV. Discussion and Conclusion

A model for predicting loudness curves of distinct auditory streams in diplophonia has been proposed. Its functionality has been demonstrated on a representative example of diplophonic phonation. The model's accuracy as compared to humanly drawn loudness curves was evaluated on a corpus of ten phonations.

The results contribute to the understanding of the auditory perception of diplophonic voice. It has been demonstrated that auditory streams of diplophonic phonation that can be segregated by humans can also be segregated automatically. For the presented example, the perceptual model predicts the perceived loudness plausibly.

The measured accuracy of the model predictions is limited. Possible reasons are twofold. First, the loudness curves obtained from the listening test may not represent the true magnitude of the percept. A limitation with regard to the listening test procedure is that more accurate approaches may exist. The direct comparison of natural stimuli with the reference 1 kHz tone is difficult and may probably be replaced by a more expensive experimental design e.g. an adjustment procedure or a staircase method. Additionally, the use of stationary synthetic stimuli may increase the precision of the loudness ratings acquired. Second, the model's validity may be limited, reasons for which are summarized as follows. The automatic segregator relies on the model assumptions that the harmonic oscillators are independent, uncorrelated and unmodulated. However, these assumptions are not totally true for natural signals. In addition, the used loudness model is a model for noise maskers and its validity may be limited for harmonic maskers. In particular, no perceptual interaction between the target and the masker is considered. An important effect that cannot be modelled is that a harmonic complex tone that is added to another one causes temporal irregularities. Those may mask the two pitches informationally and be a detection cue (i.e. beating) at the same time. Perceptual magnitudes of two added harmonic complex tones do not add up linearly, an observation of which the model is not capable.

A practical consequence that can be inferred from the described procedure is that automatically segregated sounds may be used for ear training of phoniatricians and logopedists, who have difficulties identifying individual pitches in diplophonic voice. Thus, human segregation may be trained with the aid of an automatic segregator.

References

[1] A. de Cheveigné, "Pitch Perception Models," in *Pitch*, vol. 24, C. Plack, R. Fay, A. Oxenham, and A. Popper, Eds. Springer New York, 2005, pp. 169–233.

[2] C. Plack, Ed., *Oxford Handbook of Auditory Science: Hearing*. 2010.

[3] E. Terhardt, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.

[4] P. Aichinger, "Diplophonic Voice - Definitions, models, and detection," Ph.D. dissertation, Graz University of Technology, 2015.

[5] P. Aichinger, M. Hagmüller, I. Roesner, W. Bigenzahn, B. Schneider-Stickler, J. Schoentgen, and F. Pernkopf, "Measurement of fundamental frequencies in diplophonic voices," in *Proceedings of the 9th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2015.

[6] P. Aichinger, A. Sontacchi, and B. Schneider-Stickler, "Describing the transparency of mixdowns: The Masked-to-Unmasked-Ratio," in *130th Audio Engineering Society Convention London*, 2011.

[7] B. Moore, B. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[8] B. Glasberg and B. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[9] B. Moore and B. R. Glasberg, "Modeling binaural loudness," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, p. 1604, 2007.

[10] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. 2007.

[11] P. Aichinger, "Audio examples of polynomial modeling of a diplophonic waveform." [Online]. Available: www.meduniwien.ac.at/phon/public/aichinger/thesis/chapter4.zip.

SCHIZOPHRENIA AND PROSODY. FIRST INVESTIGATIONS

E. Cresti[1] , F. M. Dovetto[2], B. Rocha[3]

[1] LABLITA – Università di Firenze, Firenze, Italia
[2] Università Federico II, Napoli, Italia
[3] CAPES Foundation, Ministry of Education of Brazil, Brasília, Brazil

[1] dovetto@unina.it, [2] emanuela.cresti@unifi.it, [3] bbruno791@gmail.com

Abstract: **This paper presents data from a research on the prosody of schizophrenic speech. Our research is based on the analysis of CIPPS corpus [13], with the criteria used being the same as that in the analysis of non-pathologic spontaneous speech data (Language into Act Theory, L-AcT [8]).**
In the schizophrenic speech under investigation the identification of utterances in the speech flow by way of prosodic terminal breaks is conserved.
On the contrary, differences between non-pathological speech and schizophrenic speech concern the pragmatic accomplishment of utterances and the organization of their information structures. Patient D presents a small variation in illocutionary types and the occurrence of information patterning is reduced and less varied in comparison with non-pathological speech. Moreover, in all patients under investigation there is a pervasive usage of *echolalia* and *palilalia*.

Keywords: **schizophrenia, prosody, CIPPS, L-AcT**

I. INTRODUCTION

The origin of the term schizophrenia dates back to the beginning of the last century [18, 5] and is used to refer to a set of illnesses characterized by psychological dissociation disorders. Bleuler in 1911 [5] defined it as a twofold kind of dissociation, both between the different parts of the Self and between ideation and affects. The literature on the topic is huge [1, 23, 3] and there is no agreement on its precise definition and extension [15].

However the literature does agree that many of its symptoms involve a linguistic component, even if what has been reported in this respect is for the most part derived from the diary notes of psychiatrists and the written documents of patients [6, 17]. To our knowledge, the analysis of large corpora of audio recordings in which the actual language performance of patients is documented is still lacking. In particular, until now we were lacking spoken corpora for schizophrenic patients with text-to-sound alignment providing both textual and acoustic dimensions, as is common in current spoken corpora compilations (e.g. Romance corpora: C-ORAL-ROM [11] ; C-ORAL-BRASIL [24]).

II. METHODS

Our research is based on the analysis of the CIPPS corpus [13]whichwascompiled in Naples (Scuola Sperimentale per la Formazione alla Psicoterapia e alla Ricerca nel Campo delle Scienze Umane Applicate – ASL Napoli 1) by Dovetto. It provides seventeen hours of dialogues between a psychiatrist and four patients diagnosed as schizophrenic, and is composed of:

- 3 sessions (2 h. 30 min.) with patient A, who has a pre-delirium condition (or *Wahnstimmung*) without hallucinations;
- 4 sessions (3 h. 58 min.) with patient B, who has paranoid schizophrenia, characterized by flight of ideas and delirium;
- 2 sessions (2h. 8 min.) with patient C, who has paranoid schizophrenia, characterized by megalomania and persecution delirium;
- 1 session (28 min.) with patient D, who has paranoid schizophrenia with delirium.

Ten hours of recordings have been transcribed according to the CLIPS format[7], and further portions have been transcribed and aligned to the sound according to the LABLITA format.The recordings comply with privacy and ethical requirements.

Generally speaking, schizophrenic speech isclassified as belonging to either thought disorders (flight of ideas, syntactic and semantic derailment, poverty of content, illogicalness, abstraction failure, redundancy) or disorders belonging to the mental content connected with the delirium.

Dovetto [13, 12] records para-etymologies, semantic manipulations, phonic associations, echolalia, and the occurrence of deictic expressions usually not reported by the literature [14].

In this research, the CIPPS corpus has been analyzed with the same criteria used for the analysis of non-pathologic spontaneous speech data. Specifically, the flow of speech has been segmented into utterances and information units based on prosodic cues [21], asin

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

the Language into Act Theory (L-AcT, [8]) framework, which is pragmatics-based and considers prosody in terms of its interfacing with linguistic content. The observations herein are taken from stretches of speech that have been aligned and analyzed using WinPitch [20] and PRAAT [4]. The results reported in this paper are mostly derived from patient D, who may be considered a symptomatic schizophrenic speaker.

III. RESULTS

In the schizophrenic speech being considered, data show that the general relation between information and prosody might be considered compatible with non-pathologic speech.

A very general aspect of speech concerns the ability to identify utterances in speech flow by way of prosodic terminal breaks. Generally speaking, the identification of the utterances (considered the reference units of speech) is preserved in all the observed CIPPS patients, while in other pathologies may not be the case (for instance in some aphasias).

This may be one reason for the assumption, reported sometimes in the literature, that there should be a similarity between schizophrenic speech and spontaneous normal speech. However, a deeper analysis carried out in accordance with L-AcT principles shows, crucially, that quantitative data and idiosyncrasies occur in the former which make it appear incompatible with normal speech.

Let see for instance (1) and (2), whose segments are transcribed according to the LABLITA format with diacritics for non-terminal prosodic breaks (/) and terminal ones (//), and tags for the different information functions (AUX for Discourse marker, TOP for Topic, COM for Comment, APC for Appendix, PAL palilalia, EMP empty unit).

(1) *PAB: io tenevo le idee // ^{COM} hai capito ? ^{COM}
quello è il problema// ^{COM}
[I had the ideas // (do) you understand ? that is the problem //]

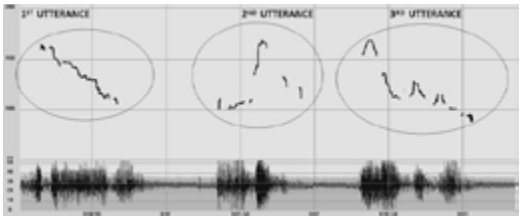


Fig. 1: Patient B utterances

(2) *PAA: come mi sento ? ^{COM} un &ope [/2] ^{EMP} e un bravo operante // ^{COM}
[how do I feel ? a &wor [/2] a good worker //]

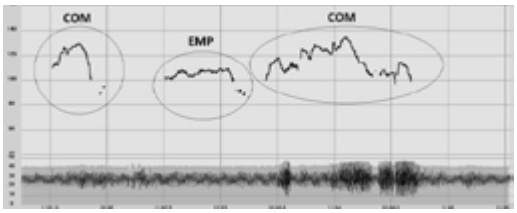


Fig. 2: Patient A utterances

If the prosodic identification of utterances may constitute a similarity with normal speech, this is not the case when the relation between prosody and both the pragmatic accomplishment of utterances and the organization of their information structure is considered.

Actually, spontaneous speech records a continuous and wide variation of communicative actions in strict correspondence with interactional needs (i.e. illocutions according to Austin [2]), but also to L-AcT [9,16]).The illocutionary variation occurs not only in the change of dialogue turns among the speakers, but crucially also within an individual turn of a speaker. Let us look at (3), where a student is wondering if the professor also needs a photocopy of something:

(3) *SUS: lei /^{TOP}gliene serve una anch‟a lei ?^{COM} una in più / o no ?^{COM} no //^{COM}lei ha questa //^{COM}
[you / (do) you need one also? one more / or not ? no // you have this one //]
%ill: [1] yes-no question; [2] alternative question; [3] self-answer; [4] ascertainment

The %ill layer indicates the different illocutions accomplished by the speaker.The link to the audio demonstrates the clear change of the communicative action, when a terminal break is perceived within the same turn of the above non-pathologic dialogue.

A second relevant aspect has to do with the information organization that may characterize the utterance. For this question we can refer to the IPIC database [22]developed by LABLITA and to the quantitative and qualitative data derived from the Italian mini-corpus therein (3544 utterances), which records that about 40% of the utterances are compound from an informational point of view, i.e. they correspond to an information composition made up of at least two information units.

The third point is that prosody is a necessary device both to express the illocutionary type of the utterance [9, 25] and to pattern its information structure [21]. Moreover, it must be stressed that the prosodic profiles expressing these functions are conventionally codified, the details of which fall outside the scope of this paper.

The rich variety of illocutions, the systematic information patterning of the utterances, and the conventionality of the prosodic profiles expressing them seem to all be lacking in patient D, or showing a

significant quantitative reduction and idiosyncrasies in the other patients.

For instance, the illocutions from the first 100 utterances of patient D correspond by more than 90% to assertive types, and are performed through idiosyncratic prosodic profiles, while normal speech records a rough average of 50% for this illocutionary class[16].Moreover, when some variations occur they do not match with interactional needs. See (5):

(5) *PAD: ‟n alber‟ viola //^{COM} ‟na casa verde //^{COM} e ‟na persona umana//^{COM}
[a purple tree // a green house // and a human person //]
%ill: [1] assertion; [2] contrast; [3] emphatic assertion

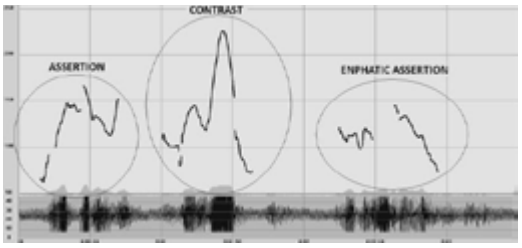


Fig. 3: Set of assertive illocutions (Assertion, Contrast, Emphatic Assertion)

In (5), D is describing his own drawing and performs a sequence of three verbless utterances, each accomplishing various kinds of assertive illocutions. They are characterized by some formal variations, but these are not motivated by any contextual conditions, as for instance in the occurrence of theContrast illocution. Moreover patient D's prosodic profiles are characterized by features typical of the schizophrenic "voice" (*paraphonia*, *hyperphonia*, *paraphasia*).

With regard to information patterning, its frequency is quantitatively reduced and less varied if compared with that employed in normal speech. For instance the Topic information unit (i.e. the basic way to organize the information structure) occurs in nearly 39% of compound utterances in normal speech, while in the first 100 utterances of patient D no Topic was found.Some Topicsappear onlyin the telling of a tale, as if they were derived from a model repeated by heart. Conversely, some examples of Topic usage are found, for instance, in patient B's dialogues and are performed with the proper conventional prosodic profile, although they are less frequent than in normal speech.

(6) *PAD: dopo quest‟incidente /^{TOP} tutto bene //^{COM}
[after this incident / everything is ok //]

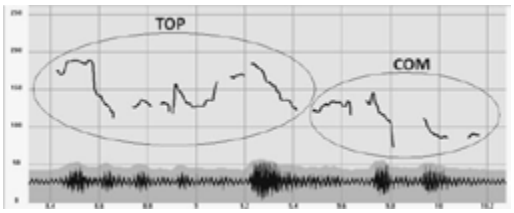


Fig. 4: Topic-Comment of patient B

Beyond these aspects, the emergence of a typical phenomenon in all patients under investigation must be stressed: a pervasive usage of *echolalia*. It is possible to find this phenomenon in connection to healthy and sane imitative learning, occurring in the first three years of life[10]. However, when this age is passed, the term *echolalia* is used to refer to one of the most significant aspects of communication disorder in psychiatry. It denotes the automatic repetition, without awareness, of short stretches of speech, implying the incorporation of another person‟s words into one's own. It happens mostly in the patient‟s answers which result in redundant speech parts taken from the question itself. This type of instance is largely present in the CIPPS corpus patients. Beyond *echolalia,* patient D systematically performs *palilalia,* i.e. the repetition of his own words. For instance, a sequence of *palilalias* may be observed in his short version of the Snow White tale:

(7) *PAD: cioè /^{AUX} Biancanev‟ /^{TOP} se trov‟ ‟in ‟na / strad‟ pien‟ ‟e neve //^{COM} neve //^{PAL}
[that is / snow white / is standing on a / street full of snow // snow //]

(8) *PAD: Biancane la &far //^{EMP} di Biancanev‟ /^{TOP} va ‟ncopp‟ /„o va al mar‟ //^{COM} ‟ncopp‟ ‟o mar‟ //^{PAL}
[Snow White it &do // of Snow White / goes to / to goes to the sea // to the sea //]

(9) *PAD: va a finire che Biancanev‟ trov‟ / ‟n ors‟ nel deserto //^{COM} ‟n orso //^{PAL}
[in conclusion / Snow White finds / a bear in the desert // a bear //]

(10) *PAD: che trov$_{PAL}$‟ ‟n alber‟ chin‟ ‟e ‟e nespole //^{COM} ‟e nespole //
[she finds a tree full of ofjapanese plums // of japanese plums //]

The prosodic performances of D‟s *palilalias* are strongly idiosyncratic, let see (11) and figure 5.

(11) *PAD: stev‟ yyyy//^{EMP} „e disegn‟ so‟ vvenut‟ duje /^{COM} ‟e sign‟ //^{APC} so‟ vvenut‟ duje //^{PAL}
[it was yyyy // the drawings have been two, the drawings // they have been two]

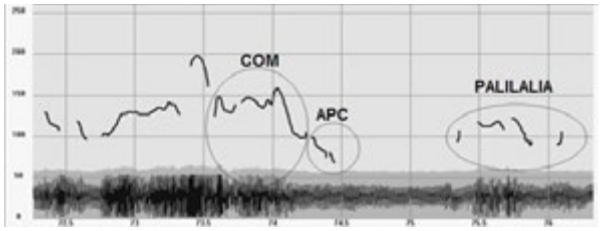


Fig. 5: An instance of *palilalia*for the schizophrenic patient D

Palilalias are performed by way of a prosodic unit clearly identified by terminal prosodic breaks and sometimes by a pause, like in this case. However, the presence of a terminal break does not match with the accomplishment of an illocution, as foreseen in normal speech. Its f_0 range is so low and its intensity so weak that no illocution can be interpreted.

Given its idiosyncratic prosodic profile and its null functional role, this phenomenon appears as a peculiar schizophreniccommunication strategy. Insteadof giving information to his addressee, the patient repeats his own words because it is necessary for him to collect his thoughts before proceeding. This may depend on the basic difficulties of the schizophrenic subject in establishing a dynamic human relationship.

REFERENCES

[1] Andreasen, N.C. 1986. *Scale for the Assessment of Thought, Language, and Communication (TLC)*, in «Schizophrenia Bulletin», 12, pp. 473–482.

[2] Austin, L.J. 1962. *How to Do Things with Words*, Oxford: Oxford University Press.

[3] Arieti, S. 2014. *Interpretazione della schizofrenia*, Roma: L"asino d"oro Edizioni.

[4] Boersma, P. &Weenink, D. 2015. *Praat: doing phonetics by computer* [Computer program]. Version 5.4.06, retrieved 21 February 2015 from http://www.praat.org/.

[5] Breuler, E. 1911. Dementia praecox oder die Gruppe der Schizophrenien, in A. Aschaffenburg (ed.), *Handbuch der Psychiatrie*, Leipzig: Franz Deuticke.

[6] Chaika, E. 1974. *A Linguist Looks at "Schizophrenic" Language*, in «Brain and Language», 1, pp. 257–276.

[7]CLIPS: http://www.clips.unina.it/it/

[8] Cresti, E. 2000. *Corpus di italiano parlato*, Firenze: Accademia della Crusca.

[9] Cresti, E. 2005. Per una nuova classificazione dell"illocuzione a partire da un corpus di parlato (LABLITA), in E. Burr (ed.), *Tradizione e innovazione: il parlato*. Atti del VI Convegno internazionale SILFI, Pisa: Cesati, pp. 233-246.

[10] Cresti, E. & Moneglia, M. 1996. Monological repetition in very early acquisition, in C. Bazzanella (ed.), *Repetition in dialogue*, Tubingen:Niemeyer, pp. 50-65.

[11] Cresti, E. & Moneglia, M. (eds.) 2005. C-ORAL-ROM. *Integrated reference corpora for spoken romance languages,* Amsterdam: Benjamins.

[12] Dovetto, F.M. 2010. Different Phenomena in Language Pathologies: A Case-Study of Schizophrenic Subjects, in M. Pettorino, S. Giannini, I. Chiari, I. & F.M. Dovetto (eds.) 2010. *Spoken Communication*, Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 113-135.

[13] Dovetto, F.M. & Gemelli, M. (eds.) 2013. [2012] *Il parlar matto. Schizofrenia tra fenomenologia e linguistica: il corpus CIPPS*. Roma: Aracne.

[14] Dovetto, F.M. 2014. *Schizofrenia e deissi*, in «Studi e Saggi Linguistici», 52, pp. 101-132.

[15] DSM-5®. 2013. *Diagnostic and Statistical Manual of Mental Disorders*.

[16] Firenzuoli, V. 2003. Le forme intonative di valore illocutivo dell"italiano parlato, PhD. Dissertation, University of Florence

[17] Fromkin, V.A. 1975. *A Linguist Looks at "A Linguist Looks at „Schizophrenic Language""*, in «Brain and Language», 2, pp. 498–503.

[18] Kraepelin, E. 1899. *Ein Lehrbuch fur Studirende and Aerzte*, Vol II, Leipzig: Verlagvon, Barth.

[19] Liddle, P.F., Ngan, E.T.C., Caissie, S.L., Anderson, C.M., Bates, A.T., Quested, D.J., White, R. e Weg, R. 2002. *Thought and Language Index: An Instrument for Assessing Thought and Language in Schizophrenia*, in «The British Journal of Psychiatry», 181, pp. 326–330.

[20] Martin, P. 2004. *WinPitch Corpus: A text to Speech Alignment Tool for Multimodal Corpora*. Lisboa: LREC.

[21] Moneglia, M. & Raso, T. 2014. Notes on Language into Act Theory (L-AcT), in T.Raso, H. Mello (eds.) *Spoken Corpora and Linguistic Studies*. 1ed. Amsterdam/Philadelphia: Benjamins, v. , p. 468-494.

[22] Panunzi, A. & Mittmann, M. 2014. The IPIC resource and a cross linguistic analysis of information structure in Italian and Brazilian Portuguese, in Raso, T. & Mello, H. (eds.) 2014, Spoken corpora and linguistic studies, Amsterdam: Benjamins, p.129-151

[23] Pennisi, A. 1998. *Psicopatologia del linguaggio*, Roma: Carocci.

[24] Raso, T. & Mello, H. (eds.) 2012. *C-ORAL-BRASIL I: Corpus de referência de português brasileiro falado informal*, Belo Horizonte: Editora UFMG.

[25] Rocha B. 2013. Metodologia empírica para o estudo de ilocuções no português brasileiro, *in «Domínios de Lingu@gem», v.7, n.2, 2013*

LOW FREQUENCY MECHANICAL RESONANCE OF THE VOCAL TRACT IN WATER RESISTANCE THERAPY

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

[1]Institute of Thermomechanics, The Academy of Sciences of the Czech Republic, Prague, Czech Republic
[2]Speech and Voice Research Laboratory, University of Tampere, Finland
jaromirh@it.cas.cz, Anne-Maria.Laukkanen@uta.fi, radolf@it.cas.cz

Abstract: **This study presents a hypothesis of a possible relation between the low frequency mechanical resonance of the human vocal tract and the frequency of water bubbling during phonation into a tube with the other end submerged in water, i.e. water resistance therapy. This relation could have either positive or negative consequences on phonation and sensations during voice production.**
Keywords: **Phonation into tube, yielding walls in vocal tract, bubbling frequency.**

I. INTRODUCTION

Recently, many studies have focused on the so-called water resistance voice therapy, where the subject phonates into a tube submerged in water, see e.g. [1, 2].

The experimental study [3] on gas bubbles formation showed that the bubbling frequency F_b increases with flow rate from $F_b \cong 0$ at nearly zero flow rate ("static bubble") to a maximum value F_{bmax} depending on the tube orifice radius R. For orifices from R=0.017 - 0.79 cm and flow rates from Q=0.01 ml/s to Q=0.25 l/s, the maximum bubbling frequency has been found to range from about F_{bmax}=25 Hz for the largest orifices to about F_{bmax}=75 Hz for the smallest orifices. It was also found that the container widths from 3 to 10 inches square had no effect and that the orifice submergences from about 2.54 cm to about 25.4 cm had negligible effect.

It is also known that a low resonance frequency of the vocal tract F_{lvt} exists, which is caused by the yielding walls of the acoustic cavities in the human vocal tract, see [4, 5]. This mechanical resonance is located in the frequency range F_{lvt}=10–50 Hz depending on the age, weight and gender. The effect of yielding walls is remarkable in speech produced by deep-sea divers at high ambient pressures. The mechanical resonance of the yielding walls causes shifts in the frequencies of the formants (acoustic resonances of the vocal tract) [4, 5].

A previous study [6] investigated the resonance properties of the vocal folds *in vivo* by means of laryngoscopy. Laryngeal vibrations were excited via a shaker placed on the neck of a male subject and observed by means of videostroboscopy and videokymography. The resonance frequencies of the aryepiglottic folds and arytenoid cartilages were suspected to be lower than 50 Hz, because when a sinusoidal excitation of 50 Hz was applied, large oscillations of the laryngeal collar, especially the aryepiglottic folds and arytenoid cartilages, were visually dominant in the stroboscopic view. The vocal folds oscillated as a unit with other laryngeal structures. Left–right and anterior–posterior phase differences in oscillation were visible among the laryngeal structures. Large vibration amplitudes at frequencies of 50 Hz caused uncomfortable sensations in the subject. Therefore the measurements for frequencies lower than 50 Hz were not performed because it was not the aim of the study to inspect these low frequency resonances in detail. We can add that the resonance frequency of the ventricular folds was found to be higher, close to 70 Hz, see [6].

Based on these earlier findings we hypothesize that the lowest mechanical resonance of the vocal tract F_{lvt} can support the effect of water bubbling in water resistance voice therapy. Consequently, a more efficient massage effect on the vocal tract and vocal fold tissues could be obtained when the bubbling frequency F_b is near or coincides with the lowest mechanical resonance of the vocal tract F_{lvt} .

II. METHODS

Two subjects (a male and a female) phonated on [u:] at comfortable pitch and loudness into a glass resonance tube (27 cm in length, 6.8 mm in inner diameter) with the outer end of the tube submerged 2 cm and 10 cm under the water, see [7]. For a comparison, the same experiment was performed on a physical model of voice production, consisting of silicone vocal folds and a hard-walled plexiglass model of the vocal tract for vowel [u:], see [8].

The set of measurements was performed for a 'soft' and 'normal' sustained phonation. The sound pressure level (*SPL*) inside the oral cavity was measured using a special microphone probe designed for measurement of

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

acoustic pressure in small cavities (B&K 4182), and the mean oral pressure (P_{oral}) was measured by a digital manometer (Gresinger Electronic GDH07AN) connected with the oral cavity by a small compliant tube. The acoustic signal outside the vocal tract was recorded using a sound level meter (B&K 2239A). The electroglottographic signal was recorded using dual-channel EGG (Glottal Enterprises). The recordings were made by the PC controlled measurement system B&K PULSE using 32.8 kHz sampling frequency.

III. Result

Figure 1 shows the measured spectra of the oral pressure during phonation, for the female subject for comfortable phonation (P_{oral}=0.88 kPa), and for the model at air flow rate Q = 0.12 *l/s* and oral pressure P_{oral}=1.24 kPa. The maximum oral pressure for the subject was about 20 dB higher at the bubbling frequency $F_b \approx 13$ Hz than at the fundamental frequency of phonation, $F_0 \approx 168$ Hz; while for the model the oral pressure level was only 6 dB higher at the bubbling $F_b \approx 16$ Hz than at the fundamental frequency of phonation, $F_0 \approx 168$ Hz.

Figure 2 shows the vertical larynx position VLP obtained with the EGG simultaneously with the oral pressure in time domain for the male subject phonating into the tube submerged 10 cm in water (P_{oral}=1.02 kPa). In all three signals it is possible to see clearly the low bubbling frequency F_b=11.5 Hz and the fundamental frequency of phonation F_0=98 Hz. The peak-to-peak (p-t-p) value of the oral pressure is about 400 Pa and the oral pressure P_{oral} varies simultaneously with the intensive vibrations of the larynx in vertical direction.

Similarly, Fig. 3 shows the results for the tube submerged 2 cm deep in water (P_{oral}=250 Pa, F_b=11 Hz, F_0=98 Hz). The VLP change is not so clear as for the tube immersion depth of 10 cm due to a more intensive boiling like effect, when the waves on the water surface are randomly changing the hydrostatic pressure.

Interesting is that neither the p-t-p amplitude of the VLP signal nor the p-t-p value of the oral pressure variation depends on the depth of submersion of the tube in water; the differences between submersion depths 2 cm and 10 cm in water were found to be very small.

IV. Discussion and Conclusion

The results suggest that there is a strong vibration of the larynx (or only vocal folds) in vertical direction during water resistance therapy. This vibration can be substantially supported by a low frequency acoustic–structural (mechanical) resonance of the vocal tract if the resonance exists in the frequency range of bubbling. The effect was not found to be dependent on the tube immersion depths between 2 and 10 cm.

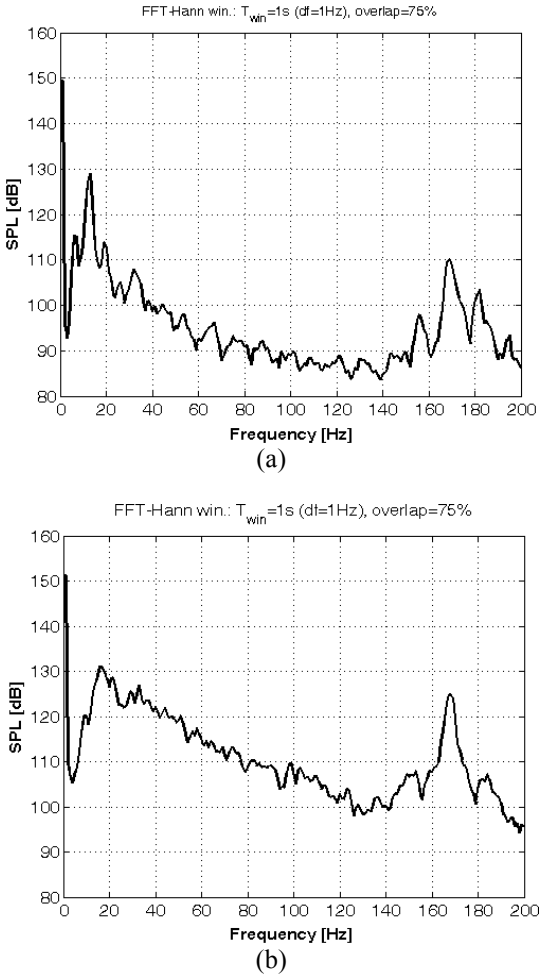


Fig. 1. Spectra of the oral pressure measured for phonation into a tube submerged 10 cm deep in water for (a) the subject and (b) the physical model.

The coalescence of the bubbling frequency and the low acoustic-structural resonance frequency of the vocal tract could potentially enhance the positive effects of the water therapy procedure. The hypothesis of enhancement of the bubbling effect by yielding walls of the human vocal tract is supported by the fact that the SPL level of the oral pressure was much higher at the bubbling frequency than at the fundamental frequency, in contrast to what can be seen in the measured spectrum for the model, see Fig.1.

On the other hand, the coalescence of the bubbling frequency and the low acoustic-structural resonance frequency of the vocal tract may result in unpleasant intensive vibrations of the laryngeal tissues which can even cause some vocal fold impairments. This is still

an open question and a more detailed investigation is needed for establishing evidence based safety recommendations for the water resistance voice therapy.

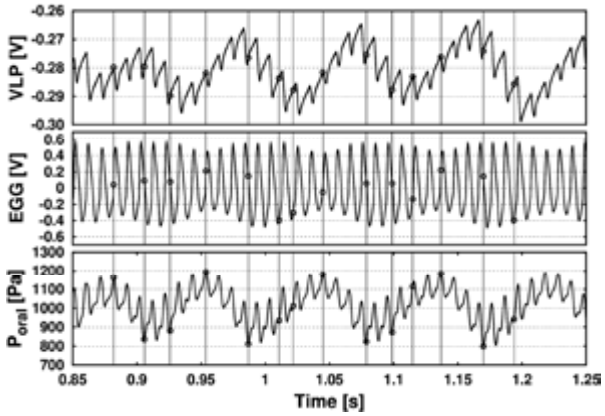


Fig. 2. Vertical larynx position (VLP), electroglottographic (EGG) and oral pressure (P_{oral}) signals captured from a male subject phonating into the tube submerged 10 cm deep in water.

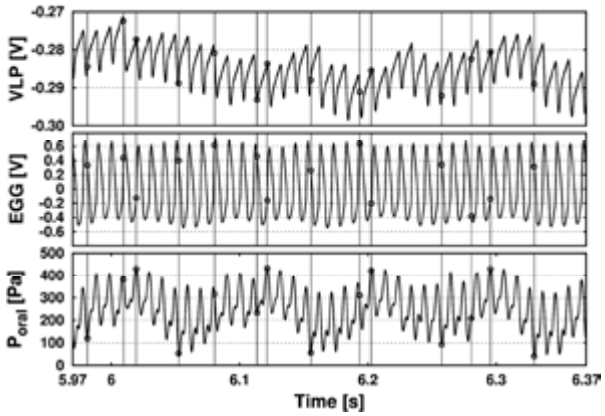


Fig. 3. Vertical larynx position (VLP), electroglottographic (EGG) and oral pressure (P_{oral}) signals captured from a male subject phonating into the tube submerged 2 cm deep in water.

Acknowledgement
The study was supported by the grant No P101/12/1306 of the Czech Science Foundation.

REFERENCES

[1] L. Enflo, J. Sundberg, C. Romedahl and A. McAllister, "Effects on Vocal Fold Collision and Phonation Threshold Pressure of Resonance Tube Phonation With Tube End in Water," *Journal of Speech, Language, and Hearing Research*, vol. 56 pp. 1530–1538, 2013.

[2] S. Granqvist, S. Simberg, S. Hertegärd, S. Holmqvist, H. Larsson, P.A. Lindestad, M. Södersten and B. Hammarberg, "Resonance tube phonation in water: High-speed imaging electrographic and oral pressure observations of vocal fold vibrations – a pilot study," *Logopedics Phoniatrics Vocology*, Early Online: 1-9, 2014.
DOI:10.3109/14015439.2014.913682.

[3] L. Davidson and E.H. Armick, "Formation of gas bubbles at horizontal orifices," *A.I.CH.E. Journal*, vol. 2, No 3, pp. 337-342, 1956.

[4] M.M. Sondhi and J.A. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP vol. 35(7), pp. 955–967, 1987.

[5] M.M. Sondhi, "Model for wave propagation in a lossy vocal tract", *Journal of the Acoustical Society of America* 55, pp. 1070-1075, 1974.

[6] J.G. Švec, J. Horáček, F. Šram and J. Veselý, "Resonance properties of the vocal folds: in vivo laryngoscopic investigation of the externally excited laryngeal vibrations," *Journal of the Acoustical Society of America*, vol. 108(4), pp. 1397-1407, 2000.

[7] V. Radolf, J. Horáček, V. Bula and A.M. Laukkanen, "Air-pressure characteristics and visualization of bubbling effect in water resistance therapy," In Fuis V. (ed.). Engineering Mechanics 2014. Brno: Brno University of Technology, 2014, pp. 528-531.

[8] J. Horáček, V. Radolf, V. Bula, 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*, vol. 21(3), pp. 193-200, 2014.

doctorVOX: A NEW DEVICE FOR VOICE THERAPY AND VOCAL TRAINING

Denizoglu I. [1,2]
[1]Director of Clinical Vocology Unit, [2]University Lecturer of Pedagogical Vocology
[1]Izmir University Faculty of Medicine, Otolaryngology Department, Izmir, Turkey
[2]Dokuz Eylül University State Conservatory Vocal Arts Department, Main Art Division of Opera, Izmir, Turkey
[2]Ege University State Turkish Music Conservatory Liberal Arts Division, Izmir, Turkey
[2]Yasar University Faculty of Art and Design, Music Department, Main Art Division of Vocal Arts, Izmir, Turkey
iilterdenizoglu@yahoo.com

Abstract: ***doctorVOX*** **is a new patented device designed by the author to provide voice therapy, vocal training and vocal humidification. This device is easy to carry and is safe to be used anywhere. It is intended to assist voice therapy and to serve as a supporting device for professional voice users.** ***doctorVOX*** **is designed to help to motor learning and cognitive processes involved in voice therapy and vocal training.** ***doctorVOX*** **provides instant humidification of the vocal folds. Additionally, herbal and medical products can also be used for inhalation.**

Keywords: ***doctorVOX,*** **voice therapy device, vocal training device, vocal humidification**

I. INTRODUCTION

Voice therapy is any kind of technique that changes voice in a behavioral way. As Aronson mentioned " Voice therapy may be defined as an effort to return the voice to a level of adequacy that can be realistically achieved and that will satisfy the patient's occupational and social needs". The main goal of voice therapy is a target voice. This is the best possible voice within the patient's anatomic and physiologic capabilities. Target voice maybe named as 'normal' or 'natural' voice but not everyone, especially not those who have irreversible neurologic or vocal fold lesions, can achieve a normal voice. In such cases, the objective is the best possible voice within the patient's anatomic and physiologic capabilities.

The method of using tubes to extend and constrict the vocal tract is still used especially in Finnish voice training and therapy for different aims. These techniques were proposed as early as 1899 by Spiess [1, 2].

Lax Vox Voice Therapy Technique (LVVT) is a direct technique which changes the vocal mechanism and it is a holistic method that includes all of the subsystems of the voice [3, 4]. LVVT is a cognitive approach which gives a multichannel biofeedback. LVVT can is suitable for almost all voice therapy patients; hypo-hyper functional voice disorders, neurologic-psychogenic dysphonias, pre-post operative phonosurgeries. LVVT is a safe method for singing voice therapy with its specially scheduled technical applications for singers.

The procedure automatically balances the functions included in voice production. It also gives biofeedback and creates holistic cognition of the vocalizing process. The main physiologic mechanism of the LVVT is altering the vocal tract inertance due to positive supraglottal pressure and artificial elongation of the vocal tract. The 'domino effect' goes with lowering the larynx and proper abdominodiaphragmatic respiration.

doctorVOX, a new device is devised and developed after a decade of clinical practice with LVVT. *doctorVOX* uses the mechanisms of LVVT for voice therapy and professional voice development. The main mechanism involves artificial elongation of the vocal tract and a secondary vibrating resistance (i.e. water bubbles) for vocal tract impedance. The artificial elongation is provided by a built-in tube which is designed nearly the same length with the human vocal tract. It is designed for rehabilitation of dysphonic patients and habilitation of the professional voice users.

Hydration and humidification of vocal folds is shown to affect the phonation threshold pressure [5]. Drinking water helps vocal fold hydration after gastrointestinal absorption and salivary hydration process. Humidity of the air during inhalation instantly affects the mucus blanket on the vocal folds. *doctorVOX* directly creates humidified air (air passes through water when the user inhales from the swan-neck shaped breathing tube) in order to humidify the vocal folds.

II. METHODS

The device is made of glass, as well as polycarbon material. It has two separable parts. Upper part is

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

designed to blow and inhale by two nested tubes with different diameters. Lower part is the water container with a neck specially designed to prevent water spillage (Fig. 1).

There are two tube openings on the top of the upper part. The swan-neck like tube indicates the breathing outlet from the container. The phonation inlet is the opening of the inner tube from which the user can blow voice into water. A silicon mouth-piece is also provided to prevent dental injury during vocal exercise. The upper part of the device is formed by two tubes mounted one inside the other. The inner tube is for blowing and phonation. The active length of the inner tube is about the same length with the human vocal tract so that standing waves form in a natural way. The bottom part of the device which is named the container is filled with water for voice therapy exercises. Maximum water height is designed to be below phonation threshold pressure. Water spillage during blowing and aspiration of water during inhalation, are prevented by two main mechanisms. First resistance to keep water in container during bubbling is the circle fold (like an inkstand) at the roof of the container. The second resistance is the enlargement in the neck part of the device.

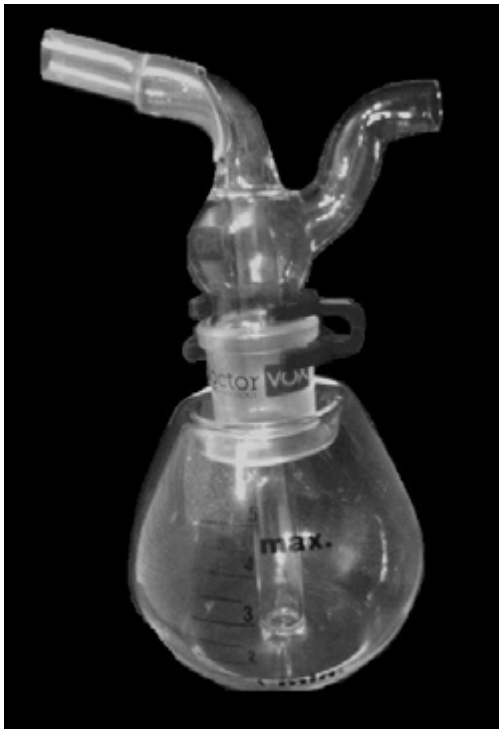


Figure 1: doctorVOX device

doctorVOX can also be used as an instant vocal fold humidifier. The user can humidify the vocal folds by inhaling water vapor (40-45°C) through the swan-neck shaped breathing tube. An optional specially designed cover for the container, functions as a thermos in order to keep the water warm for a longer time. Herbal/medical products are also able to be vaporized through bubbling and can be inhaled from breathing tube. During vocal exercise, user blows air/voice through the inner tube and takes the advantages of Lax Vox technique. During inhalation from the breathing outlet, air enters from the phonation inlet and passes through water to be humidified. The humidified air directly affects vocal fold mucosa

III. Results and Discussion

Treatment success for voice disorders depends on many factors: the disorder being treated, the clinicians treating it and the patient. Patients who are interested in their voices and are motivated to make improvements generally are likely to get the best overall long-term results from treatment. This is partly because such patients tend to follow treatment recommendations.

In the voice clinic, doctorVOX provides a holistic therapy for various functional and organic voice disorders (muscle tension dysphonias, vocal fold nodules and polyps, habitual and psychogenic dysphonias-aphonias, vocal fold paralysis, presbiphonias, pre and postoperative phonosurgery …)

doctorVOX is also a useful device for voice professionals. Singers can use it to warm-up and down and to find the 'position' of singing voice. By relaxing the unnecessary muscle groups, it increases the consciousness about the vocal mechanism. It can be used as a vocal muscle developing assistant through different vocal exercises such as *sostenuto, glissando, portamento, staccato*.

Motor learning principles render the main pathway for voice therapy applications. Task orientation, motivation and focusing the attention are the most important principles for motor learning. Exercises by doctorVOX are easy to follow, easy to teach, easy to learn and easy to perform at home. There are not so many things to remember and do at the same time. Patient uses a ‘device’. The patients accept and use the technique easily and may daily train by themselves.

IV. Conclusion

doctorVOX is a new device for voice therapy and vocal training ss well as vocal humidification. Clinical studies are needed to be done.

References

[1] Spiess, G. Methodische Behandlung der nervösen Aphonie und einiger anderer Stimmsthrungen (Methodological treatment of neurologic aphonia and

several other voice disorders). Archives of Laryngology and Rhinology 1899, 9,368-376.
[2] Simberg S, Laine A. "The resonance tube method in voice therapy: Description and practical Implementations" Logopedics Phoniatrics Vocology. 2007; 32: 165ı 170
[3] Denizoğlu İ, Sihvo M. Lax Vox Voice Therapy Technique. Current Practice in Otorhinolaryngology and Head and Neck Surgery, 2010; 6(2): 284-205
[4] Denizoğlu İ. The Lax Vox Voice Therapy: Method and Applications (Lax Vox Ses Terapi Tekniği'nde Yöntem ve Uygulamalar) . Türkiye Klinikleri J E.N.T.-Special Topics 2013; 6(2):32-40
[5] Leydon C1, Wroblewski M, Eichorn N, Sivasankar M. A meta-analysis of outcomes of hydration intervention on phonation threshold pressure. J Voice, 2010 Nov;24(6):637-43

AUTHOR INDEX

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)