Proceedings e report

54

MODELS AND ANALYSIS
OF VOCAL EMISSIONS
FOR BIOMEDICAL APPLICATIONS

6th INTERNATIONAL WORKSHOP

December 14-16, 2009
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2009

Cover: designed by CdC, Firenze, Italy.

MAVEBA
2009
Firenze, Italy

This event is sponsored by:

Ente CRF – Ente Cassa di Risparmio di Firenze
http://www.entecarifirenze.it/

COST Action 2103 – European Cooperation in the field of Scientific
and Technical research
 http://www.cost2103.eu/

ELSEVIER EDS. – Biomedical Signal Processing and Control
http://www.elsevier.com/locate/bspc

This event is supported by:

ISCA – International Speech and Communication Association
http://www.isca-speech.org/

Provincia di Firenze
http://www.provincia.firenze.it/

Associazione Italiana Scienze della Voce
http://www.aisv.it/

IEEE EMBS – IEEE Engineering in Medicine and Biology Society
http://www.embs.org/

A.I.I.M.B. – Associazione Italiana di Ingegneria Medica e Biologica
http://www.aiimb.it

I.N.F.M. – Istituto Nazionale per la Fisica della Materia
http://www.infm.it

CONTENTS

Special Session on Voice Modeling I (Chairperson and Introduction: H. Kawahara, Japan) Voice Modeling II

Obstructive Sleep Apnoea

Mechanical Models

FOREWORD

It is a great pleasure for me to introduce this 6th edition of the Proceedings of the MAVEBA Workshop, devoted to the relevant topic of voice modelling and analysis under a biomedical perspective.

MAVEBA 2009, the 6th event of this series, celebrates ten years of scientific uninterrupted success. The attendance of researchers from all over the world, that has always distinguished this event, makes me proud and stimulates in pursuing this initiative in the future.

Since its first edition in 1999, the series of MAVEBA workshops aims to fill the gap between different research fields on human voice that historically developed independently from each other. This meeting stimulates contacts between specialists active in clinical, research and industrial developments in the area of voice signal and images analysis for clinical treatment, care and rehabilitation and other biomedical applications, aiming at gathering together knowledge, experience and technology from researchers coming from a wide range of institutions.

The MAVEBA Workshop is organised every two years in Firenze, Italy. This sixth Workshop offers again the participants an interdisciplinary platform for presenting and discussing new knowledge in the field of models and analysis of voice signals and images, as far as both adults, singing and children voices are concerned, ranging from fundamental research to all kinds of biomedical applications and related established and advanced technologies. Modelling the normal and pathological voice source and the analysis of healthy and pathological voices are among the main fields of research. The aim is that of extracting the main voice characteristics, together with their deviation from "healthy conditions". This needs to result in developing accurate, objective and clinically useful methods of investigation of voice quality in patients, and of strategies for preventing occupational voice disorders in professional speakers.

Modelling is one of the hot topics in voice analysis to which the international community devotes great efforts. It has strict links with other equally important fields of research such as:

- diagnosis and classification of pathological voice
- monitoring voice quality during rehabilitation
- development of vocal prostheses and aids for disabled
- analysis of other vocal emissions (infant cry, cough, snoring, swallowing), in neurological dysfunctions, obstructive apnoea, asthma, etc.
- protocols and reliable objective parameters form images through videolaryngoscopy, videokymography, fMRI and other emerging techniques
- emotional voice as related to psychological/neurological conditions, e.g. epilepsy, autism, schizophrenia, stress etc.
- interaction with hearing impairments
- relationships among subjective-perceptive-objective voice analysis

From this long and non-exhaustive list, it appears that the need for interaction between different fields of research has become of utmost importance. The subject of voice analysis has recently gained more and more attention from the international community and is rapidly growing, and in the last ten years links and cooperation among different research fields have become effective to define and set up simple and reliable tools

for voice analysis. A deeper insight into the voice production mechanism and its relevant parameters could in fact help clinicians in improve prevention and treatment of vocal apparatus pathologies.

The interest is also demonstrated by several initiatives that have been set up all over the world that focus on voice. In 2002, April 16, the American Academy of Otolaryngology--Head and Neck Surgery founded the World Voice Day, to encourage men and women, young and old, to assess their vocal health and take action to improve or maintain good voice habits. The World Voice Day is now celebrated world-wide, jointly by the clinical and the biomedical engineering community.

Moreover, both in 2007 and in the present 2009 edition, the MAVEBA Workshop has hosted the Management Committee and Working Groups meetings of COST Action 2103 "Advanced Voice Function Assessment", a 4-years lasting (2006-2010) joint initiative of speech processing teams (engineers and physicists) and the European Laryngological Research Group (ELRG) (laryngologists/phoneticians). A main objective of COST 2103 is in fact a better understanding of the relationship between biomechanical changes of the vocal folds and alterations of the acoustical voice signal. Modelling normal and pathological voice source is an essential tool in this process.

We are definitely moving towards interdisciplinary research, made easier by worldwide fast communication capabilities. Thus great effort should also be directed towards setting up a common framework among all interested researcher and companies. This would be of great help to finalise and speed up research, enhance methodological results, increase and update the production of dedicated, user-friendly and cheap devices and, most important, sensitising people on a still underestimated subject, such as the prevention of vocal apparatus pathologies.

Within this volume of Proceedings, papers range from fundamental research to development and testing of software tools and measurement devices. Specifically, the volume includes three Special Sessions organized and given by worldwide well-known experts on:

- Newborn Infant Cry
- Voice Modelling
- Singing Voice

And other six Sessions on the following topics:
- Emotional Voice
- Voice Quality Assessment
- Voice Images
- Devices
- Obstructive Sleep Apnoea
- Mechanical Models

Some papers on the above mentioned topics are presented during the Workshop in the poster session.

From these papers, and those collected in Special Issues of international Journals: Medical Engineering & Physics (2002), Biomedical Signal Processing and Control (2006, 2009), Acta Acustica - Acustica (2006) devoted to past MAVEBA Workshops, I hope that the interested reader could find useful suggestions and further spurs to carry on research in the important and increasing field of voice analysis.

Finally, I express my gratitude to the members of the organising committee, the anonymous reviewers that helped in improving the quality of the papers, the supporters and sponsors who confidentially gave financial contribution, the administrative staff of the Department of Electronics and Telecommunications that contributed to make this Workshop a successful one.

Claudia Manfredi
Conference Chair
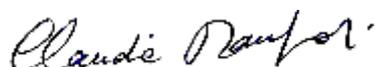
Special Session on Newborn Infant Cry

Chairperson and Introduction:
C.A. Reyes Garcia, Mexico

NEUROFUNCTIONAL SPECTROGRAPHIC ANALYSIS OF THE CRY OF BRAIN INJURED ASPHYXIATED INFANTS: A PHYSIOACOUSTIC AND CLINICAL STUDY

D. Lenti Boero MD, PhD

Facoltà di Psicologia. Università della Valle d'Aosta – Université de la Vallé d'Aoste. Chemin des Capucins 2A 11010 Aosta. Italy
d.lentiboero@univda.it
http://www.disat.unimib.it/bioacoustics/it/

Abstract: The aim of this pilot study cry analysis of brain injured asphyxiated infants, aiming to identify parameters that might predict clinical oucome. Thirteen controls and six asphyxiated subjects with MRI evident lesions were included. Spectrographic analysis of manipulation cries showed that vibrato contours were significantly more frequent in the brain lesioned group than in the controls, and prevalent in two subjects whose outcome was spastic displegy and death. F° parameters were significantly lower in infants with midbrain injuries, this finding is in contrast with previous literature.
Keywords : **infant cry, neonatal asphyxia, brain injury, physioacoustic, spectrographic analysis.**

I. INTRODUCTION

Neuroradiological examinations such as CT or MRI or cranial ultrasonography are the elective tools for early diagnosis of encephalopathy. There is extensive experimental, clinical and neuroimaging date that intrapartum hypoxic-ischemic insult is an important factor in the genesis of irreversible brain injury, especially in term newborns. However, if neuroradiology does effectively detect more evident lesions, on the other hand it might miss subtle and/or minimal brain lesions, that might be of prognostic value [1,2,3,4]. In this respect the spectrographic analysis of cry patterns might be a useful diagnostic tool in evidencing brain disorders in infants, and in providing a prognostic aid as was already found in preterm infants [1,5,6]. In fact, the spectrographic analysis of the cry of human infants affected by brain dysfunction related to neonatal asphyxia, congenital hypothyroidism, or neonatal jaundice, has demonstrated that abnormal patterns occur in the pitch contour and in both domains of time and frequency [7,8,9,10,11,12,13,14]. But, until now, few attempts were done in order to correlate objective, spectrographic and clinical results as regards as long-term outcome, in particular after neuroimaging became available to clinicians [2,4]. This is the aim of present study.

II. METHODOS

Subjects: Six male newborns affected by hypoxic-ischemic central injury, and with MRI diagnosed brain damage and a control group of 13 infants (9 M; 4 F) matched for weight were recruited after parents' informed consent in a local hospital in Milan, Italy. Within the first four days of life, cries were induced by manipulation stimuli during neurological examination performed at the same hour of the day (10.30 am), this eliciting context has the advantage of guaranteeing homogeneous state of arousal in all infants [2]. Cries were recorded by means of DAT sound recorders (Sony TCD D7 and TASCAM DAP1) and of a Sennheiser ME66 unidirectional microphone positioned between two to five cm from the mouth of the crying babies.

Sound analysis: Cries were recorded along the entire duration, sampled at 44.100 Hz, (sample size16 bits), and analysed by means of Canary 1.2 mounted on Powermac 7600 with 45 Megabytes of RAM [14] and Raven on MacBookPro computer. They were subdivided in six subsamples of four to six cry units: the first included the beginning, the last the end of the cry, the others were taken along the time axis at homogeneous intervals in order to guarantee a good representation of continuous variations [12,13]. Spectrograms were produced by means of the above mentioned softwares. Voicing was put in evidence by means of Praat. In this study we mostly concentrated on fundamental pitch whose contours were individually evaluated for melody characteristics *(i.e. rising, rising-falling, falling, and flat steady contours)*, and vibrato *(i.e. a saw-like profile)* (Appendix 1) by two independent observers, according to [15], inter-observer agreement was 97%, hence only 3% of our cry units were discarded. Quantitative time parameters were measured in sec. and frequency parameters in Hz. by means of screen cursor. We analyzed a total of 334 asphyxiated infants' cry units and 341 control infants' cry units. All infants entered a clinical follow-up and mental and motor development were assessed by means of Bayley scale (PDI and MDI) at 4th, 8th 12th month of life; at that time they underwent a second MRI. Only data at birth and at one year results are reported in this study.

III. RESULTS

Weight at birth was not statistically different between pathological infants and controls (SPSS one-way anova, $F_{1,17} = 0.019$, $P = 0.892$).
Global sample analysis.

Pitch contour. Asphyxiated infants had a significantly higher amount of voiceless and partially voiced profiles than the controls (GLIM chi-square = 9.286 df = 1, P < 0.005), (Tab. 1).

Tab. 1. Voicing in normal and pathological subjects				
	Voiceless n (%)	Partially voiced n (%)	Voiced n (%)	Total
pathological subjects	3 (1)	72 (22)	259 (77)	334
normal subjects	2 (1)	44 (12)	295 (87)	341

Brain injured infants' voiced profiles were significantly different from controls' (SPSS Cross tabs, Pearson, chi-square = 286.84, df = 6, P < 0.000), having less normal patterns (i.e. rising, rising-falling, falling, and flat steady contours): 85 (33%) and 215 (85%) respectively for brain injured and normal subjects (GLIM, chi-square = 148.80 df = 1, P < 0.0001); conversely they had an higher amount of vibrato contour: 157 (62%) and 36 (14%), respectively for brain damaged and the normal subjects (GLIM chi-square = 127.91 df = 1, P < 0.0001), in addition they uniquely showed frequency jumps and furcated profiles. The percentage amount of biphonations in both group was identical: 22% (GLIM, chi-square = 0.01, df =1, P > 0.1), we also calculated the percentage difference in length of biphonation over the total amount of voiced cries: 3% for the brain injured group and 4% for the controls (GLIM, chi-square = 0.3, df = 1, P > 0.1). *Quantitave analysis*. Length of wails was significantly shorter ($F_{1,552}$=13.51; P < 0.000) and interval between cries was significantly longer ($F_{1,417}$ = 7.30; P =0.007) in the asphyxiated group). All parameters measured on the fundamental were lower in the asphyxiated group, as show in tab. 2.
Tab. 2. Mean Hz, SD, and ANOVA for F° and peak F of brain injured subjects (1st column) and controls. MANOVA for first four parameters: Pillais trace 4,549 = 47.09; P <0.000.

	Mean	SD	Mean	SD	F_{df}	P
Start F	328.0	72.2	383.8	73.4	80.78 $_{1,552}$	≤0.000
Max. F	432.0	76.7	517.8	73.5	180.41 $_{1,552}$	≤0.000
Min. F.	279.8	71.9	323.9	75.2	49.39 $_{1,552}$	≤0.000
End F	313.8	78.5	348.3	84.8	25.30 $_{1,552}$	≤0.000
Peak F	405.6	77.2	490.5	78.7	156.93 $_{1,527}$	≤0.000
dyn g	711.8	130.2	841.7	123.0	145.55 $_{1,552}$	≤0.000

Individual analysis.
We wanted to explore at the individual level if and how those global differences mirrored for each subjects in our sample.
Pitch contour. The counting of voicing profiles at the individual level in the brain damaged subjects showed a pattern of great variation (Tab. 3).

Tab. 3. Individual voicing in brain damaged sample				
	Voiceless n (%)	Partially voiced n (%)	Voiced n (%)	Total
Ca.	1 (1)	8 (5)	146 (94)	155
Co.	0	3 (13)	21 (88)	24
Cr.	0	7 (58)	5 (42)	12
Pr.	2 (5)	14 (37)	22 (58)	38
Ta.	0	1 (5)	21 (95)	22
Tr.	0	39 (47)	44 (95)	83
Total	3	72	259	334

As regards as pitch contours, only Ca., the subject with most severe lesions vs all other brain injured subjects together (GLIM, chi-square = 11.341, df = 1, P < 0.001) had abrupt discontinuities (pitch jumps), together with an other subject he also had a significant higher amount of vibrato contours (Tab 4), Hz mean difference on F° and significance are shown in fig. 1, the fundamental frequency parameters were significantly lower in all subjects but in one.
Tab. 4. Amount of melodic contours of phonated cries in each pathological infant vs individual control matched for weight

	pitch jumps n (%)	furca tion n(%)	normal n(%)	vibrato n(%)	Normal vs vibrato Chi-square	DF	P
Ca.	11 (8)	5 (3)	31 (21)	99 (68)	102.85	1	< 0.0001
Co.	0	0	4 (19)	17 (81)	17.324	1	< 0.001
Cr.	0	0	2 (40)	3 (60)	0.402	1	> 0.1
Pr.	0	0	9 (43)	12 (57)	0.860	1	> 0.1
Ta.	0	0	13 (62)	8 (38)	2.404	1	> 0.1
Tr.	0	0	26 (59)	18 (41)	2.925	1	< 0.1

Hz mean difference on F° and significance are shown in fig. 1, the fundamental frequency parameters were significantly lower in all subjects but in one. Neurological evaluations were performed at 4th, 8th 12th mouth of life, at that time MRI was repeated. Mental and motor development were assessed with Bayley scale (PDI and MDI). Results are shown in Tab. 5.

Fig. 1. Difference (in Hz) between each single asphyxiated subject and matched

** = P < 0.01; *** = P < 0.001

☐ Ca. ☐ Co. ■ Cr. ☐ Pr. ☐ Ta. ■ Tr.
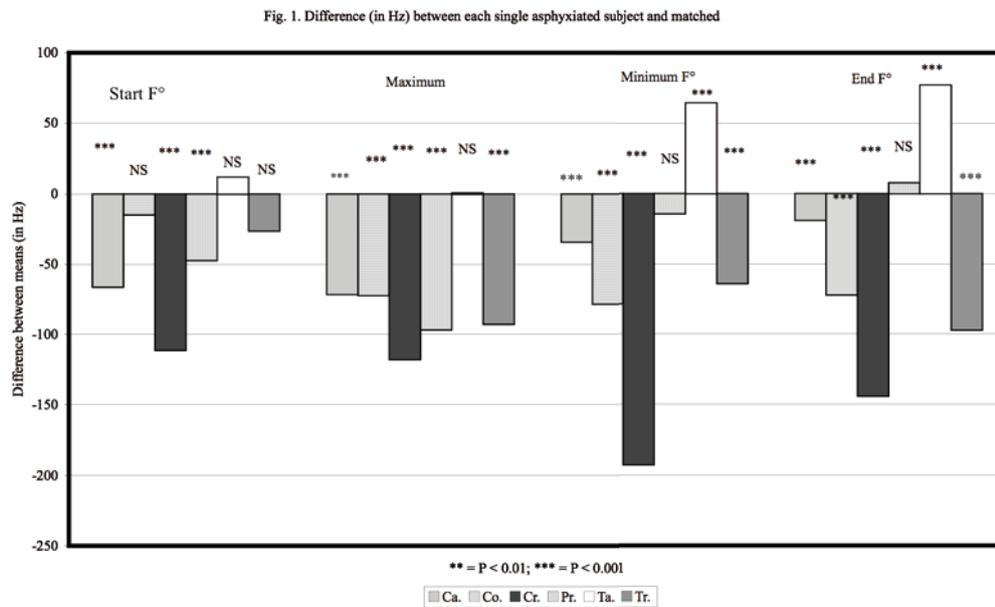
Tab.5. Synthesis of neuroradiological, clinical and cry analysis for six asphyxiated subjects and outcome at one year. Asterisks show significant difference from matched subjects. Subject in the first row deceased in his first month of li

Subjects	G. A.	Weight (gr.)	Apgar (1' 5')	MRI diagnosis (at birth)	Neurological examination (1st week)	Cry's analysis (first week)	MRI control (1 year)	Neurological control (1 year)
Ca.	41	3240	1 5	cerebral hemispheres, basal ganglia, thalamus, hypothalamus	global hypertonia hyperexitability, seizures	* higher % of vibrato contour, * lower F°		
Co.	41	2780	5 7	thalamus periventricular white matter	global hypertonia reduced alertness and poor movement seizures	* higher % of vibrato contour * lower F°	brain injury	dystonic tetraplegia important mental delay
Cr.	38+3	3180	3 7	cerebral cortex (right slight predominance); subcortical white	mild hyperexitability	* lower F°	brain injury	normal motor and mental dev.
Pr.	39+1	3080	1 2	periventricular white matter (left frontal region predominance)	mild hypertonia	no difference	brain injury	normal motor and mental dev.
Ta.	41+5	3170	2 <5	right insulo-temporal region	mild hypertonia	* lower F°	brain injury	normal motor and mental dev.
Tr.	39	4440	4 7	periventricular white matter	hypertonia hyperexitability, seizures	* lower F°	brain injury	normal motor and mental dev.

IV. DISCUSSION

The present pilot is one among the few studies trying to put together three different level of investigation (brain imaging, cry characteristics, and clinical evaluation) in order to better understand the prognostic value of infant cry in brain damaged infants long time after the insult

6

(one year). At the global level our data reflect the findings of previous studies, were an higher amount of voiceless and partially voiced cries in the asphyxiated sample was found [7,9]. However, when we performed the analysis at the individual level, the most important for the clinician, it was found that voicing profiles were different among the individual brain damaged infants, and did not correspond to the degree of brain insult, indeed, the two most sever outcome (Tab. 5) had a very low percentage of voiceless and partially voiced cries. Thus voicing per se should be considered with caution as an indicator of brain pathology due to neonatal asphyxia. In accord with other authors [5,7] we found that a significant higher proportion of vibrato contours are well correlated with major negative outcome at one year (one fatality and one tetraplegia). In an other study (Lenti Boero unp.), comparing cry output from a sample of brain damaged infants with a sample of infants affected by neonatal asphyxia but without MRI evident lesions, and a control, it was found that vibrato contours were significantly more frequent in the brain damaged group than in controls ($x^2_1 = 10.76$, $p < 0.025$), but no difference was found between the asphyxiated subjects without lesions and the normal group ($x^2_1 = 1.42$, $p < 0.25$ ns). Also F° pitch continuity seems to be of great importance: the only fatal case we had was the only one having important alteration in this parameter: frequency jumps and furcated cries (Tab. 2), this is in accord with other authors' findings [5]. Interestingly, all but one brain damaged subjects had significantly lower F° parameters than matched controls, this finding is in contrast with [7]. A good resolution for evident brain lesions was not available at the time this study was performed, and this fact underscores the necessity of having many more clinically detailed cases.

REFERENCES

[1] B.M. Lester, "Developmental outcome prediction from acoustic cry analysis in term and preterm infants,"*Pediatrics*, vol. 80(4), pp. 529-534, 1987.
[2] C. Lenti, D. Lenti Boero, C. Volpe, E. Bianchini, C. Bianchi, "Il neonato con sofferenza cerebrale: correlazioni clinico sonografiche e neuroradiologiche," *Riv. Neuroradiologia*, vol. 7, pp. 276-277, 1994.
[3] J. Blackman, "Crying in the child with a disability: the special challenge of crying as a signal," in *Crying as a sign, a symptom and a signal,* R.G. Barr, B. Hopkins, J.A. Green, Eds. London: Mac Keith Press, 2000, pp. 106-120.
 [4] J.A. Green, J.R. Irwin & G.E, Gustafson, "Acoustic cry analysis, neonatal status and long-term developmental outcomes," in *Crying as a sign, a symptom and a signal,* R.G. Barr, B. Hopkins, J.A. Green, Eds. London: Mac Keith Press, 2000, pp. 137-156.
[5] W. Mende, K. Wermke, S. Schindler, K. Wilzoplosky & S. Hock, "Variability in the cry melody and the melody spectrum as indicators for certain CNS disorders," *Early Child Dev. and Care*, vol. 65, pp. 95-108, 1990.
[6] W. Ludge, & H. Rothganger, "Jitter-index of the fundamental frequency of infant cry as a possible diagnostic tool to predict future developmental problems," *Early Child Dev. and Care*, vol. 65, pp. 145-152, 1990.
[7] K. Michelsson, "Cry analysis of symptomless low birth weight neonates and of asphyxiated newborn infants," *Acta Paed. Scan. Supp.*, vol. 216, pp. 10-45, 1971.
[8] K. Michelsson, P. Sirvio, & O. Wasz-Hockert, "Sound spectrographic cry analyses of infants with bacterial meningitis," *Dev. Med. Child Neur.*, vol. 19, pp. 309-315, 1977.
[9] P. Sirvio, & K. Michelsson, Sound spectrographic cry analysis of normal and abnormal newborn infants. *Folia Phon.*, vol. 28, pp. 151-173, 1976.
[10] O. Wasz-Hockert, J. Lind, V. Vuorenkoski, T.J. Partanen, E. Valanne, *The infant cry: a spectrographic and auditory analysis* (Spastics Int. Med. Pub.), London: Heinemann, 1982.
[11] M. Koivisto, "Cry analysis in infants with Rh haemolytic disease," *Acta Paed. Scan. Supp.,* vol. 335, pp. 1-73, 1987.
[12] D. Lenti Boero, C. Volpe, A. Marcello, C. Bianchi, C. Lenti, "Newborns crying in different situational contexts: discrete or graded signals?," *Perc. Motor Skills.*, vol. 86, pp. 1123-1140, 1998.
[13] D. Lenti Boero, G. Weber, M.C. Vigone, C. Lenti, Crying abnormalities in congenital hypothyroidism: a preliminary spectrographic study. *Journal of Child Neurology*, vol. 15 (9), pp. 603-608, 2000.
[14] R.A. Chariff, S. Mitchell & C.W. Clark, *Canary 1.2 User's Manual*, Ithaca, NY: Cornell Laboratory of Ornithology; 1997.
[15] A. Patel, J.R. Daniele, "An empirical comparison of rhythm in language and music," *Cogn.*, vol. 87, pp. 35-45, 2003.

Appendix1. Spectrogram showing vibrato contours of a 2 days old asphyxiated infant (Co.).
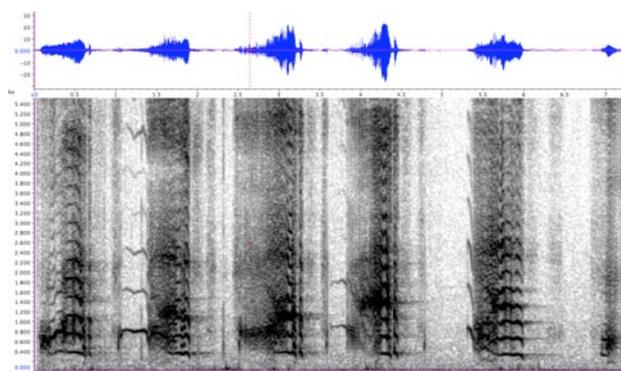
MELODY ANALYSIS OF THE NEWBORN INFANT CRIES

G. Várallyay Jr.[1], András Illényi[2], Zoltán Benyó[1]

[1]Budapest University of Technology and Economics, Dept. of Control Engineering and Information Technology, Budapest, Hungary
[2]Budapest University of Technology and Economics, Dept. of Telecommunications and Media Informatics, Budapest, Hungary

Abstract: **The melody analysis of the infant cries was performed manually in the past. Based on subjective listening, only estimations could be achieved about the real melodies. A novel method had been introduced a few years ago to categorize the melody shapes. It says that the melodies of the infant cries are combined from elementary units. The melody shapes can be classified according to the order of these elementary units. Utilizing this automatized system authors showed that there are 39 different melody shapes of the newborn infant cries, the top 15 categories cover the 93% of the analyzed 580 melodies.**
Keywords: **Newborns, infant cry, melody analysis, melody shape classification**

I. INTRODUCTION

We meet melodies in our everyday life in different circumstances, *e.g.* the cadence of the human voice to express our emotions, or the voice of singing. In general **'melody' means the changing of the fundamental frequency as a function of time**. It is well worth trying to talk about the melody of the infant cry, as it could also carry several information about the infant [1], [2]. Only a few research teams had been dealt with the analysis of the melodies of the infant cries, because it is hard to obtain, visualize and compare them. In this way most of the teams applied subjective, manual investigations with the melodies.

A quite impressive melody analysis was performed in Hungary by *Makói et al.* in 1975. They represented the melody contour on *music paper* [3] in the same way as *Gardiner* did in 1838 [4]. A fellow of their team had absolute pitch, she wrote down the melodies after listening. In 1996 *Schönweiler et al.* classified the melody shapes of crying into six categories [5], these were: *rising*, *falling*, *rising-falling*, *falling-rising*, *flat* and *glottal plosive* (see Fig. 1.). However, the elements of these groups are quite simple and easy to use, it will be shown soon that there are much more categories of the shapes of the melodies.

In 1999 *Möller* and *Schönweiler* discussed about how difficult to evaluate and compare the cries of newborn infants [6]. They suggested investigating the complexity of the cries as a theoretical solution, but they couldn't find a suitable method for that.
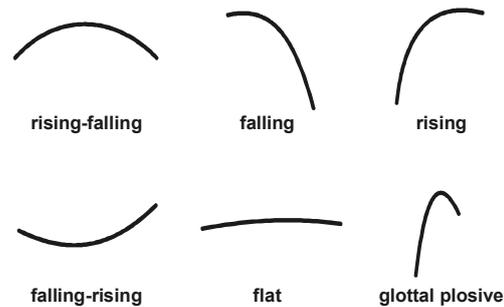
Fig. 1. The groups of the melody shapes by *Schönweiler et al.* from 1996.

Also in 1999 *Michelsson* and *Michelsson* tested the newborns' crying sounds with spectrography [7]. Nowadays spectrography is still a general tool in the analysis of the infant cry. There are several physical attributes of the cry which can be obtained from a spectrogram, as length of the cry, spectral components, etc. Although it seems that the spectrogram doesn't provide a good resolution for the analysis of the melody of the infant cry [8].

Wermke et al. published a study about the development of the melodies in 2002 [9]. They investigated the shape of the melody and the intensity of the cry together. The visualized melodies were obtained by the *MDVP* software from *Kay Elemetrics Corp.* They've found that in case of 8-9 weeks old infants there were simple melody shapes, while older infants doubled or tripled these shapes. In 2003 *Rothgänger* reported that there was a 20-30 Hz rising period in the first third of the duration of the *hunger cries* followed by a 60-70 Hz falling period in the next two third [10].

In 2007 *Várallyay* has published a new idea about the objective analysis of the melodies of the infant cry [11], [12]. From the classification of the melodies it was shown that there were 77 different categories of their shapes and 30% of the melody shapes were simple rising-falling. In another study from *Várallyay et al.* from 2007 the development of the melodies was investigated [13]. They found that in the first two months of life infants cried with mostly shorter, simpler melodies, while later the durations increased and the melody shapes got more complex.

8

In this study authors will analyze the melodies of newborn infants to compare them with the results obtained from the studies above.

II. METHODS

A. Data Collection

For this study data from 73 newborns (0-6 days old) were collected between 2001 and 2006. There were 270 boys and 310 girls with a mean age at 3.47 days. The recordings were made in several hospitals in Hungary. The typical duration of the recordings was 25-30 s, the reason of crying was *spontaneous* in most cases. All the sound recordings were made in quiet places, but not in special silence rooms. There were different recording devices applied as minidisk recorder (SONY MZ-R55), digital video camera (SONY DCRTRV25), digital dictaphone (SONY ICD-P28) and PC sound card with several microphones (SONY ECMMS907, AKG D55S) attached. The melody is such a robust attribute of the crying, that it is not impressionable by the type of the recording device.

The digitalization of the recorded crying sounds was performed at 44.1 kHz or 48 kHz (depending on the recording device), each sample was assigned to 16 bits. Finally, all the recorded sounds were saved onto PC as separate wave (*.wav*) files.

B. Signal Processing

Authors used MATLAB for the signal processing and the analysis of the recordings. The first step of the signal processing was to select the voiced crying sounds from the whole recording with the authors' *Automatic Infant Cry Detection* method [14], [15]. A total of 580 voiced crying sounds were obtained from the 73 infants, their durations were typically between 0.3 and 1.6 s.

The second step was to divide these voiced sounds into short-time (around 50 ms), non-overlapping windows [16]. The fundamental frequency (F_0) was obtained in each windows with the *Smoothed Spectrum Method*, which was developed especially for the F_0 detection of the infant cry [17]. The detected consecutive F_0 values form the melody of the voiced crying sound.

The obtained melodies were visualized with the *Five Line Method* (*FLM*) [18]. *FLM* is an objective method for visualizing the melodies of the infant cry, similar to the music paper but it applies logarithmic scaling. The frequency values of the five lines are shown and the limits of the time axis are fixed. Fig. 2. shows some examples to the obtained melodies, visualized by the *FLM*.
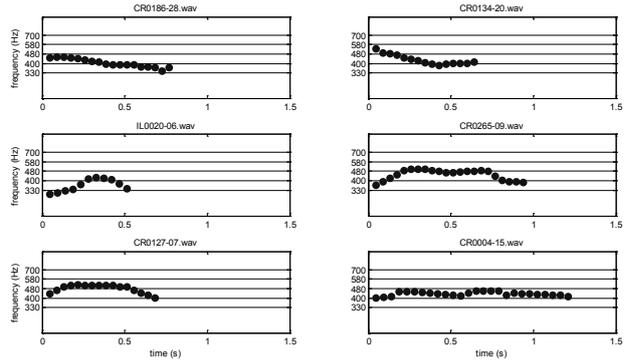
Fig. 2. Examples to the obtained melodies of cries represented by the *Five Line Method*. The method of preparation reminds to the music paper, in this way it is easier to read.

Utilizing the *Five Line Method* the melodies of the cries become more comparable. It can be easily read from the figures if the melody is low-pitched or high-pitched, short or long, simple or complex, etc.

C. Melody Analysis

After observing the shape of 580 melodies authors established that *Schönweilers'* six categories covered only the 70% of the melodies as there were many melodies having more complex shapes. For a higher efficiency the classification system of *Várallyay* from 2007 was applied. This system is based on the segmentation of the melody into **elementary shapes**: *rising* (+1), *flat* (0) or *falling* (-1), and the categories are defined by the order of these units. Every melody could be described as a kind of combination of these elementary shapes. The names of the new categories were created directly from these sequences between chevrons. For example a *rising-falling* type of melody is combined from a *rising shape* (1) and a *falling shape* (-1), in this way these kind of melodies were classified to the following category: <1 -1>.

III. RESULTS

A total of **39 different categories** were found from the *easier* ones (having only one unit) to the more *complex* ones (having 3-4 units). Out of the 39 categories there were 15 which included the 93% of the 580 melodies. The distribution of these 'top 15' categories and the schemes of their shapes are shown in Table 1.

The most typical category in this classification system was the <1 -1> category with its 178 cases (31% of the melodies). There were three more main categories found: <-1>, <1>, and <0 -1> with their 20%, 11% and 9% rate of occurrence. No significant difference was found in the

Table 1. Illustrating the 15 most common melody shapes of the 580 newborn infant cries.

code: <1 -1> cases: 178	code: <-1> cases: 114	code: <1> cases: 65	code: <0 -1> cases: 52	code: <0> cases: 37
code: <1 0> cases: 19	code: <-1 1 -1> cases: 14	code: <-1 1> cases: 10	code: <-1 0 -1> cases: 8	code: <0 1 -1> cases: 8
code: <1 -1 0> cases:7	code: <1 -1 1 -1> cases: 7	code: <1 -1 0 -1> cases: 7	code: <-1 0> cases: 7	code: <1 -1 1> cases: 6

first 6 days of life regarding to the age. In some top categories significant differences were found among the melodies regarding to the gender: <1> was 30% more typical for girls while <0 -1> was 60% more frequent for boys.

IV. DISCUSSION

The most common category was the <1 -1> which is equivalent to *Schönweilers' rising-falling* category. At the 2nd, 3rd, 5th and 8th places there were the <1>, <-1>, <0> and <-1 1> categories which ones had a *Schönweiler*-equivalent as well: *rising, falling, flat* and *falling-rising*. We can declare that *Schönweiler et al.* tried to find the most typical categories for the melody shapes of the infant cries with subjective methods and they reached around 70% effectiveness.

Comparing the obtained categories with the publication of *Várallyay* from 2007 authors found that the 0-6 days old newborns had much less melody shape types (39) than the group of 0-18 months old infants had (77). It meets the results of *Wermke et al.* from 2002 and *Várallyay et al.* from 2007 who stated that the melodies got more complex as the infants got older.

At this time the shapes of the melodies were analyzed while the duration, the intensity and the frequency range of the melodies were discarded. It is planned in a future work to respect other attributes as well and to merge some rarer groups.

V. CONCLUSION

In this study authors obtained and analyzed 580 melodies from sound recordings from 73 newborn infants. Till this time the melody analysis has not been performed simply as it was not obvious how to handle the various melody shapes. A novel method had been developed a few years ago to categorize the shapes of the melodies and to (re)start the melody analysis with objective tools. Utilizing this automatized system authors showed that there were 39 different melody shapes of the newborn infant cries.

Authors recommend for other research teams dealing with the infant cry to perform objective melody analysis as well to increase the effectiveness of their works.

REFERENCES

[1]   A. Fort, and C. Manfredi, "Acoustic analysis of newborn infant cry signals," *Med Eng Phys*, vol. 20(6), pp. 432-442, 1998.
[2]   J. Hirschberg, P. H. Dejonckere, M. Hirano, K. Mori, H-J. Schultz-Coulon, and K. Vrticka, "Voice disorders in children," *Int J Pediatr Otorhinolaryngol*, vol. 32, pp. S109-S125, 1995.
[3]   Z. Makói, Z. Szőke, L. Sasvári, P. Gegesi-Kiss, and P. Popper, "1st cry of newborn after vaginal and cesarean

10

delivery," *Acta Paediatr Hung*, vol. 16(2), pp. 155-161, 1975.

[4] W. Gardiner, *The Music of Nature*, Boston, J.H. Wilkins & R.B. Carter, 1838.

[5] R. Schönweiler, S. Kaese, S. Möller, A. Rinscheid, and M. Ptok, "Neuronal networks and self-organizing maps: new computer techniques in the acoustic evaluation of the infant cry," *Int J Pediatr Otorhinolaryngol*, vol. 38, pp. 1-11, 1996.

[6] S. Möller, and R. Schönweiler, "Analysis of infant cries for the early detection of hearing impairment," *Speech Commun*, vol. 28, pp. 175-193, 1999.

[7] K. Michelsson, and O. Michelsson, "Phonation in the newborn, infant cry," *Int J Pediatr Otorhinolaryngol*, vol. 49(1), pp. S297-S301, 1999.

[8] G. Várallyay, *Analysis of the Infant Cry with Objective Methods,* Ph.D. Thesis, Budapest University of Technology and Economics, Hungary, 1999.

[9] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," *Med Eng Phys*, vol. 24(7-8), pp. 501-514, 2002.

[10] H. Rothgänger, "Analysis of the sounds of the child in the first year of age and a comparison to the language," *Early Human Dev*, vol. 75, pp. 55-69, 2003.

[11] G. Várallyay Jr, "The Melody of Crying," *Int J Pediatr Otorhinolaryngol*, vol. 71, pp. 1699-1708, 2007.

[12] G. Várallyay Jr., and Z. Benyó, "Melody Shape – A Suggested Novel Attribute for the Biomedical Analysis of the Infant Cry," *Proc. 29th Conf. IEEE Engineering in Medicine and Biology, Lyon*, pp. 4119-4122, 2007.

[13] G. Várallyay Jr., Z. Benyó, and A. Illényi, "The development of the melody of the infant cry to detect disorders during infancy," *Proc Fifth IASTED Int Conf on Biomedical Engineering (BioMED 2007), Innsbruck*, pp. 186-191, 2007.

[14] G. Várallyay Jr., A. Illényi, and Z. Benyó, "The automatic segmentation of the infant cry," *Proc. BUDAMED '08 Conference, Budapest,* pp. 28-32, 2008.

[15] G. Várallyay Jr., A. Illényi, and Z. Benyó, "Automatic Infant Cry Detection," *Proc 6th Int Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2009), Firenze,* 4 pages, 2009, *accepted.*

[16] G. Várallyay Jr., "Analysis of the infant cry with digital signal processing (DSP)," in *Pediatric Airway – Cry, Stridor, and Cough*, J. Hirschberg, T. Szende, P. Koltai, and A. Illényi. San Diego: Plural Publishing, 2008, pp. 53-77.

[17] G Várallyay Jr, "SSM – A Novel Method to Recognize the Fundamental Frequency in Voice Signals," *Lecture Notes on Computer Sciences*, vol. 4756, 88-95, 2007.

[18] G. Várallyay Jr., Z. Benyó, A. Illényi, Z. Farkas, and L. Kovács, "Acoustic analysis of the infant cry: classical and new methods," *Proc 26th Conf. IEEE Engineering in Medicine and Biology, San Francisco,* pp. 313-316, 2004.

AUTOMATIC INFANT CRY DETECTION

G. Várallyay Jr.[1], András Illényi[2], Zoltán Benyó[1]

[1]Budapest University of Technology and Economics, Dept. of Control Engineering and
Information Technology, Budapest, Hungary
[2]Budapest University of Technology and Economics, Dept. of Telecommunications and
Media Informatics, Budapest, Hungary

Abstract: **Cry detection can be defined as a procedure where the voiced crying sounds are selected from the recording. The most difficult part of the cry detection is to recognize the inspiratory sounds and separate them from the voiced sounds. In addition, sound recordings may come from different places and recorded with several devices, in this way the method of the cry detection has to be universal. Authors created the *Extended Harmonic Product Spectrum* method to classify the spectral structure of a given signal. Based on this new method authors developed the *Automatic Infant Cry Detection* (*AICD*) system to detect voiced cry sounds in any kind of recording.**
Keywords: **Infant cry, cry detection, Extended Harmonic Product Spectrum**

I. INTRODUCTION

Not only crying sounds can be found in a recording of an infant cry. For example, the infant takes an inspiration between two voiced crying and there can be shorter-longer pauses as well. During these pauses the background noises might be heard in the recording. The recording device might have its own noise. The inspiration can be quiet or audible. It can be placed before or after the voiced crying sound. The infant can suspend the voiced sound or reduce it. The sound of crying can be high-pitched or low-pitched, nasal, veiled, reedy, woody, etc. Many further attributes could be listed in connection with the crying sound.

In a 60 s long recording 8-10 pieces of *voiced* cries can be found on an average. **Cry detection** can be defined as a procedure where the voiced crying sounds are selected from the recording. As there are many different kinds of cries, and there might be misleading sounds in the recording as well (background noises, inspiratory sounds, etc.) the cry detection was performed manually in most of the research teams as [1], [2]. For example in 1982 *Hirscberg* and *Szende*, or in 1999 *Michelsson* and *Michelsson* applied spectrographic analysis of the infant cry, and they selected the voiced crying sounds manually after determining a visual spectrogram from the recording [3], [4]. In the last decade some teams have started applying speech detection software, but generally these software can be used with limitations as the speech and cry signals have differences [5], [6].

The most difficult part of the cry detection is to recognize the inspiratory sounds and separate them from the voiced sounds. In addition, sound recordings may come from different places and recorded with several devices, in this way the method of the cry detection has to be universal. It will be shown that effective cry detection can be executed with limitations and considerations both in the time and the frequency domains.

II. METHODS

From a simplified view the goal of the speech detection is to detect the boundaries of each word, accordingly in the cry detection the start and the end points are to be found of each crying segments. A common attribute of the words that they have a relative big energy, in this way they can be detected by applying a well-chosen energy threshold [7], [8]. In case of cry recordings by seeking for the high energy parts not only the crying segments but also the inspiratory sounds, louder background noises are found.

To create an effective *Automatic Infant Cry Detection* (*AICD*) system authors recommend inspecting the *spectral content* along with the *energy content* of the recordings. While the crying segments are typically harmonic signals (*i.e.* having the fundamental frequency and its subharmonics in the spectrum), generally the noise signals (*e.g.* the inspiratory sounds) have less regular spectral structure [9]. In this study the well-known *Short-Time Energy Function* was applied to obtain the energy content of the recordings. The spectral content was determined with the extension of the *Harmonic Product Spectrum* (*HPS*) method.

A. Short-Time Energy Function

The *Short-Time Energy* (*STE*) function of an audio signal is defined as:

$$E_n = \frac{1}{N} \sum_m \left[x(m) \cdot w(n-m) \right]^2 \qquad (1)$$

where $x(m)$ is the discrete time audio signal, n is time index of the short-time energy, and $w(m)$ is a rectangle window, *i.e.*

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1, \\ 0, & otherwise \end{cases} \qquad (2)$$

It provides a convenient representation of the amplitude variation over the time [10]. It is fact that values of E_n for the unvoiced (*i.e.* coughing, silence, etc.) components are in general significantly smaller than those of the voiced (*i.e.* the real crying) components [11]. It can be used as the measurement to distinguish audible sounds from silence when the signal-to-noise ratio is high. The loudness of the crying segments is typically decreasing at the end, in this way the only analysis of the energy content would issue in losing more quiet parts of the crying segments. There are also some cases when the start and the end are louder than the mid of the crying segment, in these cases the only analysis of the energy would issue in cutting the segment to two pieces.

Moreover a main task of the effective *AICD* is to find the voiced crying sounds and to separate the inspiratory sounds from them. Authors found that in many cases the inspiratory sounds were stuck to the voiced crying sounds in this way the energy function could not distinguish between them. A subsidiary method is needed to analyze the spectral content of the recordings as well to be able to detect and cancel the inspiratory sounds.

B. Extended Harmonic Product Spectrum

The *Harmonic Product Spectrum* (*HPS*) is a robust algorithm to determine the fundamental frequency of a multimodal signal [12]. The *HPS* extracts the fundamental frequency directly from the signal spectrum by decimating the input spectrum by integer factors and computing their product (see Fig. 1.). The input parameter of the *HPS* is N, which refers how many decimated spectrums to determine for the calculation. The primary point for choosing the value of N is the expected number of the subharmonics. Authors found that $N = 9$ is an optimal value for the infant cry in general.
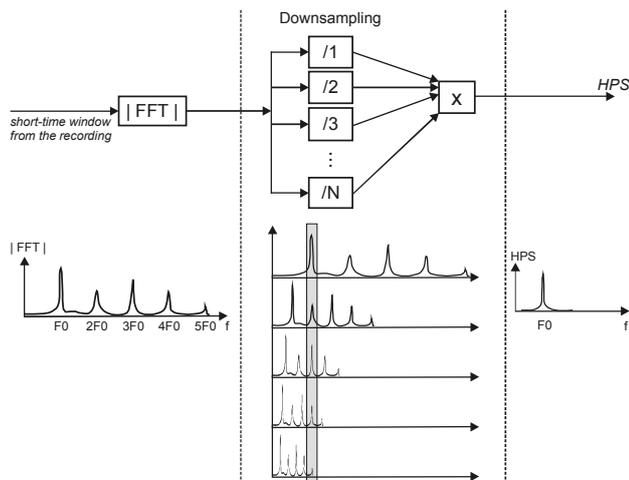
Fig. 1. Illustration of the *Harmonic Product Spectrum* (*HPS*) method.

In 2009 *Várallyay* stated that the *HPS* may be capable for describing the spectral content of crying sounds [13]. To prove this statement he utilized the following: if the spectrum is harmonically rich, the *HPS* will result one enhanced peak at the fundamental frequency [7]. Beyond the position of the *HPS* peak, some other attributes of the harmonic product spectrum might be informative. It is worthy of note that in case of noise-like signals several peaks can be expected in the harmonic product spectrum, not only one. He defined two new parameters (H_{max} and F_{width}) by *extending* the *HPS* to classify the regularity of the structure of the spectrums (*regular* structure means *harmonic* structure). H_{max} is the intensity of the biggest peak in the product spectrum, F_{width} means the bandwidth of the product spectrum at the level of $10^{-4}H_{max}$, see Fig. 2.

Várallyay found that:

- The higher the *HPS* peak was, the more regular the structure of the original spectrum was, and *vice versa*.
- The narrower the *HPS* bandwidth was, the more regular the structure of the original spectrum was, and *vice versa*.

C. Comparing the energy and the spectral methods

On Fig 3. the outputs of the *STE* and the *EHPS* are shown in case of a 12 s long recording. Every method gives information about the cry recording from different aspects. There are five voiced crying sounds in this recording between 0.1-1.3; 1.7-2.7; 3.1-3.8; 7.2-9.7; and 10.2-11.6 s. The 4th has smaller amplitude than the others have. There are audible inspiratory sounds after the 1st, the 2nd, the 4th and the 5th voiced crying sounds.

The *Short-Time Energy Function* is capable to detect loud voiced crying sounds, while the detection of quiet ones (as between 7.2 and 9.7 s) is less efficient. The Y-axis is normalized between 0 and 1. The quiet parts of the recordings have small E_n values (<0.1) while the loud inspiratory sounds and crying segments have bigger E_n values (>0.2).

H_{max} and F_{width} obtained by the *EHPS* refer to the regularity of the spectrum of a short-time crying window. As the range carrier of H_{max} could overstride more magnitude orders it is logarithmized and normalized between 0 and 1. H_{max} is at high level (>0.8) continuously in case of crying segments and having significant start
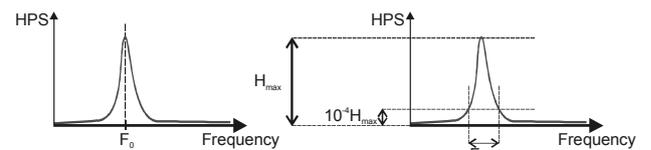
Fig. 2. The outputs of the original *HPS* (left) and the *Extended HPS* (right) methods.
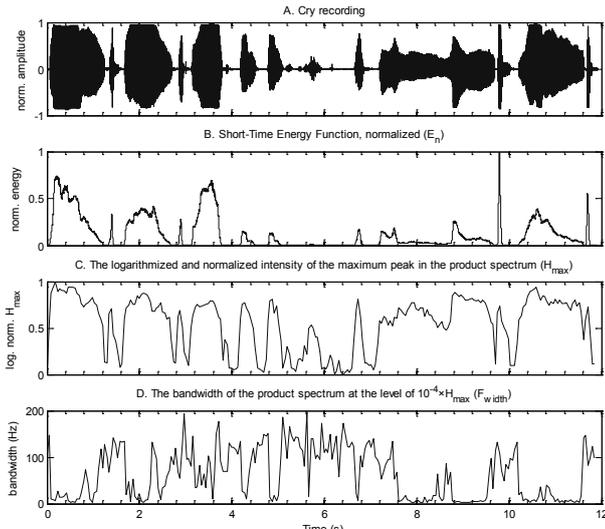
Fig. 3. Comparison of the outputs of the three methods in case of a 12 s long recording.

Fig. 4. Illustration of the main steps of the *Automatic Infant Cry Detection*.

and end points, in this way it can be applied for automatic cry detection with high efficiency. Usually F_{width} has smaller values (<15 Hz) at the place of the crying segments, but it is very sensitive to the noises in the recording, thus its curve is less continuous.

D. Guidelines for the Automatic Infant Cry Detection

To create the *Automatic Infant Cry Detection* system authors utilized the following experiments [14]:
- The voiced crying sounds have a greater duration than 250 ms.
- The distance between the inspiratory sounds and the crying segments can be even less than 100 ms, which results that the maximum window length shouldn't exceed 50 ms.
- In general the energy of the voiced crying sounds and the inspiratory sounds is greater than the background noises.
- The spectrum of the voiced crying sounds has more regular structure than the inspiratory sounds have.
- The recordings might come from different places and devices, in this way the energy and/or the spectral thresholds have to be determined separately for each recording,
- The calculation of the energy function needs less time than the *EHPS*.

According to the results from Fig. 3., the **Automatic Infant Cry Detection should be implemented with the application of H_{max} obtained from the EHPS**. Although authors recommend applying a **pre-selection by the energy function** as the first main step of the automatic cry detection, in this way the analysis of the spectral

content can be focused only on these pre-selected parts. By using the F_{width} at the last steps of the automatic cry detection it is possible to **recognize and separate the inspiratory sounds** from the detected crying segments.

E. Main steps of the Automatic Infant Cry Detection

The *Automatic Infant Cry Detection* was implemented in MATLAB. To illustrate the main steps of the *AICD* a short recording is shown which contains three voiced crying sounds (Fig. 4/A). There are two audible inspiratory sounds (before the 2nd and the 3rd voiced crying sounds), a coughing and a sudden noise.

After the DC component extraction and normalization an energy threshold was determined to focus only on the interesting parts of the recording (Fig. 4/F.). The parts shorter than 200 ms were eliminated (Fig. 4/G.). The start and end points of the remaining signals were revised (Fig. 4/J.) and the F_{width} was applied to find and cancel the inspiratory sounds which had stuck to the voiced sounds (Fig. 4/M.). The spectral contents of the remaining parts were investigated with H_{max} to find the clear, voiced crying sounds of the recording (Fig. 4/P.)

III. RESULTS

A. Detected voiced crying sounds

With the developed *AICD* authors detected 2780 voiced crying sounds from 366 recordings. The 95% of the detected sounds were between 0.3 and 0.2 s in point of their duration. The mean value of the duration of the voiced crying sounds was 0.79 ± 0.54 s and the median was 0.91 s. The total length of the recordings was 8753 s, while the total duration of the detected voiced crying sounds segments was 2535 s.

B. Exactness

To test the exactness of the developed *Automatic Infant Cry Detection* three different recordings (from different

places and devices) were listened and 24 voiced crying sounds were detected manually. The pre-selection of the *AICD* resulted 27 probable voiced crying sounds. After obtaining their spectral contents 23 voiced crying sounds were chosen by the *AICD*. The missing crying sound which was selected manually was quite hoarse, that is why the *AICD* was failed to select that sound as well.

All the detected voiced crying sounds were devoid of the inspiratory sounds.

Regarding to the boundaries of the detected sounds there were a 0.024/0.006 s difference at the start/end points between the manually and the automatically selected sounds on an average.

IV. DISCUSSION AND CONCLUSION

The recognition and the separation of the audible inspiratory sounds are critical tasks and specialty in the cry detection. These inspiratory sounds can be found both before and after the voiced crying sounds.

The *Extended Harmonic Product Spectrum* method is capable to classify the spectral content of cries, and to distinguish the crying segments from the inspiratory sounds as well. H_{max} is at high level continuously in case of crying segments and having significant start and end points, in this way it can be applied for automatic cry detection with high efficiency. Authors recommend using the F_{width} to avoid selecting the inspiratory sounds.

Highly effective automatic cry detection can be accomplished in consideration of the energy content, the spectral content and limitations according to the experienced features of the infant cry. Within these limitations authors recommend utilizing the minimal duration of the crying segments, the wide range of the amplitude of the crying segments and the minimal distance between the inspiratory sounds and the crying segments.

The implemented *Automatic Infant Cry Detection* can be downloaded from the *File Exchange* at the *MathWork's* website from December 2009 [15].

ACKNOWLEDGEMENTS

REFERENCES

[1] Z. Makói, Z. Szőke, L. Sasvári, P. Gegesi-Kiss, and P. Popper, "1st cry of newborn after vaginal and cesarean delivery," *Acta Paediatr Hung*, vol. 16(2), pp. 155-161, 1975.

[2] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," *Med Eng Phys*, vol. 24(7-8), pp. 501-514, 2002.

[3] J. Hirschberg, and T. Szende, *Pathological cry, stridor and cough in infants*, Budapest: Akadémiai Kiadó, 1982.

[4] K. Michelsson, and O. Michelsson, "Phonation in the newborn, infant cry," *Int J Pediatr Otorhinolaryngol*, vol. 49(1), pp. S297-S301, 1999.

[5] A. Fort, and C. Manfredi, "Acoustic analysis of newborn infant cry signals," *Med Eng Phys*, vol. 20(6), pp. 432-442, 1998.

[6] R. G. Barr, B. Hopkins, and J. A. Green, *Crying as a sign, a symptom and a signal*, London: MacKeith Press, 2000.

[7] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, New York: MacMillan Publishing Co., 1993.

[8] G. Gordos, and G. Takács, *Digital speech processing*, Budapest: Műszaki Kiadó, 1983.

[9] G. Várallyay Jr., "Analysis of the infant cry with digital signal processing (DSP)," in *Pediatric Airway – Cry, Stridor, and Cough*, J. Hirschberg, T. Szende, P. Koltai, and A. Illényi. San Diego: Plural Publishing, 2008, pp. 53-77.

[10] T. Zhang, and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 441–457, 2001.

[11] G. Várallyay Jr., "Future Prospects of the Application of the Infant Cry in the Medicine," *Per. Pol. Elec. Eng*, vol. 50(1-2), pp. 47–62, 2006.

[12] V. Parsa, and D. G. Jamieson, "A Comparison of High Precision F0 Extraction Algorithms for Sustained Vowels," *J of Speech, Language and Hearing Res*, vol. 42, pp. 112-126, 1999.

[13] G. Várallyay, *Analysis of the Infant Cry with Objective Methods,* Ph.D. Thesis, Budapest University of Technology and Economics, Hungary, 1999.

[14] G. Várallyay Jr., A. Illényi, and Z. Benyó, "The automatic segmentation of the infant cry," *Proc. BUDAMED '08 Conference, Budapest,* pp. 28-32, 2008.

[15] http://www.mathworks.com/matlabcentral

RECOVERY OF OXYGEN SATURATION LEVEL IN NEWBORNS

S. Orlandi[1], L. Bocchi[1], M. Calisti[1], G. Donzelli[2], C. Manfredi[1]

[1]Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy
[2]Department of Paediatrics, Children Hospital A. Meyer, Università degli Studi di Firenze, Firenze, Italy

Abstract: **With the increased survival of very preterm infants, there is a growing concern for their developmental outcomes.**
Infant cry characteristics reflect the development and possibly the integrity of the central nervous system. This study evaluates the distress occurring during cry in preterm newborn infants, as related to decrease of central blood oxygenation. A recording system has been developed, that allows synchronised, non-invasive monitoring of blood oxygenation and audio recordings of newborn infant's cry.
In the present work we evaluate the changes in the oxygen saturation levels in the central nervous system in full term and in preterm infants, and analyze possible differences between the two groups of patients.
The method has been applied to preterm and full term newborns at the Intensive Care Unit, A.Meyer Children Hospital, Firenze, Italy and at Nuovo Ospedale S.Giovanni di Dio, Scandicci, Firenze, Italy.
Results indicate that a similar decrease of central blood oxygenation occurs in both groups of patients, but the recovery time after the crying episode is more stable and faster in full term newborns than in preterm ones.
Keywords : **Oxygen saturation, preterm newborn, infant cry.**

I. INTRODUCTION

With the increased survival of very preterm infants, there is a growing concern for their developmental and socio-emotional outcomes. Infant cry characteristics reflect the development and possibly the integrity of the central nervous system.

However, in preterm and/or low-birth-weight infants it could imply an effort which may have an adverse impact on blood oxygenation. In fact, preterm new born infants have an impaired auto regulation of the cerebral blood flow [1-4]. Irregularities in the blood flow and pressure may adversely influence the development of the child [5-7]. Some studies have been performed to evaluate both cerebral and peripheral blood oxygenation in the newborn by Near InfraRed Spectroscopy (NIRS) and pulse oximetry, also as linked to other techniques [1-7].

Previous studies have shown that preterm infants and infants with neurological conditions have different cry characteristics when compared to healthy full-term infants. Research has been developed to study possible differences between full-term and preterm infants in their neuro-physiological maturity and the subsequent impact on their speech development [8]. Our previous results demonstrate that blood oxygenation level in preterm newborns is affected by stress caused by the effort required during crying [9-10]. These studies indicate that the distress effect of crying seems larger on central blood saturation than on peripheral saturation, hence here we will consider only central blood saturation as related to cry.

In this work, we extended previous studies to include a comparison of the results obtained in preterm and in full term infants.

II. METHODS

Monitoring has been performed by collecting data from two different sources: central blood saturation was measured with a NIRS device, and a microphone connected to a laptop has been used to record cry emissions.

A unidirectional microphone (Shure SM58), equipped with Tascam US-144 portable audio/MIDI interface (96 kHz/24-bit recording) has been used to record cry emissions. Audio recordings were stored on a multimedia laptop on a single channel audio track, with sampling rate Fs=44 kHz and 16 bit resolution.

Central blood saturation has been measured by means of a NIRS device (Somasensors by INVOS 5100C

Fig. 1: Experimental setup
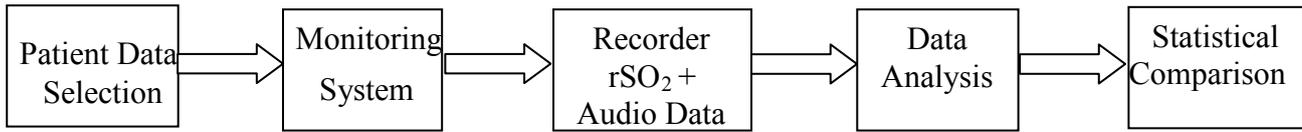
Fig. 2 Block diagram of the system

Somanetics Corp.), with sampling rate of 0.6 samples/sec. The NIRS signal is composed of up to four independent channels, each made up of two data, one containing the relative saturation of oxygen, and the other representing the quality of the signal, useful to detect possible artifacts related to patient movement or poor contact of the sensor with the patient.

Specific software has been designed and implemented to allow synchronization of the output of the two devices by means of a digital connection linking the laptop with the output of the NIRS device. The software implements simultaneous recording of the audio channel trough the US-144 board and of the NIRS signal using a RS-232 connection. Moreover, the software allow for basic management of the patient database, allowing to record anamnesis data and to manage multiple recording sessions for the same patient. The overall setup used in the experiments is described in Fig. 1. Fig. 2 shows the block diagram of the whole recording and processing system.

All subjects were recorded in a quiet room, with low background noise and stable levels of illumination, according to the NIRS device requirements. Moreover, special care has been used to assure a good contact between sensors and patient's skin thus avoiding artifacts caused by sudden movements.

Each recording lasts at least 15min, in order to include several crying episodes, with a suitable amount of time both before and after each cry episode.

A preliminary analysis of the data indicates high variability of the baseline oxygenation level, both in full term and in preterm infants. The baseline oxygenation has been considered equal to the average oxygenation level during a convenient period of time when the child was awake and calm. On the whole test set we observed an oxygenation ranging from 65% to 85%. At the same time, the average variation in the oxygenation level during each recording is approximately of the same order of magnitude. Therefore, in order to assess the change in the oxygenation level during each recording, we considered the difference between the oxygenation level during and after the crying episode and the baseline oxygenation, measured in the time interval just before the episode.

As shown in our preliminary work [9], the recordings indicate a significant difference, in the preterm infants, of the oxygenation levels before the cry episode and during the episode. As it can be expected, the oxygenation level
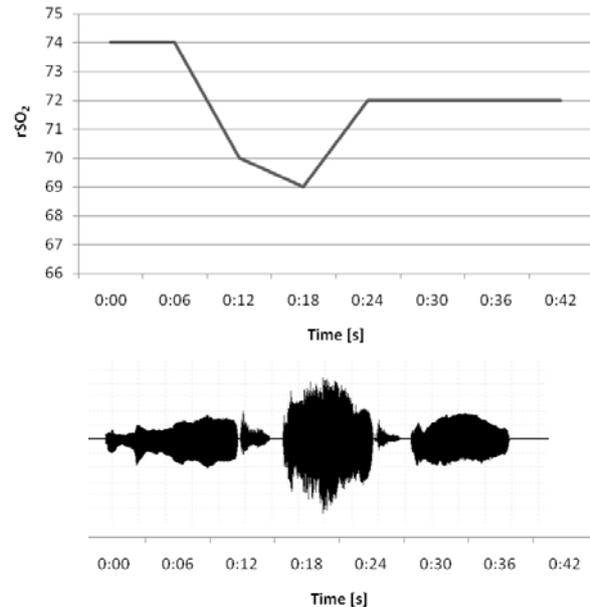
Fig. 3: Plot of the oxygenation level in a sample signal and the corresponding audio track

decreases during the cry, pointing out possible relationship between stress and cry (Fig.3).

Starting from these results, the work has been extended to include a control group composed of full-term patients. The analysis has taken into account the recovery of oxygenation level when the crying episode is over and the infant is calm (ether awake or sleeping).

Each recording has been manually analyzed, and three different crying episodes have been selected from it. Crying episodes have been selected of comparable length, and include a suitable period of rest (patient either sleeping or calm) both before and after the cry episode. Three parameters have been extracted from the oxygenation signal for each crying episode: the average saturation level before the episode (baseline level, B), the oxygenation level during the episode, and saturation after a reasonable recovery time. The baseline oxygenation level has been assumed equal to the average oxygenation over a period of 15 samples acquired before the beginning of cry. Then a "cry oxygenation" (C) has been evaluated using the average value over a time span of 18s

Group	PreCry - Cry (ΔC)	Cry - Recovery (ΔR)	PreCry –Recovery (ΔO)
Preterm	2,20442E-05	0,377462	0,000175
Full-Term	3,19572E-13	6,31E-11	0,014014

Table 1: Statistical significance (p-values) of the differences in oxygenation levels

approximately in the middle of the crying episode. A last reference value (R), related to the capability of the patient to recover the baseline oxygenation level, has been obtained by averaging the oxygenation level measured during 90s from the end of the cry episode.

Data have been analyzed in order to compare the oxygen saturation in basal condition (before the crying episode), in case of stress (during the episode), and the recovery capability of the newborn (90s after the episode). Comparison has been carried out, given the high differences in the absolute oxygenation levels, by comparing, on each episode, the variation of the oxygenation during and after the cry episode with the saturation before the episode:

$$\Delta C = C - B$$
$$\Delta O = R - B$$

We also evaluated the recovery of the oxygenation occurred during the recovery time:

$$\Delta R = R - C$$

The selected parameters have been evaluated separately on all episodes related to full term newborns and to preterm ones, and t-test has been applied to assess their statistical significance.

III. RESULTS

The analysis has been carried out on a group of 20 preterm and/or low weight infants and 28 full term infants, having a pregnancy period ranging from 23 to 42 weeks and a weight at birth between 590g and 4250g, selected by physicians among patients at the Critical Care Unit of the Children Hospital A.Meyer, in Firenze, Italy and Nuovo Ospedale S.Giovanni di Dio, Scandicci, Firenze, Italy.

Full term newborns have been recorded a day after birth, while preterm newborns could be recorded only 20-30 days after birth, due to their long staying in the incubator.

Fig. 3 reports a sample extracted from the data set. In the upper part of the figure, the NIRS track is shown, while the bottom part of the figure shows the audio track acquired in the same period of time. The behavior shown in the figure is typical of full term newborns: during the cry episode, there is a clear decrease in the saturation

level, which is promptly recovered when the crying episode is over.

We obtained about 150 cry episodes, which were analyzed by evaluating the difference between the values before, during, and after the cry episode using a paired t-test analysis. Results, summarized in Table 1, indicate there is a highly significant ($p \ll 0.01$) difference in the oxygenation level before and during the cry episode, both in the full term and in the preterm groups. A different behavior can be noticed comparing the values measured during the cry episode and the values after the recovery time: in the full term group, the t-test indicates the presence of a highly significant difference, while in the preterm group the increase is less pronounced, and is not statistically significant. The same result is confirmed by the comparison of the saturation measured before the cry episode and after the recovery time. This difference is highly significant in the preterm group, indicating that the oxygenation after recovery is noticeably lower than before the crying episode, while in the full term group the difference is only marginally significant ($0.01 < p < 0.05$), suggesting that oxygenation has been recovered, although not completely.

IV. CONCLUSION

The results of the experiments indicate that, both in the full term and in the preterm infants, a significant decrease of the oxygenation occurs during a cry episode. However, the two groups behave differently during the recovery time after the crying episode. Full term infants can recover almost completely the oxygenation levels before cry in less the 90s, while preterm infants need a longer period of time to achieve a full recovery of the oxygenation level.

REFERENCES

[1] Pryds O. & Edwards, A.D., "Cerebral blood flow in the newborn infant", Archives of Disease in Childhood: foetal and neonatal edition, 74 (1), pp. 63-69, (1996).
[2] Greisen G. "Cerebral blood flow preterm infant during the first week of life", Acta Paediatrica Scandinavica, 75, pp.43-51 (1986).

18

[3] Lou H.C., Lassen N.A. & Frii-Hansen B., "Impaired autoregulation of cerebral blood flow in the distressed new born infant", Journal of Paediatrics, 94, 118-121, (1979).

[4] Miall-Allen V.M., de Vries L.S., Whitelaw A.G. (1987). Mean arterial blood pressure and neonatal cerebral lesion. Archives of Disease Childhood, 62, 1068-1069.

[5] Van De Bor M. & Walther F.J., "Cerebral blood flow velocity regulation in preterm infant", Biology of the Neonate, 59, pp. 329-335, (1991).

[6] Friis - Hansen B., "Perinatal brain injury and cerebral blood flow in newborn infant", Acta Paediatrica Scandinavica, 74, pp. 323-331, (1985).

[7] Delpy DT., Cope MC., Cady EB, Wyatt JS.,Hamilton PA., Hope PL, Wray S. & Reynolds EO., "Cerebral monitoring in newborn infants by magnetic resonance and near infrared spectroscopy", Scandinavian Journal of Clinical Laboratory Investigation,188, pp. 9-17, (1987).

[8] Goberman, A.M., Robb, M.P., "Acoustic examination of preterm and full-term infant cries—the long-time average spectrum", J Speech Lang Hear Res 42 (1999), pp. 850–86

[9] Manfredi, C., Bocchi, L., Orlandi, S., Calisti, M., Spaccaterra L., Donzelli, G.P., "Non-invasive distress evaluation in preterm newborn infants", Proc. 30th IEEE EMBS Annual Int. Conf. Vancouver, Brit. Col., Canada, p.2908-2911, August 20–24 (2008)

[10] Manfredi, C., Bocchi, L., Orlandi, S., Donzelli, G.P., High-resolution cry analysis in preterm newborn infants, Medical Engineering & Physics, 31(5), (2009), pp 528-532

Emotional voice

VOICE PARAMETER DYNAMICS IN PORTRAYED EMOTIONS

I. Yanushevskaya, C. Gobl, A. Ní Chasaide

Centre for Language and Communication Studies, Trinity College Dublin, Ireland

Abstract: **This paper is concerned with voice source variation associated with different emotional portrayals of an utterance:** *bored, sad, happy, surprised, angry* **and** *neutral*. **The source analyses involved pulse-by-pulse inverse filtering to yield the differentiated glottal flow, and subsequent parameterisation of the source signal using the LF model. The glottal source parameters included in the analysis were F0, EE, RK, RG, RA, FA, OQ and RD. For the data set analysed, each emotion seems to have its own distinct pattern of source parameter settings. Analysis of the dynamics of the source variation illustrated here on the RD parameter suggests that to better understand source variation we need to study it in terms of the prosodic components of the utterance.**

Keywords : **Voice source, dynamics, emotion**

I. INTRODUCTION

This paper deals with voice source variation which is associated with different emotional portrayals of an utterance. Our broad concern is the study of how voice source parameters vary as a function of linguistic (prosodic and segmental) aspects of an utterance [1, 2], as well as how such source differences may signal affective states. Here we consider source variation for a single utterance produced by a male speaker, repeated so as to convey six targeted affective states: *bored, sad, happy, surprised, angry* and *neutral*. The sentence read by the speaker was 'We were aWAY a YEAR ago', and the stressed syllables are shown in capitals.

Note that we do not claim that these represent 'true emotions' as might occur in spontaneous interactions, but rather the type of portrayals one might use when, for example, reading a bedtime story to a child. As such they represent 'feigned' emotion, but we would argue that such feigned emotion is not only part and parcel of narrative reading but is also used in discourse, e.g., when a mother feigns being cross to influence a child's behaviour or when one feigns being calm while truly agitated in a stressful social encounter. In many true-life situations effective social interactions depend more on the ability to feign emotion than to reveal true underlying emotion. Further example of the use of feigned emotion

in discourse is when tone-of-voice is mismatched to the utterance as a humorous device or to express sarcasm, etc.

II. METHOD

The source analyses used involved pulse-by-pulse inverse filtering to yield the differentiated glottal flow, and subsequent parameterisation of the source signal using the LF model [3]. These techniques involve a manual interactive analysis system and are described in [4], as are the source parameters. The glottal source parameters included in the analysis were F0, EE, RK, RG, RA, FA, OQ and RD. F0 is the fundamental frequency. EE is a measure of the strength of the main glottal excitation. RK is a measure of the skew of the glottal pulse; e.g., a higher RK value indicates a more symmetrical glottal pulse. RG is the glottal frequency FG normalised to F0, where FG is the characteristic frequency of the glottal pulse during the open phase. RA and FA are related parameters capturing spectral tilt. Thus, a high FA (or low RA) value indicates a source spectrum with relatively strong higher harmonics. OQ is the duration of the glottal open phase in relation to the duration of the whole glottal period, and is linked to the strength of the lowest harmonics of the source spectrum. RD is a global wave shape parameter, and is thought to be highly correlated with voice quality variation on the tense to lax continuum [5, 6].

III. RESULTS AND DISCUSSION

A. Overall vocal parameter settings

Fig. 1 illustrates the global source parameter settings for the different affective states extending the preliminary analysis reported in [7]. Note that for the single utterance in question, depending on the emotion expressed, 81 to 120 individual glottal pulses were analysed. Parameter levels are calculated as a percentage difference relative to the *neutral*, based on mean values for the entire utterance.

The scaling allows one to see the extent to which a particular parameter deviates from the *neutral*: from -2 = [< -25% of neutral value set at 0] (very low) to +2 = [> 25% of neutral value] (very high). Note also that filled black circles show parameters demonstrating relatively high dynamic variation as indicated by the mean rate of change (delta) values. These were obtained by calculating
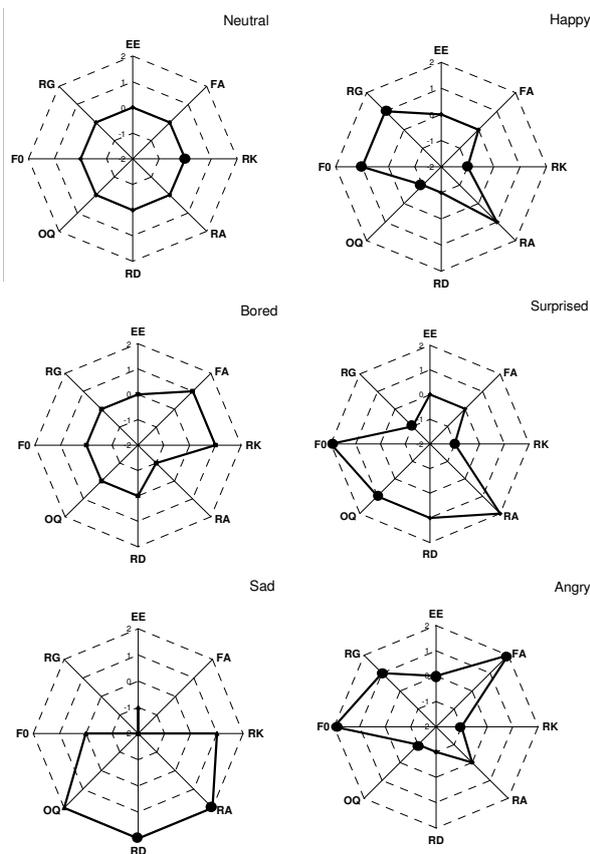
Fig. 1. *Levels for glottal parameters for different emotions: -2 = [< -25%] (very low), -1 = [-25%, -5%] (lower than neutral), 0 = [-5%, 5%] (within the neutral range), 1 = [5%, 25%] (higher than neutral), 2 = [> 25%] (very high). Filled black circles show parameters demonstrating relatively high dynamic variation.*

the first order difference from the smoothed parameter values. The smoothing, which allowed us to decrease the amount of pulse-to-pulse noise while preserving the overall parameter dynamics, involved calculating the moving average of parameter values with a three pulse frame and a one pulse frame-shift.

As evident in Fig. 1, the combination of parameter settings is different for each emotion. In these renditions, each emotion seems to have its own distinct pattern.

Sad relative to *neutral* shows an overall pattern of weak glottal pulses (very low EE), very leaky, breathy voice quality (as suggested by the very high RA, RD and OQ parameters) and large attenuation of high frequency components in the signal (very low FA).

Surprised shares some of these characteristics of the *sad* repetition. Although there is an indication of greater

breathiness (rather high OQ, RD and RA values), there is overall less weakening of the glottal pulse excitation, or of the higher frequencies in the signal (more modal-like EE and FA values). It also has strikingly high mean F0.

Angry and *happy* both show broadly opposite deviations from the *neutral* baselines, as can be deduced from the generally upward shift in parameter values. Note that *happy* and *angry* are frequently confused in perception experiments on vocal expression. The raised RG and the lowered OQ, RD and RK values suggest an overall more tense voice quality setting. *Angry*, however, differs here from *happy* in having extreme FA and F0 values, suggesting more extreme vocal tension.

Bored differs least from the *neutral* setting, showing mainly somewhat more strength in the higher frequencies (FA/RA values).

B. Source dynamics

As mentioned above, filled black circles in Fig. 1 denote dynamic variation. As can be seen from the prevalence of such circles in the case of *happy, surprised* and *angry*, there is more dynamic variation in source parameters than for the relatively low activation states of *sad, bored* and even *neutral*. This seems intuitively in keeping with what we might expect for these more aroused high activation states.

Although Fig. 1 gives some idea of the global trends for these different renditions of the utterance, it does not adequately show the considerable dynamic variation in source parameters in the course of the utterance. To illustrate this, in Fig. 2 we show the dynamic course of the RD parameter for these utterances (as represented by the smoothed parameter trajectories). Note that RD tends to be viewed as indicative of the tenseness/laxness of the voice [5, 6]. To facilitate the inspection and comparison of parameter trajectories, the time axis of each
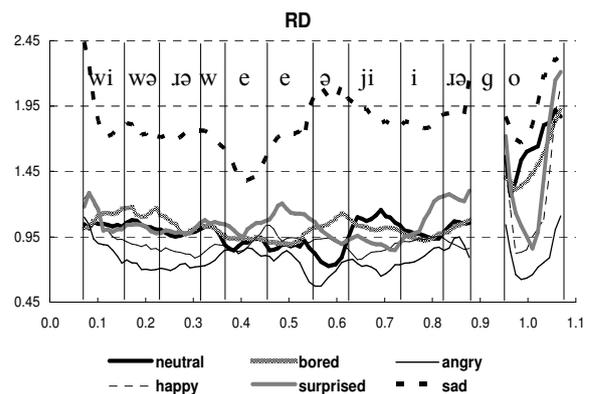
Fig. 2. *Dynamic variation of the RD parameter for different emotions across the utterance.*

emotionally coloured utterance was normalised to that of *neutral* according to a number of anchor points (shown in Fig. 2 as vertical lines). These included utterance and syllable boundaries as well as midpoints of the vowels in the accented syllables and the approximate boundary of [w] and [e] in the stressed syllable WAY. The [g] segment was excluded from the analysis as it had not been consistently realised as voiced across the utterances. For each part of the utterance between anchor points, the time axis was scaled to be of the same duration as the corresponding *neutral* one. As the utterances had a different number of pulses linear interpolation was used to plot all utterances to the same time axis points as the *neutral* utterance.

Note that in Fig. 2 RD values overall are much higher for the *sad* repetition (a very lax quality) and highest (tense) for the *angry* repetition. Also evident are complex parameter variation over time depending on the segmental characteristics of the utterance (consonants tend to lower RD values as higher degree of constriction in the vocal tract have upstream influences on vocal fold vibration). The large differentiation among the emotions in the final syllable of the utterance, as well as the rapid increase in RD values at the end of the utterance are likely to be linked to the realisation differences in the final accent as well as to the transition into breathiness as the vocal folds open prepausally. This suggests that affect related voice differences may be strongly anchored to prosodically important aspects of the utterance. In the *sad* utterance, RD dips in the accented vowels of WAY and YEAR, indicating a less breathy quality. Across all the emotions looked at here there is a distinct trend for these vowels to be associated with relatively strong glottal pulses with more stable parameter settings.

Fig. 3 provides further information on this last point. It shows the mean and standard deviation values of the RD parameter (panel A), as well as the rate of change (delta values) of RD in the course of the utterance, for each affect with an indication of the standard error (panel B). Note that while *bored* and *neutral* have similar means and standard deviations, there is overall much less dynamic variation of the *bored* RD values.

The mean and standard deviation of parameter values are shown separately for the stressed and unstressed syllables in each affect in Fig. 3 (panel C). Similarly, in panel D, the mean rate of change of the RD parameter is shown separately for the stressed and unstressed syllables together with the standard error values. This illustrates again the point made above that for this parameter, although the average values do not differ greatly in the stressed/unstressed conditions, there is considerably more dynamic variation in the unstressed than in the stressed syllables.
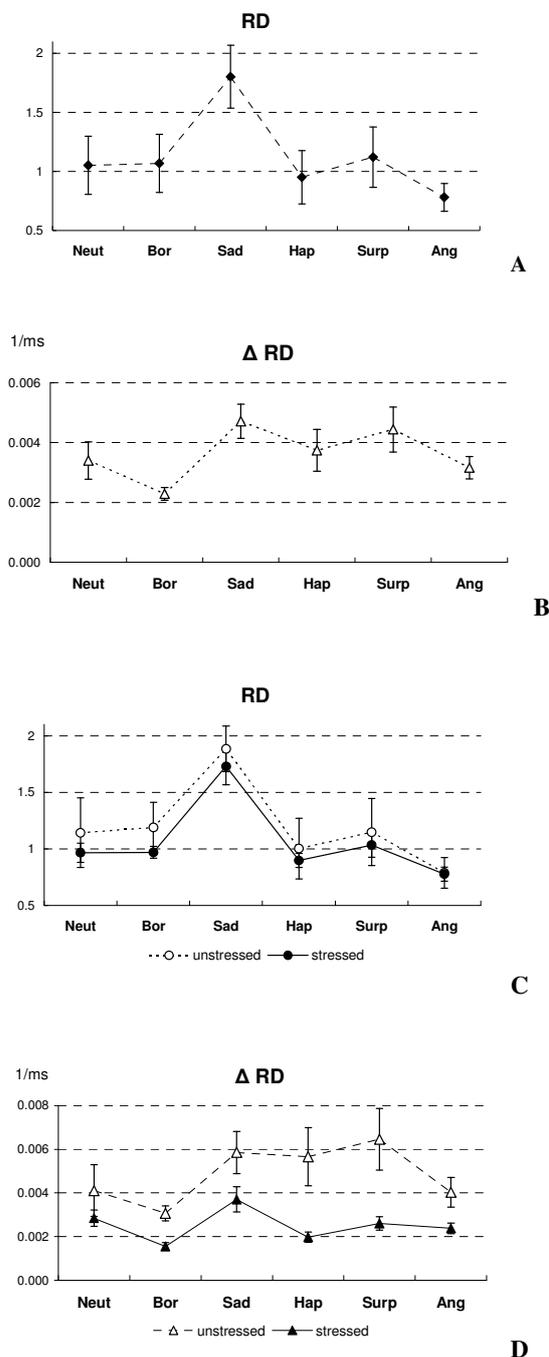
Fig. 3. *RD parameter for different emotions: A- mean values, B - mean delta (rate of change) values, C – mean values for stressed and unstressed syllables, D - mean delta values for stressed and unstressed syllables.*

24

IV. Conclusion

Although based on very limited sample utterances, we feel that a detailed study can nonetheless yield new insights and prompt us to look at what might be important in the analysis of source variation. Similarly we would argue that being able to analyse (and eventually hopefully to resynthesise) these kinds of simulated portrayed emotions might have many applications in the use of speech technology.

This illustration highlights the need to look closely at the utterance internal dynamics of source variation. We would suggest that these dynamics will be best understood if studied in terms of the prosodic components of the utterance. Differences between stressed and unstressed syllables have been pointed out and there are indicators in the present data that accentuation and in particular the nucleus and the post-nucleus material may be of particular importance. In our future work we hope to examine in greater detail the linkage between prosodic structure and voice source variation, as a basis for understanding how and where source variation may signal affect.

Acknowledgements

References

[1] C. Gobl and A. Ní Chasaide, "Voice source variation in the vowel as a function of consonantal context," in *Coarticulation: Theory, Data and Techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge University Press, 1999, pp. 122-143.

[2] A. Ní Chasaide and C. Gobl, "Voice quality and f0 in prosody: towards a holistic account," in *Speech Prosody 2004*, Nara, Japan, 2004.

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

[4] C. Gobl and A. Ní Chasaide, "Techniques for investigating laryngeal articulation (Section B: Techniques for analysing the voice source)," in *Coarticulation: Theory, Data and Techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge University Press, 1999, pp. 300-321.

[5] G. Fant, "The LF-model revisited: transformations and frequency domain analysis," *STL-QPSR* vol. 2-3, pp. 119-156, 1995.

[6] C. Gobl and A. Ní Chasaide, "Amplitude-based source parameters for measuring voice quality," in *VOQUAL'03*, Geneva, Switzerland, 2003, pp. 151-156.

[7] I. Yanushevskaya, M. Tooher, C. Gobl, and A. Ní Chasaide, "Time- and amplitude-based voice source correlates of emotional portrayals," in *Affective Computing and Intelligent Interaction: Proceedings of the ACII 2007*. vol. 4738, A. Paiva, R. Prada, and R. W. Picard, Eds. Lisbon, Portugal: Springer-Verlag, 2007, pp. 159-170.

DETECTION OF NEGATIVE EMOTIONAL STATE IN SPEECH WITH ANFIS AND GENETIC ALGORITHMS

Humberto Pérez Espinosa, Carlos Alberto Reyes García

Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica Óptica y Electrónica

Luis Enrique Erro No. 1, Tonantzintla, Puebla, México

Abstract: **In this work we present the design of an Automatic Emotion Recognizer which tries to take advantage of three soft computing techniques: Neural Networks, Fuzzy Inference Systems and Genetic Algorithms in order to indentify and classify emotions from a speech signal. The classification is done between two emotional states: Negative and Idle. The emotion recordings used for this work belongs to the FAU AIBO database, where children interacting with Sony's pet robot Aibo were recorded. We propose and analyze the use of 18 acoustic features. A classification system based on ANFIS is implemented. Genetic Algorithms are used to select features and tune the ANFIS configuration settings. System implementation and some experimental results are shown.**
Keywords: **Emotion recognition, ANFIS applications, genetic algorithms, acoustic speech features, feature selection**

I. INTRODUCTION

Emotions are prominent elements always present in the mind of human beings. By the emotions expression present during oral communication useful information about the speaker is transmitted. This information complements the information contained in the explicit exchange of linguistic messages. For over 40 years, psychologists have been studying the effect of emotions in the speech of individuals, being Paul Ekman a pioneer in the study of emotions and their relation to facial expressions. More recently computer scientists have got involved in the problem of automatic emotion recognition and have classified emotions using pattern recognition techniques. Knowing the emotional state of individuals offers relevant feedback information about the psychological state of a speaker in order to take important decisions on how a system's user should be attended. For example, in the case of a call center that provides medical support where users call asking for help [1]. These users could present different emotions such as stress, pain, fear, or panic depending on the sickness or emergency they are experiencing. Based on the classification of emotional states, incoming calls could be handled differently, giving priority to the truly urgent; routing it to the appropriate medical staff member. Another application is that of an interactive voice response system that attends patients with psychological problems [2]. The system detects if there is some degree of depression based mainly on articulatory and quality features of the patient's voice. The system alerts a human expert when it finds an alarming degree of depression. As these couple of applications show, automatic emotions recognition can improve the performance, usability and in general, the quality of human-computer interaction systems, client attention systems and other kinds of applications.

II. METHODS

Because fuzzy logic has shown good results in problems where the information is complex, for this work we chose ANFIS, which combines fuzzy logic and neural networks, to carry on the recognition of emotions from speech process. These techniques have been proved to behave well when solving complex problems. Another technique with good behavior in solutions search is the one known as Genetic Algorithms. In this work we try to take advantage of the capacity of ANFIS to implement non linear mapping from input patterns towards the corresponding emotional state, and the optimizing capability of genetic algorithms to find the best features subset and the best configuration parameters to create the most adequate ANFIS architecture. The parameters to optimize are: number of membership functions, type of input membership functions and type of output function.

III. FEATURE EXTRACTION

In [3] a study of the acoustic elements that determine the emotions in the speech is done. Three groups of features are proposed, namely; Utterance Timing, Utterance Pitch Contour, and Voice Quality. In the same work the findings of several authors are summarized in a table (see Table 1). This table shows different aspects of the voice qualified by means of objective measurements, like speech rate, and intensity, and some other subjective measurements that are usually determined by experts and not easy to calculate automatically, the quality of the voice for example. On the other hand, qualifying voices in a linguistic way, as in Table 1, is very close to the form in which fuzzy membership functions are assigned to linguistic variables, which lead us to choose a fuzzy classification model for this task.
We think that the parameters in the table sum up very well the voice aspects that are important to find emotions.

In this work we propose a set of 18 characteristics directly related with these aspects of the voice. The characteristics were obtained using PRAAT [4].

	Anger	**Joy**	**Sadness**	**Fear**	**Disgust**
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest tone	Breathy blaring	Resonant	Irregular voicing	Grumbled, chest tone
Pitch Changes	Abrupt on stressed syllables	Smooth upward inflections	Downward inflections	normal	Wide, downward terminal inflections
Articula-tion	Tense	Normal	Slurring	Precise	normal

Table 1 Summary of human vocal emotion effects

A. Utterance Timing Measures

Speech Rate is the only feature of this type that we include in our feature set. It was calculated using the techniques described in [5]. There, the syllables are detected automatically without needing a transcription. For this, peaks of intensity preceded by dips are considered as potential syllables that in a later process are confirmed or discarded. After detecting syllables, the total sounding time for every recording is calculated. Finally, the speech rate for every recording is obtained dividing the total amount of detected syllables by the total sounding time of the recording.

B. Utterance Pitch Contour Measures

Pitch was obtained from every recording through the correlation method and the following statistic measures were calculated: Pitch Average, Pitch Range, Pitch Median, Pitch Standard Deviation, Minimum Pitch Point, and Maximum Pitch Point.

C. Voice Quality Measures

Voice quality is not a feature that can be calculated directly. In medicine there are some scales for perceptual measurement of the patient voice quality. One of the most used is the scale GRBAS. The aspects that this scale focuses on a voice are: "graded", "rough", "breathy" "aesthenic, and strain." These descriptors are considered as a benchmark for evaluating pathological voices and are also related to aperiodicity descriptors of the physical vibration of vocal cords like jitter, shimmer, tremor, harmonic to noise ratio, voice breaks, etc [6]. We include the following features as voice quality indicators:

- Intensity Average: The mean of intensity contour.
- Jitter: Average absolute difference between consecutive periods, divided by the average period. It is used to detect pitch perturbations.
- Shimmer: Average absolute difference between amplitudes of consecutive periods divided by the average amplitude. This measure is used to detect intensity perturbations. Jitter and Shimmer are measures broadly used to detect pathologies in voice and to estimate its quality.
- Number of Voice Breaks: The number of distances between consecutive pulses that are longer than 1.25 divided by the pitch floor. It measures how long a voice can keep phonation in a period of time.
- Degree of Voice Breaks: Total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal.
- Harmonicity: Grade of acoustic periodicity. This measure is divided in Harmonics-to-noise ratio and Noise-to-harmonics ratio.
- Voice Articulation: Voice articulation is the process by which speech organs interact to produce voice. In order to measure the speech articulation quality of a speaker it is used to build a Vowel Space Area analyzing the first two formants. In this work we could not build the vowel space area due to the lack of phonetic level labels. However we included as articulation indicators some statistics of the first two formants. The features included were: standard deviation of formant 1, mean of formant 1, standard deviation of formant 2 and mean of formant 2.

IV. CLASSIFICATION METHODS

Recently, hybrid classifiers have shown a more robust classification since they combine the best characteristics of two or more classification methods. One of those hybrid approaches is Adaptive Neuro-Fuzzy Inference System (ANFIS) which combines fuzzy logic and neural networks techniques. This model was chosen because of its capacity to extract knowledge from a database through neural networks and tune rules for a Fuzzy Inference System (FIS) automatically, in contrast to traditional FIS where rules are specified by a human expert. ANFIS was originally proposed in [7]. As an adaptive network, the parameters in some ANFIS nodes are adapted during training, in which case the node is called adaptive node. There are also nodes whose parameters remain unchanged during training, they are called fixed nodes. ANFIS applies two techniques in updating parameters.

For premise parameters that define membership functions, ANFIS employs gradient descent to fine-tune them. For consequent parameters that define the coefficients of each output equations, ANFIS uses the least squares method to identify them. This approach is thus called hybrid learning method since it combines gradient descent and the least-squares method. On the other hand Genetic Algorithms are a family of computational models inspired by evolution. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures so as to preserve critical information. Genetic algorithms are often viewed as function optimizers, although the range of problems to which genetic algorithms have been applied is quite broad. In this work a Genetic Algorithm is applied to search the best parameters combination for the ANFIS architecture as well as to search the best feature combination.

V. EMOTION CLASSIFICATION SYSTEM

9,956 training features vectors were extracted from the FAU Aibo Emotion Corpus described in [8], one vector per each audio file. The features extracted are described in section 2. Through experimental tests it was observed that better classification results were reached when subsets of less than 5 features were used for training instead of the whole 18 features set. Also it was determined that Harmonicity and Voice Breaks features were not helping in any case to increase classification performance. So we reduced the feature set to 15 features and implemented a feature selection technique to find the combination of features that reaches the highest Classification performance. The feature set used to train the ANFIS model, as well as the ANFIS parameters settings, are coded in a binary chromosome in the following way: 4 bits to choose the first feature, 4 bits to choose the second feature, 4 bits to choose the third feature and 4 bits to choose the first fourth.

In this way we choose between all the subsets from 1 to 4 elements taken from 15 features. As we mentioned before there are several parameters to generate a FIS for ANFIS training:

- Number of Membership Functions per Input: This is a scalar value. In this work we are using the same number for membership functions.
- Type of Membership Function for each input: This is an array of string that specifies the name of the function. In our case we use the same membership function for each input.
- Output membership function type: It can be linear or constant.

These parameters are coded in the binary chromosome in the following way: 2 Bits to choose between 2 to 5

membership functions, 2 Bits to choose one type of input function from Sigmoid, Bell Curve, Gaussian Curve and Two-sided Gaussian Curve and 1 Bit to choose the output function which can be Linear or Constant. These three parameters are used to Generate an initial Sugeno-type FIS for ANFIS training. We had 21 bits chromosomes. The chromosome composition is illustrated in Fig 1.
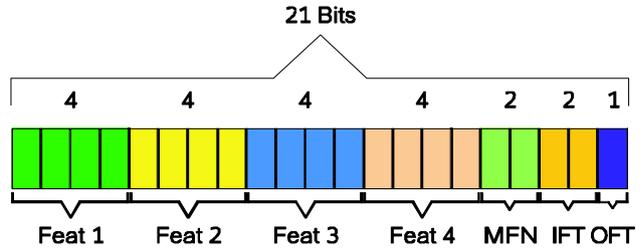
Fig. 1 Chromosome composition. Feat1-4 = Feature 1-4, MFN = Number of Membership Functions, IFT = Type of Input Function, OFT = Type of Output Function.

The genetic algorithm has the following configuration:

- Population Type: Binary
- Population Size: 20
- Crossover Fraction 0.8000
- Mutation Function: Gaussian
- Fitness Function: Unweighted Average (UA) recall resulting from the ANFIS classification.

VI. RESULTS

The final result from the genetic algorithm is the binary chromosome that reaches the best fitness score. In this case the chromosome was 0000 0100 0011 0110 01 10 0 and is illustrated in Fig 2. This means that the best configuration was:

- First Feature: No feature was selected
- Second Feature: Feature in position 4 which is Intensity Average
- Third Feature: Feature in position 3 which is Pitch Range
- Fourth Feature: Feature in position 6 which is Shimmer
- Number of Membership Functions: 5
- Type of Input Functions: Bell Curve
- Type of Output Function: Constant.

In table 2 we can see the confusion matrix obtained from the classification with the best individual.

	NEG	IDL	SUM
NEG	761	1704	2465
IDL	582	5210	5792

Table 2 Confusion Matrix
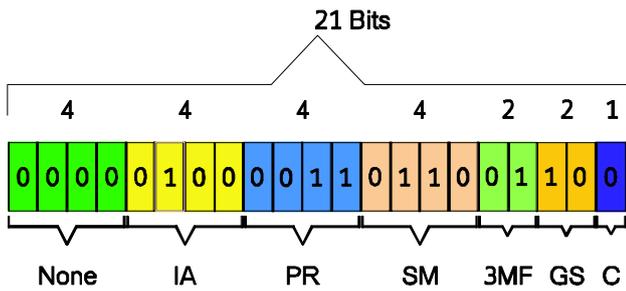
Fig. 2 The Best Chromosome Found. IA = Intensity Average, PR = Pitch Range, SM = Shimmer, 3MF = 3 Membership Function per each Input Variable, GS = Gaussian Curve Membership Function, C = Constant Type Output Function

To train the ANFIS model the FAU Aibo training set was divided using 70% as training data, 15% as checking data and 15% as testing data. We selected the 3 best individuals with the highest Unweighted Average Recall (UA Recall) to test them with FAU Aibo testing set. The predicted classes were evaluated on line through the INTERSPEECH 2009 Emotion Challenge System [9]. The other parameters for comparison are Weighted Average Recall (WA Recall), Weighted Average Precision (WA Precision) and Unweighted Average Precision (UA precision). Results are shown in table 3

WA Recall	UA Recall	WA Precision	UA Precision
72.31 %	**60.41 %**	**69.76 %**	**66.01 %**
72.07 %	59.90 %	69.41 %	65.58 %
71.90 %	59.62 %	60.41 %	65.58 %

Table 3 Results of the 3 best individuals found.

In tables 4 and 5 we present the feature combinations and ANFIS settings corresponding to the results presented in table 3. We can see that Intensity Average was selected in three cases, meaning that it is an important factor to discriminate between the two emotional states.

Feature 1	Feature 2	Feature 3	Feature 4
-	Intensity Average	Pitch Range	Shimmer
Intensity Average	-	Formant 1 Stdev	-
Formant 1 Stdev	Intensity Average	-	Maximum Pitch

Table 4 3 best feature combinations

Number of MF	Membership Function	Output Function
3	Gaussian	Constant
5	Gaussian	Constant
5	Sigmoid	Constant

Table 5 Results of the 3 best FIS initialization parameters

VII. CONCLUSION

We proposed some features to classify emotions in voice signals. These features are based on an abstraction of findings made by several authors. With the features proposed a hybrid classifier was trained, it combines neural networks and fuzzy logic. It was used a genetic algorithm to optimize the ANFIS training settings and to find the subset of features that provide the best classification results. The results indicate that the combination of characteristics: Intensity Average, Range and Pitch Shimmer, provided more information to discriminate between classes of Negative and IDLE emotions. Also we observed that both categories of attributes Utterance Pitch Contour Measures and Voice Quality Measures provide important information to classify emotional states in voice signal. Although the results were not significantly better than those presented as baseline it leads us to believe that all features studied here contain valuable information and encourage us to keep working on this feature set and to improve it. We also plan to test other classification methods based on fuzzy logic such as Type 2 Fuzzy Pattern Matching.

REFERENCES

[1] Vidrascu, L Devillers, L. Real-life emotion representation and detection in call centers data. Lecture Notes on Computer Science. v3784. 739-746. 2005
[2] González, G.M., 1999. Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in chicanos/latinos. Tech. Rep. 39, University of Michigan.
[3] Murray, I. R. and Arnott, J. L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", Journal of the Acoustical Society of America. v93 i2. 1097-1108, 1993.
[4] Boersma, P.,Weenink, D. Praat v. 4.0.8. "A system for doing phonetics by computer", Institute of Phonetic Sciences of the University of Amsterdam. February, 2002
[5] De Jong, N.H. and Wempe, T., "Automatic measurement of speech rate in spoken Dutch", ACLC Working papers, 2 (2), 49 58, 2007.
[6] Bhuta, T., Patrick, L., Garnett, J. "Perceptual evaluation of voice quality and its correlation with acoustic measurements", Journal of Voice, 18 (3): 299-304. 2004.
[7] J.-S. Roger Jang, " ANFIS: Adaptive-Network-Based Fuzzy Inference Systems," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 23, No. 03, pp 665-685, May 1993.
[8] Steidl, S.: "Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech", Logos Verlag, Berlin, 2009.
[9] Schuller, B.; Steidl, S.; Batliner, A.: "The Interspeech 2009 Emotion Challenge", Interspeech (2009), ISCA, Brighton, UK, 2009.

EVALUATION OF A PITCH ESTIMATION ALGORITHM FOR SPEECH EMOTION RECOGNITION

N. Vanello[1], N. Martini[2], M. Milanesi[3], H. Keiser[1], M. Calisti[4], L. Bocchi[4], C. Manfredi[4], L. Landini[1]

[1]Department of Information Engineering, University of Pisa, Pisa, Italy [2]Interdepartmental Research Center "E. Piaggio", University of Pisa, Pisa, Italy [3]MRI Laboratory, "G. Monasterio" Foundation, Pisa, Italy [4]Department of Electronics and Telecommunications, University of Florence, Florence, Italy

Abstract: **The analysis of parameters extracted from speech data may contribute, together with other approaches, to the analysis and classification of a subject emotional status. Pitch value and variability have been shown to carry useful information to reach this goal. However the non stationarity of running speech and the short duration of utterances represent a difficulty for the estimation of these parameters. In this work a method based on a variation of the Sawtooth Waveform Pitch Estimator (SWIPE') to estimate pitch and jitter in vowel sound, is evaluated. The performances of the approach are assessed on simulated datasets with varying signal to noise ratios and jitter values. Issues related to data length are introduced and discussed through simulations. A comparison of the approach performances with the Simplified Inverse Filtering Technique (SIFT) is presented. Preliminary results on vowels extracted from a database of emotional utterances are introduced.**

Keywords : **Pitch, jitter, swipe', emotion, vowels**

I. INTRODUCTION

The development of automatic methods to estimate subjects' psychological status has drawn the attention of the research community. The achievement of such information has several positive outcomes on fields such as psychology, for development of tools for patients monitoring or for improving occupational safety. To reach this goal multiparametric approaches have been proposed as those based on the acquisition of vital signs related to the activity of the autonomous and the central nervous systems and on the analysis of speech. As regards the latter approach several features have been proposed as those based on speaking rate, spectral characteristics and prosody [1][2]. Pitch related variables have been proposed as F0 level, range, contour and jitter. In particular F0 mean values and variability was found to be larger for angry and happy speech rather than neutral or sad speech [1].

The estimation of pitch represents a challenging task in running speech given the short duration of sounds and due to the noise [3]. Moreover the non stationarity of speech signals requires the use of short analysis windows thus allowing to estimate the changes of pitch across time.

In this work a method based on a variation of the Sawtooth Waveform Inspired Pitch Estimator, namely SWIPE' algorithm [5], is introduced for pitch and jitter estimation in vowels sounds. The approach is tested by using synthesized vowels and results are compared with those obtained by the Simplified Inverse Filtering Technique (SIFT) [6]. Application for classification of vowels as extracted by emotional utterances is introduced.

II. METHODS

Synthetic data were obtained by an autoregressive moving average exogenous (ARMAX) model. The parameters of the model were estimated from an healthy male /a/ vowel, with model orders for the AR, MA and X part equal to 16, 4 and 2 respectively. The model input for synthesis purposes was obtained with an impulse train sequence, whose distance between two successive pulses was modulated to produce the desired jitter. The amount of the imposed jitter was changed across different simulations ranging from a minimum of 0 to a maximum of 2 percent. The signal to noise ratio (SNR) of the simulated vowels was modified by inserting additive Gaussian noise at the model output.

Real dataset consisted of vowels extracted from a German database of emotional utterances [7]. Ten different sentences are repeated by different actors and labelled according to perceived emotional content, respectively as neutral, anger, fear, joy, sadness, disgust and boredom. Vowels were extracted from the sentences according to dataset labels and segmentation provided with the datasets. Dataset labeling and segmentation is based on auditive judgement supported by visual analysis of oscillogram and spectrogram, as described in [7].

SWIPE' algorithm measures pitch by estimating average peak to valley distance at harmonic locations. This goal is achieved by comparing the spectrum of the signal with that of cosine based kernel functions, thus weighting the pitch candidate and its harmonics according to a $1/\sqrt{f}$ law. This choice matches the decay trend of harmonics relative to vowels sounds. To avoid subharmoics of pitch being estimated as the real pitch, non prime harmonics, except the first one, are removed

from the kernel. This algorithm uses a window size related to the pitch to be estimated: in particular a Hann window size of length $T=4/f_0$ is chosen. We applied SWIPE' algorithm to a sliding window of length T seconds. A pitch value is estimated at every step. The algorithm results are evaluated for two different values of the window time shifts, namely T seconds, obtaining non overlapping windows, and T/4 seconds. This approach requires a two-step process, the first being a preliminary estimation of the pitch value. The time window length is then determined as $T=4/(\alpha f_0)$ where $\alpha<1$ is used to guarantee a sufficient time window length in the case of pitch underestimation after the first step. In this work $\alpha=0.9$ has been used. The pitch value is estimated as the mean value across windows. Jitter was estimated according to the following formula

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}|F_{i+1}-F_i| \Bigg/ \frac{1}{N}\sum_{i=1}^{N}F_i \qquad (1)$$

where F_i is the estimated pitch at the i-th window. As a comparison, the same procedure was applied using the well known SIFT algorithm [7]. SIFT algorithm is based on inverse filtering of the speech data, where the filter is obtained by inverting a low order linear predictor that models the vocal tract. The pitch is then estimated by computing the autocorrelation function of the residuals that are related to the exciting source of the vocal tract.

III. RESULTS

A. Simulated Data

In Fig. 1 (upper window) the percentage error of the estimated pitch, with respect to actual pitch, is shown as a function of SNR by using the SWIPE' based approach.
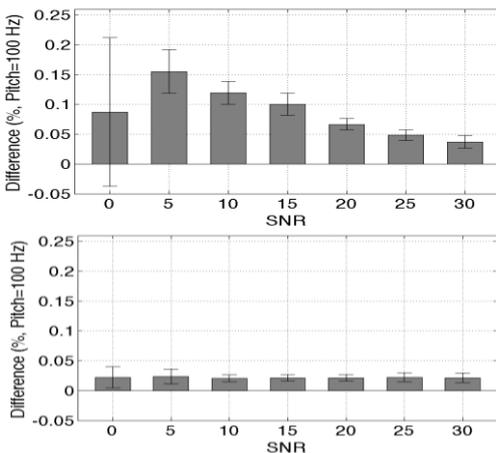
Figure. 1 Percentage error of estimated pitch at 100 Hz for different SNR values using SWIPE' and SIFT based approach (upper and lower window respectively).

Jitter estimation results for SNR=18 dB are reported in Fig. 2 to Fig. 4. The mean and the standard deviation of the estimated jitter are shown with respect to imposed jitter. For each value of the imposed jitter 20 different data segment were analyzed, each 300 ms long. In Fig. 2 the results obtained by using non overlapping windows are shown, for SWIPE' and SIFT based approach. The two algorithms yield similar results with small differences: in particular SWIPE' based approach is more accurate than SIFT based at lower jitter values and less accurate for higher jitter values.
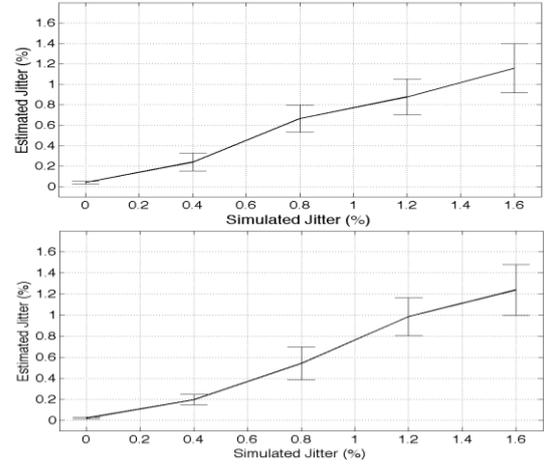
Figure 2 Estimated against imposed jitter, using SWIPE' (upper window) and SIFT (lower window) based approach and non overlapping windows. Total signal length: 300 ms.

In Fig. 3 the results obtained by using overlapping windows are shown. In this case the estimated jitter standard deviation is smaller than that obtained by employing non overlapping windows. The mean value of the estimated jitter found is always monotonically increasing with the imposed jitter. The results obtained with overlapping windows are less accurate than those obtained with non overlapping windows, resulting in an underestimation of the jitter.
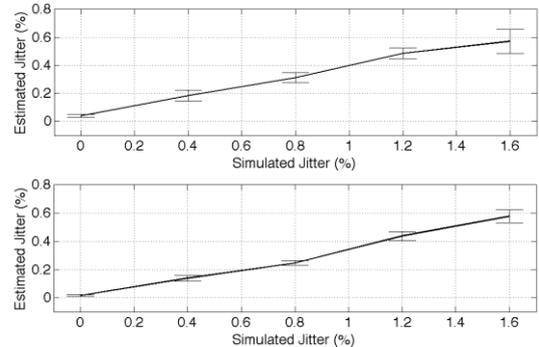
Figure 3 Estimated against imposed jitter, using SWIPE' (upper window) and SIFT (lower window) based approach and T/4 overlapping windows. Total signal length: 300 ms.

In Fig. 4 the results obtained by applying the proposed approach to a simulated vowel, 80 ms long, are shown. For lower jitter values the results obtained with the SWIPE' based approach are slightly more accurate. For higher jitter values the standard deviation of the estimated jitter is high. In particular SWIPE' based approach results in a very high standard deviation with respect mean jitter value.
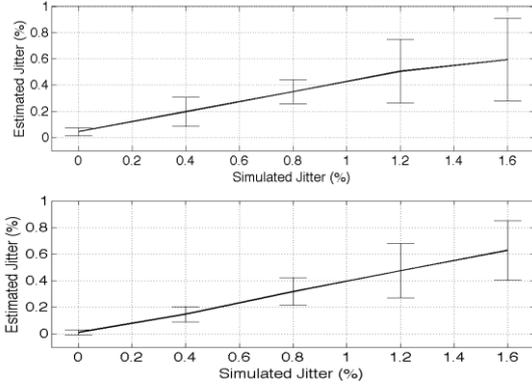
Figure 4 Estimated against imposed jitter, using SWIPE' (upper window) and SIFT (lower window) based approach and T/4 overlapping windows. Total signal length: 80 ms.

The standard deviation of the shown results is given by two sources, the estimation error of the algorithm and the trial by trial changes in the actual simulated jitter. In fact the jitter is simulated by imposing the standard deviation of the intervals between two successive pulses, given as input of the ARMAX model. By analyzing short duration windows, the actual jitter may be significantly different from the average jitter. The evaluation of the jitter from the pulse sequence used for the simulations, in fact resulted in standard deviation values equal to 0, 0.04, 0.09, 0.15 and 0.22 % for average jitter equal to 0, 0.4, 0.8, 1.2 and 1.6 % respectively. These values were estimated from the ARMAX input pulses using an 80 ms data segment and overlapping windows. In Fig. 5 the relationship between the expected jitter and the jitter estimated by the proposed approach, using SWIPE' (left) and SIFT (right) algorithm respectively, are shown.
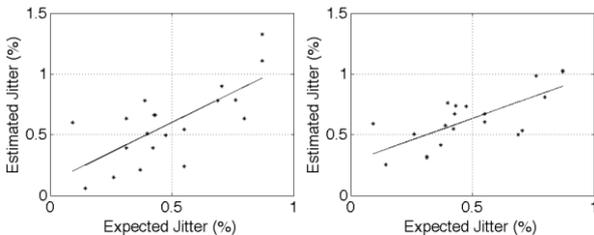
Figure 5 Regression model between estimated and expected jitter (imposed jitter=1.6%) using SWIPE' (left) and SIFT (right) based approaches (T/4 overlap).

These results pertain 20 data segment, 80 ms long, obtained imposing jitter 1.6% (see Fig. 4). The SWIPE' based approach in this simulation outperforms SIFT based one, in estimating actual jitter. A regression model is estimated describing the relationship between estimated and expected jitter. The r^2 statistic, the intercept and the slope for the linear regression models shown in the right (SWIPE' based) and the left (SIFT based) of Fig. 5 are respectively $r^2=0.51$, a=0.11, b=0.98 and $r^2=0.53$, a=0.28, b=0.71. The significance of the regression model was found to be reduced for lower imposed average jitters, resulting in lower variance estimates.

B. Real utterances

Results obtained from real dataset are shown in tables 1 and 2, for SWIPE' and SIFT based methods respectively. The results were obtained by applying the two approaches to the same vowels. Each vowel was extracted from a different subject. The results shown in Tables 1 and 2 are obtained using overlapping windows. Given the short time course of real utterances the proposed approach shows some limitations given the window length needed for pitch estimation. These preliminary results on real datasets highlight an increase of pitch values of vowels scored as anxiety and anger with respect to boredom and neutral. As regards jitter values an increase was observed predominantly in vowels scored as anger.

Table 1. Pitch (p in Hz) and jitter values (%) as estimated from real utterances using SWIPE' and overlapping windows.

Vowel	Neutral	Anxiety	Anger	Boredom
/e/	p=136 jitt=1	p=170 jitt=0.45	p=220 jitt=1.89	p=132 jitt=0.5
/i/	p=109 jitt=0.1	p=144 jitt=0.63	p=250 jitt=2.15	p=118 jitt=2.9
/u/	p=140 jitt=0.8	p=250 jitt=1.6	p=248 jitt=1.3	p=113 jitt=0.5
/a/	p=115 jitt=0.6	p=125 jitt=0.9	p=156 jitt=0.27	p=140 jitt=0.16

Table 2. Pitch (p in Hz) and jitter values (%) as estimated from real utterances using SIFT and overlapping windows.

Vowel	Neutral	Anxiety	Anger	Boredom
/e/	p=138 jitt=0.87	p=170.2 jitt=0.6	p=217 jitt=2.27	p=134 jitt=0.59
/i/	p=110 jitt=0.49	p=144.8 jitt=1.03	p=247 jitt=2.24	p=117.9 jitt=1.75
/u/	p=140 jitt=0.8	p=250 jitt=1.6	p=248 jitt=1.3	p=113 jitt=0.5
/a/	p=115.9 jitt=0.67	p=123 jitt=1.27	p=155 jitt=0.51	p=141 jitt=0.22

More severe problems are found using the approach with non overlapping windows, since the jitter estimation in case of low pitch short utterances may not be possible. The results obtained with SWIPE' and SIFT based approach are similar as regards pitch estimation, while jitter could not be estimated in most of the cases (data not shown).

IV. DISCUSSION

Pitch estimation with the proposed method was achieved with an error smaller than 0.2% in the worst case (SNR=0) and improves considerably at higher SNRs. The comparison with SIFT based approach highlights that that the latter approach offers better results. However comparison in a wider frequency range is not explored in this work. We have to stress that a comparison of the original SWIPE' algorithm with other approaches can be found in [5], while the results in this work are related to a different approach, described in the Methods section, that may take advantage of SWIPE' as well as other pitch estimation algorithms. Given this observation this work does not aim at evaluating the SWIPE' algorithm per se but to evaluate a SWIPE' based approach for pitch and jitter estimation in short time vowels.

Jitter estimation resulted in an estimated mean value monotonically increasing with imposed jitter. The jitter values are underestimated and the average slope of the obtained results reduces in the case of overlapping windows. These results are in good agreement with those expected, given the fact that pitch is estimated by using $4/f0$ seconds long windows. This choice implies an average of the pitch changes across 4 glottal cycle repetitions resulting in a systematic underestimation of the real jitter value. In the case of overlapping windows smaller changes in pitch estimation are to be expected resulting in smaller value calculated as in (1). Since our final aim is to look at possible changes of these values with respect to the expressed emotions, this issue may not represent a limitation. Moreover by using overlapping windows it is possible to give an estimate of jitter value in shorter utterances. The need for the SWIPE' algorithm to have a 4 period long time window in order to have an optimal estimate, may impose severe limits to jitter estimation for short utterances characterized by a low pitch value. In fact for 100Hz mean pitch value, a data window 40 ms long is needed. In this conditions, if the overlapping windows approach is used, a 60 ms data length would allow to estimate 3 pitch values.

By analyzing the relationship between expected and estimated jitter for higher values of the imposed jitter, a linear regression model was found to be significant or close to significance. This result shows that a large part of the jitter variance in Fig. 5 can be explained as trial by trial jitter variance in short simulated dataset.

Furthermore, as it could be drawn from results in Fig. 5 the proposed approach is more robust than expected from results shown in Fig. 4 in tracking jitter chnages. The analysis confirmed however that a significant portion of variance may be related to estimation error. This result should be taken into account when analyzing real data. From the results here shown, no strong significant differences were highlighted between the SWIPE' based and SIFT based approaches as regards jitter estimation.

The preliminary results on real dataset seem to indicate a significant jitter difference in vowel scored as anger. The proposed approach could be applied only using overlapping windows given the short duration of extracted vowels. Moreover the estimation of jitter values for short duration, low pitch vowels may not be possible or it may results in bad estimates given the small number of pitch periods available. An analysis of the time profile of pitch and jitter was out of the scope of this work, that was motivated by the need of characterizing the approach on short, quasi stationary vowels. Future work should take into account the analysis of long sentences. However the interpretation of the results on real datasets may take into account these considerations.

V. CONCLUSION

The proposed approach allows estimating pitch with good performances. Simulated data results show that an index proportional to jitter value can be estimated as well, allowing to employ this method for classification purposes. Preliminary results on real dataset indicate the potential application to running speech albeit with some limitations in the case of short utterances at low pitch.

REFERENCES

[1] Bulut, M, Narayanan, S. On the robustness of overall F0-only modifications to the perception of emotions in speech. J Acoust Soc Am 2008;123(6):4547-4558

[2] Tao, J, Kang, Y, Li, A. Prosody Conversion From Neutral Speech to Emotional Speech. IEEE Trans Audio Speech Lang Processing 2006;14(4):1145-1154

[3] Vasilakis, M, Stylianou, Y. Spectral jitter modeling and estimation. Biomedical Signal Processing and Control 2009; Available on-line

[4] Boyanov B, Hadjitodorov S. Acoustic analysis of pathological voices. IEEE Eng Med Biol Mag 1997;16(4):74–82

[5] Camacho, A, Harris, J. A sawtooth wave form inspired speech estimatore for speech and music. J Acoust Soc Am 2008; 124(3):1638–1652

[6] Markel, JD, The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Audio Electroacoust 1972; 20: 367-377

[7] Burkhardt, F, Paeschke, A, Rolfes, M, Sendlmeier, W, Weiss, B. A Database of German Emotional Speech, Proc. Interspeech, Lisbon, Portugal, 2005; 1517–1520

PROSODY FEATURES ANALYSIS

Jana Krutišová, Jana Klečková
Department of Computer Science and Engineering,
University of West Bohemia, Pilsen, Czech Republic

Abstract: **Prosody is a set of non-verbal features by which a speaker exhibits his attitude, his idea and his emotions. In a verbal communication it helps to understand one another. A production of a speech, and therefore also the prosodic characteristics depend on the physiological characteristics of a voice organ. These characteristics depend not only on sex or age, but also on the quality of a voice organ. This paper describes an idea to use suitable tools for storage and multidimensional analysis of prosody features. This work has been partly supported by the Ministry of Education of the Czech Republic in the National Research Programme II- project 2C06009.**

Keywords : **verbal communication, multidimensional modeling, prosody**

I. INTRODUCTION

A speech is a basic means and a form of human communication. In recognition and understanding of a spontaneous speech is important, what a speaker says, but also how he says it. People do not speak monotone word by word, but they use elements, that affect and accent the interpretation of the verbal utterance. Every speaker represents his attitude and his emotion, for example joy, anger, sadness, surprise or fear due to many variations in melody, intonation, pauses and accent of his speech.

Due to these characteristics both the same speaker may modify the speech, although it is the same message in term of the semantic and syntax aspect, both semantically and syntactically the same speech uttered by different speakers shows different values of monitored characteristics, depending on the context of utterances, expressed the attitude of speaker, the environment in which the speech is made. The melody of speech expresses different emotions; intonation reveals the origins, geographical factors, social status of speaker, etc.

The set of these non-verbal attributes is called prosody. Prosody is a natural part of the verbal communication. Prosodic features have a key role in a speaker recognition.

The production of a speech, and therefore also the prosodic characteristics depend on the physiological characteristics of voice disorders. Speech oscillations are produced in the cooperation of the vocal cords, throaty cavity - oral and nasal, soft and hard palate, teeth and tongue, with the support of the lungs and respiratory muscles. The frequency of vocal cord oscillations F0 characterizes the fundamental tone of a human voice and it is one of the fundamental prosodic characteristics. The frequency range F0 is affected by age, sex, but also depends on the sentence melody, emotion, or fatigue, and speaker´s health.

A communication may be significantly affected, impaired or even made impossible by certain pathological changes of the voice organ. The changes influence the actual production of the voice and the voice disorders affect the speech. The quality of a voice organ then becomes one of the fundamentals affecting prosodic characteristics.

II. METHODOS

The characteristic appearance of a voice disorder is hoarseness. Physical fundamentals of hoarseness are irregular oscillations of the vocal cords and incorrect closing of the glottis.

A phoniatric investigation of persons help to determine diagnosis and it repeats during the treatment in the pre-defined time period. Reference values of tests are obtained as outputs of the investigative-diagnostic tools and methods that have the character of an objective assessment, the result of subjective-expert assessment of voice quality is a diagnosis.

For another group, for example future teachers or singers, the voice investigation is carried out by reason of their professional interests. One of the methods for testing the state of the voice system is multi-dimensional analysis of the voice. The method determines numerous qualitative parameters of voice and compare with the normal, i.e., parameters measured in healthy subjects with normal voice. Another method is a determination of the Voice Range Profile, which is used to measure quantitative parameters of the voice. The measured values are subsequently stored and may be evaluated during the observation, therapeutics and the post-surgery rehabilitation. The voice profile of the patient is recorded before surgery and then monitored over a period usually two weeks, two months, six, twelve and twenty-four months after surgery. From the experimental point of view, it is important to identify what affects the reference values and in what reciprocal context these influences operate. For the purpose the records obtained in examinations should be appropriately categorized.

III. RESULTS

The experimental data was acquired during the five years at the Otorhinolaryngology Clinic of the University Hospital in Pilsen. The set of the data includes more than 1,600 records from the investigations of voice apparatus of persons of a different age. There are about 274 men and 277 women monitored at different times - before the surgery, after the surgery of vocal cords, during the voice rehabilitation, some records concerning the healthy subjects before and during the study at the faculty of education.

The concept of structure for data storage is based on the assumption that the doctor performs the investigation in defined method which commonly would result in n-dimensional structure of values to be evaluated and assessed. The current set of patients can be analyzed not only by sex, age representation, diagnosed disorders, but also by other keys, such as a period of their rehabilitation. In the future there are another interesting views for analysis, for example groups of professional speakers, or whether patient is a smoker or a non-smoker.

From the standpoint of multidimensional modeling terminology is the determination of dimensions. Tables of dimensions, together with the table of facts create the basis of a multidimensional schema. Dimension tables are expressed segmentation or categorization. There are measured values contained in the table of facts, which are dependent on the dimensions.

IV. DISCUSSION

For a small data set and especially small number of aspects on which the data are evaluated, can be used 3D contingency table. Data can be presented graphically too.
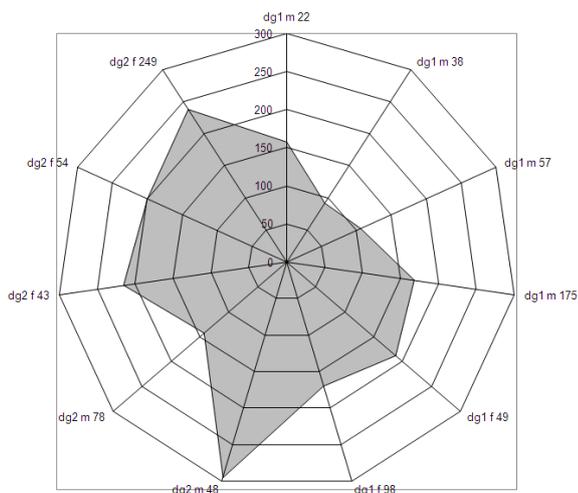
Fig. 1 Fundamental frequency of selected patients

For example we can consider a sample of patients and measured values of their fundamental frequency. For each patient we have information on sex (f, m) and diagnosis (dg1, dg2). A graphic presentation of this case looks like a spider graph on the figure *Fig. 1*.

There are more important aspects for analysis of measured values. Therefore in the area of multi-dimensional and heterogeneous data is appropriate to use the principles of multidimensional modeling and OnLine Analytical Processing (OLAP) tools, which are available for the analytical data processing and decision support. In this particular case, the results of the analysis, for example, could contribute to design scheme of patients' aftercare and their rehabilitation.

V. CONCLUSION

Computer data processing significantly facilitates such like operations in which it is necessary to summarize the data, according to the requirements of various data combinations and categorization. However, it is accepted that information obtained by the treatment is to some extent subjective. The quantity of the derived information is dependent on the recipient, his ability to use the information for a future decision making. The same can be expected from a case of medical data, where the resulting information extracted from the measured values analyzing is supported by subjective evaluation of an expert, his experience and knowledge. In this area it must be consider that the analyzed object, every patient is to some extent an individual one with his unique properties. Therefore it is important to create certain groups or categories under which data could be assessed. This idea leads to the use of multidimensional modeling and tools that support an approach to answer multidimensional analytical queries.

REFERENCES

[1] Nový, P., Vávra, F., Pešta, J., Marek, P.: *Parameter Identification from Phoniatrical Examinations*, Summer School DATASTAT 06, Proceedings, (221-234), Brno: Masaryk University, 2007, ISBN 978-80-210-4493-7.

[2] Vávra, F., Nový, P., Mašková, H.: *Voice range profile and problem of fundamental frequency,* Summer School DATASTAT 03, Proceedings, (257-271), Brno: Masaryk University, 2003, ISBN: 80-210-3564-1.

[3] Inmon, W. H.: *Building the Data Warehouse – Second Edition,* JOHN WILEY & SONS, Inc. New York, 1996.

Voice quality assessment I-II

A CLINICAL WORKSTATION SOFTWARE FOR VOICE QUALITY ASSESSMENT

J. Cai[1], A. Alpan[1], T. Dubuisson[2], I. Verduyckt[3], F. Grenez[1], J. Schoentgen[1, 4]

[1] Laboratory of Images, Signals & Telecommunication Devices, Université Libre de Bruxelles, Brussels, Belgium

[2] Service de Théorie des Circuits et Traitement du Signal, Faculté Polytechnique de Mons, Mons, Belgium

[3] Service ORL-ORLO, Cliniques Universitaires UCL de Mont-Godinne, Yvoir, Belgium

[4] National Fund for Scientific Research, Belgium

Abstract: **This paper presents the design and implementation of a clinical workstation software for analyzing voice disorders. The software is developed by using Java technology and MySQL database system. A variety of vocal cues, e.g. jitter and shimmer, that describe irregularities of speech cycles in sustained vowels can be automatically derived by the system. For assessing voice disorders in connected speech, a vocal cue called signal-to-dysperiodicity ratio is evaluated by carrying out a generalized variogram analysis. In the development, special attention has been paid to software engineering conventions and the principles of architectural design of software structures to achieve good quality attributes such as developmental simplicity and modifiability. Preliminary tests have shown that the system provides satisfactory usability and performance for clinical applications.**

Keywords: **Pathological voice assessment, disordered voice analysis, software engineering, Java application**

I. INTRODUCTION

Disordered voice timbres are usually caused by improper vibrations of the vocal folds, as a consequence of pathological changes of the larynx. Recently, voice disorders have been observed more frequently and more extensively than before because of an increasing number of professional voice users. Therefore, reliable and efficient means of evaluating pathological voice quality are required for the assessment and prevention of laryngeal problems.

In clinical voice evaluation, acoustic assessment methods have been used to facilitate the clinical documentation of vocal problems because previous experiments have established that vocal cues exist that are clinically relevant [1, 2]. Furthermore, these acoustic-based methods have the advantages of non-invasiveness and quantitativeness. In this paper, the design and implementation of a clinical workstation software for analyzing pathological voice signals are presented. A variety of vocal cues such as jitter, shimmer, and harmonics-to-noise ratios in temporal and spectral domains, which describe irregularities of speech cycles in voiced speech can be automatically obtained for sustained vowels by means of the system. In addition, connected speech quality assessment is also included. The reason is that in clinical practice, people consider connected speech to be more informative than sustained vowels. Moreover, the perceptual evaluation of voice quality is likely to be based on both connected speech and sustained vowels uttered by the same patient. Variogram-based analysis is carried out to track dysperiodicities in connected speech, and thus a signal-to-dysperiodicity ratio value is obtained as a vocal cue of voice disorders [3].

Graphical means, e.g. spectrogram, phonetogram, and spider charts, are available to visualize the analysis results. Java technologies have been utilized to build the application system, mainly for the purpose of facilitating portability on different operating-system platforms. Software engineering conventions and the principles of architectural structures design have been used to guide the design and development of the system, to achieve developmental simplicity, modifiability, and other quality attributes.

II. METHODS

A. Disordered Voice Analysis

Because of technical feasibility, voice analysis is usually performed by providing different vocal cues of voice disorders for sustained vowels only. In this clinical software, the voice disorder analysis is versatile in terms of that not only sustained vowels but also connected speech segments can be assessed. For sustained vowels, features such as mean and standard deviation of fundamental frequency, jitter, shimmer, and harmonics-to-noise ratios in both time and spectral domains [4, 5] can be obtained to describe voice disorders.

A technique called speech sample salience analysis [6] has been used to perform voice cycle detection. Conventional voice cycle detection relies on the selection of signal peaks from several candidates. The selection technique usually assumes that the signal peaks are regularly spaced in time so that they can be determined one by one based on an a priori estimation of the typical

fundamental period. This assumption does not apply to disordered voices, though it is valid for modal ones.

Speech salience analysis can be performed without an a priori knowledge of the typical cycle length. Thus, it is well suited for tracking vocal cycles in pathological voices. For a speech signal $v(n)$ of M samples, the salience $S(k)$ of the kth sample ($0 \le k < M$) is defined as the length of the longest interval within which that sample is a maximum. Based on a sliding analysis window technique, which eliminates the bias related to the arbitrary position of the signal origin, salience analysis is performed sample by sample to determine a speech cycle sequence which minimizes the standard deviation of the durations of all cycle candidates. Based on this sequence, jitter, the vocal cue for describing the small random perturbation of voice cycle lengths, can be calculated.

For assessing the voice quality in connected speech, the segmental signal-to-dysperiodicity ratio (SDR) [7, 8], which is based on generalized variogram analyses, is used to summarize vocal perturbations. For a stationary signal $x(n)$, the variogram, which is defined in Eq. (1), is a measure of the departure from periodicity over an interval of length N:

$$v(T) = \sum_{n=0}^{N-1} \left[x(n) - x(n-T) \right]^2 , \quad (1)$$

where the variable lag T satisfies $-T_{max} \le T \le -T_{min}$ and $T_{min} \le T \le T_{max}$. T_{min} and T_{max} are, in number of samples, the shortest and longest acceptable glottal cycle lengths. They are fixed to 2.5 ms and 20 ms, respectively (i.e., 50 Hz $\le F_0 \le$ 400 Hz). For voiced speech sounds, the lag T_{opt}, which minimizes (1), is interpreted as a multiple of the speech cycle length.

Speech signals are expected to be locally stationary at best. A weighting coefficient can be inserted to account for slow changes in signal amplitude. Therefore, the variogram computation defined in Eq. (1) can be modified as follows:

$$v(T) = \sum_{n=0}^{N-1} \left[x(n) - \alpha\, x(n-T) \right]^2 \quad (2)$$

The coefficient α is defined so as to equalize the signal energies in the current and shifted analysis windows:

$$\alpha = \sqrt{\frac{E}{E_T}} , \quad (3)$$

where E and E_T are the signal energies of the current and the lagged frames. The frame length N and frame shift length are set to the value of 2.5ms, which guarantees that each signal frame is included exactly once in the analysis. The instantaneous value of the dysperiodicity is estimated as follows:

$$e(n) = x(n) - \alpha\, x(n-T_{opt}), \; 0 \le n \le N-1 \quad (4)$$

where T_{opt} is equal to the lag which minimizes the variogram for the current frame position. For a given fragment of connected speech, the analysis interval is divided into K blocks of length M and the SDR of each block of length 20 ms can be computed as follows:

$$SDR(k) = 10\log \frac{\sum_{n=Mk}^{Mk+M-1} x^2(n)}{\sum_{n=Mk}^{Mk+M-1} e^2(n)}, \; 0 \le k \le K-1 \quad (5)$$

B. Software Architecture Design

The objective is to develop a workstation application system which runs on a normal PC platform and supports pathological voice quality assessment. Modifiability, usability, and portability are the software quality attributes which are emphasized throughout the development of this system. To achieve portability, Java programming language and MySQL database system (MySQL Community Server) are selected as the development infrastructure because of their cross-platform availability.
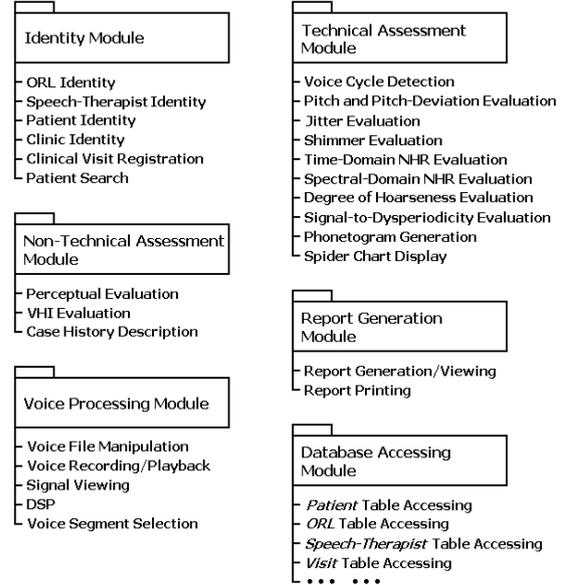
Fig. 1 Module decomposition View of the System

Architecture Tradeoff Analysis Method [9] is employed to devise the architectural structures of the software. To guarantee the clinical usability of the system, different stakeholders in the system's development, including developers of disordered voice analysis methods, otolaryngologists, speech therapists, and software engineers, have been involved in designing the functionalities. Based on the use case views [9] which specify the behavior of both the system and the users, a module decomposition view (see Fig. 1) is determined to assign the functionalities to different modules of the system. All the functionalities of a particular data-processing type are organized in a specific module. For

instance, the computation and graphical display of different vocal cues of voice disorder are grouped into the Technical Assessment Module.

A layered architecture is adopted to allocate the functional modules into two tiers, shown in Fig. 2. The data access logic implemented by the Database Accessing Module is contained in a separate Data Tier which is dedicated to database communication. Above the tier, there is the Application Component Tier that packs all the other modules carrying out the presentation logic and application-relevant logic. This layered architecture forms the basis for the development project's organization; i.e., the source code files are organized into packages according to the module decomposition and the layered architecture. The module decomposition also ensures that possible changes of the system are localized to only one or a few small modules, enabling a large part of the system's modifiability.
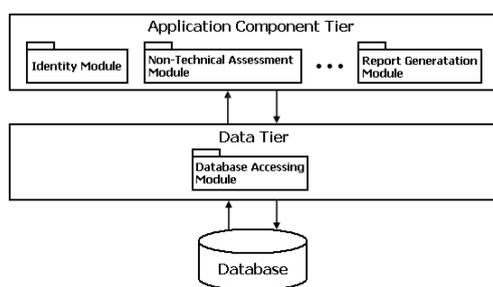
Fig. 2 Two-Tier Software Architecture

C. Functionality Design

The Identity Module is used to maintain the identity information of ORLs, speech therapists, and patients. Also, the information related to each clinical visit, e.g. the diagnosis, the type and state of the therapy, can be recorded by using this module. Fig. 3 shows the user interface for visit registration. Since a clinical assessment is typically performed at each clinical visit, each recorded voice and the relevant assessment results are uniquely associated with a specific visit. Therefore, the registration of visit information is designed as the first step which triggers the other events in the assessment, as well as creates a data slot in the database to store the visit information and the assessment results. To achieve flexible data management, a Patient Search sub-module is designed to enable the user to find the clinical data of patients by using different search conditions such as name, pathology, or a certain ORL.

A specific module, the Non-Technical Assessment Module, is designed to facilitate the patient interview and perceptual voice quality evaluation, which constitutes the beginning of the assessment process. To inquire about the case history, a questionnaire format [10] has been chosen to perform the interview. A VHI (Voice Handicap Index)

Questionnaire sub-module is designed to facilitate the assessment of the patient's perception of discomfort, handicap, and distress resulting from voice difficulties. For singing voice and speaking voice, different sets of questionnaires have been designed to derive the VHI score.
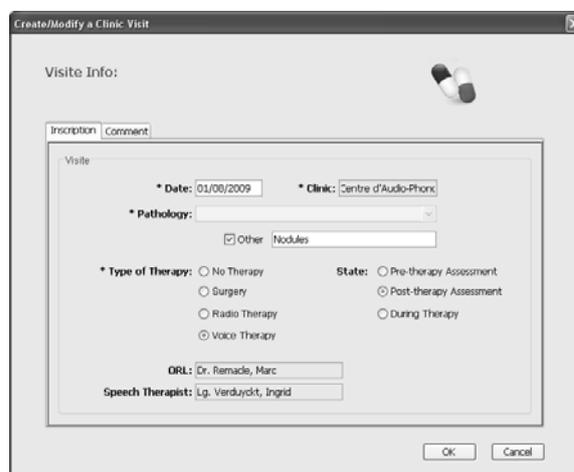
Fig. 3 User Interface for Visit Registration

The Voice Processing Module enables the user to record patients' voices, including sustained vowels and connected speech fragments, with a high quality microphone and a built-in sound card which performs 16 bit, 44.1 kHz sampling. All the voice signals can be stored as files in WAV format. With the toolbox in this module, the user also can playback the recorded voice and view its waveform or spectrogram. Voice samples can be selected for subsequent computation of the vocal cues of voice disorder.

All the vocal cues are computed in the Technical Assessment Module. For a particular voice sample, seven vocal cues can be computed the values of which can be visualized by means of spider charts. The software can display the spider charts for two different voice samples in the same axes system in an overlapped manner. Fig. 4 depicts such an example. The spider chart of the voice sample for the pre-therapy assessment of a patient is plotted in dark color, while the spider chart of another voice sample related to the post-therapy assessment of the same patient is plotted in a bright color. By using overlapped spider charts, it is easy to compare the post-therapy vocal cue values with the pre-therapy ones for a certain patient, and to compare the vocal cue values of two voices from different patients. Therefore, this kind of display provides a useful means of evaluating the effectiveness of therapy. Besides the spider chart, another graphical tool – the phonetogram – is implemented in the system to depict the dynamic ranges of both the pitch and the intensity of the voice.
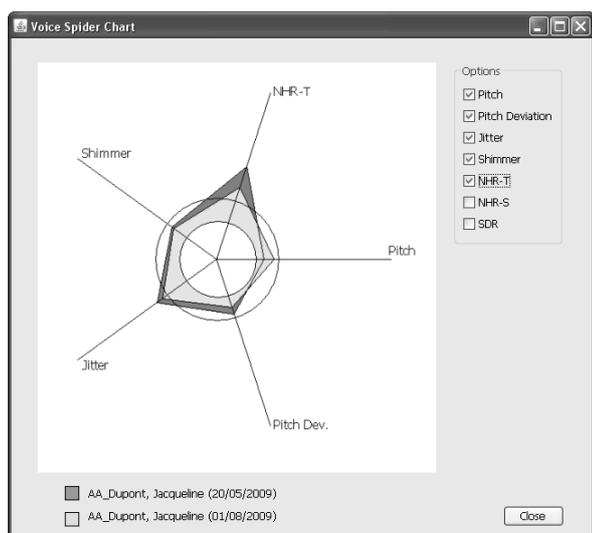
Fig. 4 Overlapped Spider Charts

III. RESULTS

Java programming language has been used to build the software for the sake of its portability to embedded systems. MySQL Community Server has been chosen as the database management system because of its speed, flexibility and reliability. The system is organized in a layered structure which therefore supports the modifiability of the system and the work assignment in the development team. The ergonomic design of the system makes the user interface easy to use owing to the conventional sequencing of the tasks which clinicians are expected to perform when assessing laryngeal function. All the acoustic cues of disordered voices and the relevant diagrams can be generated in real-time or quasi-real-time by the software running on a normal PC platform. Preliminary tests have shown that the system has a satisfactory usability and performance, though further clinical tests and development remain to be carried out to establish its suitability for pathological voice assessment.

IV. CONCLUSION

The Java-based voice assessment workstation software can be deployed on almost all PC platforms. A variety of vocal cues of voice disorders can be provided by the system for both sustained vowels and connected speech to support the clinical voice quality assessment. By using the principles and methods of architectural structures design in the development of the system, the quality attributes of the software, such as the developmental simplicity, modifiability, and usability can be well achieved.

REFERENCES

[1] J. Schoentgen, Vocal Cues of Disordered Voices: An Overview", *Acta Acustica united with Acustica*, vol. 92(5), pp. 667-680, 2006.
[2] S. Hadjitodorov, and P. Mitev, "A Computer System for Acoustic Analysis of Pathological Voices and Laryngeal Diseases Screening", *Medical Engineering & Physics*, vol. 24(6), pp. 419-429, 2002.
[3] A. Kacha, F. Grenez, and J. Schoentgen, "Estimation of Dysperiodicities in Disordered Speech", *Speech Communication*, vol. 48(10), pp. 1365-1378, 2006.
[4] S. N. Awan, and M. L. Frenkel, "Improvements in Estimating the Harmonics-to-Noise Ratio of the Voice", *Journal of Voice*, vol. 8(3), pp. 255-262, 1994.
[5] Y. Qi, and R. E. Hillman, "Temporal and Spectral Estimations of Harmonics-to-Noise Ratio in Human Voice Signals", *J. Acoust. Soc. Am.*, vol. 102(1), pp. 537-543, 1997.
[6] C. Mertens, F. Grenez, and J. Schoentgen, "Speech Sample Salience Analysis for Speech Cycle Detection", (To be published, INTERSPEECH 2009)
[7] A. Kacha, F. Grenez, and J. Schoentgen, "Frame-Based Acoustic Cues of Vocal Dysperiodicity in Connected Speech", in *Proc. of ICASSP 2006*, vol. 1, pp. I385-I388, 2006.
[8] A. Alpan, A. Kacha, F. Grenez, and J. Schoentgen, "Assessment of Vocal Dysperiodicities in Connected Disordered Speech", in *Proc. of INTERSPEECH 2007*, pp. 1178-1181, 2007.
[9] L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice*, 2nd ed., Addison Wesley, 2003, pp. 99-124, 153-170.
[10] J. C. Stemple, L. E. Glaze, and B. G. Klaben, *Clinical Voice Pathology, Theory and Management*,3rd ed., Singular, 2000, pp. 156-175.

MODULATION SPECTRAL FEATURES FOR OBJECTIVE VOICE QUALITY ASSESSMENT: THE BREATHINESS CASE

Maria Markaki[1], Yannis Stylianou[1,2]
[1]Computer Science Dept, University of Crete, Greece
[2]Institute of Computer Science, FORTH, Crete, Greece

Abstract: **In this paper, we employ normalized modulation spectral features for objective voice quality assessment regarding breathiness. Modulation spectra usually produce a high-dimensionality space. For classification purposes, the size of the original space is reduced using Higher Order Singular Value Decomposition (SVD). Further, we select most relevant features based on the mutual information between the degree of breathiness and the computed features, which leads to an adaptive to the classification task modulation spectral representation. The adaptive modulation spectral features are used as input to a Naive Bayes (NB) classifier. By combining two NB classifiers based on different feature sets a global classification rate of 79% for breathiness was achieved.**
Keywords: Objective voice quality assessment, breathiness, modulation spectrum, mutual information, SVD

I. INTRODUCTION

Objective voice quality assessment has been introduced to assist the perceptual evaluation of dysphonic voice quality used by the clinicians. The most common systems of pathological voice description refer to the degree of "hoarseness" [1]. Hoarseness is perceptually related to the noise generation during phonation. The degree of voice hoarseness can be quantified according to the GRASB (grade, roughness, asthenicity, strain and breathiness) scale proposed by Hirano [1].

The definition of these quantifiable perceptual dimensions (GRASB parameters) is related to a set of descriptive parameters for acoustic phenomena. The perceived voice abnormality is assumed to originate at the vocal source rather than resulting from abnormalities in the vocal tract configuration. Hence, many studies have focused on parameters such as pitch perturbation quotient (PPQ), jitter, shimmer, harmonics to noise ratio, etc. [2, 3, 4]. Acoustic measures that highly correlate with voice alterations can be associated then with a classification system to provide an automatic decision.

In this work we investigate the correlation of modulation spectral features [7, 8] to the degree of breathiness

(B) of pathological voices. Dysphonic voices are characterized by frequency-band dependent, time-varying amplitude fluctuations [5]. Modulation spectral features can capture a class of source mechanism characteristics related to voice qualities (glottal source differences) [5]. Breathiness typically refers to the voice quality related to the audible turbulence generated at the glottal level; this turbulence acts as a noise source to the vocal tract (see [9] and references therein). This paper pursues a previous work in which the authors presented an automatic dysphonia recognition and classification system built on modulation spectral representations [10].

The paper is organized as follows: In Section II we briefly describe the dataset, modulation spectral features and their normalization, and the method of dimensionality reduction and feature selection we use. Specifically, the initial representation is first transformed to a lower-dimensional domain using Higher Order SVD [11]. Projection of modulation spectral features on the principal axes with the higher energy in each subspace results in a compact set of features with minimum redundancy. We further estimate the relevance of these projections to dysphonic voice characterization based on their mutual information to breathiness class variable. Section III describes the experiments we conducted on breathiness classification using a combination of two naive bayes (NB) classifiers based on different feature sets [13]. Finally in Section IV we summarize features of our approach and discuss next steps.

II. METHOD

A. *Dysphonic voice corpus*

We used a database provided to us by Universidad Politécnica de Madrid, which is referred to as Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [14]. Similar to MEEI, PdA contains recordings of sustained vowels (/a/) and was developed for voice function assessment purposes. The voices of 201 dysphonic and 209 normal subjects have been classified according to the B parameter (breathiness) of the Hirano's GRASB

scale. A four-point scoring system is used to rate each subject along the B dimension: 0 denotes no breathiness, 1 means a slightly breathy voice, 2 refers to moderate breathiness, whereas 3 describes a severely breathy voice. For the following experiments, we selected 200 dysphonic subjects (74 men and 126 women, aged 11 to 76) affected by nodules, polyps, oedema, etc, as well as 24 subjects with normal voice (7 men and 17 women, aged 17 to 54). Specifically, we used 26 dysphonic (plus 24 normal) voices with zero breathiness, 50 voices with $B = 1$, 119 with $B = 2$ and 3 voices with $B = 3$. Due to the very small number of subjects with a breathiness rating equal to 3, these were joined with the subjects with a rating 2 breathiness.

B. Modulation Spectra

The most common modulation frequency analysis framework [8] for a discrete signal $x(n)$, initially employs a short-time Fourier transform (STFT) $X_k(m)$

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_K^{kn}, \quad (1)$$
$$k = 0, \dots, K - 1,$$

where $W_K = e^{-j(2\pi/K)}$ and $h(n)$ is the acoustic frequency analysis window with a hop size of M samples (m denotes time). Mel scale filtering can be employed at this stage in order to reduce the number of frequency bands. Subband envelope detection - defined as the magnitude $|X_k(m)|$ or square magnitude of the subband - and their frequency analysis with Fourier transform are performed next:

$$X_l(k, i) = \sum_{m=-\infty}^{\infty} g(lL - m)|X_k(m)|W_I^{im}, \quad (2)$$
$$i = 0, \dots, I - 1,$$

where $g(m)$ is the modulation frequency analysis window and L the corresponding hop size (in samples); k and i are referred to as the "Fourier" (or acoustic) and "modulation" frequency, respectively. Tapered windows $h(n)$ and $g(m)$ are used to reduce the side lobes of both frequency estimates.

A modulation spectrogram representation then, displays modulation spectral energy $|X_l(k, i)|$ (magnitude of the subband envelope spectra) in the joint acoustic/modulation frequency plane. In order to enable cross-database portability of the classification system, feature subband normalization has been employed according to [15].

C. Normalized modulation spectra

The distribution of envelope amplitudes of voiced speech has a strong exponential component. Hence we cal-culate modulation spectra using a log transformation of the amplitude values $|X_k(m)|$ and subtracting their mean log amplitude before windowing in (3):

$$\hat{X}_k(m) = \log|X_k(m)| - \overline{\log|X_k(m)|} \quad (3)$$

where $\{\bar{\cdot}\}$ denotes the average operator over m. This is analogous to the cepstral mean subtraction approach, which is commonly employed to compensate for convolutional noise in the case of MFCC features. Next, we normalize every acoustic frequency subband with the marginal of the modulation frequency representation:

$$X_{l,sub}(k, i) = \frac{X_l(k, i)}{\sum_i X_l(k, i)} \quad (4)$$

Previous work [15] has shown that this subband normalization is important when there is a mismatch between training and testing conditions, or in other words, when the detection system is employed in real (testing) conditions.

D. Dimensionality reduction and Feature Selection

We used a generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [11] to reduce dimensions in acoustic and modulation frequency subspaces separately. HOSVD enables the decomposition of tensor $\mathcal{D}$ to its $n-$mode singular vectors (or, principal components). Ordering of these $n-$mode singular values implies that the "energy" of tensor $\mathcal{D}$ is concentrated in the singular vectors with the lowest indices. Each singular matrix containing the $n-$mode singular vectors, can be truncated then by setting a predetermined threshold so as to retain only the desired number of principal axes in each mode.

Projection of modulation spectral features on the principal axes with the higher energy in each subspace results in a compact set of features with minimum redundancy. We further selected features which were more relevant to the given classification task using mutual information (MI). That is, relevance is defined as the mutual information $I(x_j; c)$ between feature x_j and class c. *Maximal relevance* (MaxRel) feature selection criterion simply selects the features most relevant to the target class c [12]. Through a sequential search, which does not require estimation of multivariate densities, the top m features in the descent ordering of $I(x_j; c)$ were selected.

Fig. 1 depicts the mutual information of the original normalized modulation spectral features for the classification of the dysphonic phonations of the vowel /AH/ in PdA in 3 scores of breathiness (B0, B1 and B2). Modulations localized lower than ~ 1600 Hz on the acoustic frequency axis seem to be more relevant; this is consistent with previous experimental results on pathological voice assessment where frequencies lower than 3000Hz led to an homogeneous discrimination between voices compared with higher frequencies [6].
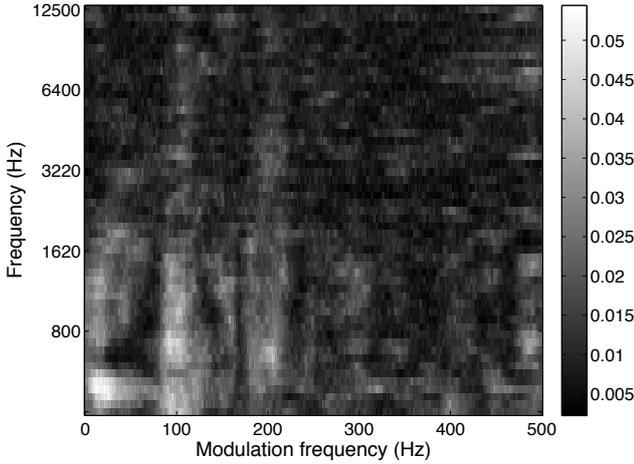
Figure 1: Mutual information of the original normalized modulation spectral features for the classification of breathiness in 3 grades (B0, B1 and B2) for the dysphonic phonations of sustained vowel /AH/ in PdA.

Mutual information to the class variable was estimated twice: for pathological subjects only, as well as when 24 normal subjects were also included. Two different feature sets were thus defined according to the sorted MI values. We conducted some preliminary experiments on breathiness classification using two naive bayes (NB) classifiers built on top of these features. We used leave-one-out cross-validation to select the top m features in every set; further the two classifiers were combined as we describe next [13].

E. Classifiers Combination

We performed a Bayes combination of the $L = 2$ classifiers, D_1 and D_2, based on the different feature sets, according to [13]. D_1 and D_2 label each data point $\mathbf{x}$ as belonging to one of $c = 3$ classes. In our case, the number of training data points is $N = 224$; from these, 50 (26 dysphonic plus 24 normal voices) are from ω_1 ($B = 0$), 119 from ω_2 ($B = 1$), and 55 from ω_3 (50 subjects with $B = 2$ and 5 subjects with $B = 3$). For each classifier D_i, a $c \times c$ confusion matrix CM^i is calculated by applying D_i to the training dataset. The (s, k) entry of this matrix, $cm^i_{s,k}$, denotes the number of elements of the dataset whose true class label was ω_k, and were assigned by D_i to class ω_s. By N_s we denote the total number of subjects from class ω_s. Taking $cm^i_{s_i,k}/N_k$ as an estimate of the probability $P(s_i|\omega_k)$, and N_k/N as an estimate of the prior probability $P(\omega_k)$ the support for class ω_k is equivalent to:

$$\mu_k(\mathbf{x}) \propto \frac{1}{N_k^{L-1}} \prod_{i=1}^{L} cm^i_{s_i,k}. \qquad (5)$$

Accordingly, subject $\mathbf{x}$ will be assigned to class ω_k if $\mu_k(\mathbf{x})$ has the highest value.

III. RESULTS

Modulation spectra were computed in a frame-by-frame basis using long windows in time (262 ms) which were overlapped by 50%. We used Mel scale filtering with 53 bands while the size of the Fourier transform for the time-domain transformation was set to 257 (up to π). Therefore, each modulation spectrum consisted of $I_1 = 53$ acoustic frequencies and $I_2 = 257$ modulation frequencies, resulting therefore in an 53×257 "image" per frame. The modulation spectra computed in each frame were mean subtracted and then they were stacked to produce a third order tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the number of frames in the training dataset. After applying the High Order SVD algorithm, we kept the principal axes (PCs) of features contributing more than 0.1% to the "energy" of $\mathcal{D}$; i.e., the first 43 PCs in the acoustic frequency and the first 29 PCs in the modulation frequency subspace. This resulted in a reduced space of $43 \times 29 = 1247$ features. Next, the features which were more correlated to the breathiness assessment were selected using the Maximal Relevance criterion (MaxRel). For details about the application of the MaxRel criterion on this task please refer to [10].

Two different feature sets were defined according to the sorted MI values. The first set included the most relevant features when MI estimation also involved voices from (24) normal subjects with zero breathiness. The second feature set was selected using the dataset of dysphonic only voices. We used leave-one-out cross validation to select the top m features for every NB classifier built on top of each feature set. NB classifier built on top of the $m = 100$ most relevant features of the first set was optimum for discriminating class $B = 0$. For classes $B = 1$ and $B = 2$, the optimum NB classifier was obtained by considering the top $m = 230$ most relevant features of the second set. By combining the NB classifiers based on these different feature sets [13], a global classification rate of 79.02% was achieved. Table 1 presents the confusion matrix from the automatic classification of the dysphonic voices into scores of breathiness. This classification has been compared with the original perceptual judgement in the PdA corpus. In Table 2 the performance per breathiness score in terms of correct classification rate is presented. We can observe that the worse performance corresponds to the B0 class which includes 26 dysphonic and 24 normal subjects. However, we note that 21 out the 24 normal speakers have been correctly classified in the B0 class (corresponding to a CCR of 87.50% for normal only voices). We conclude then that the breathiness of dysphonic only voices has been overestimated in the case of the $B = 0$ class.

Table 1: Confusion matrix between scores of breathiness given by the automatic classification system (S-B0, S-B1, S-B2) and the perceptual judgement of phonations (P-B0, P-B1, P-B2).

	P-B0	P-B1	P-B2
S-B0	**33**	4	0
S-B1	16	**99**	10
S-B2	1	16	**45**

Table 2: Performance per breathiness score in terms of correct classification rate (CCR %) of phonations in PdA [14].

Score 0	Score 1	Score 2	Total
66.00	83.19	81.82	79.02

IV. DISCUSSION

In this paper we have proposed a method for objective assessment of breathy voice quality, based on modulation spectra. We used a method for dimensionality reduction and feature selection on a database of sustained vowels. Using mutual information, we could locate the most relevant frequency bands at the "formant zone", i.e. lower than 3000 Hz. Based on different feature sets, two NB classifiers were tested and found to be optimal in the discrimination of different classes. By combining them, a global classification rate of 79.02% was achieved.

Future work will address additional GRASB parameters using a database of reading text. We will explore the discriminative ability of consonant classes as well in the objective assessment of different voice qualities. In addition, benchmarking against more standard approaches like those used for the automatic speaker recognition [6] will be performed.

ACKNOWLEDGEMENTS

REFERENCES

[1] M. Hirano, "Objective evaluation of the human voice: clinical aspects", *Folia Phoniatr.*, vol. 41, pp. 89-144, 1989.

[2] S.B. Davis, "Computer evaluation of laryngeal pathology based on inverse filtering of speech", *SCRL Monograph Number 13*, Speech Communications Research Laboratory, Santa Barbara, CA, 1976.

[3] R.A. Prosek, A.A. Montgomery, B.E. Walden and D.B. Hawkins, "An evaluation of residue features as correlates of voice disorders", *Journal of Communication Disorders*, vol. 20, pp. 105-117, 1987.

[4] V. Parsa and D.G. Jamieson, "Identification of pathological voices using glottal noise measures", *J. Speech, Lang., Hear. Res.*, vol. 43(2), pp. 469-485, 2000.

[5] N. Malyska, T.F. Quatieri and D. Sturim, "Automatic Dysphonia Recognition Using Biologically Inspired Amplitude-Modulation Features", *Proc. ICASSP'05*, pp. 873-876, 2005.

[6] G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio and A. Giovanni, "Frequency study for the characterization of the dysphonic voices", in *Proc. Interspeech'07*, Antwerp, 2007.

[7] H. Hermansky, "Should recognizers have ears?", *Speech Communication*, vol. 25, pp. 3-27, 1998.

[8] L. Atlas and S.A. Shamma, "Joint Acoustic and Modulation Frequency", *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668-675, 2003.

[9] D. Mehta and T.F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2005.

[10] M. Markaki and Y. Stylianou, "Using Modulation Spectra for Voice Pathology Detection and Classification", *Proc. IEEE EMBC'09*, 2009.

[11] L. De Lathauwer, B. De Moor and J. Vandewalle, "A multilinear singular value decomposition", *SIAM J. Matrix Anal. Appl*, vol. 21, pp. 1253-1278, 2000.

[12] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27(8), pp. 1226-1238, 2005.

[13] L.I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, John Wiley & sons, 2004.

[14] J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz and Ramírez- Calvo, C., "Acoustic Analysis of Voice using WPCVox: a comparative study with Multi Dimensional Voice Program", *European Archives of Otolaringology*, vol. 265 (4), pp. 465-476, 2008.

[15] M. Markaki and Y. Stylianou, Y., "Normalized Modulation Spectral Features for Cross-Database Voice Pathology Detection", in *Proc. Interspeech'09*, Brighton, 2009.

VOICE PATHOLOGY GRADING BY GAUSSIAN MIXTURE MODELS: STUDY CASES

P. Gómez-Vilda[1], R. Fernández-Baíllo[1], V. Rodellar-Biarge[1], J. I. Godino-Llorente[2]

[1] Grupo de Informática Aplicada al Procesado de Señal e Imagen
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
[2]Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid,
Ctra. de Valencia km 7, 28031, Madrid, Spain
e-mail: pedro@pino.datsi.fi.upm.es

Abstract: **The purpose of the present paper is to study how the statistical dispersion of distortion and biomechanical parameters may be used in producing an objective evaluation of voice quality. For such, the behaviour of the same GMM classifiers used in the detection of pathology will be exploited. The work will show specific cases derived from a database of normal and pathological voice, set into contrast against a Universal Background Model built from the population of normal subjects. Results will be contrasted against classical subjective scoring and a proposal for automatic voice quality evaluation in terms of the most relevant parameters will also be discussed.**

Keywords: **GRBAS, Voice Pathology Grading, Gaussian Mixture Models**

I. INTRODUCTION

The purpose of the present paper is to produce a score to grade voice pathology from unnoticed to extreme from the acoustic properties of the produced signal. The proposed score can be contrasted against subjective evaluations as GRBAS [2][7]. For such, the properties of the same GMM classifiers used in the detection of pathology will be exploited as described in Section II, devoted to discuss the materials and methods used in the study over a broad data collection specifically balanced by pathology and gender. Section III presents the results obtained for the proposed scoring. Section IV discusses the relevance of the study of a pre-post surgery case. Section V is devoted to present the conclusions derived from the study.

II. MATERIALS AND METHODS

The materials used in the present study consist in a set of 200 recordings of vowel /a/ extracted from the database in [8]. This data set is organized in 4 subsets equally balanced: MN (50 male, normophonic), MP (50 male, pathologic), FN (50 female, normophonic), FP (50 female, pathologic). The processing steps are as follows:

- Produce a set of observation parameters from each record including *jitter, shimmer, HNR,* a set of parameters derived from glottal source spectral profile, a set derived from biomechanical correlates and another set derived from the estimation of the glottal source time-domain cycle [1]. The parameters used are described in [4] and [5].

- Divide the set into control (MN+FN) and control (MP+FP) subsets independently for the male and female subsets.

- Evaluate the covariance matrices for the sets of normal male and female control sets. Use the matrix eigenvectors for principal component analysis transformation of the original parameter spaces to principal component parameters.

- Model the principal component parameter matrices by a GMM system following [6] using the control subsets. Two Universal Background Models must be created: one for male and one for female subjects.

- Produce the following estimates for each subject in the test sets:

$$\gamma_i = \log_{10}\left[\operatorname{int}\left(\frac{g_i}{\hat{g}} \right) \right] + 1 \qquad (1)$$

where:

$$g_i = \operatorname{int}\left[\left(\mathbf{y}_{ti} - \boldsymbol{\psi}_n \right)^T \mathbf{C}_n^{-1} \left(\mathbf{y}_{ti} - \boldsymbol{\psi}_n \right) \right] + 1 \qquad (2)$$

$\mathbf{y}_{ti}$, $\boldsymbol{\psi}_n$, and $\mathbf{C}_n$ being respectively the principal component parameter vectors, the centroids of the Universal Background Model GMM's and the Covariance Matrices of the observation sets.

- Evaluate the pathological condition of the subject by estimating the odds of its membership to the distribution of normophonics Γ_{nm} as:

$$p(\mathbf{y}_{tm,fi}/\Gamma_{nm,f}) =$$
$$\frac{1}{(2\pi)^{Q_m/2}\left| \mathbf{C}_{nm,f} \right|^{1/2}} e^{-1/2\left(\mathbf{y}_{tm,fi} - \boldsymbol{\psi}_{nm,f} \right)^T \mathbf{C}_{nm,f}^{-1} \left(\mathbf{y}_{tm,fi} - \boldsymbol{\psi}_{nm,f} \right)} \qquad (3)$$

where *m* and *f*, stand for the distributions of normophonic male and female. The subject score is usually given as a Log-Likelihood Ratio of the odds:

$$\Lambda(\mathbf{y}_{tmi}) = \log\{p(\mathbf{y}_{tmi}/\Gamma_{nm})\} - \log\{p(\mathbf{y}_{tmi}/\Gamma_{\bar{n}m})\} \quad (4)$$

This score is based on distance metrics as shown in Figure 1, and it may be used for detecting the pathological condition of the subject using classical ROC-DET procedures. Depending if $\Lambda(y_{tmi}/\Gamma_{nm})>\theta$ or $\Lambda(y_{tfi}/\Gamma_{nf})<\theta$ the voice of the subject under test is considered normal or pathological. As the results of the grading are defined by the cluster of normophonics expressed in the distribution-compensated distance in the exponent of the probability function in (3), a functional to estimate the pathology grade could be defined as:

$$\delta_{osi} = \begin{cases} 0; & |\gamma_{oi} - \gamma_{si}| \le 1 \\ 1; & |\gamma_{oi} - \gamma_{si}| > 1 \end{cases}; D_{os} = \sum_i \delta_{osi} \quad (5)$$

where γ_{oi} and γ_{si} refer respectively to the objective and subjective grading scores (GO and GS) assigned to subject i, and D_{os} is the cumulative score along the set considered (male or female). This means that a deviation of 1 is considered irrelevant between both scoring signals, whereas larger deviations are taken into account, to cope with the subjectivity implicit in GRBAS.
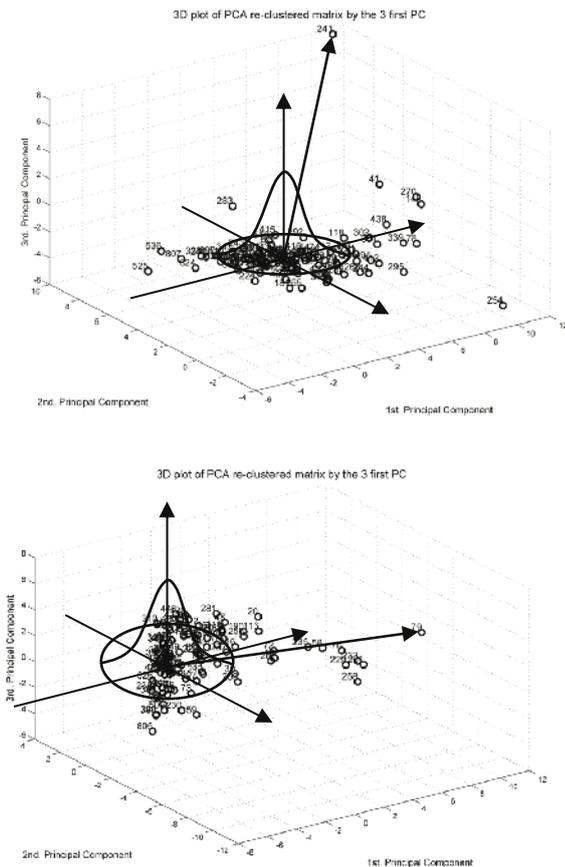
Figure 1. Top: Male cluster set with joint normophonic and pathologic distributions. The distance to the normal distribution may serve as a measure of the pathology grade. Bottom: Idem for the female cluster set.

The figures give an idealized idea on how each respective GMM quantifies the attribution probability of each subject with respect to its respective control set (Universal Background Model) plotted in terms of the three principal componens for each observation set. The normalized distance of each subject to the respective GMM centroid is used as a voice quality evaluation factor for each individual (g_i). This distance is pointed by an arrow for the two cases more far apart for each subset.

III. RESULTS

The application of the methodology described before to grade the 100 pathological cases formed by the collection of MP+FP taken from the Database MAPACI is given in Table 4 at the end of the paper. The confusion curve associated to the evaluation is given in Figure 2. The total results from applying the method to both normophonic and pathologic speakers is given in Table 1.
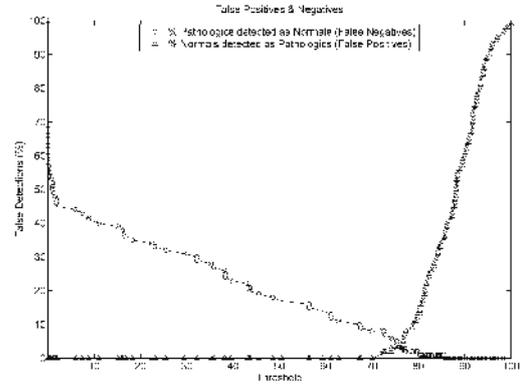
Figure 2. Confusion curve giving the percentage of False Negatives and False Positives in terms of the normalized threshold.

Table 1. Confusion matrix after grading the detection results for the male and female sets

	Normophonic	Pathologic
MN+MP	0	16
FN+FP	0	12

It may be seen that the degree of accuracy between the subjective expert evaluation given by the GRBAS index as contrasted against the one produced by the automatic system is consistent with a difference of 1 step or less for most of the male and female cases, within reasonable limits (normophonic ones are all well classified, whereas the differences in grading for pathologic ones is relatively small in most of the discrepancies found).

To better understand the underlying principles of the classification methodology proposed a contrastive evaluation of a pre-post-surgery case is given extending those presented in a previous study [3]. The recordings

of the pre-surgery case correspond to trace #232 whereas the post-surgery case (3 months later) corresponds to trace #732. A blind grading gives the results in Table 2.

Table 2. Comparison of the proposed grading to records obtained from the same person before and after surgery.

File	Log Likelihood Ratio	Grade
RegVoz-232.wav	-302,409187	2
RegVoz-732.wav	-87,6466868	0

It may be seen that the results of the automatic evaluation are well in consonance with the physiological evaluation of the patient, considering that the threshold for Grade 0 is around -94. The parameter profiles most affecting the re-classification of the patient (#732) as normophonic after surgery are given in Table 3 in contrast with the same ones before surgery (#232) in averages and standard deviations.

Table 3. Comparison between resolutive parameters before and after surgery for cases #232 and #732.

Parameter	Pre-surgery	Post-surgery
2. av. jitter (%)	3.80	0.35
5. av. shimmer (%)	2.80	1.13
8. av. HNR (%)	6.22	7.01
18. av. 1st min. (dB)	-23.22	-30.65
19. av. 2nd max. (dB)	-15.90	-27.17
33. av. 1st min slend. (%)	14.4	14.7
35. av. body mass (mg.)	13.4	14.3
37. av. body stiffness (g.sec^{-2})	21711	16433
38. av. body mass unb. (%)	5.93	0.44
41. av. cover mass (mg.)	12.7	11.4
43. av. cover stiffness (g.sec^{-2})	25828	13622
44. av. cover mass unb. (%)	7.48	1.91
2. std. jitter (%)	3.07	0.31
5. std. shimmer (%)	1.95	0.88
8. std. HNR (%)	1.49	0.21
18 std. 1st min. (dB)	1.72	2.11
19. std. 2nd max. (dB)	1.27	1.02
33. std. 1st min. slend. (%)	1.47	1.89
35. std. body mass (mg.)	0.66	0.007
37. std. body stiffness (g.sec^{-2})	2264	130
38. std. body mass unb. (%)	5.49	0.54
41. std. cover mass (mg.)	0.70	0.23
43. std. cover mass unb. (%)	5.57	1.06
44. std. cover stiffness (g.sec^{-2})	6264	706

IV. DISCUSSION

The study of the results for the pre-post surgery case is very illustrative of the way in which pathology classification is established by the set of parameters used. First of all, it confirms that classical distortion parameters as jitter and shimmer (2,5) have a role to play in determining the degree of pathology. The case with HNR (8) may not be the same, as the definition of this parameter is rather controversial, and its estimation is not as trivial as could be assumed. The biometrical power spectral density of the glottal source is also sensitive to pathology as revealed by parameters 18 and 19, which are well below the reference value after surgery, revealing a reduction in tenseness. This is also in agreement with the evaluation of body and cover stiffness (37 and 43). The body and cover mass unbalance (38 and 44) suffer important reductions when comparing before and after surgery values. Each estimation standard deviation is also given in Table 3 as a validation statistics for the estimation of each parameter. It may be seen that this control parameter is much better in most post-surgery cases indicating the smaller dispersion range of the affected parameter after treatment.

V. CONCLUSION

The most important conclusion derived from the study is that the grading methodology exposed is well in agreement with expert judgment within a reasonable extension, having in mind the subjectivity implied in expert judgment. The second conclusion is that the consistency in the estimates is availed by differential pre-post-surgery cases as the one presented. The relevance and reliability of the parameters used in the study is clearly availed by the results exposed in Table 3. The materials used in the study, as well as a free copy of the software employed in parameter extraction and contrast are publicly available from [8] on demand. The next challenge to be faced is that of pathology classification in terms of parameter "color", as well as it use in speaker identification and verification applications.

REFERENCES

[1] Alku, P., "Parameterisation Methods of the Glottal Flow Estimated by Inverse Filtering", Proc. of VOQUAL'03, Geneva, 2003, pp. 81-87.

[2] Baken, R. J. and Orlikoff, R., *Clinical measurement of speech and voice*, 2 ed., Singular Pub. Group, 2000.

[3] Gómez. P. et al., "Detecting Pathology in the Glottal Spectral Signature of Female Voice", Proc. of MAVEBA07, Firenze, Italy, December 2007, pp. 183-186.

[4] Gómez. P. et al., "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters", *J. Voice*, Vol. 21, No. 4, 2007, pp. 450-476.

[5] Gómez. P. et al., "Glottal Source Biometrical Signature for Voice Pathology Detection", *Speech Communication*, Vol. 51, 2009, pp. 759-781.

[6] Gómez. P. et al., "PCA of Perturbation Parameters in Voice Pathology Detection", Proc. of INTERSPEECH'05, 2005, pp. 645-648.

[7] M. Hirano et al., Acoustic analysis of pathological voice. Some results of clinical application, Acta Otolaryngologica, Vol. 105 (5-6), 1988, pp. 432-438.

[8] Project MAPACI: http://www.mapaci.com.

Table 4. Comparison of the results from the proposed methodology vs the GRBAS evaluation by 3 experts on 100 pathologic cases (50 MP + 50 FP). LR: Likelihood Ratio. GO: Grade, Objective. Evaluation: Subjective GRBAS from experts. GS: Grade, Subjective, from experts. δ: Difference between GO and GS. Cases showing a disagreement in δ are shown in Italics. The two extreme cases for male (#241) and female (#79) correspond to the ones pointed out in Figure 1.

Male #	LR	GO	Evaluation	GS	δ	Female #	LR	GO	Evaluation	GS	δ
21	*-151,755102*	*1*	*Extreme*	*4*	*1*	11	-118,219122	1	Moderate	0	0
41	*-743,746925*	*3*	*Mild*	*1*	*1*	12	-280,427775	2	Severe	3	0
62	-85,0760409	0	Mild	0	0	16	-184,148152	1	Mild	1	0
67	-154,041072	1	Mild	1	0	*17*	*-743,746925*	*3*	*Moderate*	*0*	*1*
74	*-455,225703*	*2*	*Extreme*	*4*	*1*	20	-743,746925	3	Severe	3	0
78	-743,746925	3	Extreme	4	0	*25*	*-743,746925*	*3*	*Moderate*	*0*	*1*
118	-522,371723	2	Severe	3	0	*30*	*-743,746925*	*3*	*Moderate*	*0*	*1*
148	-743,746925	3	Moderate	2	0	32	-96,0778838	1	Moderate	0	0
181	*-725,265419*	*2*	*Normal*	*0*	*1*	*33*	*-369,545985*	*2*	*Moderate*	*0*	*1*
221	-265,84043	2	Severe	3	0	42	-108,329266	1	Moderate	0	0
224	-114,202468	1	Mild	1	0	*43*	*-319,288792*	*2*	*Moderate*	*0*	*1*
231	-119,007622	1	Mild	1	0	48	-243,147773	2	Mild	1	0
240	*-110,213884*	*1*	*Severe*	*3*	*1*	*58*	*-743,746925*	*4*	*Moderate*	*0*	*1*
241	**-743,746925**	**4**	**Extreme**	**4**	**0**	59	-428,507786	2	Severe	3	0
243	*-88,8242377*	*1*	*Severe*	*3*	*1*	60	-149,768546	1	Moderate	0	0
244	*-97,3724014*	*1*	*Severe*	*3*	*1*	*70*	*-517,390726*	*2*	*Moderate*	*0*	*1*
246	*-110,532015*	*1*	*Severe*	*3*	*1*	72	-743,746925	3	Severe	3	0
254	-743,746925	4	Severe	3	0	73	-151,355549	1	Moderate	0	0
255	*-86,559584*	*0*	*Moderate*	*2*	*1*	76	-743,746925	4	Extreme	4	0
264	-266,615095	2	Severe	3	0	**79**	**-743,746925**	**7**	**Severe**	**3**	**1**
265	-291,522052	2	Severe	3	0	81	-176,560913	1	Mild	1	0
267	-115,504944	1	Mild	1	0	111	-743,746925	2	Moderate	2	0
268	-99,5074735	1	Mild	1	0	*113*	*-743,746925*	*4*	*Moderate*	*2*	*1*
270	-743,746925	3	Severe	3	0	116	-132,598113	1	Moderate	2	0
283	-743,746925	2	Severe	3	0	119	-743,746925	2	Mild	1	0
291	-89,0212634	1	Mild	1	0	120	-96,1015064	1	Mild	1	0
292	-126,40482	1	Patología	0	0	179	-105,673576	1	Mild	1	0
294	-110,072433	1		0	0	190	-743,746925	3	Moderate	2	0
295	*-242,594658*	*2*	*Extreme*	*4*	*1*	213	-731,570724	2	Mild	1	0
296	-87,441914	1	Normal	0	0	215	-743,746925	3	Moderate	2	0
299	-104,103068	1	Mild	1	0	222	-136,410801	1	Mild	1	0
301	*-176,928237*	*1*	*Severe*	*3*	*1*	225	-743,746925	3	Severe	3	0
302	*-552,203323*	*2*		*0*	*1*	226	-743,746925	2	Moderate	2	0
328	-322,517671	2	Mild	1	0	227	-191,757581	1	Moderate	2	0
330	*-115,542813*	*1*	*Severe*	*3*	*1*	*229*	*-743,746925*	*5*	*Severe*	*3*	*1*
336	-92,1508474	1	Mild	1	0	230	-167,433623	1	Mild	1	0
339	-423,839846	2	Mild	1	0	232	-743,746925	3	Moderate	2	0
369	-250,639609	2	Moderate	2	0	*233*	*-743,746925*	*4*	*Moderate*	*2*	*1*
398	-146,349367	1	Moderate	2	0	235	-743,746925	4	Severe	3	0
401	-88,1088171	1	Moderate	2	0	239	-95,0254838	1	Moderate	2	0
403	-99,4242089	1	Mild	1	0	249	-87,412922	1	Normal	0	0
412	-115,72244	1	Mild	1	0	253	-743,746925	2	Moderate	2	0
415	-451,791895	2	Mild	1	0	*256*	*-743,746925*	*3*	*Mild*	*1*	*1*
424	-91,9636509	1	Moderate	2	0	258	-743,746925	4	Severe	3	0
438	-699,415794	2	Moderate	2	0	259	-743,746925	2	Severe	3	0
447	-122,726323	1	Moderate	2	0	261	-743,746925	5	Extreme	4	0
468	*-82,2694054*	*0*	*Moderate*	*2*	*1*	269	-269,036479	2	Moderate	2	0
474	*-122,663689*	*1*	*Severe*	*3*	*1*	276	-743,746925	2	Severe	3	0
491	-239,492397	2	Moderate	2	0	279	-256,000675	2	Severe	3	0
516	*-137,704816*	*1*	*Extreme*	*4*	*1*	281	-586,006678	2	Severe	3	0

ESTIMATION OF HOSPITALIZATION PROGRESS FOR PATIENTS WITH STROKE WITH USING OF VOICE ANALYSIS

Damian Krzesimowski[1], Zygmunt Ciota [1]

[1]Department of Microelectronics and Computer Science, Technical University of Lodz, Łódź, Poland

Abstract: **Modern medicine calls for new diagnostic methods. Emphasis is placed on non-invasive methods. In addition, they should be characterized by high efficiency, which is a combination difficult to predict. In this area signal processing offers the greatest potential. It is used in many branches of medicine. This article presents one of the possible uses of signal processing, focused on the pathologies of voice, resulting from brain damage caused by vascular problems. Group of 41 patients neurology branch was recorded, with indications of ischaemic stroke, or hemorrhagic. The results clearly indicate the possibility of using the selected voice signal processing algorithms.**
Keywords : **Signal processing, voice pathology, stroke, vocal track filter**

I. INTRODUCTION

Exploiting of non-invasive method for diagnosis purpose is frequently more popular in medical environment presently. Also an increasing of diagnosis accuracy and speed of results obtainment and simplicity of evaluation has been observed. Operations of these kinds impose miscellaneous demand in relation to length and qualities of samples data. Probably, non-invasiveness is the most desirable feature for neurologists. It results from serious danger of injury during invasion operations on most important human organ – cerebrum. That's why authors have suggested considerable expansion of existing method of patient's condition estimate and monitor hospitalization progress, on base of voice parameters. Existing non-invasive methods of diagnosis are magnetic resonance and tomography. Nevertheless, most often they are execute only one time, during acceptance of patient on ward. Besides, analyzing of voice is fast and convenient.

II. METHODS

Presently, majority applicable method does not allow exact results getting properly, because of susceptibility of algorithm on errors in progress of recording emerged, and come of lack or scarce correction mechanisms, which could be adaptable to external conditions. Besides, it belongs to take into consideration approximation errors. The simplest methods of voice quality evaluation are based on experienced phoniatrist opinion. The subjective classification of the voice requires experience and intuition, and cannot be applied commonly, particularly in comparative investigations led through the various medical centers.

Objective acoustic analysis is perfect technique of estimate of voice quality definitely. Spectrographic, sonographic and the temporary analyses of the signal of the speech are useful in objective acoustic methods of the voice measure. Computer technology leaves across these requirements offering speed and convenience of computing.

Speech signal can be regarded as a dynamic object. Systems that track the volatility of stocks of such plants make use of the linear recursive estimation. These tools are for tests used to evaluate both the signal input and output characteristics, which are the result of actions processing functions. There are two methods here: the method of least squares and the minimal-mean-square method. Using the first one, the average signal of *N* samples can be estimated. It can be written as:

$$\bar{\hat{x}}_N = \frac{1}{N}\sum_{n=1}^{N}x(n) \qquad (1)$$

This model can be written in the form:

$$\bar{\hat{x}}_N = \frac{1}{N}x(N) + \frac{1}{N}\sum_{n=1}^{N-1}x(n) = \frac{1}{N}x(N) + \frac{N-1}{N}\left(\frac{1}{N-1}\sum_{n=1}^{N-1}x(n)\right) = \qquad (2)$$

$$\frac{1}{N}x(N) + \frac{N-1}{N}\bar{\hat{x}}_{N-1} = \bar{\hat{x}}_{N-1} + \frac{1}{N}\left[x(N) - \bar{\hat{x}}_{N-1}\right]$$

That is, to present a new estimate of the average of N points as the sum of the old estimate, calculated on the basis of *N-1* points, and its correction after taking into account the new n-th sample *x(n)*. The adjustment is calculated as the weight value of the error between the value of a new sample, and an old estimate of a mean value. This pattern is the current standard in adaptive recursive estimation of the parameters:
new estimate = prognoses + correction
correction = amplification · (measurement – prognoses of measurement)
where one of parameter is measured and another, related with it, is estimated.

The function of the quality of least-squares estimation is defined as:

$$J = (\underline{z} - H\,\underline{\hat{x}})^T (\underline{z} - H\,\underline{\hat{x}}) \qquad (3)$$

where: $\underline{z}$ is a measured vector, $\hat{\underline{x}}$ is estimate of vector generated by physical object, and H is matrix of measurement system. Estimate can be appoint as:

$$\hat{\underline{x}} = (H^T H)^{-1} H^T \underline{z} \qquad (4)$$

In the case of the minimal-mean-square method quality function becomes:

$$J = E\left[(\underline{x} - \hat{\underline{x}})^T (\underline{x} - \hat{\underline{x}})\right] \qquad (5)$$

where $E[.]$ is the expected value in statistical terms. In this case, the model boils down to two equations: the process model and measurement model. In the case of estimation of power spectral density function of signals to the disposal are several methods. Basis, used in the study, is periodogram, which is the square of the module of discrete Fourier transform of N samples for analysed signal $x(n)$, divided by N:

$$\hat{P}_x^{Per}(e^{j\Omega}) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n) e^{-j\Omega n}\right|^2, \qquad \Omega = 2\pi \frac{f}{f_{pr}} \quad (6)$$

Estimators of power spectral density function in the majority basis of the signal on the autocorrelation function estimation $R_{xx}(m)$ of the analysed signal $x(n)$. In the case of speech, where the signal is the sum of sinusoidal components, and broadband white noise, part of their vectors depends primarily on the components of harmonic signals, and the other only from the noise. Their vectors are orthogonal and unfasten two complementary areas: signal and noise. For at least one known vector lying in the noise, you can take advantage of the fact ortogonality to each of the vector space of signals and thus set a frequency. Autocorrelation function of the signal can be written in the form:

$$R_{yy} = \sum_{k=1}^{p}(\lambda_k + \sigma_s^2) v_k v_k^H + \sum_{k=p+1}^{M} \sigma_s^2 v_k v_k^H \qquad (7)$$

where: λ_k - values of matrix of autocorrelation function, v_k - vectors of this matrix, M - number of values of matrix R_{xx}.

In the classical Pisarenko method:

$$H(z) = \frac{1}{1 + \sum_{m=1}^{2p} a_m z^{-m}} \qquad (8)$$

The noise has zero average and is a non-correlated signal, so autocorrelation function is:

$$R_{yy}a = \sigma_s^2 a \qquad (9)$$

It follows that the searched vector of coefficients a is a vector of matrix R_{yy}, associated with its value σ_s^2.

In the investigation a method derived from Pisarenko - MUSIC (Multiple Signal Classification) was used. In this method, the frequency is estimated on the basis of arguments maximum of functions:

$$P_{MU}(e^{j\Omega}) = \frac{1}{\sum_{k=p+1}^{M}\left|e^H v_k\right|^2}, \qquad e = \left[1, e^{j\Omega}, e^{j2\Omega}, ..., e^{j(M-1)\Omega}\right]^T \quad (10)$$

In addition, the calculations of spectroscopic estimation were carried out by Welch method. The analysis of variable frequency signals using non-time-frequency representation of signals was used. Used Fourier transform STFT (Short-Time Fourier Transform) and the Wigner-Ville transform. The first can be interpreted as non-discredited in time and frequency Gabor transform. Used a description:

$$STFT_x^T(t, f) = \int_{-\infty}^{+\infty} x(\tau)\gamma*(\tau - t) e^{-j2\pi f\tau} d\tau \quad (11)$$

$$STFT_x^F(t, f) = e^{-j2\pi ft} \int_{-\infty}^{+\infty} X(v)\Gamma*(v - f) e^{-j2\pi vt} dv \quad (12)$$

where $\gamma(t)$ is a function of the time window of observation, $\Gamma(f)$ is the Fourier spectrum that acts as a window.

In the case of Wigner-Ville transform should be noted that it perfectly reflects in the time-frequency linear change of frequency. By definition:

$$S_x^{W(V)}(t, f) = \int_{-\infty}^{+\infty} x\left(t + \frac{\tau}{2}\right)x*\left(t - \frac{\tau}{2}\right)e^{-j2\pi f\tau} d\tau \quad (13)$$

$$S_X^{W(V)}(t, f) = \int_{-\infty}^{+\infty} X\left(f + \frac{v}{2}\right)X*\left(f - \frac{v}{2}\right)e^{-j2\pi vt} dt \quad (14)$$

This representation is characterized by the highest concentration of energy in the time-frequency space that means that has the best resolution.

In addition, in our investigation a vocal track filter transform is determined, witch maximum on the characteristics of time-frequency results from a formant diagram. The purpose of the designation of the fundamental frequency is to detect the first signal by the maximum value on the axis and cut the higher frequency harmonics. Then a maximum for each step are determined again in the given range. Used here with autocorrelation functions, which turned out to be more accurate than cepstral and timing method to determining the basic frequency. Also spectrogram of the input signal is appointed for power of the harmonics tracking.

III. RESULTS

Form of obtained speech sound is determined by specificity of executive voice apparatuses. Match of vocal strings generates sound and it subjects modulation during proceeding by vocal track. It depends on programming action of the central nervous system and the condition of the broadcast of stimuli in cortial-subcortial area, in the trunk of the cerebrum, nerves and nervous-muscular

synapses. That's why authors are of an opinion that a large capability exists to diagnose and estimate of injury of cerebrum on base of analysis of patient voice. Additionally, it is possible to get information of progress of treatment in a fast and simple manner. For our study several patients with most commonplace injury of central nervous system have been included, namely strokes hemorrhagic and ischaemic. These patients have problems with speaking out, which is defined as aphasia. However, even at patients who are good speaking out, possibility of changes detection in course of chosen characteristics has been checked. This paper contains results of present evaluations, in cooperation with neurologists of one of hospital in Lodz.

The investigation confirmed the usefulness of the signal analysis of speech in monitoring abnormalities the sound. The fundamental frequency was disturbed in almost all patients. This applies primarily to patients with ischaemic stroke of any cerebrum hemisphere. Patients with hemorrhagic stroke, and haematoma in the right hemisphere had little change in the fundamental frequency.
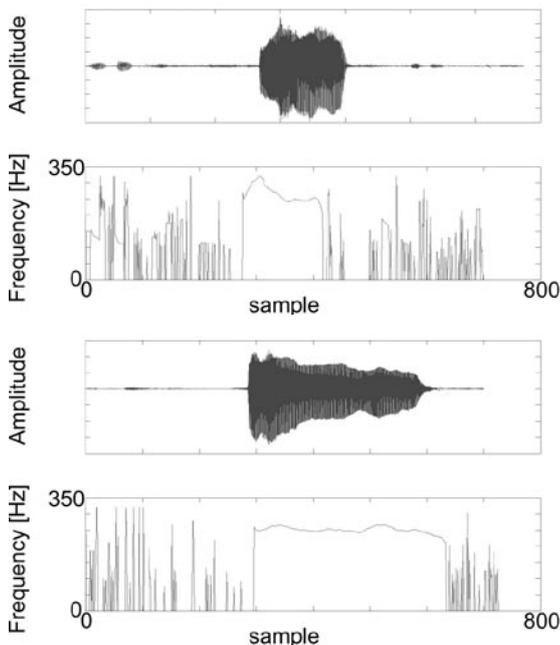
Fig. 1. Fundamental frequency for patient with ischaemic stroke in left hemisphere in the first (top) and last (bottom) day of hospitalization.

In some patients noted a significant increase in the fluctuations periodogram, with the progress of hospitalization. Probably this is related to the clean tone and the power generated by the larynx and depends on the vocal track in lesser extent. Interestingly, these changes were recorded for the case ischaemic stroke in left hemisphere and hematoma in the left and right hemisphere.
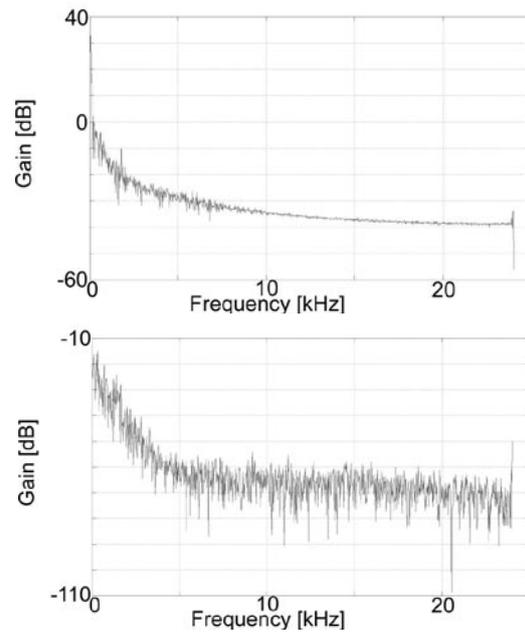
Fig. 2. Periodogram for patient with ischaemic stroke in the left hemisphere in the first (top) and last (bottom) day of hospitalization.

In both cases, the visibility was noted significant differences in the characteristics of vocal track filter. In the first days of the charts are very flat, and differences appear along with the progress of treatment. Each of the peaks corresponds to the formant graph, which means that they are clearer. Here was shown also another case of ischaemic stroke in the left cerebral hemisphere in right-handed person.
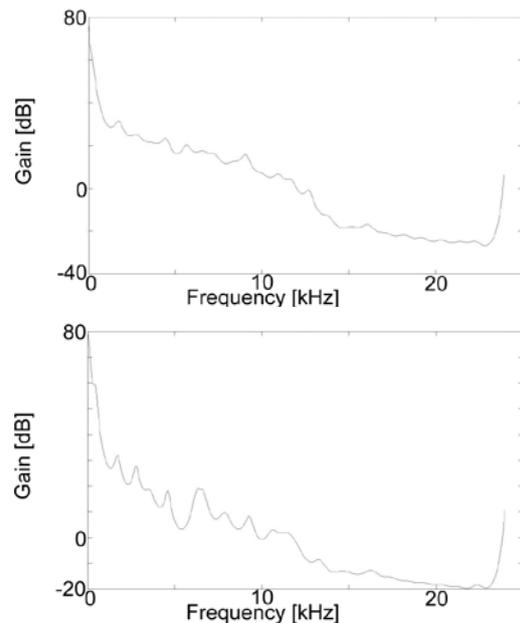
Fig. 3. Vocal track filter for patient with traumatic haematoma in the left and right hemisphere in the first (top) and last (bottom) day of hospitalization.

For patients with ischaemic stroke in the left hemisphere and aphasia one noted that with the progress of treatment the second formant is distinguished. This is independent of gender.

In addition, the Welch method of spectral estimation shows similar changes as the vocal track filter. In addition to this there is a more detailed, and thus signal changes can be analyzed in a more precise manner. Approximation should carry out, in order to remove the disturbance of signal and noise.
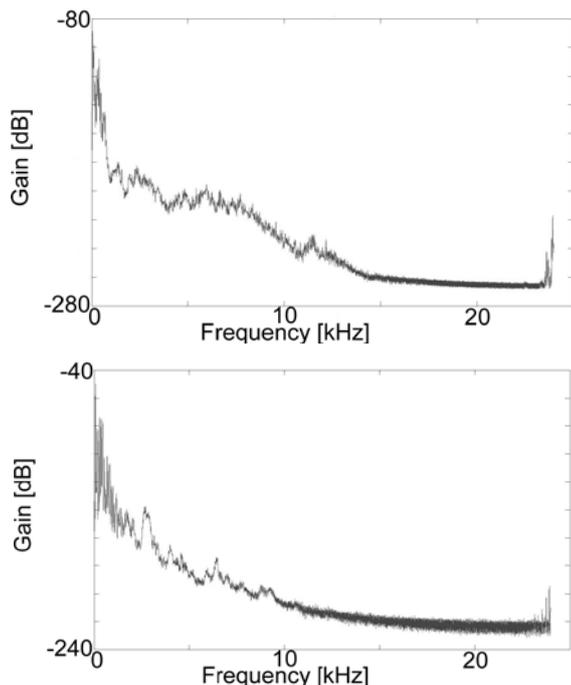
Fig. 4. Graph of Welch estimation method for patient with ischaemic stroke in the left hemisphere in the first (top) and last (bottom) day of hospitalization.

In the case of hemorrhagic stroke in the right hemisphere to increase the altitude of the peaks on the characteristics of the vocal track filter was observed. The changes are compatible with the emergence of a first and second formant. The power of the other formants remained unchanged.

In right-handed person with ischaemic stroke in right hemisphere we also noted changes in the characteristics of the vocal track filter and fundamental frequency progress. It reflects on the first formant in terms of increased clarity of this formant. And in the case of the other formants one noted a slight decrease in their gain.

At the same time, the results of other types of analysis does not give unambiguous answers to their usefulness in the diagnosis of disorders of speech. This may be due to too small amount of material inaccuracies in the test or imprecise calculation. In the future, it will be developed and tested new types of analysis on the held and the newly gained samples. Statistical functions will be designed also to computer evaluation of results.

IV. CONCLUSION

In accordance with mentioned limitations, using technique relying on more advanced signal processing for obtainment more precise results, we expect to open new capabilities in non-invasion research of patients with cerebrum injury with blood vessel problems. Authors expect, that information obtaining by using this technique will be helpful at diagnosis and for estimation of hospitalization progress for persons with mentioned above ailment. It can be useful supplement for conclusions of tomography and magnetic resonance images, as well as for monitoring condition of patient health in assigned period of time.

V. REFERENCES

[1] P.Gómez, F.Díaz, C.Lázaro, K.Murphy, R.Martínez, V.Rodellar & A.Álvarez, "Principal Component Analysis of Spectral Perturbation Parameters for Voice Pathology Detection", CBMS, Dublin, 2005, pp. 41-46
[2] J.C.Stemple, L.E.Glaze & B.Klaben Gerdemann, Clinical Voice Pathology Theory and Management, 3rd Edition, New Jersey, Thomson Delmar Learning, 2000
[3] J.Schoentgen,"Towards a Classification of Phonatory Features of Disordered Voices", MAVEBA, Florence, 2005, pp. 19-22

ON THE MUTUAL INFORMATION OF GLOTTAL SOURCE ESTIMATION TECHNIQUES FOR THE AUTOMATIC DETECTION OF SPEECH PATHOLOGIES

T. Dubuisson, T. Drugman, T. Dutoit

TCTS Lab, Faculté Polytechnique de Mons, 31 Boulevard Dolez, 7000 Mons, Belgium

Abstract: **This paper focuses on the automatic detection of speech pathologies by exploiting the estimation of the glottal source. Three methods of estimation are compared and time and spectral features are extracted. The relevancy of these features is assessed by means of information theory-based measures. This allows an intuitive interpretation in terms of discrimination power and redundancy between the features. It is discussed which features are informative or complementary for detecting voice pathologies and the glottal source estimation methods are compared.**
Keywords: **Voice Pathology, Glottal Source, Mutual Information**

I. INTRODUCTION

Perceptive evaluation performed by clinicians suffers from the dependency on the experience of the listener and the inter- and intra-judges variability. There is thus a need to develop objective tools. For this, a part of research in speech processing has focused on the detection of speech pathologies from audio recordings. Indeed it could be useful to detect disorders when perturbations are still weak, to prevent the degradation of the pathology, or to measure the voice quality before and after surgery [1]. As video recordings of the vocal folds show that their behavior is linked to the perception of different kinds of voice qualities, including pathologies, isolating and parametrizing the glottal excitation should lead to a better discrimination between normal and pathological voices. Such parametrizations of the glottal pulse have already been proposed both in time and frequency domains ([2], [3]).

This paper pursues the work presented in [4], in which it was shown that features respectively extracted from the vocal tract and glottal contributions (estimated by the IAIF algorithm [5]) are synergic and can lead together to an efficient discrimination of voice disorders. The present study addresses the comparison between IAIF and two other methods for the same problem. As in [4], the performance of classification is assessed by computing information theory-based measures in order to provide an intuitive interpretation in terms of discrimination power, redundancy and synergy between the features.

The paper is structured as follows. In Section 2, the different methods of glottal source estimation are presented. Section 3 defines the features extracted from the glottal source. Section 4 reviews the mutual information-based measures that are used in this work and highlights their interpretation for a classification problem. Experiments and results are detailed in Section 5. It is discussed which features are informative for the detection of voice disorders and which ones are complementary. Finally Section 6 concludes.

II. GLOTTAL SOURCE ESTIMATION

Three methods of glottal source estimation are considered here: the Complex Cepstrum Decomposition (CCD) [6], the Iterative Adaptative Inverse Filtering (IAIF) [5] and the Closed Phase Inverse Filtering (CPIF) technique [7]. The application of these three methods on a fragment of a normophonic sustained vowel /a/ is presented in Fig. 1.

A. Complex Cepstrum Decomposition

It has been recently shown that complex cepstrum can be efficiently used for glottal source estimation [6].This method aims at separating the minimum and maximum-phase components of the speech signal. Indeed it has been shown previously [8] that speech is a mixed-phase signal where the maximum-phase (i.e. anti-causal) contribution corresponds to the glottal open phase, while the minimum-phase component is related to the vocal tract transmittance (assuming an abrupt glottal return phase). Isolating the maximum-phase component of speech then provides a reliable estimation of the glottal source, which can be achieved by the complex cepstrum.

B. Iterative Adaptative Inverse Filtering

The IAIF technique [5] (publicly available in the Aparat Toolkit [9]) iteratively estimates the vocal tract contribution from the speech signal using a Discrete All Pole model whose order is different for the successive iterations. The glottal source is estimated by filtering the speech signal by the inverse of the filter modeling the contribution of the vocal tract.

C. Closed Phase Inverse Filtering

The CPIF technique exploits the fact that the glottal cycle consists of two phases, during which the vocal folds are respectively open and closed [7]. The key idea of this technique is to estimate the vocal tract transmittance during the closed phase, when it is assumed to be almost free of any excitation. Linear prediction is thus applied on

the speech signal during the closed phase and the glottal source is estimated by inverse filtering of the speech signal.
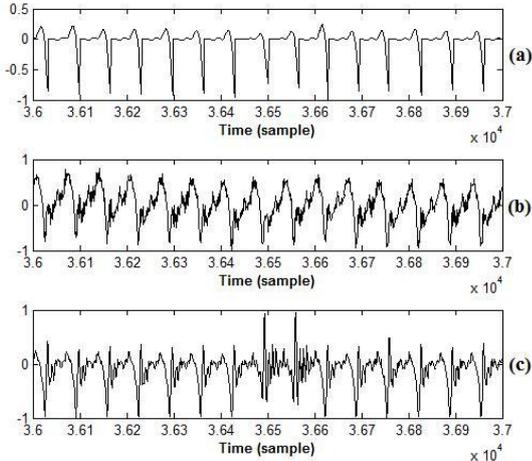
Fig. 1. Comparison of the three glottal source estimations (a: CCD; b: IAIF; c: CPIF) for a normophonic sustained vowel.

III. Feature extraction

Features are here extracted from glottal pitch-synchronous frames in voiced parts of speech. These frames are two-pitch period long, centered on the glottal closure instant (GCI) and weighted by a Blackman window. Pitch and voicing decision are computed using the Snack library [10] while GCIs are located according to the method described in [11].

A. Spectral Features

The amplitude spectrum of a voiced glottal source generally presents a low-frequency response called *glottal formant* produced during the open phase [3]. This formant is here characterized both by its frequency F_g and bandwidth B_w.

The spectral content of the glottal source spectrum is summarized by computing characteristics describing the repartition of its energy. The global repartition of spectral energy is captured in the spectral center of gravity CoG. A finer distribution of energy is quantified by considering an approach similar to [12] but using the perceptive mel scale. For this, the power spectral density is weighted by a mel-filterbank consisting on 24 triangular filters equally spaced along the whole mel scale. Three perceptive spectral balances are then computed:

$$Bal_1 = \frac{\sum_{i=1}^{4} PE(i)}{\sum_{i=1}^{24} PE(i)} \quad Bal_2 = \frac{\sum_{i=5}^{12} PE(i)}{\sum_{i=1}^{24} PE(i)} \quad Bal_3 = \frac{\sum_{i=13}^{24} PE(i)}{\sum_{i=1}^{24} PE(i)} \quad (1)$$

where PE(i) denotes the cumulated weighted power spectral density for the ith filter.

B. Time Features

In many studies, the glottal flow and its derivative (called here *glottal source*) have been used to characterize voice quality [2]. Two parameters are computed here for characterizing the amplitude and duration of the open phase of the glottal cycle. The Normalized Amplitude Quotient (*NAQ*) [2] is defined as the ratio of glottal flow amplitude and the minimum peak of glottal flow derivative, normalized by the length of the glottal cycle. The Quasi-Open Quotient (*QOQ*) [2] is defined as the duration during which the glottal flow is 50% above the minimum flow. Unlike the other parameters, *NAQ* and *QOQ* are computed for one glottal cycle instead of a two-period long frame centered on the GCI. Furthermore, we observed in [4] that the discontinuity at the GCI is generally more significant in case of normal voice than in case of pathological voice. The minimum value at the GCI (*minGCI*) of energy-normalized glottal source frames is thus also considered here.

IV. Information theory-based measures

The problem of automatic classification consists in finding a set of features X_i such that the uncertainty on the determination of classes C is reduced as much as possible [13]. For this, Information Theory [14] allows to assess the relevance of features for a given classification problem, by making use of the following measures (where *p(.)* denotes a probability density function):

- The entropy of classes C is expressed as:

$$H(C) = - \sum_c p(c) \log_2 p(c) \quad (2)$$

and can be interpreted as the amount of uncertainty on their determination.

- The mutual information between one feature X_i and classes C:

$$I(X_i; C) = \sum_{x_i} \sum_c p(x_i, c) \log_2 \frac{p(x_i, c)}{p(x_i) p(c)} \quad (3)$$

can be viewed as the information the feature X_i conveys about the considered classification problem, i.e. the discrimination power of one individual feature.

- The joint mutual information between two features X_i, X_j, and classes C can be expressed as:

$$I(X_i, X_j; C) = I(X_i; C) + I(X_j; C) - I(X_i; X_j; C) \quad (4)$$

and corresponds to the information that features X_i and X_j, when *used together*, bring to the classification problem. The last term can be written as:

$$I(X_i; X_j; C) =$$
$$\sum_{x_i} \sum_{x_j} \sum_c p(x_i, x_j, c) \log_2 \frac{p(x_i, x_j) p(x_i, c) p(x_j, c)}{p(x_i, x_j, c) p(x_i) p(x_j) p(c)} \qquad (5)$$

An important remark has to be underlined about the sign of this term. It can be noticed from expression of $I(X_i, X_j; C)$ that a positive value of $I(X_i; X_j; C)$ implies some **redundancy** between the features, while a negative value means that features present some **synergy** (depending on whether their association brings respectively less or more than the addition of their own individual information).

V. EXPERIMENTS

A. Database

A popular database in the domain of speech pathologies is the MEEI Disordered Voice Database [15]. This database contains sustained vowels and reading text samples, from 53 subjects with normal voice and 657 subjects with a large panel of pathologies. Here, all the sustained vowels of the MEEI Database resampled at 16 kHz are considered.

B. Mutual Information Computation

To evaluate the significance of the proposed features, the following measures are computed:

- the relative intrinsic information of one individual feature $I(X_i; C) / H(C)$, i.e. the percentage of relevant information conveyed by the feature X_i,
- the relative redundancy between two features $I(X_i; X_j; C) / H(C)$, i.e. the percentage of their common relevant information,
- the relative joint information of two features $I(X_i, X_j; C) / H(C)$, i.e. the percentage of relevant information they convey together.

For this, equations presented in Section IV are calculated. Probability density functions are estimated by a histogram approach. The number of bins is set to 50 for each feature dimension, which results in a trade-off between an adequately high number for an accurate estimation, while keeping sufficient samples per bin. Since features are extracted at the frame level, a total of 32000 and 107000 examples is available respectively for normal and pathological voices. Mutual information-based measures can then be considered as being accurately estimated. Class labels correspond to the presence or not of a voice disorder.

C. Results

The values of the measures detailed in the previous section for the three methods are presented in Fig. 2. For each table, the diagonal indicates the percentage of relevant information conveyed by each feature. It can be observed that QOQ is the most informative feature for CPIF and CCD methods (respectively 31.5% and 32.8%) while F_g is slightly more informative (25.9%) than QOQ in the case of IAIF method. The top-right part contains the values of relative joint information of two features. When used together, the combination of QOQ and F_g brings, for the three methods, the most important information about the classification problem, with a maximum value for the CPIF method (63.8%). The bottom-left part shows the values of relative redundancy between two features. For CCD and CPIF methods, F_g is synergic ($I(X_i; X_j; C) < 0$) with all the features, including QOQ, while this latter is less synergic and in some cases redundant with the other features.

The results show that applying the CCD technique gives generally better results than other methods in terms of intrinsic discrimination power. The synergy for the CDD technique is also the highest for most of features pairs. Moreover, using the combination of QOQ and F_g computed by CCD is the most interesting for the distinction between normal and pathological voices. Indeed, their mutual information is high, each feature brings its own information in the combination and is not redundant with the information conveyed by the other.

For the three methods, the highest amount of information conveyed by the combination of two features is about 60%. This means that there is a need of other information to distinguish normal and pathological voices. For this, it was shown in [4] that combining only one vocal tract-based and one glottal feature allows explaining 81% of the difference between normal and pathological voices.

VI. CONCLUSION

This paper focused on the problem of automatic detection of voice pathologies from the speech signal. The goal was to compare the classification performance of the features extracted from the glottal source estimated by three different methods (CCD, IAIF, and CPIF). These features were assessed through mutual information-based measures. It turned out that CCD technique generally provides features that convey higher intrinsic, mutual information and synergy. It was also shown that the couple of features (QOQ, F_g) has the highest mutual information (63.8%) and is also characterized by a high synergy, meaning that their association brings more than the addition of their intrinsic information.

CCD	Fg	Bw	CoG	Bal1	Bal2	Bal3	MinGCI	Naq	Qoq
Fg	12,0	42,0	47,8	35,7	41,8	43,6	42,3	52,0	60,0
Bw	-15,1	14,9	42,8	33,4	36,6	39,2	34,6	44,9	49,8
CoG	-17,0	-9,0	18,9	32,9	35,9	34,6	36,7	48,4	54,1
Bal1	-10,6	-5,3	-0,9	13,2	29,5	35,0	33,2	43,5	48,0
Bal2	-7,2	0,9	5,6	6,3	22,6	36,2	37,4	43,1	48,8
Bal3	-10,4	-3,0	5,4	-0,7	7,6	21,2	36,1	43,9	48,7
MinGCI	-11,6	-0,9	0,9	-1,3	3,9	3,8	18,7	41,1	40,9
Naq	-15,6	-5,6	-5,1	-6,0	3,9	1,7	2,1	24,4	52,8
Qoq	-15,3	-2,0	-2,5	-2,0	6,5	5,3	10,6	4,4	32,8

IAIF	Fg	Bw	CoG	Bal1	Bal2	Bal3	MinGCI	Naq	Qoq
Fg	25,9	52,6	42,1	43,3	38,9	40,5	37,2	49,6	60,3
Bw	-10,7	16,0	36,4	36,6	30,7	33,3	28,5	44,2	52,9
CoG	7,8	3,6	24,0	30,0	26,7	25,1	32,3	41,6	48,9
Bal1	1,2	-2,1	12,5	18,5	29,1	27,7	26,6	40,1	48,0
Bal2	0,2	-1,5	10,5	2,6	13,2	24,4	22,0	34,4	39,8
Bal3	5,5	2,7	18,9	10,9	8,8	20,1	28,0	39,1	46,2
MinGCI	-3,7	-4,9	-0,8	-0,5	-1,3	-0,3	7,5	24,3	32,9
Naq	-2,1	-6,6	3,9	0,0	0,3	2,5	4,8	21,6	46,1
Qoq	-9,7	-12,2	-0,2	-4,8	-2,0	-1,4	-0,6	0,1	24,7

CPIF	Fg	Bw	CoG	Bal1	Bal2	Bal3	MinGCI	Naq	Qoq
Fg	18,9	49,9	35,2	34,7	33,7	31,5	27,6	45,5	63,8
Bw	-16,6	14,4	28,5	27,5	26,5	25,7	21,9	38,6	58,9
CoG	-8,4	-6,2	7,9	17,3	17,5	21,6	17,2	25,2	46,4
Bal1	-13,0	-10,3	-6,7	2,7	14,2	12,9	11,8	27,1	48,7
Bal2	-11,9	-9,2	-6,9	-8,7	2,8	12,1	8,2	25,0	45,2
Bal3	-9,3	-8,0	-10,5	-7,0	-6,1	3,2	12,7	21,9	44,2
MinGCI	-4,9	-3,6	-5,5	-5,2	-1,6	-5,6	3,8	20,7	36,7
Naq	-13,4	-11,0	-4,1	-11,2	-9,0	-5,5	-3,7	13,2	46,0
Qoq	-13,4	-13,0	-7,1	-14,5	-10,9	-9,5	-1,4	-1,4	31,5

Fig.2. Mutual information-based measures for the proposed features. *On the diagonal*: the relative intrinsic information. *In the bottom-left part*: the relative redundancy between two considered features. *In the top-right*: the relative joint information of the two considered features

REFERENCES

[1] P. Gomez-Vila, R. Fernandez, V. Rodellar, V. Nieto, A. Alvarez, R. Mazaira, and J. L. Godino, "Glottal source biometrical signature for voice pathology detection," *Speech Comm*., vol. 51, pp. 759-781, 2008.
[2] M. Airas, and P. Alku, "Comparison of multiple voice source parameters in different phonation types," *Proc. of Interspeech 07*, pp. 1410-1413, 2007.
[3] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "A method for glottal formant frequency estimation," *Proc. of ICSLP 2004*, 2004.
[4] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," *Proc. of Interspeech 09*, 2009.
[5] P. Alku, "Glottal wave analysis with pitch-synchronous iterative adaptative inverse filtering," *Speech Comm*., vol. 11, no 2-3, pp. 109-118, 1992.
[6] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation," *Proc. of Interspeech 09*, 2009.
[7] D. Veeneman, and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. on Signal Proc*., vol. 33, pp. 369-377, 1985.

[8] B. Bozkurt, and T. Dutoit, "Mixed-phase signal modeling and formant estimation using differential phase spectrums," *Proc. of VOQUAL' 03*, pp. 21-24, 2003.
[9] TKK Aparat website, http://aparat.sourceforge.net.
[10] K. Sjölander, and J. Beskow, "Wavesufer − an open source speech tool," *Proc. of ICSLP 2000*, vol. 4, pp. 464-467, 2000.
[11] T. Drugman, and T. Dutoit, "Glottal Closure and Opening Instant Detection from Speech Signals,"*Proc. of Interspeech 09*, 2009.
[12] J.B. Alonso, J. de Leon, I. Alonso, and A.M. Ferrer, "Automatic detection of pathologies in the voice by HOS based paramerers," *EURASIP Journal on Applied Signal Processing*, 2001:4, pp. 275-284, 2001.
[13] L. Huan, and H. Motoda, *Feature selection for knowledge discovery and data mining*, The Springer International Series in Engineering and Computer Science, vol. 454, 1998.
[14] T. Cover, and J. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, New York, 1991.
[15] Kay Elemetrics Corp. "Disordered Voice Database Model (version 1.03)", Massachusetts Eye and Ear Infirmary Voice and Speech Lab, 1994.

ACOUSTIC ANALYSIS OF VOWEL SEGMENTS FOR CLINICAL PURPOSES: PRELIMINARY OBSERVATIONS

O. Amir, S. Ziv, N. Amir

Department of Communication Disorders, Tel-Aviv University, Israel

Abstract: Sustained phonations of the vowels /a/ and /i/ were recorded from 89 patients with voice disorders, who were divided into five pathological subgroups. In addition, recordings were made from a control group of 23 normophonic participants. All recordings were segmented into onset, steady state, and offset, and analyzed acoustically. Results revealed the following findings: 1) only minor differences were found between the acoustical analyses based on the steady state versus the entire vowel; 2) the tested acoustic measures could discriminate between test and control groups, but not among the different pathological groups; 3) the contribution of the data gathered from the onset and/or offset of the vowel did not contribute significant information beyond that which was provided by the steady state.

Keywords: voice pathology, sustained phonation, acoustic analysis, vowel segment

I. INTRODUCTION

Acoustic analysis of voice for clinical purposes has become a common practice. While it is agreed that the selection of the stimuli (test utterance) greatly affects the results of the acoustic analysis [1], various acoustic studies have examined different test utterances. Several studies have performed acoustic analyses of continuous speech under the assumption that it better represents "natural" voice production and vocal dynamics. Therefore, continuous speech can be considered more likely to correlate better with perceptual evaluation of voice quality [2, 3, 4].

Nonetheless, the majority of studies in the field of acoustic analysis of voice have examined isolated sustained vowels, under the assumption that they elicit a stationary process in the vocal folds vibration [1]. Because acoustic analysis of voice is, in many cases, based on measurements of fundamental frequency, the robustness of the analysis is dependent on the accuracy of the automatic extraction procedure of fundamental frequency [5]. Thus, isolated vowels are favored, since the extraction of fundamental frequency is more reliable for vowels than for speech. Whether the selection of sustained vowels as the test utterance is justified clinically, or it is merely simpler and more reliable for analysis, the fact is that sustained vowels are the most common selection for stimuli in voice analysis studies.

Acoustic analyses of vowels are usually based on a segment extracted from the so-called "steady state" of the vowel, thus the onset and offset of the vowel are discarded [6]. The rationale for discarding the onset and offset of the vowel lays in the assumption that these segments do not provide crucial diagnostic information, or at least such information that could be identified using acoustic analysis. Furthermore, these segments are (by definition) less stationary than the steady state, and therefore present a challenge for automatic extraction of fundamental frequency.

In contrast, a few studies have performed acoustic analysis on the complete vowel, without discarding the onset and offset from the analysis [7, 8]. These studies, however, did not provide the rationale for their selection of stimuli. Analyses of complete vowels are simpler, as they do not require segmentation of the vowels and identification of the boundaries of the steady state. In addition, perceptual evaluation of vowels in clinical setting requires the listener to evaluate all parts of the vowel, and not only the steady state. It was also suggested that valuable information on vocal folds' function may be revealed in the non-stationary segments of the vowel [5]. Therefore, including the onset and offset of the vowel in the analysis is more similar to the natural perceptual task of evaluating voice quality of vowels in the clinical setting.

It is logical to expect that values of perturbation parameters would be higher for complete vowels compared to analyses based on steady states. It is not clear, however, if acoustic analysis of complete vowels has clinical merit, and whether it could differentiate between pathological and healthy voices and between specific pathological groups. The purpose of the present study was, therefore, to learn whether a basic acoustic analysis of different segments of sustained vowels (i.e., onset, steady state, offset and the complete vowel) would yield observable differences of clinical value between voices of patients with different voice pathologies and controls.

II. METHODS

A. Subjects

Two groups of subjects were included in the study. The first group consisted of 89 patients from the voice clinic in Sheba Medical center, who were diagnosed with different laryngeal pathologies (37 Men and 52 Women, mean age 43.3). The second group (control) consisted of 23 singing students with no laryngeal pathology. All participants underwent a stroboscopic examination. The pathological group was divided into five subgroups:

1. **Benign mass lesion:** nodules, polyps or cysts;
2. **Neurogenic disorders:** e.g., Spasmodic disphonia, uni-lateral vocal fold paralysis, or paresis;

3. **Mucosal irregularity**: e.g., scarring of the epithelial tissue, ectasia of vocal folds' capillaries;
4. **Inflammations:** e.g., Edema or chronic laryngitis;
5. **Functional:** disordered voice with no observed laryngeal pathology.

Eight patients were discarded from the initial pool of 97 potential participants, due to noisy or interrupted recordings, or due to lack of periodicity in the voice signal. Table 1 summarizes the information of the subjects in the different study groups.

Table 1 – Division of subjects into groups

Group	N	Gender	Age (years, *SD*)
1. Mass lesion	20	15 Women	35.1 *(12.3)*
		5 Men	36 *(15.8)*
2. Neurogenic	18	6 Women	55.5 *(12.2)*
		12 Men	59.2 *(10.4)*
3. Mucosal irr.	11	6 Women	42.8 *(8.0)*
		5 Men	44.8 *(16.9)*
4. Inflammation	20	10 Women	46.2 *(15.0)*
		10 Men	47.4 *(17.8)*
5. Functional	20	15 Women	35.3 *(16.1)*
		5 Men	31 *(11.1)*
10. Control	23	11 Women	23.2 *(3.1)*
		12 Men	24.5 *(2.8)*

B. Procedure

Recordings: Recordings were performed individually in a quiet room, using a Sennheiser PC150 headset microphone, connected directly to a computer soundcard, sampling at 48kHz, 16 bits. Each subject was instructed to produce sustained phonations of /a/ and /i/ for 2-5 seconds. Each vowel was recorded six times, of which two were analyzed in the present study.

Segmentation: All recorded vowels were divided manually, by an experienced clinician, into three segments: onset, steady state and offset. Criteria for determining the boundaries between segments were based mainly on stability of the intensity contours. For this purpose, the F0 contours were inspected simultaneously, but for most recordings they did not provide sufficiently clear indications to be used for setting boundaries. Intrajudge and interjudge reliability were evaluated for 20% of the data, using Pearson correlation. High correlation coefficient values were obtained for both reliability measures ($0.88 < r < 0.99$, $p < 0.0001$; and $0.89 < r < 0.98$, $p < 0.0001$, respectively).

It was interesting to observe that both onset and offset segments could be classified into three groups, based on the shape of the intensity contour. Onsets were either (1) a rapid ramp-up to the steady state; (2) a rapid ramp-up followed by a small decay down to steady state levels; (3) a segment of chaotic changes in intensity before stabilizing to a steady state level. The intensity of the offset segments presented similar patterns: (1) an abrupt decay; (2) a small rise before an abrupt decay; (3) a chaotic decay.

Acoustic Analysis: All recordings were subjected to acoustic analysis. This analysis was performed once over the entire recording, and then over each of the three segments separately. The following seven acoustic features were calculated using Praat software:

1) **Duration**, 2) **Mean F0**, 3) **F0 std**, 4) **Jitt er** (Local), 5) **Shimmer** (Local), 6) **Autocorrelation** and 7) **HNR**.

Additionally, a set of 200 acoustic features were calculated using ad-hoc software written in Matlab. These are described in detail in Kessous et al. [9]. Based on prior experience with this set of measures, nine were selected for inclusion in this study:

8) **LTAS std** - standard deviation,
9) **Bark-std** – Standard deviation of the mean energy of Bark scale spectral decomposition,
10) **Bark-high** - Ratio of mean energy in the higher half of spectrum versus lower half of spectrum in bark scale decomposition,
11) **Bark-2/4** - Ratio of mean energy in the second quarter of the spectrum to the entire spectrum of bark scale decomposition,
12) **Fourier-freq** - Energy in the 25-50Hz band of the Fourier decomposition of the F0 contour,
13) **Fourier-int** - Energy in the 25-50Hz band of the Fourier decomposition of the intensity contour,
14) **MicPr-freq** - Regression error of a 3rd order polynomial fit to the F0 contour,
15) **MicPr-int** - Regression error of a 3rd order polynomial fit to the intensity contour,
16) **Fourcin** - A measure of F0 irregularity based on the work of Fourcin & Abbertion [10].

F0 contours were also calculated by Praat and corrected manually for octave errors by an experienced research assistant. Overall, these acoustic measures can be divided into three groups: (a) F0 based features (2, 3, 4, 6, 12, 14, 16); (b) Intensity based features (5, 13, 15) and (c) spectral features (7, 8, 9, 10, 11).

III. RESULTS

MANOVA with repeated measures was applied to the data. In this analysis, pathology group and gender were defined as fixed variables, while vowel (/a/, /i/) and segment (onset, steady state, offset, all) were treated as repeated measures.

Table 2 summarizes the Group main effect results for all acoustic measures, and ad-hoc group contrasts

Table 2 – Summary of MANOVA main effects for Group, and ad-hoc groups contrasts (groups are numbered based on table 1)

Measure	Group main effect	Contrasts*
Duration	$F^{\dagger}$=38.33, p<0.0001	10>1,2,3,4,5
Mean F0	$F^{\dagger}$=2.03, p=0.08	NA
F0-std	$F^{\dagger}$=4.81, p=0.0006	10<1,2
Jitter	$F^{\dagger}$=3.02, p=0.01	10<2
Shimmer	$F^{\dagger}$=0.77, p=0.58	NA
Autocorr.	$F^{\dagger}$=1.39, p=0.23	NA
HNR	$F^{\dagger}$=4.34, p=0.0014	10<5
LTAS-std	$F^{\dagger}$=8.13, p<0.0001	10<1,4,5; 2<1,5
Bark-std	$F^{\dagger}$=0.84, p=0.52	NA
Bark-high	$F^{\dagger}$=2.94, p=0.0164	NA
Bark-2/4	$F^{\dagger}$=0.80, p=0.52	NA
Fourier-freq	$F^{\dagger}$=5.77, p<0.0001	10<1,2; 5<2
Fourier-int	$F^{\dagger}$=3.58, p=0.0052	10<5
MicPr-freq	$F^{\dagger}$=6.00, p<0.0001	10<1,2,3; 4<1
MicPr-int	$F^{\dagger}$=10.48, p<0.0001	10<1,3,4,5; 2<1
Fourcin	$F^{\dagger}$=4.03, p=0.0024	10,5<1

(†df=5,92; *adjusted p<0.05)

Generally, in the parameters that yielded a significant Group effect, a significant contrast was found between one or more pathology group and the control group (#10). Very few significant contrasts were found among the five pathological groups. These contrasts were not consistent using the different parameters.

Table 3 - Summary of MANOVA main effect for Segment (Ons=onset, Sts=steady state, Offs=offset, All=complete vowel), and ad-hoc group contrasts

Measure	Segment main effect	Contrasts*
Duration	$F^{\dagger}$=2945.53, p<0.0001	All>Sts>Ons>Offs
Mean F0	$F^{\dagger}$=1.10, p=0.35	NA
F0-std	$F^{\dagger}$=30.70, p<0.0001	Ons>All>Sts,Offs
Jitter	$F^{\dagger}$=3.83, p<0.0001	Offs>Ons>All,Sts
Shimmer	$F^{\dagger}$=47.56, p<0.0001	Offs>Ons>All,Sts
Autocorr.	$F^{\dagger}$=47.90, p<0.0001	Sts,All>Offs,Ons
HNR	$F^{\dagger}$=80.91, p<0.0001	Sts,All>Offs,Ons
LTAS-std	$F^{\dagger}$=280.7, p<0.0001	Ons>All,Sts>Offs
Bark-std	$F^{\dagger}$=33.00, p<0.0001	Ons,All>Offs>Sts
Bark-high	$F^{\dagger}$=7.13, p<0.0001	Ons>All,Sts,Offs
Bark-2/4	$F^{\dagger}$=24.25, p<0.0001	Ons>All,Sts>Offs,
Fourier-freq	$F^{\dagger}$=2.31, p=0.13	NA
Fourier-int	$F^{\dagger}$=41.28, p<0.0001	All>Sts
MicPr-freq	$F^{\dagger}$=36.31, p<0.0001	Ons,All>Sts,Offs
MicPr-int	$F^{\dagger}$=218.74, p<0.0001	All>Ons>Sts>Offs
Fourcin	$F^{\dagger}$=17.11, p<0.0001	Ons>Offs>All,Sts

(†df=3,293; *adjusted p<0.05)

Table 3 summarizes the results of the main effect for Segment, and post hoc significant contrasts. As can be seen, most parameters yielded a significant main effect for Segment. Nonetheless, in most cases there was no significant difference between the values obtained from the steady state (Sts) and from the complete vowel (All).

A significant interaction between Group and Segment was found for nine of the parameters tested. These included Duration ($F_{(15,293)}$=61.55, p<0.0001), Mean F0 ($F_{(15,293)}$=1.92, p=0.02), Jitter ($F_{(15,293)}$=2.34, p=0.0036), Autocorrelation ($F_{(15,293)}$=2.67, p=0.0008), HNR ($F_{(15,293)}$=3.30, p=0.001), Bark-std ($F_{(15,293)}$=4.36, p<0.0001), MicPr-freq ($F_{(15,293)}$=2.40, p=0.0027), MicPr-int ($F_{(15,293)}$=7.01, p<0.0001) and Fourcin ($F_{(15,293)}$=1.87, p=0.026). Table 4 presents the significant group contrasts within the four segment categories, only for these parameters. Data show that the majority of the group contrasts were found either in the steady state or in the complete vowel. Only a few group contrasts could be found based on the analyses of the onset and offset of the vowels. Similar results were found when analyses were based on the steady state and on the complete vowel.

Table 4 – Significant group contrasts (adjusted p<0.05) within the four segments, for the acoustic measures that yielded a significant Group X Segment interaction (groups are numbered based on table 1)

Measure	Ons	Sts	Offs	All
Duration	NA	10>1,2,3,4,5	NA	10>1,2,3,4,5
Mean F0	2>10	2>10 3>4	3>4	2>10 3,10>4
Jitter	1>3,4,10	2>4,10	NA	2>4,10,5
Autocorr.	NA	4>10	5>10	4>10
HNR	3>10	1,3,4,5>10 3,4,5>2	4,5>10	1,3,4,5>10 3,5>2
Bark-std	NA	NA	10>1,5	NA
MicPr-freq	1>4,10	1,2>4,5,10	1>4 2>10	1>4,5,10 2,3>10
MicPr-Int	4>10 1>5,10	1>4,10 5>2,4,10	NA	1,3,4,5>10 5,4>2;1>3,5
Fourcin	1>2,3,4,5,10	2>1,5,10 3>5,10	NA	1,3>10

A significant difference between the two vowels was found for the following measures: Duration ($F_{(1,95)}$=7.28, p=0.0083), Mean F0 ($F_{(1,95)}$=86.18, p<0.0001), Autocorrelation ($F_{(1,95)}$=17.66, p<0.0001), HNR ($F_{(1,95)}$=118.28, p<0.0001), Bark-high ($F_{(1,95)}$=144.41, p<0.0001), Bark-2/4 ($F_{(1,95)}$=1169.11, p<0.0001), Fourier-freq ($F_{(1,95)}$=4.22, p=0.04), Fourier-int ($F_{(1,95)}$=38.03, p<0.0001) and MicPr-int ($F_{(1,95)}$=116.9, p<0.0001).

Gender differences were found for the following measures: MeanF0 ($F_{(1,92)}$=122.6, p<0.0001), Bark 2/4 ($F_{(1,92)}$=6.73, p=0.01), Fourier-freq ($F_{(1,92)}$=9.18, p=0.0001), Fourier-int ($F_{(1,92)}$=17.1, p<0.0001) and MicPr-freq ($F_{(1,95)}$=11.66, p<0.0001).

IV. DISCUSSION

Results demonstrate that most acoustic measures differentiated significantly between the control group and one or more of the pathological groups. However, in most cases, differences among the five pathological groups failed to reach statistical significance. This is reminiscent of previous studies [3, 4]. It indicates that the acoustic measures tested can serve as an indication of pathological voice, but not differentiate among pathologies.

Some significant differences were found between the acoustic features obtained from the different vowel segments. For the majority of features, no significant differences were found between the values obtained for the steady-state and for the entire vowel. This can be interpreted as weakening the clinical justification for extracting the steady-state of the vowel for acoustic analysis, as analysis of the complete vowel provides similar results.

Similar clinical contrasts (i.e., group differences) were obtained for analyses based on the steady state and on the complete vowels. On the other hand, analyses based on the onset and offset of the vowel revealed only a limited number of group contrasts. Furthermore, in contrast to our expectations, the analyses of the onset and offset did not reveal additional group differences that were not revealed by the analyses of the steady-state alone or the complete vowel.

Apparently, the onset and offset of the vowel did not provide additional acoustic cues for the comparison among pathological groups, which could be identified using the acoustic measures included in the present study. This can be interpreted as further support for the use of acoustic analysis of complete vowels for clinical comparison between pathological groups and control, in a parallel manner to the use of the steady state.

It should be noted, though, that different results might have been obtained if the allocation of patients to the pathological subgroups was done differently, or if additional acoustic measures had been analyzed. Different acoustic measures, which are specifically designed to examine the onset/offset of the vowel, could provide additional valuable clinical information that might assist in differentiating among pathologies.

REFERENCES

[1] Titze, I.R.(1995) *Workshop on acoustic voice analysis*; Summary statement. Denver: National Center for Voice and Speech.
[2] de Krom, G. (1994). Consistency and reliability of voice quality rating for different types of speech fragments. *Journal of Speech and Hearing Research 37*, 985-1000.
[3] Eadie, T.L., & Doyle, P.C. (2005). Classification of dysphonic voice: Acoustic and auditory-perceptual measures. *Journal of Voice, 19,* 1-14.
[4] Fourcin, A. (2009). Aspects of voice irregularity measurement in connected speech. *Folia Phoniatrica et Logopaedica, 61,* 126-136.
[5] Vasilakis, M., & Stylianou Y. (2009). Voice pathology detection based on short-term jitter estimations in running speech. *Folia Phoniatrica et Logopaedica, 61,* 153-170.
[6] Carding, P.N., Steen, I.N., Webb, A., Mackenzie, K., Deary, I.J., & Wilson, J.A. (2004). The reliability and sensitivity to change of acoustic measures of voice quality. *Clinical Otolaryngology, 29*, 5538-544.
[7] Campisi, P., Manoukian, J.J., Schloss, M.D., Pelland-Blais, E., & Sadeghi, N. (2002). Computer-assisted voice analysis. *Archives of Otolaryngology Head and Neck Surgery, 128,* 156-160.
[8] Amir, O., Biron-Shental, T. & Shabtai, E. (2006). Birth control pills and non-professional voice: Acoustic analyses. *Journal of Speech, Language and Hearing Research, 49, (5),* 1114-1126.
[9] Kessous, L., Amir, N., Cohen, R. (2007). Evaluation of time/frequency representations for automatic classification of expressive speech. *Paraling 2007, Saarbrucken.*
[10] Fourcin, A., Abberton, E. (2007). Hearing and phonetic criteria in voice measurement: clinical applications. *Logopedics Phoniatrics Vocology*, 1-14

VOICE QUALITY EVALUATION USING CAPE-V AND GRBAS IN EUROPEAN PORTUGUESE

Luis M. T. Jesus[1,2], Anna Barney[3], Pedro Sá Couto[4], Helena Vilarinho[1], Ana Correia[1]

[1]Escola Superior de Saúde da Universidade de Aveiro (ESSUA), Universidade de Aveiro, Aveiro, Portugal
[2] Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), Universidade de Aveiro, Aveiro, Portugal
[3]ISVR, University of Southampton, Southampton, UK
[4] Research Unit Mathematics and Applications, Departamento de Matemática, Universidade de Aveiro, Portugal

Abstract: **In this study, the voice quality of 40 patients was assessed, with the Universidade de Aveiro's Voice Evaluation Protocol. The sample included 40 patients with a variety of clinical diagnoses. A number of acoustic parameter were extracted including: median F0, mean F0, F0 std deviation, Jitter and Shimmer and HNR. Analysis of the correlation between corresponding parameters of the CAPE-V and GRBAS scales was made. The perceptual parameters grade (global in CAPE-V), roughness and breathiness were also compared individually with the objective acoustic parameters.**

Keywords : **Voice, assessment, acoustic parameters**

I. INTRODUCTION

Studies in the area of voice assessment still lack objectivity in the description and evaluation of many aspects of vocal quality. In therapeutic practice in Portugal, the assessment of vocal pathologies is not uniform, because each Speech and Language Therapist (SLT) uses individual and diverse assessments to evaluate. The patient is assessed with scales that try to include the most important and appropriate parameters for each patient, based on existing protocols [1-11] which have normally been developed for languages other than European Portuguese (EP) and which, therefore, may not capture or account for language specific features of voice. Although voice quality measures using GRBAS [11] are normally considered to be language independent, the performance of this scale for assessment of EP has not, to date, been systematically tested. It is not the scale we expect to be language dependent, but the material used to carry out the test (e.g., if the text is English some EP vowels will be missing and some non-EP vowels might be included).

The aim of this project was to develop the first standardised and evaluated protocol for subjective voice assessment in EP: Universidade de Aveiro's Voice Evaluation Protocol [12]. It is intended as a working tool for Speech and Language Therapists (SLTs), which brings together a range of essential information, thus preparing patients for a therapeutic intervention. SLTs involved in future studies will use the same evaluation instrument to acquire data that is comparable, thereby normalising the practice and nomenclature for this area of intervention and allowing also better inter-professional communication. A pilot study to test the reliability of the protocol including the analysis of inter-rater correlation, using a group of patients with various vocal disorders, has been reported elsewhere [12].

The full voice evaluation protocol includes a wide range of parameters for assessing vocal function (see [12] for a complete description) but we concentrate here on parameters for subjective assessment of voice quality and on objective measures derived from the acoustic signal. Perceptual analysis of voice quality includes the application of an EP version of the CAPE-V [10] developed as part of this project and of the GRBAS [11] scales. The EP version of the CAPE-V [12] includes six new EP sentences designed to stimulate production of every oral vowel in EP, easy onset with /s/, voiced-only phonemes, hard glottal attack, nasal phonemes and voiceless stops, as described in [10].

Perceptual analysis is based on sustained productions of /a, i, u, O/, CAPE-V sentences and reading the EP version of the "The North Wind and the Sun" passage, recently proposed as a standard text for "Advanced Voice Assessment" [13].

In our previous study [12], the reliability of the voice quality protocol was tested, using two independent raters, who evaluated, coincidentally, a group of patients who exhibited some change in voice quality.

The protocol parameters severity, roughness, breathiness, change of loudness (CAPE-V), grade, breathiness and strain (GRBAS), presented high reliability and were highly correlated (with good inter-rater agreement and a high value of correlation [12], similar results to assessments of other languages [1, 4, 14]). Values for the overall severity and grade were similar to those reported in the literature.

II. METHOD

In this study, the voice quality of 40 patients was assessed, with the Universidade de Aveiro's Voice Evaluation Protocol [12] (with full ethics committee approval). These patients had been admitted to the Department of Otolaryngology of the Hospital de São João, Porto, Portugal. The sample included several clinical diagnoses: nodules, polyps, hypotonia of the

vocal folds, Reinke's oedema, musculo-skeletal syndrome and dysfunctional dysphonia. The diagnosis for the sample was made by an experienced SLT and an Otorinolaringology consultant.

The speech tasks were recorded directly onto a PC, using Praat 5.1.10 [15] in a quiet environment.

The majority of recordings used a Sony F-V220 microphone and a SoundMAX Digital Audio internal soundcard (16 bits and 22050 Hz sampling frequency). A small number of recordings were made instead with an external sound card Edirol UA-25, set to 16 bits and 44100 Hz sampling frequency, and a Sennheiser e815S microphone. During the recordings, the microphone was held on a tripod placed 25-30 degrees to the left of the patient's mouth, at a distance of 30-40 cm.

Various acoustic parameters were extracted from the audio signal using Praat 5.1.10 [15] including: F0 Hz (median), F0 Hz (mean), F0 (std deviation), Jitter% (ppq5 – five-point frequency perturbation quotient equivalent to MDVP's PPQ) and Shimmer% (apq11 – eleven-point amplitude perturbation quotient equivalent to MDVP's APQ) and HNRdB (mean Harmonics-to-Noise Ratio).

A single sustained sample selected from each speaker's productions of vowel /a/ was used to extract the acoustic parameters using the following criteria: one token "considered as perceptually closest to the subject's natural voice" [16, p. 23] and produced with a "comfortable pitch and volume" [16, p. 23]; 100 consecutive cycles taken 200 ms after phonation onset [17, p. 1261] were used for analysis.

Analysis of the correlation between corresponding parameters of the CAPE-V and GRBAS scales was made. The perceptual parameters grade (global in CAPE-V), roughness and breathiness were also compared individually with the objective acoustic parameters.

Acoustic parameter data and scale evaluation scores were compared to find statistically significant differences between males and females using the Mann-Whitney U Test. Correlation analysis (Spearman correlation test) between the perceptual parameters in CAPE-V and GRBAS scales (global in CAPE-V and grade in GRBAS, roughness and breathiness) were also evaluated with the acoustic parameters: median F0, mean F0, standard deviation F0, jitter ppq5, shimmer apq11, and mean HNR. Finally, analysis of the correlation (Spearman correlation test) between corresponding parameters of the CAPE-V and GRBAS scales was made. All statistical analyses were conducted using SPSS 13.0 and a p value of less than 0.05 was considered significant. All data presented are given in mean ± standard deviation (S.D.)

III. RESULTS

In the tables below, the variable *total* represents the case when male and female data is combined.

Table 1 shows correlation analysis between CAPE-V and GRBAS scales. Statistical significances are found between the perceptual subscale grade from GRBAS and subscales global and roughness from CAPE-V, roughness in GRBAS and global in CAPE-V, and breathiness in GRBAS and in CAPE-V. The correlation values are good, ranging from 0.60 to 0.87, with the exception of the correlation value between the subscale roughness in GRBAS and the subscale global in CAPE-V for the *total* value. The results found for males alone can be ascribed to the smaller sample size.

Table 2 presents the acoustic parameter data and scale evaluation scores. The sample consists of 9 males (mean age 56.11±3.55) and 31 females (mean age 43.29±2.36). Statistically significance differences between males and females are found in age, median F0, mean F0, jitter ppq5, mean HNR and in several parameters in the CAPE-V scale (global and roughness) and in the GRBAS scale (grade and roughness). Such differences in the acoustic parameters are to be expected. The differences in scales can be explained due to the smaller number of males.

Table 3 shows the correlation analysis between the CAPE-V and GRBAS scales with the selected acoustic parameters. Statistical significances are found for median F0 and mean F0 with the perceptual subscales global and roughness for CAPE-V and grade and roughness for GRBAS. However, the correlations are weak, with values ranging from -0.38 to -0.60. No significant differences are found when we consider either only the male sample or the other acoustic parameters.

IV. DISCUSION AND CONCLUSIONS

The two scales (GRBAS and CAPE-V) have been previously used simultaneously [4], with results showing a strong correlation between the two rating systems (Spearman's correlation coefficients from ranging 0.89 to 0.95) for: GRBAS grade vs. CAPE-V global; GRBAS roughness vs. CAPE-V roughness; GRBAS breathiness vs. CAPE-V breathiness. Our results have also shown a good correlation except for roughness, because the term used in EP and Brazilian Portuguese for Grade is "grau de rouquidão", which (see Table 1), appears to have been erroneously related to the CAPE-V EP term "rouquidão" (roughness).

This issue will be addressed in the future with a further validation of the EP version of CAPE-V that will use the procedures presented in [18], including the production of a CD-ROM with voice samples to be evaluated, voices used for training and samples of voices that represent specified grades of severity.

The lack of success in finding hypothesised correlations between acoustic and perceptual measures have long been known [19, pp. 75-80], and do not seem to be related to language specific characteristics, as our results have shown, even when we limit our set of

acoustic and perceptual parameters as in [20]. Different factors contribute to a failure to find consistent correlations: deficiencies in the theoretical framework; incoherencies in the definitions of parameters; limitations in estimation techniques [19, pp. 75-80].

Ongoing and future work will extend this study to a larger number of patients, especially by increasing the number of males analysed, so the protocol can be used with more confidence. The pilot protocol presented in [12] has a large number of parameters and it is intended to evaluate these further to derive, if possible, a best set of parameters for EP voice quality assessment.

V. ACKNOWLEDGEMENTS

REFERENCES

[1] B. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G. Berke, "Comparing Internal and External Standards in Voice Quality Judgments," Journal of Speech Hearing Research, vol. 36, pp. 14-20, 1993.

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

[3] J. Piccirillo, C. Painter, D. Fuller, A. Haiduk, and J. Fredrickson, "Assessment of two objective voice function indices," Ann. Otol. Rhinol. Laryngol., vol. 107, pp. 396-400, 1998.

[4] M. Karnell, S. Melton, J. Childes, T. Coleman, S. Dailey, and H. Hoffman, "Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient- Based (V-RQOL and IPVI) Documentation of Voice Disorders," Journal of Voice, vol. 21, pp. 576-590, 2007.

[5] J. Muñoz, E. Mendoza, M. Fresneda, G. Carballo, and I. Ramirez, "Perceptual analysis in different voice samples: agreement and reliability," Journal Percept Mot Skills, vol. 94, pp. 1187-9, 2002.

[6] P. Carding, E. Carlson, R. Epstein, L. Mathieson, and C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom," Logopedics Phoniatrics Vocology, vol. 25, pp. 133-138, 2000.

[7] K. Peppé, "Assessment of Prosodic Ability in Atypical Populations," Department of Clinical Language Sciences, Reading University. , 2007.

[8] D. Boone, The Boone Voice Program for Adults, 2nd Ed ed., 1982.

[9] W. Haynes, M. Moran, and R. Pindzola, "Voice Disorders," In: Haynes, Moran and Pindzola "Communication Disorders in the Classroom: An Introduction for Professionals in Schools Settings," 2006, pp. 267-269.

[10] ASHA, CAPE-V Form and Procedures. ASHA Special Interest Division 3, Voice and Voice Disorders, 2006.

[11] M. Hirano, "Clinical Examination of Voice," Vienna: Springer-Verlag, 1982.

[12] L. Jesus, A. Barney, R. Santos, J. Caetano, J. Jorge and P. S. Couto, "Universidade de Aveiro´s Voice Evaluation Protocol," in Proceedings of InterSpeech, Brighton, UK, 2009.

[13] M. Pedersen and K. Munck, "Advanced Voice Assessment," in Proceedings of MAVEBA, Florence, Italy, 2007, pp. 61-64.

[14] I. Bele, "Reliability in Perceptual Analysis of Voice Quality," Journal of Voice, vol. 19, pp. 555-573, 2004.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," in Glot International, 2001, pp. 341-345.

[16] P. Yu, R. Garrel, R. Nicollas, M. Ouaknine and A. Giovanni, "Objective Voice Analysis in Dysphonic Patients: New Data Including Nonlinear Measurements," Folia Phoniatr. Logop., vol. 59, pp. 20-30, 2007.

[17] R. Scherer, V. Vail and C. Guo, "Required Number of Tokens to Determine Representative Voice Perturbation Values," Journal of Speech and Hearing Research, vol. 38, pp. 1260-1269, 1995.

[18] R. I. Zraick, "Results of the CAPE-V Validation Study," in Oral Presentation at ASHA Convention, Boston, 2007.

[19] P. Dejonckere, J. Martens, H. Versnel and M. Moerman, "The Effect of Visible Speech on Perceptual Rating of Pathological Voices, and on Correlation Between Perception and Acoustics," in Proceedings of MAVEBA, Florence, Italy, 2007, pp. 21-24.

[20] J. Kreiman and B. Gerratt, "Measuring Vocal Quality," In: Kent and Ball, "Voice Quality Measurement," 2000, pp. 73-101.

Table 1. Correlation analysis between CAPE-V and GRBAS scales.

CAPE-V	GRBAS		
	Grade	Roughness	Breathiness
Global			
Male	0.00	**0.86***	0.24
Female	**0.65***	0.20	0.26
Total	**0.60***	**0.42***	0.20
Roughness			
Male	0.23	-0.14	-0.15
Female	**0.79***	0.14	-0.19
Total	**0.75***	0.26	-0.20
Breathiness			
Male	-0.18	**0.67***	**0.63***
Female	0.09	-0.10	**0.87***
Total	0.04	0.08	**0.80***

* Spearman correlation test, statistical significance ($p < 0.05$).

Table 2. Acoustic parameters and scale evaluation scores (Gender: Male=9, Female=31; Age: Male=56.11±3.55*, Female=43.29±2.36*, Total=46.18±13.62).

Acoustic parameters	Mean ± S.D.	Scales	Mean ± S.D.
		CAPE-V	
Median F0 (Hz)		Global	
Male	**123.88±7.59***	Male	**0.62±0.07***
Female	**184.93±6.11***	Female	**0.41±0.04***
Total	171.19±40.76	Total	0.46±0.21
Mean F0 (Hz)		Roughness	
Male	**124.09±7.62***	Male	**0.63±0.04***
Female	**184.86±6.10***	Female	**0.42±0.05***
Total	171.19±40.65	Total	0.46±0.24
Std Dev F0 (Hz)		Breathiness	
Male	2.50±0.42	Male	0.37±0.10
Female	2.63±0.51	Female	0.36±0.04
Total	2.60±2.57	Total	0.36±0.25
Jitter ppq5 (%)		GRBAS	
Male	**0.85±0.28***	Grade	
Female	**0.34±0.04***	Male	**2.33±0.17***
Total	0.45±0.49	Female	**1.55±0.15***
Shimmer apq11 (%)		Total	1.73±0.85
Male	3.68±0.56	Roughness	
Female	2.99±0.48	Male	**0.89±0.31***
Total	3.15±2.46	Female	**0.32±0.13***
Mean HNR (dB)		Total	0.45±0.81
Male	**12.28±1.56***	Breathiness	
Female	**17.88±0.84***	Male	1.22±0.32
Total	16.63±5.20	Female	1.22±0.14
		Total	1.23±0.83

* Mann-Whitney test, statistical significance ($p < 0.05$).

Table 3. Correlation analysis between the CAPE-V and GRBAS scales with the selected acoustic parameters: median F0, mean F0, standard deviation F0, jitter ppq5, shimmer apq11, and mean HNR.

Scales	Acoustic parameters					
	Median F0 (Hz)	Mean F0 (Hz)	Std Dev F0 (Hz)	Jitter ppq5 (%)	Shimmer apq11 (%)	Mean HNR (dB)
CAPE-V						
Global						
Male	-0.08	-0.08	0.43	0.22	0.34	-0.29
Female	**-0.41***	**-0.39***	-0.09	0.08	0.00	0.03
Total	**-0.49***	**-0.48***	0.00	0.20	0.13	-0.18
Roughness						
Male	0.13	0.13	-0.06	0.16	0.26	-0.16
Female	**-0.55***	**-0.53***	0.04	0.19	-0.01	-0.10
Total	**-0.60***	**-0.58***	0.03	0.29	0.12	-0.27
Breathiness						
Male	0.00	0.00	0.38	0.10	0.12	-0.05
Female	0.17	0.20	0.04	0.10	0.11	0.00
Total	0.12	0.14	0.12	0.08	0.10	0.00
GRBAS						
Grade						
Male	-0.09	-0.09	-0.09	-0.18	-0.37	0.00
Female	**-0.52***	**-0.48***	0.10	0.10	-0.06	-0.02
Total	**-0.58***	**-0.56***	0.09	0.17	0.02	-0.18
Roughness						
Male	-0.26	-0.26	0.13	-0.21	-0.16	-0.19
Female	-0.22	-0.24	0.02	0.21	0.26	-0.22
Total	**-0.38***	**-0.39***	0.05	0.19	0.19	-0.30
Breathiness						
Male	0.29	0.29	0.03	0.24	0.24	0.16
Female	0.20	0.24	-0.07	0.05	0.09	0.03
Total	0.18	0.20	-0.05	0.06	0.13	0.08

* Spearman correlation test, statistical significance ($p < 0.05$).

VOICE RELATED QUALITY OF LIFE IN SPASMODIC DYSPHONIA : A DETAILED VHI-ANALYSIS BEFORE AND AFTER BOTULINUM TREATMENT

K. J. Neumann[1], P. H. Dejonckere[2,3,4],

[1]J.W. Goethe Universität, Frankfurt-am-Main, Germany
[2]Utrecht University, Utrecht, The Netherlands
[3]Federal Institute of Occupational Diseases, Brussels, Belgium
[4]Catholic University of Leuven, Leuven, Belgium
COST-Action 2103 Advanced Voice Function Assessment

Abstract: **The Voice Handicap Index (VHI) is a widespread instrument for measuring the psycho-social handicapping effect of a voice disorder over 3 domains, the Physical (P), the Emotional (E) and the Functional (F) domain. It is a disease specific quality of life instrument and consists of 30 items/statements (10 in each domain), which are to be scored from 0 to 4 with a maximum score of 120. The higher the score, the more there is a handicapping effect caused by the voice disorder. An abridged version (10 out of the 30 statements : VHI10) has been proposed and validated.**

Spasmodic Dysphonia (SD) patients *(adductor type)* are known to report in average extremely high VHI-scores. A detailed analysis is necessary to get better understanding of this phenomenon, particularly in the scope of therapy effects with Botulinum Toxin injections.

I. MATERIAL AND METHODS

28 VHI forms were filled in and analyzed : 24 are originating from 12 patients diagnosed with adductor SD, and investigated (just) pre- and (a few weeks) post treatment. 3 patients had no post-treatment self-evaluation. 1 patient had 2 pre- self-evaluations at different moments, with a time interval of several months. There were 9 females and 6 males. Mean age was 60,6 (+/- 9,3) years.

II. RESULTS & DISCUSSION

The average pre-therapy score is 64.17 (+/- 21.98), and is reduced to 48.75 (+/- 22.54) after treatment. A reduction of 15,41 points may be considered as clinically relevant for a group design.

A paired comparison pre-/post also demonstrates a significant improvement in voice-related quality of life ($p = .039$). The effect size is to be considered as medium to large (Cohen's d = .7). The median value for the VHI-score in the general population is 6 with an asymmetrical distribution (p 25 = 2; p75 = 12; p90 = 23; p95 = 32.8). None of our patients originally scores within the p95 range of the general population, but 33% shift to this range after treatment. No clear age or gender related effect is observed. Factor and principal component analysis identifies clusters of statements, but these differ from the P, E and F domains as defined by the original authors. Clusters of statements can be ranked according to their sensitivity to changes induced by therapy. Scores of the total VHI are also correlated with those of the VHI10.

III. CONCLUSION

Patients with SD report a strong impact of their voice disorder on their quality of life, but VHI (and VHI10) are sensitive to therapeutic changes. Clustering of statements is possible, but these clusters differ from the original 'domains'. Ranking these clusters according to their sensitivity to changes induced by therapy provides interesting insights in the background of self-assessment.

LONG TERM FOLLOW-UP OF PATIENTS WITH SPASMODIC DYSPHONIA

P.H. Dejonckere[1], J.P. Martens[2], M. Moerman[3]

[1] Utrecht University, Catholic University of Leuven, Federal Institute of Occupational Diseases, Brussels
[2] Ghent University, Computer Science
[3] Utrecht University & Jan Palfijn Hospital, Ghent
COST Action 2103 Advance Voice Function Assessment

Abstract: **'Adductor spasmodic dysphonia' (SD) is a focal laryngeal dystonia mainly resulting in a strained voice quality with spastic voice breaks and frequency shifts, perturbing fluency and intelligibility. It is well known that SD-patients report unusually high scores on the VHI, as they experience their disease as seriously impairing their quality of life. The standard treatment is Botulinum Toxin (BT) injection in the thyroarytenoid muscles, in order to interfere with the perturbed sensory feedback loop of kinetic muscle tension regulation. The mode of action of this toxin is at cholinergic nerve terminals where it inhibits the release of acetylcholine. However, the globally favourable effects are only temporary, in part because of the formation of remodeled neuromuscular junctions after a few months, but the Botulinum injections can be repeated. There is a lack of information about long term effects.**

I. MATERIALS AND METHODS

In the current study, long term evolution is analysed in 19 patients having been injected with BT between 4 and 18 times over periods of 3 to 16 years. Our approach is based on
(1) a differentiated perceptual panel rating, including conventional and dedicated parameters
(2) a computerized program for signal analysis that is suited for irregular voices, and that mainly deals with voicing and aperiodicity criteria. Material is a phonetically selected constantly voiced sentence.
(3) a patient' self evaluation on 2 visual analog scales: voice quality itself and social/occupational handicap.
(4) a quantification of side effects : temporary breathiness and aspiration

II. RESULTS

Moments of treatment clearly determine a saw teeth effect in most parameters, particularly those self evaluated by the patient. Over time the acoustic parameters just before a new injection become significantly less deviant, without reduction of time delay between the injections. This differs from the patient's self evaluations : the pre-treatment scores worsen with time, while the best scores between consecutive injections remain remarkably stable.

III. CONCLUSION

Repeated BT injections remain active even at long term, but Spasmodic Dysphonia cannot be cured with Botulinum. There seems to be an individual ceiling effect for the achievable functional result. Objective measurements demonstrate stability, even a slight improvement over time. Patient's self evaluations worsen over time. Side effects do not grow worse.

PRINCIPAL COMPONENT ANALYSIS FOR HMM-BASED PATHOLOGICAL VOICE DETECTION

M. Sarria-Paja[1], G. Castellanos-Domínguez[2], N. Gaviria-Gómez[3]

[1]Inteligent Machines and patern recognition Group, Instituto Tecnológico Metropolitano, Medellín-Colombia
[2]Control and Digital Signal Processing Group, Universidad Nacional de Colombia, Manizales-Colombia.
[3]Facultad de Ingeniería, Universidad de Antioquia, Medellín-Colombia

Abstract: **This paper presents a methodology for feature selection in dynamic problems based on the analysis of the variation of linear components in acoustic features combined with an estimation of the ratio between a compactness measure to the separation measure. The methodology is applied to the automatic detection of voice disorders by means of stochastic dynamic models; results showed a significant reduction in the number of features, 96.6% of accuracy, and a 62.2% of computational cost reduction.**
Keywords: **Dynamic features, HMM, PCA, feature selection, pathological voice, clustering.**

I. INTRODUCTION

During phonation of sustained vowels, the normal voice is a regular and periodic signal; however changes in its waveform can be appreciated if some disorders arise. Moreover, the classical distortion measures based on fluctuations of acoustic measures may be complemented with dynamic features obtained from its contours, as pointed out by other studies [1]. One of the key properties that make dynamic features useful is that they consider changes in the temporal structure of the excitation signal. Short term features combined with dynamic classifiers (e.g. Hidden Markov Models - HMM), have been used in the classification of pathological voices [2]. But, it is no clear whether gathering of dynamic features should lead to an improved representation capability, and hence to higher performance of the dynamic classifier. Namely, in voice recognition where the training data are labeled, a projection is often required to emphasize the discrimination between the clusters. Therefore, a more detailed study should be conducted to assess the relevance of dynamic features that describe pathologies, which could be used in data analysis and evaluation to support diagnose by automatic dynamic classifiers.

Performance in training of pattern recognition systems to detect pathologies can be increased, if proper feature extraction is done. Training procedures usually deal with a high number of features, nevertheless a high dimension input space means significant processing time, higher cost of the collected biosignal records since more observations are needed, and the well known curse of dimensionality phenomena [3].

The aim of this paper is to assessment an approach combining PCA with HMM for pathological voice diagnosis based on a concrete cluster validity measure. For this purpose, the feature selection methodology presented in [2], is adopted, and by incorporating both measures of cluster separability and cluster compactness, it is showed that one can provide analysis of clustering scatter for groups with varying populations. The method is tested on the voice disorders database developed by The Massachusetts Eye and Ear Infirmary Voice Laboratory (MEEIVL).

This paper makes a contribution to the effort to make an automatic discrimination between pathological and normal voice.

II. METHODS

A. Hidden Markov Models

Hidden Markov models are double-layer stochastic processes, composed of a hidden layer that controls the time evolution of spectral characteristics of an observable layer. A hidden Markov model has N distinct states and each state is uniquely defined by an observation (or output) probability density, usually a mixture Gaussian density (continuous case) or a discrete density (Discrete case), that provides a likelihood for a given vector having been generated by the state. The transition from a state is governed by the state transition probabilities and influenced by the current observation vector. The state observation and transition probabilities provide a probabilistic mechanism for association of a time sequence of vectors with a given HMM model [4].

B. Principal Component Analysis in dynamic features

Widely known approaches, like PCA [5,6,7] and sequential search methods, have been customized as eature selection methods for the use with a HMM classifier. Assuming that the input contour data are highly correlated, linear transformation methods such as PCA tries to exploit the correlation present in the data by projecting the data onto a new space where the axes are orthogonal to each other.

Let $\xi_{ij}[k], k=1,\ldots,m$ be the j-th dynamic feature belonging to i-th observation, where $j=1,\ldots,p$, $i=1,\ldots,n$; being n the number of observations and p the number of features, which change over time k. Each

vector observation ξ_i can be represented by a supervector of size $mp \times 1$:

$$\xi_i = [\xi_{i1}[1], \xi_{i1}[2], \cdots, \xi_{i1}[m], \xi_{i2}[1], \xi_{i2}[m], \cdots, \xi_{ip}[m]]^T$$

The respective covariance matrix, after centering each one of the observation supervectors is computed as:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^0\xi_i^{0T} = \frac{1}{n}\mathbf{G}\mathbf{G}^T \qquad (1)$$

Where $\mathbf{G}$ stands for matrix $\mathbf{G} = \begin{bmatrix} \xi_1^0 & \xi_2^0 \cdots \xi_n^0 \end{bmatrix}$. In most cases, we are far away from computing the eigenvectors $\mathbf{v}$ and eigenvalues λ of such a huge matrix. Nevertheless, the rank properties of $\mathbf{G}$ can be used, in special, the one that state that $\mathbf{G}\mathbf{G}^T$ has the same non-null eigenvalues than $\mathbf{G}^T\mathbf{G}$ and the advantage of $n \ll pm$, as given in [8]:

$$\mathbf{G}^T\mathbf{G}\hat{\mathbf{v}}_i = \lambda\hat{\mathbf{v}}_i \qquad (2)$$

being $\hat{\mathbf{v}}_i$ the eigenvectors of $\mathbf{G}^T\mathbf{G}$, so that, $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i$. Therefore, the eigenvectors corresponding to non-zero eigenvalues of $\mathbf{S}$ are $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i / \|\mathbf{G}\hat{\mathbf{v}}_i\|$. The eigenvectors associated with the r largest eigenvalues of $\mathbf{S}$ are selected as Principal Directions [9], which span an orthonormal basis for a subspace containing most of the information given by observations. Trying to reproduce the observation in the original space as a linear combination of the r principal directions,

$$\hat{\xi}_i^0 = \sum_{k=1}^{r} w_k \mathbf{v}_k^T \qquad (3)$$

so, from (3) the reconstruction weights $w_k = \mathbf{v}_k^T \xi_i^0$ can be though as the new set of features, and taking advantage of the orthonormality property of the basis, observations can be recognized using geometric criteria to partition the subspace off.

On the other hand, this method allows identify and choose those dynamic features that influence the most. The magnitudes of the entries of the eigenvectors that span the representation basis, tell us the variables to be choose. Let ρ be the vector expressed as; $\rho = \sum_{k=1}^{r} \lambda_k |\mathbf{v}_k|$, so, that its larger values are the most significant windows from the dynamic features, this sum of absolute values is an approximation due to the equivalence of norms in finite subspaces (L^1, and L^2). Rearranging ρ in the following manner:

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1m} & \rho_{21} & \cdots & \rho_{2m} & \cdots & \rho_{p1} & \cdots & \rho_{pm} \end{bmatrix}^T \qquad (4)$$

$$\Rightarrow \mathbf{P} = \begin{bmatrix} \rho_{11} & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{p2} \\ \vdots & \vdots & & \vdots \\ \rho_{1m} & \rho_{2m} & \cdots & \rho_{pm} \end{bmatrix}$$

it is possible to obtain the scalar $\hat{\rho}_j = \sum_{k=1}^{m}\rho_{jk}, j=1,\ldots,p$ which is the sum of the elements of each column j from $\mathbf{P}$ matrix. In consequence, the main assumption is that the largest values of $\hat{\rho}_j$ point out to the best input attributes since they exhibit higher overall correlations with principal components.

C. Definition of a clustering validity measure

Th Given a set of n observations in a p-dimensional input training space, $\mathbf{X} \in \{\mathbf{x}_i \in R^p : i=1,\ldots,n\}$, the main goal of a partitioned clustering is to determine an assignment $\mathbf{b}_i = \{b_{ij} \in \{0,1\} : j=1,\ldots,K\}, i=1,\ldots,n$, such that a given cost function is minimized, where $b_{ij}=1$ if observation $\mathbf{x}_i$ is assigned to the j-th partition, and $b_{ij}=0$ otherwise; K is the number of clusters (in this case, specified by the user). Mainly, the cost function is defined $C(\mathbf{b};K)$ as a weighted average, i.e., $C(\mathbf{b};K) = E\{\bar{d}(\mathbf{x};\mathbf{b}_i)\} = \sum_{i=1}^{n}\sum_{j=1}^{K}b_{ij}d(\mathbf{x}_i,\mathbf{m}_j)$, where $\mathbf{m}_j = E\{\mathbf{x};\mathbf{b}_i\} = \sum_{i=1}^{n}b_{ij}\mathbf{x}_i / \sum_{j=1}^{n}b_{ij}$ (named the center of the j-th cluster). Notation $d(\mathbf{x}_i,\mathbf{x}_j)$ stands for a distance metric between two observation vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. Because it is a reliable metric for early stages of training, the most commonly used distance is the Euclidean metric, $d_\varepsilon(\mathbf{x}_i,\mathbf{m}_j) = (\mathbf{x}_i - \mathbf{m}_j)^T(\mathbf{x}_i - \mathbf{m}_j)$. Though, other distance measures, such as Mahalanobis, can also be used in the clustering criterion to take care of hyper ellipsoidal-shaped clusters, and which is defined as $d_\mu(\mathbf{x}_i,\mathbf{m}_j) = (\mathbf{x}_i - \mathbf{m}_j)^T \sum_j^{-1}(\mathbf{x}_i - \mathbf{m}_j)$, where $\sum_j^{-1}$ is the inverse of the $p \times p$ covariance matrix of the observation set belonging to the jth cluster.

The intraclass distance, denoted as δ_i, is defined as overall statistical distance between the data points inside clusters $\mathbf{b}_i : \delta_i = \text{var}\{d(\mathbf{x}_j,\mathbf{m}_i),$ $\forall j,k=1,\ldots,n, i=1,\ldots,K.$, The overall clustering compactness measure is defined as

$$\delta = \max_{\forall i} \{\delta_i\} \qquad (4)$$

The interclass distance, denoted as $d(\mathbf{b}_i, \mathbf{b}_j)$, is the distance between the elements in cluster $\mathbf{b}_i$, and those in cluster $\mathbf{b}_j$: $d(\mathbf{b}_i, \mathbf{b}_j) = \min\{d(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \in \mathbf{b}_i, \mathbf{x}_j \in \mathbf{b}_j\}$. Then, overall clustering separation measure is defined as

$$\Delta = \min_{\forall i,j} \{d(\mathbf{b}_i, \mathbf{b}_j)\} \qquad (5)$$

Assuming $d(\mathbf{b}_i, \mathbf{b}_j)$ as Euclidean metric, value Δ represents the minimum Euclidean distance between clusters. Inspired by previous definitions, the following goodness-of-clustering measure J relating the cluster separability and cluster compactness can be used. Namely, the validity measure is defined as the ratio of the compactness measure δ to the separation measure Δ, i.e.,

$$J = \Delta/\delta \qquad (6)$$

It is expected that clusters should be as dense as possible and the distance between clusters should be as large as possible. Therefore, the compactness measure δ is expected to be small and the separation measure Δ to be large. A bigger J means a more compact and separate cluster configuration and, hence, a better validity measure for the clustering.

C.1 Estimate of clustering validity measure
The sample estimate of the covariance matrix for the i-th class is given by $\hat{\Sigma}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_j - m_i)(x_j - m_i)^T$, where x_j denotes the j-th recording, m_i the mean estimate, and n_i the number of patterns belonging to the i-th class.

The pooled estimate from all classes of the within matrix can be obtained as, $\mathbf{S}_w = \frac{n_i}{n} \sum_{j=1}^{C} \hat{\Sigma}_j$, being n the size of the whole sample, and C the number of classes. The between covariance matrix is computed as $\mathbf{S}_b = \frac{n_i}{n} \sum_{j=1}^{n_i} (m_j - m)(m_j - m)^T$, with m as the overall mean of the entire sample. Finally, the separability measure is derived from the following expression:

$$J = |\mathbf{S}_b + \mathbf{S}_w|/\mathbf{S}_w \qquad (7)$$

It is important to note that this criteria is similar to the multivariate fisher score, which in the two class case is given by the largest eigenvalue of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

III. EXPERIMENTAL SETUP

A. Database
The Massachusetts Eye and Ear Infirmary Voice Laboratory (MEEIVL). 48 features were computed.

These features correspond to 16 measures and its first and second derivatives. These measures are: 12 Mel Frequency Cepstrum Coefficients (MFCC) the Harmonics to Noise Ratio (HNR), the Glottal to Noise Excitation Ratio (GNE), the Normalized Noise Energy (NNE), and the Energy of the frame. A total of 48 features were taken in account [2].

B. Feature selection strategy
For each observation are taken j $(j = 1, 2,..., 48)$ dynamic features. These features were selected using the relevance measure presented in section II, then we use the Eq. (7) to estimate the cluster separability, this process generates a curve and each point in the curve is obtained by a incremental representations; that is, a larger set is obtained by adding features to the preceding one.

For feature selection we use the first k features, such that $J_k \leq \varepsilon \cdot \max(J)$, where $1 \leq k \leq 48$ and $0 < \varepsilon \leq 1$. In this case $\varepsilon = 0.99$.

The accuracy was measured using a *k-folds* cross validation strategy. In particular, 11 folds have been used, splitting the 70% of the files for training the classifier, and the remaining 30% for validating. These sets were randomly chosen.

III. RESULTS

As a result of the relevance analysis carried out above, a set of weights for the features was obtained. Fig. 1 shows the weights for each one of the 48 features.

The estimated values for relevance weight of dynamic features shown in Fig 1 are the starting point for selection and subsequent reduction of features; this analysis showed that the most significant features are the instantaneous measures, without the first and second derivatives, since the most weighted are the first 16.
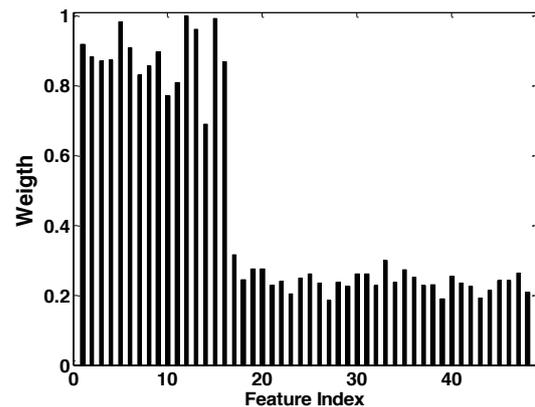
Fig. 1 Relevance of dynamic features.

Fig. 2 shows *Cluster separability vs. number of features*, the measure is done on the original variables. It is possible to observe how a gradual increment of the number of features (which were chosen according to Fig.

1) reflects on increasing separation between clusters. Nonetheless, this behavior holds up to certain number of features (the most relevant features).
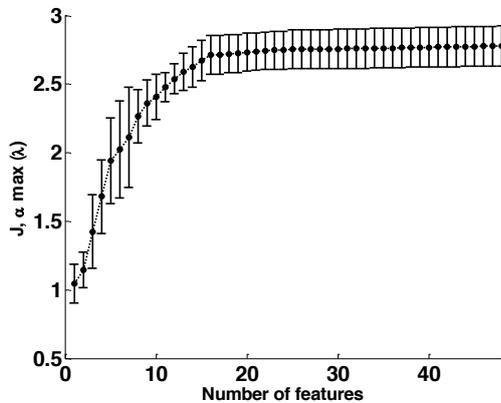
Fig. 2 Cluster separability vs. number of features.

[2] for books or book chapters [3]. Only items published or accepted for publication may be cited in the reference list.

The accuracy results are calculated employing continuous HMMs and the complete set of features. Several values of mixtures (NG=2, NG=3, and NG=4) and states (S=2, S=5, and S=10) were tested. The best results were obtained with NS=2, and NG=3. The results obtained are shown in the Table 1.

| NG | Number of states (NS) | | |
| | 3 | 5 | 10 |
	Accuracy	Accuracy	Accuracy
2	94.1±1.1	94.2±2.6	84.7±3.3
3	94.6±1.8	91.1±1.8	82.3±2.3
4	91.5±1.8	90.4±2.8	81.0±3.0

Table 1. Accuracy results using continuous HMMs

The accuracy results are recalculated employing NS=2, NG=3 and applying the feature selection strategy. In the Table 2 we compare these results using the accuracy, the area under the ROC curve (AUC), the number of features (NF) and time per iteration in the training phase.

Complete feature set			
Accuracy (%)	AUC	NF	Time (s)
94.6±1.8	0.9604±0.029	48	9.24
Feature selection strategy			
96.6±1.3	0.9758±0.01	19	5.75

Table 2. Accuracy results using a feature selection strategy

V. CONCLUSION

An approach combining PCA with HMM for pathological voice diagnosis has been tested based on cluster validity measure. For this purpose, a ratio of the compactness measure to the separation measure is carried out.

The proposed methodology for reducing the number of dynamic features in the identification of pathological voices proved to be useful for the experiments carried out. As a result was obtained an adequate performance while employing a considerably reduced feature set. The presented way of training shows that for the automatic detection of pathological voices is better to use a good set of features than a complex stochastic dynamic training model, because the later may have lower generalization capabilities.

REFERENCES

[1] P. Gómez, J. I. Godino, F. Rodríguez, F. Díaz, V. Nieto, A. Álvarez, V. Rodellar, 2004. Evidence of Vocal Cord Pathology From the Mucosal Wave Cepstral Contents. *Acoustics, Speech, and Signal Processing, vol 5, pp 437 – 440.*

[2] G. Daza-Santacoloma, J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Dynamic feature extraction: An application to voice pathology detection," Intelligent Automation and Soft Computing, vol. 15 (4), pp. 665–680, 2009.

[3] J. Lee, A. Lendasse, and M. Verleysen, "Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis," Neurocomputing, vol. 57, p. 49–76, 2004,

[4] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[5] L. Rankinea, M. Mesbaha and B. Boashash. 2007. IF estimation for multicomponent signals using image processing techniques in the time–frequency domain. Signal Processing, vol. 87, pp. 1234-1250.

[6] G. Stemmer, C. Hacker, E. Noth and H. Niemann. 2001. Multiple Time Resolutions for Derivative s of Mel-Frequency Cepstral Coefficients. Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on, pp. 37-40, 2002.

[7] C. Navin-Gupta, R. Palaniappan, S. Rajan, S. Swaminathan and S.M. Krishnan. 2005. Segmentation and Classification of Heart Sounds. CCECE/CCGEI, IEEE, pp. 1674-1677.

[8] M. Turk and A.Pentland. 1991. Eigenfaces for recognition. Cognitive Neuroscience, vol. 3, no. 1, pp.71-86.

[9] I. Jolliffe, 2002. Principal Components Analysis, Springer. 2nd edition

IDENTIFICATION OF FUNCTIONAL VOICE DISORDERS BY BIOMECHANICAL ANALYSIS

R. Fernández-Baillo, P. Gómez,
Laboratorio de Comunicación Oral, Universidad Politécnica de Madrid,
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
roberto@junipera.datsi.fi.upm.es

Abstract: Present work is focused in the study of the functional alterations classed as hypofunctional and hyperfunctional, with the aim to describe the dynamic of the vocal folds in each of these manners of phonation, which allows getting accurate data to its discrimination from the non-pathological voices. Preliminaries results were gotten using records from 20 subjects with non-pathological voice and 20 with functional pathology (10 hypofunctional and 10 hyperfunctional). The normal and functional alteration condition was based on the results obtained from the assessment, image and voice capture after a medical and acoustic study. The inclusion of the subjects with functional disorders inside their correct group was based on criterions related to the glottal closure and the vocal quality. The results show that the rate of amplitude between open/close and the starting point of the open phase are decisive. The data allow us to offer a new classification system of the functional voice disorders in which each group (hypofunctional and hiperfunctional) includes several subclasses, giving decisive information for the voice treatment.

Keywords: **Voice disorders, Hypofunctional, Hyperfunctional, Biomechanical Analysis, Glottal Source.**

I. INTRODUCTION

The study of the functional disorders is one of the bigger difficulties found by the clinical voice analysis. Normally functional dysphonia is defined how that in which there is not organic lesion in larynx but there is a voice alteration affecting one of principal components of the voice: pitch, intensity and/or timbre. It can find many classification systems of the voice pathologic and all of them are distinguishing four principal groups: hypofunctional, hyperfunctional, pubertal voice and psychological disorders. This work will be focus in the first two.
Hyperfunctional dysphonia is also named as *Muscular Tension Dysphonia* [1]. It is generally seen in younger to middle aged people with extensive voice use in stressful situations [1]. External features include visible and palpable muscular tension around the larynx. The tightness increases the pitch and is accompanied by an observable rise of the larynx in the neck. The voice acoustically is characterized by a pitch raised considering age and gender and monopitch, high energy formants, glottal attack, glottal fry, harshness, and a noise component lower.
Hypofunctional voice disorders are not so described and generally are referred in association with psychological factors like depression, distress, etc. Although, sometimes this could be true, it is important to consider hypofunction like a biomechanical alteration by itself. The voice is characterized acoustically by lower pitch, instability, weak formants and a noise component raised.

According to the described above is easy to see that hypofunctional and hyperfunctional disorders differ in the biomechanical pattern used to get the glottal closure. Thus, the study of the dynamic of the vocal folds and specifically the ratio between open and close phases is a useful parameter which will allow us to discriminate sort of functional disorders.

II. METHODOS

This work was carried out using records from. 30 female subjects (*Table 1*) between 25 to 45 years age extracted from de MAPACI database [2]. Subjects were classified after evaluation and discussion as non-dysphonic or dysphonic based on videostroboscopy evidence, acoustic analysis, and electro-glottographic trace inspection. Later the functional disorders were differentiated as hypofunctional and hyperfunctional basing in criterions related with glottal closure type and acoustical features of the voice.

The voice recording protocol included three utterances of vowel /a/ with duration not shorter than 3 sec. for each emission. Segments of 0.2 sec. were produced from the recording central parts for the analysis.
The extraction and analysis of the glottal source and mucosal wave correlate in this work was done by the software GLOTTEX® [2]. This tool provides you all the singular points that can be obtained from the glottal source profile and that allow you to study temporal phases.
The analysis of the glottal source in the time domain allows the evaluation of the normal or non-normal phonation conditions depending on the resulting profile. For such the following singular points during the open-close phases of a glottal excitation with period given by T

have to be determined as (*See Figure 1*): return interval ($T_r=t_r$), closed interval ($T_c=t_o-t_r$), open interval ($T_o=t_{cl}-t_o$) and closing interval ($T_{cl}=T-t_{cl}$) [3]. Also it allows extracted singular points related to the close-open phases. One of them is the point of the starting of the open phase (Px_o) [4]. This point is decisive to discriminate functional voice disorders.
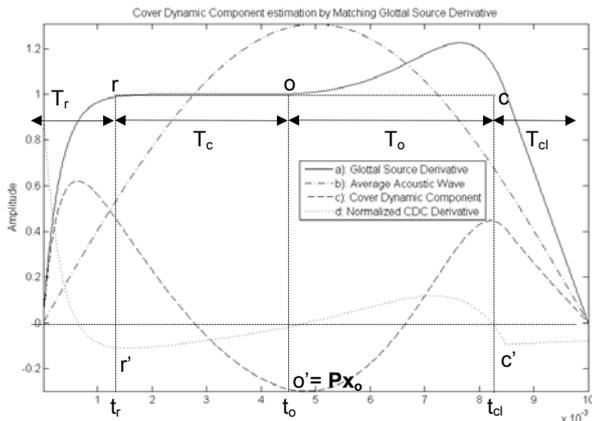
Figure 1. *Singular points in the opening-closing phases of a phonation cycle according to the L-F model.*

The beginning point of the open close (Px_0), also we can refer it how the ending point of the close phase, is decisive for the study of the functional disorders.

The graphic in *Figure 2* represented a distribution of the sample, normal and functional disorder voice, based on the value of the Px_0 (*See Table 2*). It can be observed how the traces belonging to normal voices are grouped close, setting a normality threshold.

Considering the normality threshold is possible to identify the pathological samples which will be those located outside. Generally, without organic lesion, in hyperfunctional voice the closure take place earlier. Thus all the subjects are located beneath the normal threshold. Hypofunctional, in same conditions, are place above the normal threshold, because its closure needs more time to be ending.

Besides, through the distribution graphic in *Figure 2,* it can be observed how inside a same pathologic group there are samples nearer to the threshold than other. This could be suggest that through the isolated analysis of functional group will be possible discriminated the grade of pathology.

S	R	Px_o	D	S	R	Px_o	D	S	R	Px_o	D
1	RegVoz 16	2,06	Hyper	11	RegVoz 374	2,4	Nor	21	RegVoz 11	5,17	Hypo
2	RegVoz 32	1,84	Hyper	12	RegVoz 365	2,14	Nor	22	RegVoz 58	4,15	Hypo
3	RegVoz 304	1,77	Hyper	13	RegVoz 332	2,29	Nor	23	RegVoz 70	4,69	Hypo
4	RegVoz 348	1,29	Hyper	14	RegVoz 322	2,31	Nor	24	RegVoz 137	2,86	Hypo
5	RegVoz 371	1,38	Hyper	15	RegVoz 288	2,59	Nor	25	RegVoz 142	4,08	Hypo
6	RegVoz 395	1,24	Hyper	16	RegVoz 286	2,43	Nor	26	RegVoz 214	3,24	Hypo
7	RegVoz 433	1,96	Hyper	17	RegVoz 180	2,54	Nor	27	RegVoz 230	2,74	Hypo
8	RegVoz 440	1,32	Hyper	18	RegVoz 47	2,27	Nor	28	RegVoz 252	3,61	Hypo
9	RegVoz 487	1,22	Hyper	19	RegVoz 40	2,22	Nor	29	RegVoz 275	3,38	Hypo
10	RegVoz 540	1,32	Hyper	20	RegVoz 521	2,45	Nor	30	RegVoz 528	2,81	Hypo

Table 1. *Distribution of the sample: S (Subject Number), R (Register Number), **Px_o** (Point of the starting of the Open Phase), D (Diagnosis).*

III. RESULTS

Normal Voice / Functional Voice Disorders. Functional voice disorders are characterized by an alteration of the manner in which the vocal folds get a glottal closure, without being organic lesion. The best way to know how this is happening is through the analysis of the open-close phases from a phonation cycle. The temporal analysis of the glottal source profile allows us getting several singular points which could be useful for the estimation of the open-closes phases [5].

Diagnosis	Means	SE
Hypofunctional	1,48	0,32
Normal Voice	2,35	0,14
Hyperfunctional	3,51	0,83

Table 2. *Means and Standard Error (SE) of the sample based on the value of Px_o*
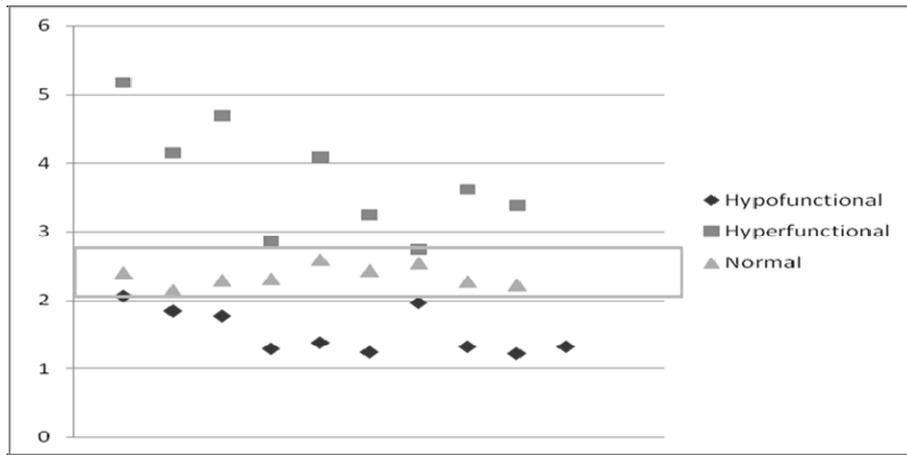
Figure2. *Distribution of normal voice and functional disorders according to the value of Px_o.*

Grade of pathology in functional alterations. The isolated study of each functional group allows to discriminate that all subjects have not the same grade of pathology, and therefore they should have an individually approach in the treatment. Now we make a new distribution considering the value of Px_o and only a functional group, the result will be a more accurate distribution. In this distribution distinguishes two grades for each group:

a) The functional disorders of *Grade I*. It could be considered as a particular mode of phonation, but not yet as pathology. Some people are characterized have a harsh or strained voice, others by a lightly breathiness or weakness voice, but neither of them are pathologic cases.

b) The functional disorders of *Grade II*. It could be considered as a type of pathology. Subjects with this sort of phonation usually derive to organic injuries. In these cases required an early rehabilitation treatment.

IV. CONCLUSION

In this work the results show a clear distinguishing between the most of the subjects in relation to the non-pathological or functional condition. The results show that the rate of amplitude between open-close and the starting point of the open phase are decisive. The data allows offer a new classification system of the functional voice disorders where each group of functional disorders (hypofunctional and hiperfunctional) including several subclass which are given decisive information for the voice treatment.
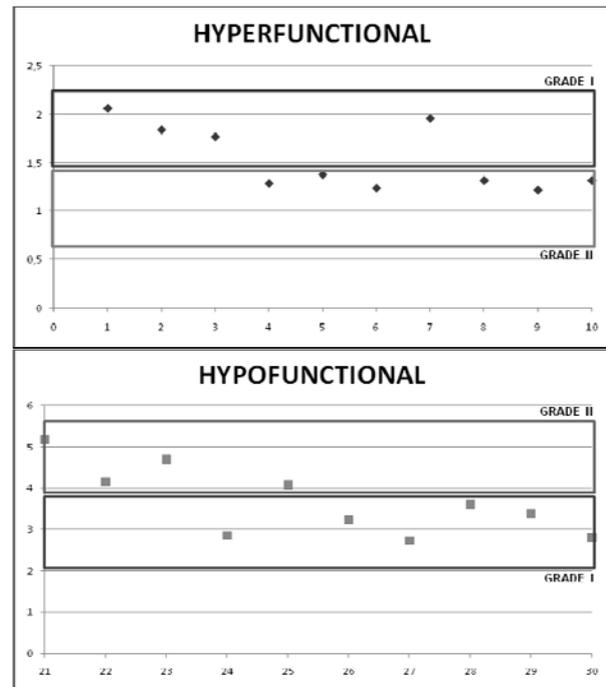
Figure 3. *Distribution of the functional disorders according to the value of Px_o for each group.*

REFERENCES

[1] Murray D. Morrison, MD. Diagnostic criteria in functional dysphonia. Laryngoscope 94: January 1986.
[2] Project MAPACI: http://www.mapaci.com
[3] Fant G., Liljentcrants J., Lin Q., "A four-parameter model of glottal flow", STL-QSPR, Vol. 4, 1985, pp 1-13. Reprinted in *Speech Acoustics and Phonetics: Selected Writings*, G. Fant, Kluwer Academic Publishers, Dordrecht 2004, pp. 95-108.

76

[4] Fernández-Baillo R., Gómez Vilda P. "Nuevo método para el estudio de la patología vocal basado en el perfil de la onda glótica y la estimación de parámetros biomecánicos". Proyecto Nets. Madrid. 2009.

[5] Fernández-Baillo R., Gómez Vilda P. "Clinical voice description based on the glottal source profile". The Voice Foundation's 37th Annual Symposium: Care of the Professional. Philadelphia. 2008.

ELECTROGLOTTOGRAPHY AND MICROPHONE SIGNALS ASSESSED BY APPROXIMATE ENTROPY IN NORMAL AND DYSPHONIC SUBJECTS

G. Sparacino[1], W. De Colle [2], D. De Luca[1], E. Arslan [2]

[1] Department of Information Engineering, University of Padova, Padova, Italy
[2] Department of Audiology and Phoniatrics, University of Padova and Ca' Foncello Regional Hospital, Treviso, Italy

Abstract: **Approximate Entropy is a method which provides a model independent nonlinear measure (the index ApEn) of the "regularity" of the process generating a time-series. In recent years, ApEn has been vigorously employed in the study of several biological signals, but only a few applications in the analysis of vocal disorders have been proposed. Here, we investigate the potential usefulness of ApEn in the study of electroglottography and microphone signals in normal and dysphonic subjects. Results show that statistically significant ApEn differences between the two groups can be found, more easily detectable in the microphone signal case.**

Keywords: **Chaos, time-series, signal processing, vocal disorders**

I. INTRODUCTION

Voice is the main vehicle of communication among human beings and its analysis is crucial for the differential diagnosis and follow-up of several pathologies. The sound signal (MIC), which can be picked up in a straightforward fashion by a microphone, brings information on several aspects related to voice generation, from vocal fold biomechanics to aerodynamic variables. In a slightly more sophisticated experimental setting, the so-called laryngograph allows the acquisition of the electroglottogram (EGG) which, by measuring the translarynx electrical impedance variation, permits the investigation of the vibration of the vocal folds.

In order to analyze MIC and EGG signals in both normal and pathological states, several tools have been proposed. In clinical practice, in particular, approaches which can be traced back to linear spectral analysis are most commonly used, with some classical parameters, such as jitter (fundamental frequency variation), and shimmer (amplitude variation), which evaluate perturbation contents. Recently, the use of these perturbation measures has been questioned. In particular, it has been suggested that linear approaches cannot reliably analyze strongly aperiodic signals and that jitter and shimmer are sensitive to several experimental and methodological settings [1]. In fact, spectral analysis does not handle cycles whose timing is inherently irregular, which can be a common situation in voice disorders, and cannot easily detect changes in the pattern of the signal which can characterize some patho-physiological states.

As a consequence, the use of nonlinear time series methods, such as correlation dimension and Lyapunov exponents, has been proposed for the study of vocal disorders [1, 2, 3]. Approaches based on the use of entropy measures have been also investigated [4]. These approaches are significantly appealing because they are able to condense the entire history of the signal into a single number, which can be relatively simple to interpret for clinical purposes.

Approximate Entropy is a method developed in the early nineties to provide a model independent measure of the "regularity" of the underlying secretion process by calculating the logarithmic likelihood that patterns in the time-series that are similar remain similar on the next incremental comparison [5]. Notably, such a notion of regularity is quite different from that usually considered in engineering, where, for a signal, regularity is meant as a synonymous of smoothness. The Approximate Entropy algorithm summarizes the time-series into a single nonnegative number, ApEn: the higher is the value of ApEn, the more irregular is the process. Approximate Entropy is not intended to replace more classic techniques such as spectral analysis, but is complementary to them. In fact, Approximate Entropy focuses on the similarity between patterns within the signal, thus relaxing the spectral analysis requirement of a dominant set of frequencies at which some patterns within the time-series are repeated.

In recent years, ApEn has been vigorously employed in the study of several biological signals, e.g. endocrine-metabolic time-series, electroencephalogram, heart rate variability, and found capable of successfully identifying pathological or pre-pathological states characterized by an enhanced signal irregularity. A few ApEn applications in the analysis of vocal disorders have been also proposed [6, 7, 8].

Here, we investigate the potential usefulness of ApEn in the study of MIC and EGG signals in normal and dysphonic subjects. The aim is to determine if statistically significant differences occur between the two groups and also to assess if such differences are more easily detectable in MIC or in EGG.

II. METHODS

Data Base. 60 subjects have been classified in two groups, normal (10 males and 10 females) and dysphonic (19 males and 21 females), according to the independent

perceptual evaluation of speech and language therapists. In all subjects, synchronous MIC and EGG recordings of the sustained Italian vowel /a/, kept at similar intensity and pitch for at least 4 seconds, were provided. Data recordings were made in a quiet room with the subject comfortably seated. The electroglottography system (Laryngograph Ltd, London, UK) employed a pair of electrodes attached on either side of the thyroid alae and held in place by a collar. The vocal signal was captured by a dynamic directional microphone (Prologue Shure, USA), placed at a constant distance of less than 5 cm from the mouth and at an angle of 45°. The MIC and EGG signals were acquired at 50 kHz, with 16 bits of amplitude resolution, by a commercial software (CSL 4300B, Kay Elemetrics, USA). In each subject, the middle, stationary appearing, segment of 1 s of data (correspondent to 50000 original samples), was considered for ApEn calculation.

The ApEn index. Briefly, let $\{u(k)\} = \{u(1), u(2), ..., u(N)\}$ denote the N-size time-series from which we want to calculate the ApEn index. Let r (a real) and m (an integer) be two given positive parameters. In order to compute ApEn, first form the sequence of vectors x(1) through x(N-m+1), where each x(i) is defined by $x(i)=[u(i), u(i+1), ..., u(i+m-1)]$. Vector x(i) contains m consecutive samples of the time-series $\{u(k)\}$, commencing with the i-th point. Having defined the distance d[x(i), x(j)] between vectors x(i) and x(j) as the maximum difference in their respective scalar components, compute, for each i≤N-m+1, the number $C_i^m(r)=\{$number of x(j) such that $d[x(i), x(j)]\leq r\}/(N-m+1)$. This values measures, within a tolerance r, the frequency, or regularity, of patterns similar to a given pattern of window of length m. Next, define $\Phi^m(r)$ as the average value of $\ln C_i^m(r)$. Finally, define the ApEn index as $ApEn= \Phi^m(r)- \Phi^{m+1}(r)$. It is possible to demonstrate that ApEn measures the logarithmic likelihood that runs of patterns that are close (within a tolerance r) for windows of m observations remain close for windows of m+1 observations. The greater the likelihood of remaining close (i.e. the regularity), the lower the value of ApEn.

III. RESULTS

Tuning of ApEn parameters. In order to speed up calculations, signals were downsampled at 10KHz. Starting from the recommendations of the author of the method, who suggested to determine m such that 10^m is of the order of the sampling points and r as a suitable value between 0.1 and 0.25 of the SD of the signal (depending on the signal-to-noise ratio), we have obtained the best values of m and r (m=4, and r=10% and 20% of the signal SD, respectively for MIC and EGG)

after retrospective analysis of the results arising from several trial values. Of note is that these parameters should be reassessed, should the original 50KHz sampling be considered.

ApEn outcome. No statistically significant ApEn differences have been found between males and females. Average values (± SD) of ApEn for the MIC signals are 0.2838 (±0.0418) and 0.4196 (±0.1662), for normal and dysphonic subjects, respectively. For the EGG signals, ApEn values are 0.1153 (±0.0525) and 0.3867 (±0.4643). Notably, ApEn in MIC signals is higher than in EGG signals, as it could be expected from the higher complexity of the system generating the sound signal. The MIC signals have ApEn values which, in both the groups, are significantly less dispersed (i.e. lower SD) than the EGG signals. Finally, even if ApEn differences between the two groups are statistically significant in both cases, the error probability using MIC (10^{-4}) is lower than using EGG (10^{-2}).

IV. DISCUSSION

ApEn is a simple, easy-to-implement, technique which, in the literature, was found useful in the nonlinear analysis of several biological signals. As stressed by the author of the method, ApEn is not intended to replace approaches resorting to spectral analysis, but is complementary to them. In fact, ApEn discerns changes in the signal behavior that are not reflected e.g. in changes in frequency and amplitudes of periodic components. Therefore, the application of ApEn to the study of vocal signals can deserve some consideration. Here, we have shown that, consistently with expectations, ApEn on EGG results smaller than ApEn on MIC, according to the fact that the laryngograph recordings are unaffected from the complicated mechanisms introduced by the vocal tract resonance, which converts the effects of fold vibrations into the sound delivered from the mouth. Also, our study shows that both EGG and MIC signals present average ApEn values which, in the dysphonic subjects, are higher than in the normal subjects. The difference is statistically significant with both signals, but seems more pronounced in the MIC case.

V. CONCLUSION

In this work, ApEn has been assessed of potential usefulness in the study of EGG and MIC signals in normal and dysphonic subjects. In particular, ApEn differences between the two groups are statistically significant and more easily detectable in the MIC case.

Although our analysis is preliminary and further studies are required to draw any conclusion on safe grounds, results suggest that the vocal tract plays a quantitatively important role in the alteration of vocal signals.

Further development of the present work should comprise the possible relationships between the optimal m and the signal sampling frequency (sampling higher than the 10KHz considered here could yield to a larger value of m) and the comparison of ApEn results with those of well consolidated linear approaches based on classic indexes such as jitter and shimmer. In fact, it is worthwhile reminding that ApEn does not replace spectral analysis techniques, but is complementary to them, as widely discussed by the author of the method in several papers e.g. in [9].

REFERENCES

[1] Zhang, Y, Wallace SM, and Jiang JJ, "Comparison of nonlinear dynamic methods and perturbation methods for voice analysis," *J Acoust Soc Am* 118: 2551–2560, 2005.

[2] Herzel H, Berry D, Titze IR, and Saleh M, "Analysis of vocal disorders with methods from nonlinear dynamics," *J Speech Hear Res* 37: 1008-1019, 1994.

[3] Matassini L, Hegger R, Kantz H, and Manfredi C, "Analysis of vocal disorders in a feature space," *Med Eng Phys* 22: 413-418, 2000.

[4] Scalassara PR, Dajer ME, Maciel CD, Capobianco GR, and Pereira JC, "Relative entropy measures applied to healthy and pathological voice characterization," *Applied Mathematics and Computation* 207: 95-108, 2009.

[5] Pincus SM, "Approximate entropy as a measure of system complexity," *Proc Natl Acad Sci* USA 88: 2297-2301, 1991.

[6] Moore C, Manickam K, Willard T, Jones S, Slevin N, and Shalet S, "Spectral pattern complexity analysis and the quantification of voice normality in healthy and radiotherapy patient groups," *Med Eng Phys* 26: 291-301, 2004.

[7] Moore C, Shalet S, Manickam K, Willard T, Maheshwari H, and Baumann G, "Voice abnormality in adults with congenital and adult-acquired growth hormone deficiency," *J Clin Endocrinol Metab* 90: 4128-4132, 2005.

[8] Aghazadeh BS, Khadivi H, and Nikkhah-Bahrami M, "Nonlinear analysis and classification of vocal disorders," in *Conf Proc IEEE Eng Med Biol Soc 2007* : 6200-6203, 2007.

[9] Pincus, SM, "Assessing serial irregularity and its implications for health," Ann NY Acad Sci. 954:245-67, 2001.

Special Session on Voice Modeling I

Chairperson and Introduction:
H. Kawahara, Japan

Voice Modeling II

SPEECH MORPHING BASED ON
BIOLOGICALLY RELEVANT SIGNAL REPRESENTATIONS

Hideki Kawahara

Auditory Media Laboratory, Faculty of Systems Engineering, Wakayama University, Wakayama, Japan

Abstract: **Voice morphing based on a high fidelity VOCODER is a unique strategy to explore attributes which are closely related to biological states of speakers. The method is based on a temporally stable power spectral representation and spectral envelope recovery based on a new formulation of the sampling theory. The morphing algorithm itself is re-formulated to enable extrapolation without introducing perceptual and objective breakdown. It also extended to make temporally-variable multi-aspect morphing possible. GUI (graphical user interface) based tools are implemented to handle complexities introduced by these extensions. For characterizing voicing, a bottom-up local repetition detector, a residual-based irregularity detector and a group delay-based acoustic event detector with multi-resolution analysis are prepared.**

Keywords:— **Spectrum, periodicity, speech perception, voicing, morphing**

I. INTRODUCTION

Repetitive structures [1] play important roles in biological systems from animal calls to voiced sounds in human speech. However, usual short term Fourier based analysis methods including cepstrum and LPC analyses, suffer from interferences caused by this repetitive structure. Recently, a simple method for calculating interference-free power spectra [2] and new formulation of sampling theory [3] led to an invention of a speech analysis, modification and synthesis procedure called TANDEM-STRAIGHT [4]. TANDEM-STRAIGHT consists of a new bottom-up procedure to scoop all local repetitive structure based on this power spectral representation. TANDEM-STRAIGHTe was also applied to extend speech morphing procedure [5] and yielded a temporally variable multi-aspect morphing procedure [6]. These new set of procedures are integrated with visualization and GUI tools [7] to provide a strong basis for investigating biomedical aspects of voice emission and perception.

II. POWER SPECTRUM OF PERIODIC SIGNALS

Assume that a repetitive signal $x(t)$ has a fundamental period T_0 and its short term Fourier transform $S(\omega, t)$ is calculated using a time window $w(t)$. With a mild conditions on the window function $w(t)$, the following power spectrum $P_T(\omega, t)$ does not have temporal variations due to periodicity (repetition).

$$P_T(\omega, t) = |S(\omega, t - T_0/4)|^2 + |S(\omega, t + T_0/4)|^2. \quad (1)$$

$P_T(\omega, t)$ is called TANDEM spectrum afterwards. This operation does not have impact on frequency resolution of the original time windowing. Typical selection of the time window is a Blackman window having its duration set $2.5T_0$.

A. Periodic variations in the frequency domain

Variations of $\log(P_T(\omega, t))$ in the frequency domain is closely approximated by an additive sinusoid with a period $\omega_0 = 2\pi f_0$, where f_0 represents fundamental frequency (F0). It is completely eliminated applying a frequency domain smoother that has zeros at $n\omega_0$ on its spatial frequency transfer function. The simplest one is a rectangular smoother with ω_0 for its width. Let this smoother $h_1(\omega; \omega_0)$. Reasonable windows convolved with this smoother also have zeros at the same place and can be used for the same purpose. Let's call this F0 adaptively smoothes power spectrum "smoothed spectrum" $P_S(\omega, t)$.

B. Consistent sampling

Unfortunately, $P_S(\omega, t)$ does not precisely agree with Fourier transform of the (hypothetical) unit waveform that is repeated. There are two sources of smearing for the unit waveform. One is frequency response of the time window and the other is the smoothing function. Consistent sampling provides a way to solve this problem. A correction digital filter $Q(z)$ in the frequency domain can be designed using convolution of $h_1(z)$ and $W(z)$, where they are $h_1(\omega; \omega_0)$ and $w(t)$ represented in terms of z transform.

$$Q(z) = \frac{1}{a(z)}, \quad (2)$$

where $a(z)$ is the convolution of $h_1(z)$ and $W(z)$. Please note that the polynomial $Q(z)$ has infinite number of coefficients, because $a(z)$ is effectively a function with a finite support. Approximation errors between harmonic frequencies is dependent on the effective interpolating function that is convolution of $h_1(z)$ and $W(z)$, this case. In other words, there is a room for improvement in designing the smoothing function, actually a triangular function that is convolution of h_1 and h_1 is a better smoother for speech sounds.

C. STRAIGHT spectrum

Taking into account of the fact that $\log(1 + x) \approx x$ when $|x| \ll 1$ and absolute value of coefficients of $Q(z)$ decreases very rapidly, smoothed spectrum that preserves values at harmonic frequencies $P_{TST}(\omega, t)$ is calculated using the following equation.

$$P_{TST}(\omega) = e^{(\tilde{q}_1(L(\omega - \omega_0) + L(\omega + \omega_0)) + \tilde{q}_0 L(\omega))}, \quad (3)$$

where $L(\omega) \equiv \log(P_S(\omega))$ and $\tilde{q}_0, \tilde{q}_1$ are truncated and adjusted version of the coefficients of $Q(z)$. Please note that variable t is not represented in this equation to make appearance simple. Afterwards, whenever not confusing, the same practice applies. This spectrum does not have trace of periodicity while it preserves spectral values at harmonic frequencies. This spectrum is called STRAIGHT spectrum, because it is virtually identical to the spectrum calculated using the legacy-STRAIGHT [1].

III. LOCAL PERIODICITY DETECTOR

Since $P_S(\omega)$ is the periodicity eliminated version of $P_T(\omega)$ and the effect of periodicity is multiplicative, dividing $P_T(\omega)$ by $P_S(\omega)$ leaves a constant c_0 and the periodic component $P_C(\omega)$.

$$P_C(\omega) = \frac{P_T(\omega)}{P_S(\omega)} - c_0. \quad (4)$$

Ideally, Fourier transform of $P_C(\omega)$ has a unique peak at $\tau = T_0$. However, in practice, low S/N in lower frequency region and FM side-bands in F0 varying speech (that is usually the case) direct application of Fourier transform on $P_C(\omega)$ yields erroneous and noisy results.

To investigate vocal fold vibration, it is better to select the base-band frequency region using a frequency domain weighting function. The frequency weighting function can be located anywhere depending on aspects to be investigated, for example, to investigate regularity of glottal closure instant, the function can be centered around 3 kHz. The weighting function for selecting base-band

region $w_{\omega 0, N}(\omega)$ has the following form and defined in $[-N\omega_0, N\omega_0]$.

$$w_{\omega 0, N}(\omega) = c_1 \left(1 + \cos\left(\pi\omega/N\omega_0\right)\right), \quad (5)$$

where c_1 is a normalization constant. Then, Fourier transform of the windowed version of the periodic component has a less noisy peak at $\tau = T_0$.

$$A(\tau; T_0) = \int_{-\infty}^{\infty} w_{\omega 0, N}(\omega) P_C(\omega) e^{-j\omega\tau} d\omega, \quad (6)$$

where the assumed period is explicitly denoted in , $A(\tau; T_0)$. Increasing N sharpens the peak and makes it tolerant to background noise while makes it susceptive to FM and AM meaning that there is a trade-off relation.

The designed detector is specialized to the assumed T_0. By assuming periods $T_{0k}, (k = 1, \ldots, M)$ systematically on the logarithmic lag axis, they are combined to cover periodicity range of interest using the following equations.

$$\bar{A}(\tau) = c_2 \sum_{k=0}^{M} w_{LAG}(\tau; T_{0k}) A\left(\tau; T_{0k}\right), \quad (7)$$

$$w_{LAG}(\tau; T_{0k}) = 1 + \cos\left(\pi \log_2(\beta\tau/T_{0k})\right), \quad (8)$$

where a constant c_2 is adjusted for $\bar{A}(\tau)$ to have a value 1 for periodic signals. The weighting function $w_{LAG}(\tau; T_{0k})$ defined in $[-T_{0k} < \beta\tau < T_{0k}]$ is use to suppress spurious peaks in $A(\tau; T_{0k})$ by adjusting selectivity using β. Typical selection of T_{0k} has the following form.

$$T_{0k} = T_L 2^{-k/N_c} \quad (9)$$

where N_c determines the number of specialized detectors in one octave and T_L represents the longest period to be investigated.

Peaks of periodicity measure $\bar{A}(\tau)$ represent local repetitions of waveform using the best time-frequency resolution in each period scale. It is a bottom-up exhaustive periodicity detection system to be used to characterize repetitive structures in speech. Therefore, this method is called XSX (eXcitation Structure eXtractor) [8].

IV. APERIODICITY REPRESENTATION

The XSX is able to extract several types of aperiodicity such as jitter and shimmer as timing fluctuations and amplitude fluctuations of each excitation event respectively. However, there still remains other types of deviations from precise repetition. Linear prediction residuals from around a repetition period apart (both forward and backward), calculated on two types of time axes, are

used to represent aperiodic component that cannot be represented by XSX. Two types of time axes are as follows. The first one is the usual time axis. The second one is a warped time axis that is stretched in proportion to its instantaneous frequency corresponding to the repetition period. This selection of the second time axis makes apparent repetition period constant. The smaller residual of these two predictions is used as an index to represent aperiodicity in each time-frequency band region. Octave division of frequency band is used in the current implementation with keeping the narrowest bandwidth wider than 500 Hz.

For diagnostic applications, an acoustic event and group delay based representation [9] is also used. However, it still is one of the future topics to integrate this event based representation into TANDEM-STRAIGHT and morphing system.

V. TIME-VARIABLE MULTI-ASPECT MORPHING

Speech morphing was originally designed [5] based on linear interpolation and extrapolation of parameter values. This definition was found fragile when parameters are extrapolated. A new definition that enables time-variable multi-aspect morphing was proposed by re-defining morphing, based on linear interpolation in the logarithm of derivative domain [6].

Using this formulation, let $T_{Am}(x_A)$ represent a morphing transformation of a parameter x_A of example A to parameter x_m on the morphing axis m. A temporally variable morphing rate for the parameter $r_{AB}(t)$ is defined to have the value 0 when the morphed result is equivalent to example A and to have the value 1 when the morphed result is equivalent to example B. A new morphing definition is introduced and described using this notation.

To alleviate breakdown in explorative morphing, morphing is redefined based on a logarithm of the derivative of mapping functions. This new definition of morphing also makes the morphing procedure simpler as follows:

$$
\begin{aligned}
T_{Am}(x_A) &= \int_0^{x_A} \exp\left(\log\left(\frac{dT_{Am}(\lambda)}{d\lambda}\right)\right) d\lambda \\
&= \int_0^{x_A} \exp\left((1 - r_{AB}(\lambda))\log\left(\frac{dT_{AA}(\lambda)}{d\lambda}\right) \right. \\
&\quad \left. + r_{AB}(\lambda)\log\left(\frac{dT_{AB}(\lambda)}{d\lambda}\right)\right) d\lambda \\
&= \int_0^{x_A} \left(\frac{dT_{AB}(\lambda)}{d\lambda}\right)^{r_{AB}(\lambda)} d\lambda \,, \quad (10)
\end{aligned}
$$

because logarithmic conversion of the identity mapping vanishes. This formulation assures monotonicity of T_{Am}
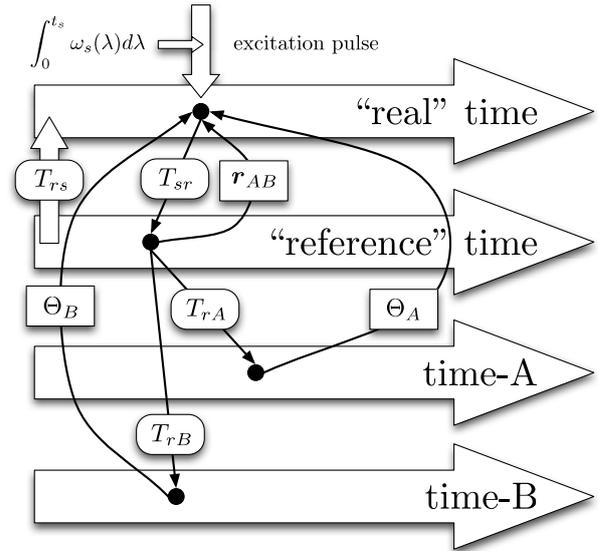
Fig. 1. Morphing procedure with a "reference" time axis for defining temporally variable morphing rates. Θ_A and Θ_B represent grouped STRAIGHT parameters of examples A and B respectively. r_{AB} represents multi-aspect morphing rates. $\omega_s(t)$ represents morphed F0 represented by angular instantaneous frequency.

if the coordinate conversion T_{AB} from speaker A to B is monotonic.

Two morphing algorithms are formulated based on this new definition of morphing: real-time morphing and off-line morphing. In the case of real-time morphing, the morphing rates are incrementally supplied and used to update morphed parameters incrementally. This formulation is useful for interactive applications.

In the case of off-line morphing, the morphed time axis, which is also the time axis for the morphed signal, is calculated for the first time. Then, other morphed parameters are calculated using the morphing rate on this new reference axis. This formulation is necessary for psychophysical stimuli preparation and biomedical diagnostic applications. Fig. 1 illustrates the synthesis procedure using these parameters and transformations.

VI. GUI TOOLS

GUI tools are equipped with analysis tools and visualization interface. Fig. 2 shows GUI for F0 extraction. The default F0 extractor is XSX. In other words, it is not a mere F0 extractor. It is a visualization tool for excitation
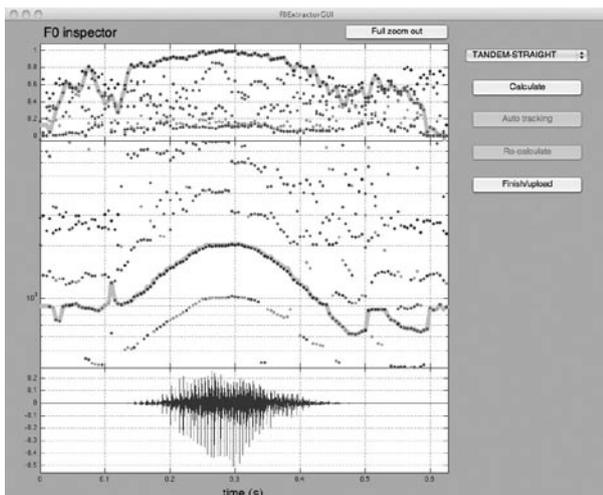
Fig. 2. GUI tool for XSX analysis. Top panel shows periodicity index for each repetition structure. The middle panel shows extracted local repetition structure in terms of frequency. The bottom panel shows the waveform. The sample is /hai/ (Yes, in English) spoken with strong anger. (In originally figure, marks are color coded in order of repetition salience.)

structure analyses. [1]

The middle plot of the display shows localized repetitions in terms of frequency. This plot can be zoomed both in time and frequency and can be dragged to center interesting regions. The thick gray line (cyan in color) represents the most salient periodicity that corresponds to F0. Application of XSX to Noh voice analyses illustrated interesting subharmonic structure and non-classical transition of the most salient periodicity [8].

VII. DISCUSSION

The GUI tools and underlying algorithms is designed to promote exploratory research strategy for investigating phenomena they are difficult to be categorized *a-priori*. By using morphing to generate a set of stimulus continuum combined with after effect, auditory adaptation in voice perception was discovered [10]. In other words, a stimulus continuum generated using morphing provides means to objectively quantify non- or pre-categorical percept and phenomena. One prospective example is evaluation of healing process from vocal fold surgery.

---

[1]This visualization tool is user customizable. By adding a line for the F0 extractor description to an extractor menu definition table and prepare an interface function using the template file, the user created extractor can be integrated into this GUI tool.

VIII. CONCLUSION

A temporally-variable multi-aspect morphing method based on a temporally stable representation of periodic signals combined with a bottom-up repetitive structure extractor and a residual-based aperiodicity extractor are introduced. This algorithm and a set of dedicated GUI tools provide a strong basis for exploratory research on biomedical aspects of voice emission and perception.

REFERENCES

[1] Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigné. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, 27(3–4), pp.187–207, 1999.

[2] M. Morise, T. Takahashi, H. Kawahara, and T. Irino, "Power spectrum estimation method for periodic signals virtually irrespective to time window position," *Trans. IEICE*, J90-D(12), pp.3265–3267, 2007, [in Japanese].

[3] M. Unser, "Sampling – 50 years after Shannon," *Proceedings of the IEEE*, 88(4), pp. 569–587, 2000.

[4] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, Hideki Banno, "TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation," *Proc. ICASSP 2008*, Las Vegas, 2008, pp.3933–3936.

[5] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," *Proc. ICASSP 2003*, I, Hong Kong, 2003, pp.256–259.

[6] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP 2009*, Taipei, 2009, pp.3905–3908.

[7] H. Kawahara, T. Takahashi, M. Morise, H. Banno, "Development of exploratory research tools based on TANDEM-STRAIGHT," *Proc. APSIPA 2009*, Sapporo, 2009.

[8] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu and M. Morise, "Noh Voice Quality", *J. Logopedics Phoniatrics Vocology*, 04 June 2009. (doi:10.1080/14015430903002288)

[9] H. Kawahara, Jo Estill and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA 2001*, Firentze Italy, 13-15 Sept, 2001.

[10] S. R. Schweinberger, C. Casper, N. Hauthal, J. M. Kaufmann, H. Kawahara, N. Kloth, and D. M. Robertson, "Auditory adaptation in voice perception," *Current Biology*, 18, pp.684–688, 2008.

TRACKING FORMANTS, EXTRA-FORMANTS AND ANTI-FORMANTS IN NON-MODAL SPEECH BY MEANS OF A SPECTRAL POLE-ZERO MODEL

J. Schoentgen[1,2], F. Grenez[1]

[1]Dpt of Image and Signal Processing & Telecommunication Devices, Université Libre de Bruxelles, Brussels, Belgium
[2]National Fund for Scientific Research, Belgium

Abstract: **The presentation concerns a preliminary investigation of a spectral pole-zero model that is fitted directly to observed log-magnitude spectra. The parameters of pole-zero models are interpretable in terms of (anti-) formant frequencies and bandwidths that may thus be tracked over time. The speech corpus has comprised connected speech tokens with prominent formant/anti-formant pairs owing to hyper-nasality in many speech frames. Results show that the direct fitting of spectral models is feasible. The quality of fit of the spectral contour by a model transfer function is comparable to the quality of fit obtained via cepstral smoothing with an effective number of cepstral coefficients equal to the number of independent model parameters.**

Keywords: **Spectral pole-zero models, formant and anti-formant tracking, hyper-nasality.**

I. INTRODUCTION

Formants are the spectral effects of the resonances of the vocal tract. The acoustic description of a majority of speech sounds rests on their formant frequencies. They therefore play a central role in many models of speech production or perception in psychology, phonetics and phonology as well as clinical phonetics or singing.

However, measuring formants reliably and automatically remains an unsolved problem at present. Possible reasons are that adjacent formants may fuse into a single spectral hump or relevant formants may be unobservable because the excitation signal lacks energy in that frequency band or because they are masked by neighboring anti-formants.

Alternatively, extra formants that together with anti-formants are the effects of side-cavities of the main vocal tract may appear. Extra formants may be phonetically relevant or not. Nasal formants and anti-formants mark the distinction between nasal and oral vowels in French, for instance.

Automatic formant extractors rely either on peak picking or linear models that involve poles and zeros as parameters. Peak picking identifies and marks as formant candidates prominent spectral humps in smoothed spectra. Pole-zero models are fitted to recorded speech signal frames. Their parameters, which are complex pole or zero frequencies and radii, are turned into candidate formant or anti-formant frequencies and bandwidths. Generally speaking, raw formant or anti-formant candidates must be post-processed to smooth their trajectories and remove unlikely candidates or insert missing ones.

By far the most popular model is the so-called linear predictive model, which is an all-pole model [1]. Its popularity rests on the linearity and simplicity of the method that fits the model to observed signals. Other models exist that involve poles as well as zeros [2]. They are therefore able to discover formants as well as anti-formants. They are used less often because they are more difficult to implement and the additional complexity is not always offset by increased reliability and validity of the formant and anti-formant candidates.

Known problems with formant extractors that rely on linear pole-zero models are the following [1]. (a) The user must fix the number of poles and zeros a priori. (b) Most often, the distinction between real and complex-conjugate parameters is beyond the control of the user because it is made automatically (and often erroneously) by the method that fits the model to the speech frame. Real poles and zeros mimic the spectral slope that is the effect of the acoustic source. (c) Often, models cannot be mathematically guaranteed to be stable. That is, the fitted model may comprise resonances or anti-resonances with negative bandwidths, which are physically impossible. (d) Often, degrees of freedom of pole-zero models are miss-used to represent the "glottal" formant as a genuine formant. The glottal "formant" is a prominent hump in the interval from 0 to 200 Hz, which is an effect of the glottal source and that is observed in many speech spectra. (e) In high-pitched voices, harmonics are far apart and models may identify individual harmonics as formant candidates when the pole bandwidths are allowed to be arbitrarily small.

This presentation reports the development and test of a spectral pole-zero model the fit of which avoids problems (b) to (e) listed above. It is directly fitted to the log-magnitude spectrum. The adjustment of the parameters of the model rests on an all-purpose optimizer that enables taking hard constraints into account.

The model is tested on a corpus of connected speech produced by a male French speaker [3]. Many speech frames of this speaker display prominent formant/anti-formant pairs (in vowel quality [a] mostly). They have been assumed to be a consequence of hyper-nasality owing to allergy ("hay fever"). In addition, the speaker's

voice has been high-pitched (> 300 Hz) and breathy in several frames of each speech token. These properties make that the corpus is suitable for testing the ability of a spectral pole-zero model to fit accurately the contour of non-modal speech frame spectra.

II. METHODS

A. Corpus

The corpus has comprised V_1V_2 and V_1CV_3 tokens, with V_1 and V_3 designating French vowels [a][i] and [u], V_2 designating French vowels [a][i][u][e][ɛ][o][ɔ] and C French consonants [p][t][k] and [f][s][ʃ]. The total number of tokens has been equal to 79, which corresponds to a total of 3183 analysis frames. Before analysis, the tokens have been down-sampled to 8 kHz.

B. Model

The spectral model (1) is a conventional pole-zero model involving M real poles, N complex-conjugate poles, K real zeros and L complex-conjugate zeros.

$$H(z) = \frac{\prod_{i=1}^{K}(z-r_i)\prod_{j=1}^{L}(z-z_j)(z-z_j^*)}{\prod_{m=1}^{M}(z-r_m)\prod_{n=1}^{N}(z-z_n)(z-z_n^*)} \tag{1}$$

Symbols z_k and r_k are the complex and real roots that are adjusted. The transfer function of (1) is calculated by replacing complex number z by $e^{i\theta}$ and varying Θ between 0 and $+\pi$.

C. Analysis

The length of the analysis frame has been equal to 25 ms and the frame hop has been half a frame. Each frame has been multiplied by a Hamming window. The spectrum has been obtained via a conventional discrete Fourier transform [4]. Model (1) has been fitted to the log-magnitude spectrum of each analysis frame.

In a first experiment, the number of complex-conjugate pole pairs of model (1) has been fixed to 5, the number of real poles to 2 and the number of complex-conjugate zero pairs to 1. This choice has been based on the observation that the signal bandwidth has been 4 kHz and the log-magnitude spectra of many speech fragments of the corpus have been characterized by one extra formant and one anti-formant. The total number of independent model parameters has therefore been equal to 14 (two per complex-conjugate pole or zero pair and one per real pole).

To test the validity of the fitted model, cepstrally-estimated spectral contours have been used as a reference to which the model transfer functions have been compared. The cepstra have been obtained by means of a direct Fourier transform of the log-magnitude spectrum, followed by zeroing of the high-quefrency cepstral coefficients and an inverse Fourier transform [5].

The low-quefrency cepstral coefficients have been windowed by means of a half-Hamming window before inverse transforming. The purpose has been the removal of ripples in the cepstrally-estimated spectral contour owing to the zeroing of finite-sized cepstral coefficients. Slow contour ripples may indeed be mistaken for extra formants.

The number of cepstral coefficients has been fixed to 28, that is, twice the number of independent parameters of model (1). The aim has been to equate the effective number of cepstral coefficients and the number of model parameters. Indeed, the number of cepstral coefficients is assumed to be effectively halved given that the non-zeroed cepstral coefficients are windowed and the cepstral coefficients near the edge of the window do not contribute much to the spectral contour.

In a second experiment, the validity of model (1) has been tested indirectly by fitting an all-pole model (in place of the pole-zero model) with the same number of parameters. The all-pole model is expected to fit the log-magnitude spectra less well than the pole-zero model, given the observation that for that corpus many speech frames display at least one prominent anti-formant.

D. Fitting

The optimizer has been the differential evolution algorithm, which is a population-based, stochastic function minimizer [6]. Differential evolution handles floating-point variables directly, which here are the model parameters. It does not request that they are encoded binarily. The optimizer involves three parameters that must be fixed by the user and which are application-dependent. These parameters are the size of the population, the mutation factor and the recombination rate. The size of the population of candidate solutions has been fixed to 15 times the number of independent model parameters. The mutation factor has been fixed to 0.5 and the recombination rate to 1.

E. Cost function

To fit model (1), the optimizer has decreased the value of a cost function, which has been the total sum of squares of the sample-by-sample differences between the target log-magnitude spectrum and the log-magnitude transfer function. The averages of the log-magnitude spectrum and log-magnitude transfer function have been offset to 0.

The optimization stopped when one of the following criteria had been met: a) the smallest cost function value

in the population had decreased below a threshold; b) the standard deviation of the cost function values of the population was below a threshold so that any further evolution was unlikely; or c) the number of generations had been larger than 7000.

F. Hard constraints

At each generation, all individuals (i.e. arrays of model parameters) of the population of evolving candidate solutions have been tested whether they complied with a set of hard constraints. Solutions that did not comply have been discarded and replaced by new random candidate solutions that did. The constraints have been the following. Pole and zero relative frequencies have to be comprised between 0.02 and 0.5 and pole and zero radii between 0.0 and 0.96. These constraints guarantee stable solutions. They also strongly favor solutions that do not cling tightly to individual harmonics and do not match the glottal "formant".

G. Testing

The validity of model (1) and the ability of the differential evolution algorithm to fit the model to log-magnitude spectra have been tested by computing a relative difference (2) for each frame.

$$100 \times \sqrt{\frac{\sum_i (target_i - contour_i)^2}{\sum_i target_i^2}} (\%) \qquad (2)$$

The *target* has been the observed log-magnitude spectrum, including noise and harmonics. The *contour* has been the transfer function of model (1) or the spectral contour estimated via cepstral smoothing.

The values of relative difference (2) cannot be interpreted as modeling errors per se because of the harmonics and noise in the spectrum that are not taken into account by the spectral contours. The quality of fit of model (1) has therefore been judged indirectly by comparing differences (2) when the contour has been obtained cepstrally or by fitting log-magnitude transfer function (1). When both differences are similar, one may conclude that model (1) and truncated cepstrum recover the spectral contours with equal fidelity.

The results are reported for all analysis frames including or excluding silent frames that occur during tract closure or precede and follow speech onset and offset. In practice, silent frames have been weak and noisy. They have been removed by computing the root mean square amplitude of each frame and discarding those frames the effective amplitudes of which have been smaller than 10 % of the maximum frame amplitude of each token.

The linear and rank correlations between fitted log-magnitude transfer functions and cepstrally-estimated log-magnitude spectral contours are also reported.

III. RESULTS

Experiment 1

Figures 1 and 2 illustrate the cepstrally-estimated and modeled contours for fragments [a] and [i] in token [ai].

Tables 1 and 2 report the quartiles of quantities informing about the ability of transfer function model (1) to represent the contours of log-magnitude spectra. Table 1 reports these quantities for all frames, including silent and weak ones. Table 2 reports the same quantities when weak and silent frames are removed.

The quantities reported are relative differences (2) between the log-magnitude spectrum and the model transfer function (second column) or the cepstrally-estimated contour (third column). The linear and rank correlations of the fitted transfer function with the cepstrally-estimated spectral contour are reported in the fourth and fifth columns.

Table 1 shows that relative differences (2) in percent are similar for the fitted pole-zero model and the cepstrally-estimated spectral contour. The latter appears to match the log-magnitude spectrum slightly better. The difference is 1 – 2 % in favor of the cepstrally-smoothed contour. This is expected because half-windowing the cepstral coefficients does not really half their number and the cepstrally-obtained contour is free to match spectral features such as glottal "formants". Matching glottal "formants" is forbidden to model (1) because the lowest relative pole frequency must be ≥ 0.02. As an example, the log-magnitude spectrum in Figure 1 displays a prominent glottal formant that is hugged by the cepstrally-estimated contour, but ignored by the log-transfer function of model (1).

The quartiles of linear and rank correlations suggest that cepstrally-estimated and modeled log-contours are similar in shape. The averages and standard deviations are indeed 0.97±0.02 and 0.96±0.03 for the linear and rank correlations respectively.

Finally, comparing Tables 1 and 2 shows that discarding frames that are weak and noisy, which may be the case for roughly one third of the total number of frames, enables decreasing differences (2) by one percent for the modeled and cepstrally-estimated contours.

Experiment 2

Table 3 shows that, as expected, replacing the pole-zero model by an all-pole model with the same number of model parameters increases differences (2) between the log-magnitude spectrum and the log-magnitude transfer function (column 1). The explanation is that the all-pole model is not able to match prominent spectral zeros. A visual examination of modeled contours suggests that the lack of a modeled spectral zero also precludes fitting the

extra-formant adjacent to the anti-formant because the abrupt transition from spectral peak to spectral through cannot be tracked by means of an all-pole model even when excess poles are available.

But, the increase of difference (2) is small, i.e. approximately 1 %. The explanation is that only a fraction of the speech sounds displays prominent anti-formants. In these sounds most of the formants are modeled correctly, which keeps the overall error small even when anti-formants and extra-formants are not matched accurately.

	Target vs Model (%)	Target vs Smoothed Contour (%)	Lin. corr.	Rank corr.
Min	25	23	0.60	0.55
1. Quartile	49	48	0.97	0.95
Median	55	54	0.98	0.97
3. Quartile	62	60	0.99	0.98
Max	95	94	1.00	1.00

Table 1: Differences (2) between log-magnitude spectra and pole-zero modeled and cepstrally-estimated contours; linear and rank correlations between model log-magnitude transfer functions (1) and cepstrally-estimated spectral contours for all analysis frames.
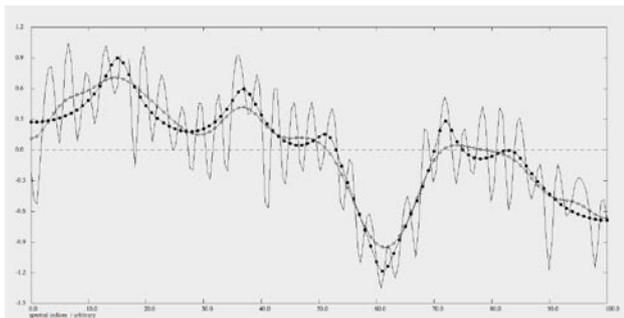
Figure 1: Log-magnitude spectrum of frame 10 ([a]) in [ai]. Overlaid are the cepstrally-estimated spectral contour (white circles) and the fitted transfer function of model (1) (black circles). The horizontal axis is the frequency ranging from 0 to 4 kHz in number of samples. The model-estimated formant frequencies for that frame have been equal to 647 Hz, 1437 Hz, 2059 Hz, 2900 Hz and 3305 Hz. The anti-formant frequency has been equal to 2497 Hz. The anti-formant is visible at that frequency as a deep valley.

IV. DISCUSSION AND CONCLUSION

Model-based spectral contour fitting enables recovering parameters that have an interpretation in terms of (anti-) formant frequencies and bandwidths. Preliminary results suggest that fitting spectral models directly to log-magnitude spectra is feasible. Direct fitting offers more flexibility with regard to the choice of models and constraints. Disadvantages are that spectral fitting via differential evolution is time-consuming and model poles or zeros may be attracted by single harmonics in high-pitched voices, which is a known problem with this kind of models.
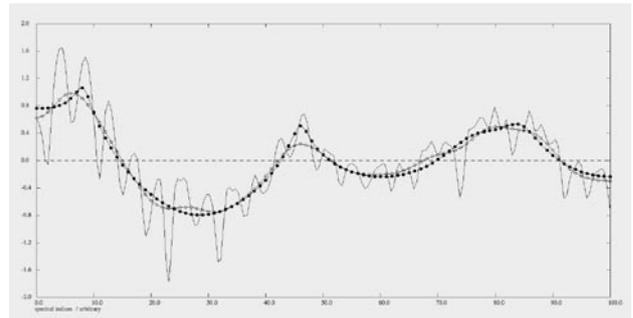
Figure 2: Log-magnitude spectrum of frame 27 ([i]) in [ai]. Overlaid are the cepstrally-estimated spectral contour (white circles) and the fitted transfer function of model (1) (black circles). The horizontal axis is the frequency ranging from 0 to 4 kHz in number of samples. The model-estimated formant frequencies for that frame have been equal to 312 Hz, 1873 Hz, 3102 Hz and 3376 Hz.

	Target vs Model (%)	Target vs Smoothed Contour (%)	Lin. corr.	Rank corr.
Min	25	23	0.67	0.66
1. Quartile	48	46	0.97	0.96
Median	54	52	0.98	0.97
3. Quartile	61	59	0.99	0.98
Max	95	87	1.00	1.00

Table 2: Differences (2) between log-spectra and pole-zero modeled and cepstrally-estimated contours; linear and rank correlations between model transfer functions (1) and cepstrally-estimated spectral contours when silent and weak analysis frames are omitted.

	Target vs Model (%)	Target vs Smoothed Contour (%)	Lin. corr.	Rank corr.
Min	25	23	0.44	0.44
1. Quartile	50	48	0.97	0.96
Median	56	54	0.98	0.97
3. Quartile	63	60	0.99	0.98
Max	99	94	1.00	1.00

Table 3: Differences (2) between log-spectra and all-pole modeled and cepstrally-estimated contours; linear and rank correlations between model transfer functions (1) and cepstrally-estimated spectral contours for all analysis frames.

REFERENCES

[1] J. D. Markel, A. H. Gray, Linear Prediction of Speech, Springer Verlag, Berlin, 1976.
[2] J. Stein, Digital Signal Processing, a Computer Science Perspective, Wiley, New York, 2000.
[3] Audio-Visual to Articulatory Speech Inversion, European FET Project no. 2005-021324.
[4] S. D. Stearns, R. A. David, Signal Processing Algorithms in Fortran and C, Prentice Hall, Englewood Cliffs, 1993.
[5] A. V. Oppenheim, Digital Processing of Speech, in Applications of Digital Signal Processing, A. V. Oppenheim (Ed.), Prentice Hall, Englewood Cliffs, 1978.
[6] K. V. Price, R. M. Storn, and J. A. Lampinen Differential Evolution: A Practical Approach to Global Optimization, Springer, New york, 2005.

AUTOMATIC PARAMETERISATION OF THE GLOTTAL WAVEFORM COMBINING TIME AND FREQUENCY DOMAIN MEASURES

John C. Kane, Christer Gobl

Phonetics and Speech Laboratory, Centre for Language and Communications Studies
Trinity College Dublin

Abstract: **This paper describes a new technique for automatically parameterising the inverse filtered speech waveform by exploiting frequency domain measures and amplitude measures in the time domain. The technique is motivated by the difficulties posed by time domain analysis and by the consequent risks of inconsistencies on the part of both researchers and time based algorithms. The results demonstrate that the system can obtain accurate measurements on synthetic source signals. Analysis was also carried out on short utterances of three male speakers producing tense, modal and breathy voice qualities. Perception tests which involved comparing different resynthesised utterances provide evidence that the new technique is at least as good as our manual method for modal and tense voices. For breathy voice qualities, however, the system needs further development to include aspects like the noise component to provide a more breathy percept.**

Index Terms: **voice source, parameterisation, LF model**

I. Introduction

Despite the attention voice source analysis has received researchers are still seeking to make improvements in terms of accuracy and robustness of parameterisation. Many applications require very accurate and consistent characterisation of the voice source. Recently researchers have started exploring the possibility of including a more sophisticated source model in HMM based speech synthesis in an attempt to reduce 'buzziness' [1, 2]. This is an example of one application which requires high accuracy as well as consistency throughout the analysis. For the purpose of analysing subtle changes in pathological voices accurate parameterisation is also required. Parameter measurement by automatic algorithms, however, tends not to be robust enough particularly across different voice qualities.

Typically the parameterisation of the voice source first requires some type of inverse filtering. This source-filter decomposition is an attempt to remove the effect of vocal tract filtering on the voice source. This is essentially the reverse of the speech production process, as described in [3]. It is done by getting an estimate of the transfer function of the speaker's vocal tract. The speech signal is then filtered using the inverse of this transfer function which produces an estimate of the speaker's voice source. Automatic inverse filtering systems exist (e.g., [4] or those described in [5]), but from our experience there is high risk of incomplete cancellation of formant oscillations. As the purpose of the paper is to test a parameterisation system we require good estimates of the source signal and, hence, have opted to inverse filter small amounts of speech data manually (as described in [6]).

Once a speech signal has been inverse filtered it can then be parameterised. This can be done by marking certain timepoints in glottal waveform or by fitting a model to the pulse. The most documented voice source model, and the one to be used in this study, is Liljencrants-Fant (LF) model [7]. The LF model is a four parameter model of differentiated glottal flow (see Fig. 1). The shape of the model can be described using the parameters R_a, R_k and R_g. The differentiated glottal flow is essentially the residual after inverse filtering as the effect of lip radiation has not been removed. The model is thought to be able to characterise a wide range of phonation types, however as with any model a certain amount of error will exist.
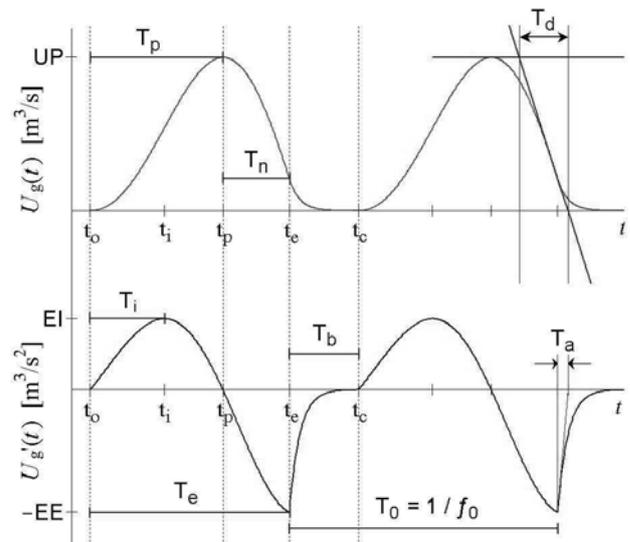
Figure 1: *Examples of LF pulses (bottom) and corresponding glottal pulses (top) (taken from [8])*

When parameterising the source signal most methods involve marking specific timepoints. The precise location of these timepoints can at times be quite unclear and can lead to errors as well as inconsistencies. These difficulties are heightened in the cases of non-modal phonations, e.g., in breathy voices, where a timepoint, for instance the point of glottal opening, can involve very subjective measuring.

A further difficulty with analysing and synthesising breathy voice is that the source signal contains both a periodic voice component and an aperiodic noise component [9]. The LF model parameters are used to characterise only the periodic as-

pect of the voice source. If the signal is not decomposed into periodic and noise components parameter measurements may also include the influence of the aspiration noise and, hence, may not effectively characterise the periodic component. In this study no noise analysis is carried out. We wish to include a noise analysis and synthesis system, perhaps similar to that in [10], in future algorithms.

Frequency domain analysis is thought to have a better mapping to the perception of speech than time domain analysis. In the time domain even minor errors in model fitting can have major perceptual effects. The power spectrum also bypasses any phase distortions which can upset time domain parameterisation. A complete frequency domain approach would, hence, lessen the need for high fidelity recordings which preserve phase linearity, and would allow for the analysis of a far greater range of speech data. This is a current direction of our research and work on a full frequency domain parameterisation system is underway. For the present study only the return phase parameter R_a will be measured from the source spectrum.

The remaining source parameters can be calculated from amplitude based measures in the time domain. Such measures are said to be more robust than marking time instances, especially in automatic systems [11]. It is hoped that this novel combination of frequency domain measures and time based amplitude measures can avoid some of the pitfalls of purely time point measurements and provide a robust and automatic analysis of the inverse filtered signal.

II. Method

This sections outlines the different methods applied in our system in order to arrive at a full set of LF model parameters. As the study is mainly concerned with parameterisation we have analysed small amounts of speech data that have been carefully inverse filtered.

A. Inverse Filtering

We have opted for a manual inverse filtering approach for this study, as in [6]. The software first uses an linear predictive coding (LPC) method for estimating formant frequencies and bandwidths. The user then fine-tunes formant frequencies and bandwidths manually by utilising time and frequency domain displays to ensure complete formant cancellation. This fine-tuning is done for each pulse and the final output is an estimation of the differentiated glottal waveform.

B. Amplitude-based measurements

The first amplitude measurement is the negative peak of the differentiated glottal waveform, E_e. This parameter is simply measured by the new automatic system as the maximum negative amplitude of each glottal pulse. The next parameter, which is again straightforward to measure, is the peak positive amplitude of the differentiated glottal pulse, E_i, see Fig. 1 (bottom).

The measurement of the maximum amplitude of the undifferentiated flow is complicated by the occurrence of zero line drift. The LF model is designed to have equal area above and below the zero axis which means when you integrate, the signal sits neatly on the zero axis, see Fig. 1, top. Real speech, however, does not maintain this exact property and as a result the integrated waveform drifts off the zero axis, see Fig. 2 (top).

To adjust for this our system marks the major negative points for each pulse. A line is drawn from the origin then to each of the negative peaks, see the dashed line in Fig. 2 (top). Then at

each sampling point the distance between the dashed line and the zero axis is added to the signal at that sampling point which results in the signal being lifted onto the zero axis, see Fig. 2 (bottom). The system can now easily measure the maximum amplitude of each pulse, our U_p value.
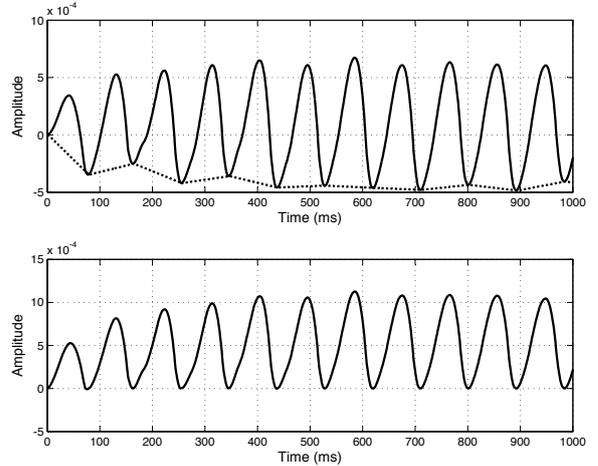
Figure 2: *The integrated source waveform (top) and the same source waveform adjusted for zero-drift (bottom)*

C. Calculating $\mathbf{R}_k$ and $\mathbf{R}_g$ from amplitude measures

Aside from f_0 and E_e the LF model can be described using three further shaping parameters. We use the parameters R_a, R_k and R_g which can be calculated from time instance measurements, as in equation 1. The positions of these time instances can be seen in Fig. 1 (bottom).

$$R_a = \frac{T_a}{T_0} \quad R_k = \frac{T_n}{T_p} \quad R_g = \frac{T_0}{2T_p} \quad (1)$$

The parameters R_k and R_g can also be estimated using our three amplitude measures (E_e, E_i and U_p). [12] gives a detailed description of how voice source parameters can be arrived at using amplitude measurements. Equation 2 shows how one can obtain an amplitude representation for R_k and R_g, as described in [12].

$$R_{ka} = \left(\frac{2}{\pi}\right)\left(\frac{E_i}{E_e}\right) \quad R_{ga} = \frac{(\frac{1}{\pi})(\frac{E_i}{U_p})}{f_0} \quad (2)$$

D. Frequency domain analysis

With R_k and R_g already estimated one parameter remains to configure the LF model. The parameter R_a characterises the return phase of the source waveform and it is, perhaps, considered to be the most important LF parameter [13]. However, getting accurate estimations of R_a is thought to be a challenging task [14]. Our approach derives a value for R_a from the frequency domain and the process is presented graphically in Fig. 3.

We define a set of possible R_a values, e.g., 1% to 20% (the values here refer to the percentage the return phase is of the pulse duration). The lowest is the minimum possible R_a value and the highest is the maximum possible R_a value and 50 linearly spaced values in between. The system takes a section of one pulse length from the signal and gets the spectrum (the dark

line in Fig. 3). We then generate 50 LF pulses for each of the R_a values but with all other parameter values remaining fixed. Spectra are then taken of each of the pulses (the grey lines in Fig. 3). The system then uses a Euclidean distance measure to choose the LF configuration with the closest match to the source signal. The R_a value used in this configuration is chosen as the optimal value. The system then proceeds to the next pulse and the process is repeated. This continues until the end of the signal is reached.
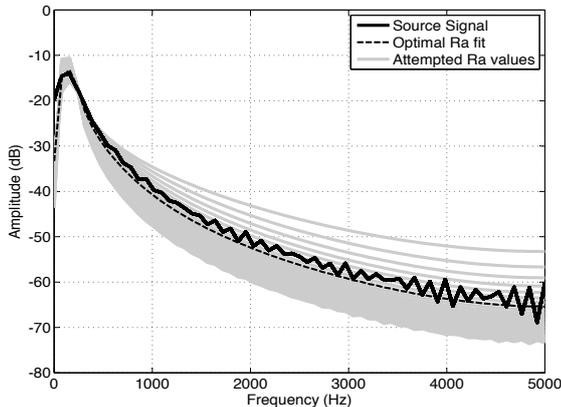
Figure 3: *Spectrum of a single pulse from the voice source signal (dark line). The grey lines show the range of LF configurations produced and the dashed line shows closest matching configuration.*

III. Evaluation

The evaluation process for the system is two-pronged. The first testing stage is on synthetic source pulses. The reason for this is that the correct source parameter values are known and the actual size of the error can be measured. The second stage involves analysing real source signals and for this the correct parameter values are not known. To provide an evaluation for real speech we have opted to apply the current system and the manual parameterisation method, as described in [6]. The extracted parameters for each method were then used to make resynthesised utterances. The resynthesised utterances were then used in perception tests where participants chose the stimulus which sounded most like the original speech utterance. This preliminary evaluation should provide evidence for whether this system provides good parameterisation for real speech. A more rigorous evaluation is required and is planned for future work.

A. Recordings

Three male English speakers were recorded producing an [a] vowel in tense, modal and breathy phonation modes. These categories of voice quality are those as described by [15] and participants listened to samples of each phonation type and practised them several times before recording. Each utterance was around half a second long and was recorded using a Pearl CC30 condenser microphone in a semi-anechoic room. Speech segments were digitised at 44.1 kHz, then downsampled to 10 kHz and high pass filtered at 40 Hz. The choice of microphone and filter ensured phase linearity was maintained. Using the method described in [6] utterances were manually inverse filtered to minimise errors.

B. Parameterisation

Each utterance was parameterised both using the new system and a manual parameterisation method, described in [6]. The manual method involves the user manually fitting the LF model to each source pulse by varying one amplitude based and four time based markers in an effort to achieve the optimal match. Although the fitting is done in the time domain, a frequency domain display is also available to the user to facilitate a more thorough fit.

C. Synthetic source analysis

Nine synthetic source signals were generated using static parameter settings. The signals were made by constructing an LF model with particular parameter settings and concatenating 10 identical pulses. In every signal E_e was set to 1. The values of R_a, R_k, R_g and f_0 were chosen so as to have a range of source signals corresponding to modal, tense and breathy voice qualities. Previous analysis of these voice qualities aided the choice of value for the above parameters. The nine signals were analysed using only the new parameterisation system.

D. Perception Tests

Perception tests were used as a method of comparing the parameterisation of the new automatic method with the manual method for real speech. 18 volunteers participated in a two-part perception test in a quiet room using high quality loudspeakers.

In test one participants were presented with 45 groups of three stimuli. In each group the first two stimuli were resynthesised versions of the original utterance, one from parameters obtained from the automatic method and one from the manual method. Resynthesis was carried out using a cascade formant synthesiser. The two synthetic signals in each group used the same f_0, EE and formant frequencies and bandwidths. They differed only in the R_a, R_k and R_g parameters which were extracted from the two methods. The third stimulus was the original speech sound and participants had to choose which of the synthesised sounds they deemed closest to the original. The assumption here being that a closer sounding resynthesis demonstrates more accurate voice source parameterisation.

Part 2 of the test was a standard ABX discrimination task where the participants were presented with the two different resynthesised utterances and then a third sound which was a randomly chosen one of the previous two. Participants had to choose which sound had been repeated. Again there were 45 groups of stimuli. This test was chosen to demonstrate whether the differences in the parameter measurements by both methods produced two differentiable sounds.

In both parts of the perception test the order of the resynthesised stimuli in each group, as well as the order of the 45 groups, were randomised.

IV. Results and discussion

Table 1 summarises the testing of the parameterisation system on the nine synthetic source signals. The mean and standard deviation of the error (i.e. the difference between the actual source parameter values and those extracted by the system) as well as the range of values used are presented. Encouragingly, the error size for all three parameters is reasonably low. This is evidence that the amplitude based representations of R_k and R_g can indeed provide good estimates of those parameters for a wide range of settings. These results also demonstrate that the

novel method of estimating R_a is effective, at least for synthetic source pulses.

Table 1: *Summary of parameterisation error for R_a, R_k and R_g. The range of values used to generate the signals and the mean and standard deviation of the differences between the extracted values and the actual parameter values are shown*

	$\mathbf{R}_a$	$\mathbf{R}_k$	$\mathbf{R}_g$
Range	1.8%-7%	28%-39%	79%-137%
Mean Error	0.9%	3.9%	6.1%
St Dev of Error	0.7%	3.1%	0.4%

The results of the perception tests are presented in Table 2. Test 1, the first row, shows the percentage of instances participants chose the automatic method to be closer to the original (i.e. over 50% shows preference for the automatic method while under 50% shows preference for the manual method). Overall, at 50.3%, we can see that participants showed no preference for either method in terms of closeness to the original. The second row contains the results from test 2 which show the percentage of instances participants correctly identified the repeated synthesised stimuli.

Table 2: *Test 1: the percentage of instances participants believed the synthesised stimuli from the automatic method to sound closer than the manual method to the original speech utterance. Test 2: the percentage of correctly identified synthesised stimuli*

Test	Modal	Tense	Breathy	Overall
1	49.8%	61.2%	40%	50.3%
2	54.5%	66.7%	72.9%	65.1%

For modal voice qualities 54.5% for test 2 suggests that participants were largely unable to discriminate between the two resynthesised utterances. Test 1 results, 49.8%, show that participants believed neither synthesised utterances to be closer to the original. For tense voice qualities participants slightly favoured resynthesised versions which used the new system's parameter values (61.2% of participants showing preference for the automatic method) and they were reasonably able to discriminate the two synthesised sounds.

Breathy voice qualities, as expected, proved more difficult than the other voice qualities in both the inverse filtering and parameterisation stages. It was found that participants showed slight preference for the manual method, with 40% stating that they preferred the automatic method. Participants also demonstrated a reasonable ability to differentiate the two sounds, at 72.9%. It should be noted that the synthesised versions of breathy voice qualities overall were of poorer quality than the modal and tense versions. This suggests that the LF model alone does not provide enough source information to convey a breathy percept.

V. Conclusion

Overall evidence from the evaluation appears to be encouraging for the new system described here. Analysis of synthetic signals confirms that R_k and R_g can be estimated with good accuracy from amplitude measurements alone. The analysis also demonstrates the effectiveness of the new method of R_a estimation in the frequency domain. We hope to include this method in our forthcoming all-frequency domain parameterisation system.

Results from the perception tests suggest that the automatic system is at least as effective as the manual method for modal to tense voice qualities. This inference, however, comes solely from the fact that resynthesised utterances were judged to sound closer to the original speech utterances and does not rigorously demonstrate accuracy of parameterisation.

For breathy voice qualities we hope that by implementing analysis and synthesis of the noise component, perhaps similar to that in [10], we can provide a more perceptually breathy sound. The issue of finding further methods of demonstrating accuracy of parameterisation for breathy voice qualities and for voiced speech in general requires further attention.

VI. Acknowledgements

VII. References

[1] Cabral, J. P., Renals, S., Richmond, K. and Yamagishi, J., "Glottal spectral separation for parametric speech synthesis", Proc. of Interspeech, 2008.

[2] Raitio, T., "Hidden markov model based Finnish text-to-speech system utilizing glottal inverse filtering", Masters Thesis, 2008.

[3] Fant, G., *The acoustic theory of speech production*, Mouton, Hauge (2nd Edition 1970).

[4] Airas, M., "Methods and studies of laryngeal voice quality analysis in speech production", Ph.D. Thesis, 2008.

[5] Pfitzinger, H. R., "Influence of differences between inverse filtering techniques on the residual signal of speech", DAGA München, 2005.

[6] Gobl, C. and Ní Chasaide, A., "Techniques for analysing the voice source" in *Coarticulation: Theory, Data and Techniques* edited by Hardcastle, W. and Hewlett, N., pp 300-320, Cambridge University Press, 1999.

[7] Fant, G. and Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 26(4):1–13, 1985.

[8] Gobl, C., "The voice source in speech communication - production and perception experiments involving inverse filtering and synthesis.", Ph.D. thesis, KTH, 2003.

[9] Mehta, D. and Quatieri, T., "Synthesis, analysis, and pitch modification of the breathy vowel", IEEE Workshop on applications of signal processing to audio and acoustics, 2005.

[10] Gobl, C., "Modelling aspiration noise during phonation using the LF voice source model", Interspeech, Pittsburgh, 2006.

[11] Alku, P., Bäckström, T. and Vilkman, E., "Normalized amplitude quotient for parameterization of the glottal flow", Journal of the accoustical society of America, 112, pp. 701-710, 2002.

[12] Gobl, C. and Ní Chasaide, A., "Amplitude-based source parameters for measuring voice quality", VOQUAL'03, 2003.

[13] Fant, G. and Gustafson, K., "LF-frequency domain analysis", STL-QPSR, 2, 1996.

[14] Strik, H., "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses", Journal of the Acoustical Society of America, 103(5), pp. 2659-2669, 1998.

[15] Laver, J., *The phonetic description of voice quality*, Cambridge University Press, 1980.

SYNTHETIC HOARSE VOICES: A PERCEPTUAL EVALUATION

S. Fraj[1], F. Grenez[1], J. Schoentgen[1,2]
[1]Laboratory of Images, Signals & Telecommunication Devices, Université Libre de Bruxelles,
50, Avenue F. D. Roosevelt, 1050 Brussels, Belgium
[2]National Fund for Scientific Research, Belgium

Abstract: **The presentation concerns the evaluation of a synthesizer of disordered voices. The objective is the perceptual assessment of the ability of the synthesizer to simulate disordered voice timbres. Three perceptual experiments, based on a pairwise comparison paradigm, have been carried out. The first involved jitter, the second breathiness and the third a combination of both. Results of the first two experiments show that the perceptual ranking accords with the synthesis parameters as well as measured speech jitter, speech shimmer and Harmonics-to-Noise ratios. For the third experiment, which involved jitter as well as additive noise, a two-dimensional multidimensional scaling analysis shows that for lower levels of additive noise, increased jitter and additive noise are perceived as distinct disordered voice timbres.**

Keywords: **Synthesis of disordered voice timbres, perceptual evaluation.**

I. INTRODUCTION

The presentation concerns the evaluation of a synthesizer of abnormal voice timbres with respect to its ability to simulate hoarseness. The synthesizer, which includes a shaping function model [1], has been presented earlier [2] and its ability to mimic modal voices has been evaluated [3].

Few studies have been devoted to the perceptual assessment of hoarseness in synthetic disordered voices. One reason is the scarcity of models which enable simulating voice disorders. Motivations for developing synthesizers of disordered voices are the discovery of speech cues that are relevant to the perception of abnormal voices, the preparation of reference stimuli in the context of the perceptual assessment of disordered voices, the training of speech therapists in the auditory evaluation of dysphonic speakers as well as the testing of the reliability or validity of acoustic cues of disordered speech.

The objective of this study is to evaluate the capacity of a synthesizer to simulate disordered voices. Increased jitter and additive noise are known causes of abnormal voice qualities. A first experiment has therefore involved jitter, a second additive noise and a third jitter and additive noise combined. For the first and second experiments, judges have been instructed to select the

hoarsest stimulus within a pair in the framework of a pairwise comparison paradigm. For the third, they have been instructed to indicate whether stimuli within a pair have been similar or not with regard to hoarseness. Results for the first and second experiments enable ranking vowels from the less to the most hoarse and correlating with acoustic features. Results of the third experiment have been investigated in the framework of a multidimensional scaling analysis and the underlying perceptual factors interpreted in terms of synthesis parameters and measured acoustic cues.

II. METHODS

A. Synthesizer

The synthesizer (*Fig.1*) rests on a template of the glottal area, which is based on nonlinear memory-less shaping functions that transform two driving harmonics into the desired glottal area [1]. The reason for opting for a shaping function model of the glottal area is that the length of a glottal area cycle and its spectral slope and amplitude may be controlled via the instantaneous frequency and amplitude of the harmonic driving functions.
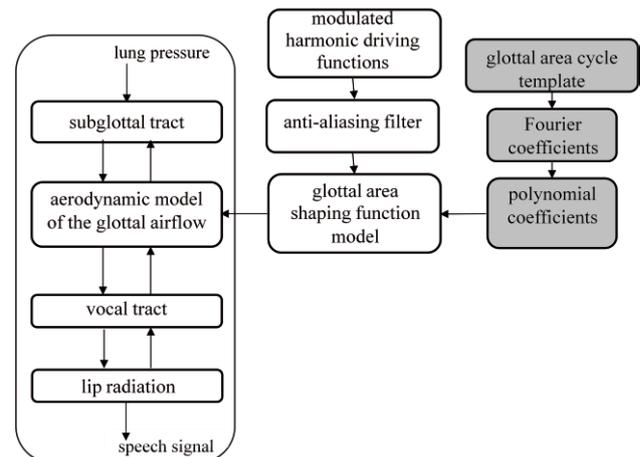
Fig.1: Block diagram of the synthesizer

The synthesizer comprises, in addition, models of the glottal airflow and the supra and infra-glottal tracts. The glottal airflow depends on the glottal area, the incident components of the infra- and supra-glottal acoustic pressure waves as well as physical constants [4]. The incoming and outgoing components of the pressure waves

96

have been obtained via the temporal simulation of the acoustic wave propagation in the vocal tract.

Trachea and vocal tract have been mimicked by means of a concatenation of cylindrical tubelets of identical length but different cross sections. Losses have been taken into account via models of the wall vibration, heat conduction and viscous friction as well as lip and glottal radiation.

Several types of perturbations, such as vocal jitter [5], vocal tremor, additive noise, diplophonia, biphonation and random vibrations have been simulated and tested separately. The focus of this presentation is the assessment of synthetic disordered vowels with increased vocal jitter and breathiness.

B. Synthesis of jitter

Vocal jitter designates small random perturbations of the glottal cycle lengths. Here, the instantaneous frequencies of the harmonic driving functions have been perturbed by white noise e_n. Different levels of speech jitter and shimmer have been obtained by changing the variance $P_{jit}^2 . \Delta$ of the white noise given by (1). Symbol Δ is the time step.

$$e_n = P_{jit} . \sqrt{\Delta} \begin{cases} +1, p = 0.5 \\ -1, p = 0.5 \end{cases} \quad (1)$$
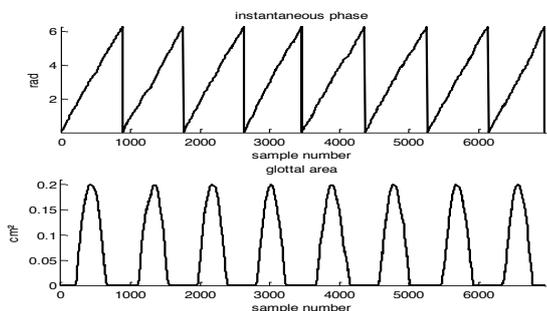
Fig.2: Simulation of vocal jitter: instantaneous phase (top); glottal area function (bottom). Perturbations are exaggerated to increase their visibility.

C. Synthesis of additive noise

Additive noise is mimicked by means of Brownian noise, the amplitude of which is modulated via an affine function (2) of the rate of flow. The modulated noise is delayed by one millisecond and added to the airflow rate. Different levels of breathiness are obtained by changing the coefficients of affine function (2). Symbols $Nois_1$ and $Nois_2$ are coefficients that are fixed by the user.

$$Nois_1 . (flow \ rate) + Nois_2 \quad (2)$$

III. PERCEPTUAL EVALUATION

A. Experiment 1

The topic of the first experiment has been synthetic vocal jitter [5]. Ten vowels [a] have been simulated by modifying the size of the white noise perturbing the instantaneous frequency of the model driving functions, yielding different levels of speech jitter and shimmer. Speech shimmer has been generated from glottal jitter via modulation distortion in the vocal tract [6]. All the other parameters have been kept constant under the threshold of pathology. Values of jitter measured by PRAAT [7] have been in the interval from 0.1% to 2.4%.

Each vowel has been one second long. A two by two comparison experiment has been carried out by three expert and seven naive listeners. Stimuli have been presented pairwise to each listener in a random order. The task has been to designate the item of each pair that was perceived as the hoarsest. The judges could also report both items as identically hoarse. Based on the judge's combined comparisons, the synthetic stimuli have been ranked from the least to the most hoarse.

B. Experiment 2

The topic of the second experiment has been synthetic breathiness mimicked by additive noise. Ten vowels [a] have been simulated by varying coefficient $Nois_1$ of the affine function that modulates the Brownian noise, yielding different levels of breathiness. All the other parameters have been kept constant under the threshold of pathology. The harmonic to noise ratios measured by PRAAT have been in the interval from 14 to 24 dB.

Each vowel has been one second long. A two by two comparison experiment has been carried out by the same listeners as in Experiment 1. The stimuli have been presented pairwise to each listener in a random order. The task has been to designate the item of each pair that was perceived as the hoarsest. The judges could also report both items as identically hoarse. Based on the judge's combined comparisons, the synthetic stimuli have been ranked from the least to the most hoarse.

C. Experiment 3

The objective of the third experiment has been the investigation of perceived dissimilarities between vowels with different values of speech jitter and shimmer or breathiness. Sixteen vowels [a] have been simulated by changing both jitter and additive noise. The sixteen vowels have combined four different levels of additive noise and four different levels of glottal jitter (*Tab. 1*). All the other parameters have been kept constant under the threshold of pathology. Measurements by PRAAT

have shown that the least perturbed stimulus involved 0.2% of jitter and a harmonic-to-noise ratio of 24 dB and the most perturbed stimulus 1.8% of jitter and a harmonic-to-noise ratio of 10 dB. In Tab. 1, parameters of additive noise refer to the coefficient $Nois_1$ of the affine function that multiplies the glottal airflow. Coefficient $Nois_2$ has been kept constant. Glottal jitter in Tab.1 refers to the coefficient P_{jit}.

Each stimulus has been one second long. A two by two comparison experiment has been carried out by the same 10 listeners as in Experiments 1 and 2. Stimuli have been presented pairwise to each listener in a random order. The task has been to indicate whether the items of each pair were equally hoarse or not.

A multidimensional scaling program PROXSCAL [8] has been applied to a global matrix in which each cell represented the number of times the two items of a vowel pair had been perceived as different with respect to hoarseness. The output of PROXSCAL is a low-dimensional space in which the stimuli are represented as points. Two vowels that are perceived as similar with regard to hoarseness are represented by two points that are close, while two vowels that are perceived different in terms of hoarseness are represented by two points that are distant.

Tab.1: Values of Jitter (local) and Mean Harmonics-to-Noise ratio obtained by PRAAT in Experiment 3 for different combinations of parameters P_{jit} and $Nois_1$.

Vowels			$Nois_1$			
			0.10	0.15	0.20	0.35
P_{jit}	4	Jitter (%)	0,26	0,23	0,27	0,30
		HNR (dB)	24,83	23,52	21,24	17,37
	12	Jitter (%)	0,74	0,89	0,87	0,81
		HNR (dB)	17,90	16,90	16,37	14,47
	20	Jitter (%)	1,26	1,18	1,39	1,29
		HNR (dB)	13,98	13,91	12,73	12,35
	28	Jitter (%)	2,05	2,10	2,02	1,85
		HNR (dB)	10,58	10,09	10,39	10,66

IV. RESULTS AND DISCUSSION

A. Experiment 1

Results (not reported here) show that listener's responses are highly correlated (>0.9). The ranking obtained via perceptual two by two comparison accords with the measured vocal jitter (*Tab.2*).

Tab.2: Average ranks obtained by 10 listeners when classifying vowels with different levels of jitter. Speech jitter, speech shimmer and HNR are measured by means of PRAAT.

Glottal jitter P_{jit}	0	4	8	12	16
Speech jitter (%)	0.09	0.28	0.56	0.89	1.03
Speech shimmer (%)	1.35	1.51	2.49	3.23	4.44
HNR (dB)	31.48	25.97	21.27	17.66	15.73
Average ranks	0.8	1.1	1.6	2.9	4.1
Glottal jitter P_{jit}	20	24	28	32	36
Speech jitter (%)	1.31	1.64	2.03	2.23	2.41
Speech shimmer (%)	6.22	6.05	8.30	8.25	9.12
HNR (dB)	14.13	12.30	10.69	9.80	9.48
Average ranks	5.4	6.5	6.5	7.1	7.9

B. Experiment 2

Results (not reported here) show that listener's responses are highly correlated (>0.9). The ranking obtained via perceptual two by two comparison accords with the measured Harmonic-to-Noise ratios (*Tab.3*).

Tab.3: Average ranks obtained by 10 listeners when classifying vowels with different levels of additive noise. Speech jitter, speech shimmer and HNR are measured by means of PRAAT.

Additive glottal noise $Nois_1$	0.05	0.10	0.11	0.125	0.15
Speech jitter (%)	0.36	0.32	0.32	0.38	0.32
Speech shimmer (%)	1.93	2.32	2.49	2.49	3.28
HNR (dB)	24.59	23.68	23.33	22.20	22.19
Average ranks	0.2	1.8	2.2	2.8	3.7

Additive glottal noise $Nois_1$	0.16	0.2	0.25	0.33	0.5
Speech jitter (%)	0.35	0.34	0.32	0.37	0.37
Speech shimmer (%)	3.35	3.53	4.69	6.75	7.72
HNR (dB)	21.17	20.47	19.06	17.16	14.74
Average ranks	3.8	6.2	6.7	7.8	8.9

C. Experiment 3

The correlation between listener's responses has been in the interval 0.60-0.85. Multidimensional scaling analysis of results of Experiment 3 suggests that the synthetic stimuli may be represented meaningfully in a two-dimensional space. This 2D representation shows that when additive noise was the highest (HNR = 10 dB), different levels of jitter were not perceived as separate qualities.

However, for lower levels of additive noise, different levels of jitter and different levels of breathiness were perceived as distinct timbres. Continuous black lines have been drawn by hand to guide the reader. They suggest iso-additive noise curves. Alternatively, iso-glottal jitter lines could be obtained connecting iso-P_{jit} values in the graph.
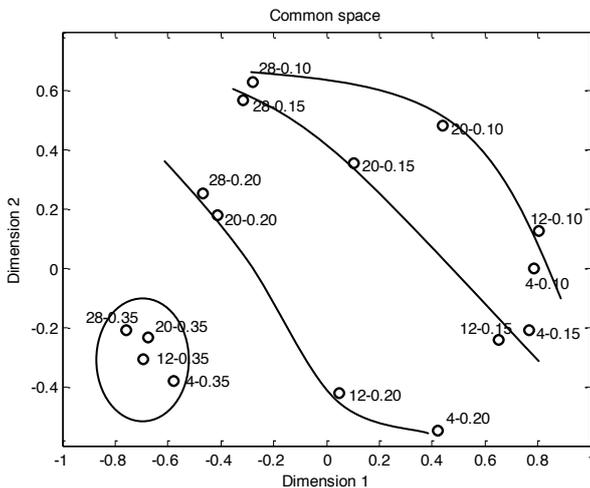
Fig.3: Multidimensional scaling analysis. The labels report P_{jit} and $Nois_1$ parameter values.

Multidimensional scaling analysis of the distances between listeners (instead of stimuli) does not suggest that the judges cluster into sub-groups (experts versus naives, for instance)

IV. CONCLUSION

Three perceptual experiments, involving synthetic disordered vowels, suggest that the synthesizer is able to simulate different disordered voice qualities. For the first and second experiments, the average ranks with respect to hoarseness evolve with the values of the synthetic parameters and measured speech jitter, shimmer and harmonics-to-noise ratio. In addition, correlations of listener's responses exceed 0.9 for both experiments. Results of the multidimensional scaling analysis of the third experiment show that when additive noise is the highest, judges are unable to distinguish different levels of jitter. However, for lower levels of additive glottal noise, jittered and noisy stimuli are perceived as distinct timbres.

REFERENCES

[1] J. Schoentgen; "Shaping Function Models of the Phonatory Excitation Signal", JASA 114(5), November 2003, pp. 2906-2912.
[2] S. Fraj, F. Grenez, J. Schoentgen; "Evaluation of a Synthesizer of Disordered Voices", 3rd Advance Voice Function Assessment International Workshop, Madrid, 18-20 May 2009, pp. 69-72.
[3] S. Fraj, F. Grenez, J. Schoentgen; "Perceived naturalness of a synthesizer of disordered voices", Proceedings INTERSPEECH 2009, Brighton, UK, 6-10 September 2009.
[4] Titze, I.R., The Myoelastic Aerodynamic Theory of Phonation, National Center of Voice and Speech, USA, 2006, pp. 265.
[5] Schoentgen, J., "Stochastic models of jitter", JASA, 109(4), April 2001, pp.1631-1650.
[6] Schoentgen, J., "Spectral models of additive noise and modulation noise in speech and phonatory excitation signals", JASA, 113(1), January 2003, pp.553-562.
[7] www.praat.org
[8] SPSS Statistics 17.0.0

SALIENCE ANALYSIS FOR GLOTTAL CYCLE DETECTION IN DISORDERED SPEECH

C. Mertens[1], F. Grenez[1], L. Crevier-Buchman[2,3], J. Schoentgen[1,4]

[1]Laboratory of Images, Signals and Telecommunication Devices, Université Libre de Bruxelles, Brussels, Belgium
[2]Laboratoire de Phonétique et de Phonologie, CNRS/Sorbonne-Nouvelle, Paris, France
[3]Hôpital Européen Georges Pompidou, Paris, France
[4]National Fund for Scientific Research, Belgium

Abstract : The presentation concerns the evaluation of a temporal method for tracking cycle lengths in voiced speech. The speech cycles are detected via the saliences of the speech signal samples. The method does not request that the signal is locally periodic and the average period length known a priori. The cycle length extraction is applied to the analysis of dysphonic speakers affected by amyotrophic lateral sclerosis (ALS). Results suggest that salience analysis is able to track reliably glottal cycles in the speech signal. SLA speakers are characterized by higher vocal tremor depths and tremor frequencies than normophonic speakers.

Keywords : vocal frequency, vocal tremor, speech salience analysis

I. Introduction

In clinical applications of speech analysis, speech cycles are detected to measure their lengths and amplitudes with a view to investigating slow (vocal tremor) and fast (vocal jitter and shimmer) perturbations of vocal frequency and speech cycle amplitude. Often, such analyses are frame-based and the cycle detection rests on the iterative selection of speech signal peaks that occur in the vicinity of the instants of maximal acoustic excitation. To facilitate this selection, one often assumes that voiced speech segments are pseudo-periodic so that the peaks can be determined one by one on the base of a prior estimation of the typical fundamental period. The assumption of quasi-equal spacing is, however, valid for modal voices only and not for pathological ones, which may be characterized by large cycle-to-cycle fluctuations. Cycle insertion or omission errors may therefore occur, which bias the acoustic cues of cycle regularity.

Here, we propose to track speech cycles via a multi-scale analysis that assigns a salience to each signal peak. The salience of a speech signal peak designates the time interval over which this peak is a maximum. A signal peak is a signal sample whose left and right neighbours are smaller. The speech cycle tracking based on peak saliences does not rest on the assumptions that the speech signal is locally periodic and the average period length known a priori.

II. Method

A. Preprocessing

The speech signal has been band-pass filtered by means of a finite response (FIR) filter with cut-off frequencies equal to $60Hz$ and $1000Hz$ to remove low-frequency hum, additive noise owing to turbulence as well as high-frequency formants.

The speech signal has been upsampled to $Fs = 200kHz$ to guarantee a precision of the peak positions requested by the size of vocal jitter expected in modal speech.

B. Speech sample salience analysis

The salience s_f of a signal sample is defined as the length of the longest temporal interval over which the signal sample is a maximum. A property of the salience is that a sample with a large salience has not necessarily a large amplitude and vice versa. In voiced speech fragments, speech cycles are often characterized by a prominent signal peak that is the effect of the glottal excitation. The salience of that peak is expected to be high irrespective of the evolving signal amplitude. Here, a salience analysis based on a sliding window is carried out. The windowed salience analysis algorithm has been presented in [1] and [2]. The window-based approach enables decreasing boundary effects at the beginning and end of the analysis interval and reducing computation time. Each signal sample is thus assigned a salience, but only the saliences of signal peaks are kept for further processing.

Fig. 1, illustrates, for instance, the peak salience values obtained for a fragment of vowel [a]. One observes that the prominent signal peaks that are due to the glottal excitation have a higher salience value than other secondary peaks that are due to tract resonances.

C. Speech cycle tracking based on peak salience

For speech cycle tracking, no assumptions are made with regard to the regularity of the cycle lengths. One assumes that the vocal frequency is comprised between
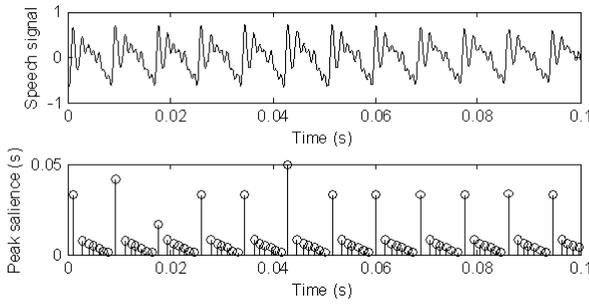
Fig. 1 : Example of signal sample saliences

$F_{0min} = 60Hz$ and $F_{0max} = 400Hz$, which corresponds to an inter-peak spacing in samples comprised between N_{min} and N_{max}.

The first step consists in ranking the signal peaks according to decreasing salience and keeping those peaks the salience values of which are greater than or equal to twice the length of the shortest possible cycle length (i.e. $\geq 2N_{min}$). The initial number of peaks is therefore in excess of the number of expected cycles.

The second step consists in finding several candidate cycle length time series by means of the retained peaks and keeping the length series that has the smallest cycle duration perturbations. Each candidate cycle length series is found by extracting iteratively a signal peak sub-sequence on the base of the local inter-peak durations and the peak salience values, assuming that prominent speech cycle peaks owing to the glottal excitation are characterized by large salience values. The difference between two candidate cycle sequences consists in two different initializations of the iterative peak search.
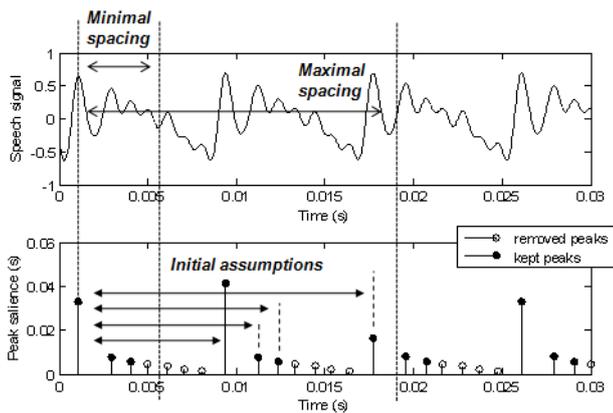
Fig. 2 : Initialization of cycle lengths tracking

The initialization consists in determining all candidate inter-peaks distances at the beginning of the signal. Let g' the position of the first peak in the peak sequence,

with $g' < N_{max}$. Starting from g', one determines all the distances between this peak and the next peaks. Accounting for the range of the vocal frequency, the search interval starts at $g' + N_{min} - 1$ and stops at $g' + N_{max} - 1$. Let h' be a candidate for the second peak. The initial cycle duration then is $d_{(g',h')}$. Fig. 2, illustrates all possible initial cycle durations. In addition, a third peak i' is required, which satisfies condition (1) :

$$\alpha \cdot d_{(g',h')} < d_{(h',i')} < \beta \cdot d_{(g',h')} \qquad , \quad \begin{array}{c} 0.5 < \alpha < 1 \\ \beta > 1.5 \end{array} \tag{1}$$

Assuming that the required initial distance hypotheses have been made, a sequence is built up comprising peaks regularly spaced in time and secondary intrusive peaks. Peak picking is based on an heuristic similar to the one explained in the previous paragraph. The problem is to detect and remove intrusive peaks. Fig. 3, illustrates the four possible situations when the algorithm deals with a new peak i, assuming previous peaks g and h given.
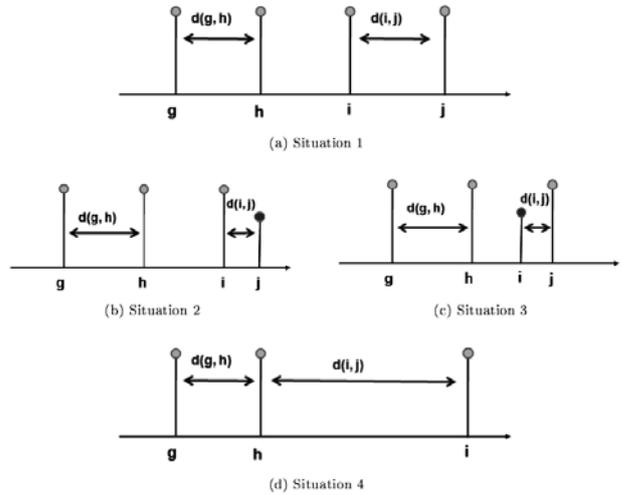
Fig. 3 : Peak selection

1) In the first situation, the inter-peak durations are quasi-identical. No peak must be removed and the algorithm proceeds to the next peak.
2) In the second and third situations, an intrusive peak is present. If condition (2) is met, the algorithm must decide which peak (i or j) is intrusive.

$$d_{(i,j)} < \alpha \cdot d_{(g,h)} \tag{2}$$

The decision is based on two factors. The first is the spacing (3) between consecutive peaks :

$$\delta_i = \left| d_{(g,h)} - d_{(h,i)} \right| \qquad \delta_j = \left| d_{(g,h)} - d_{(h,j)} \right| \tag{3}$$
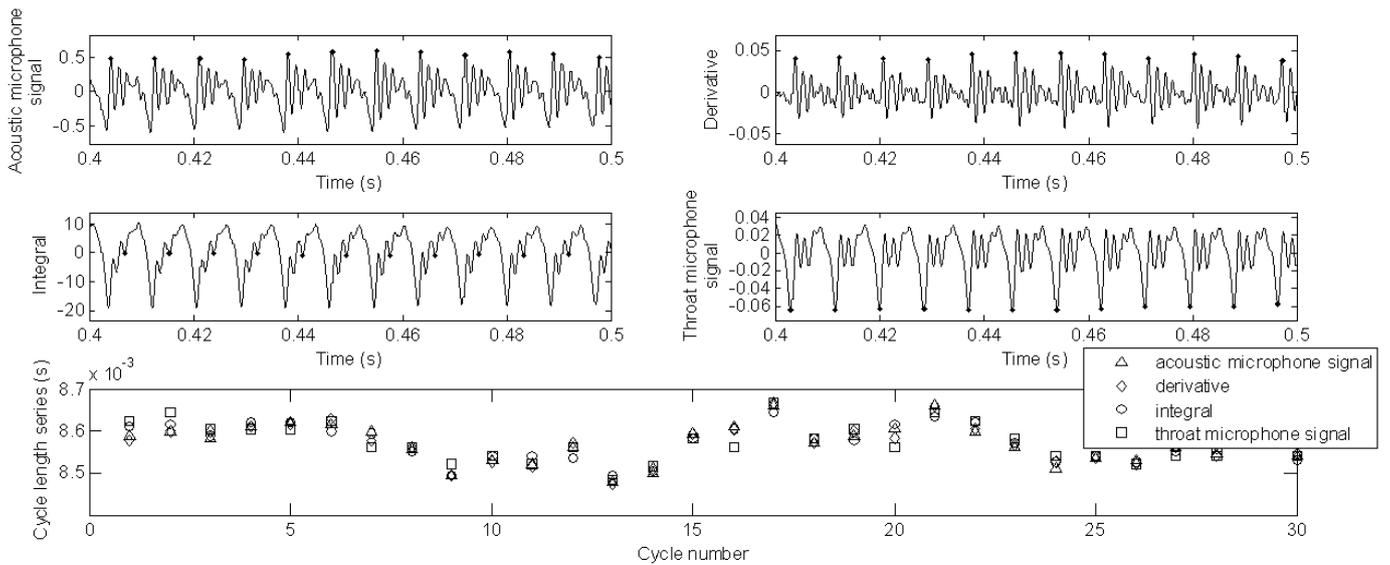
Fig. 4 : Fragments of vowel [a] and time cycle length series (speaker 3)

The second are the saliences s_f assigned to peaks i and j. Distances and saliences are combined in local costs (4) :

$$c_i = \frac{1 + \delta_i}{s_f(i)} \qquad c_j = \frac{1 + \delta_j}{s_f(j)} \qquad (4)$$

The peak the local cost of which is minimal is kept. If peak i is kept, peak j is omitted and the following peak is considered. If j is kept, it becomes the current active peak $(j \rightarrow i)$.

3) In the fourth situation, peak i is situated too far from the previous peak. If condition (5) is met, that situation would lead to a peak ommission error. The sub-sequence would not be relevant. Therefore, the extraction of the current peak sub-sequence is canceled and a new peak sub-sequence is extracted on the base of a new initial distance hypothesis.

$$d_{(h,i)} > \beta \cdot d_{(g,h)} \qquad (5)$$

Several peak sub-sequences are so obtained. To determine the relevant one, the standard deviation of the inter-peak durations is computed. The peak subsequence giving rise to a minimal standard deviation (i.e. overall smallest perturbations) is kept. The sample salience analysis and the cycle detection heuristic are carried out once for each polarity of the signal. The polarity giving the smallest overall perturbation is retained.

D. Corpora

Cycle length time series have been obtained via speech sample salience analysis of French vowel [a] sustained by normophonic and dysphonic speakers. The corpus of normophonic speakers has comprised 8 subjects for which both the acoustic and contact microphone signals have been recorded simultaneously. The integral and derivative of the acoustic signal have also been obtained. The corpus of dysphonic speakers has comprised 72 patients with amyotrophic lateral sclerosis (ALS), a neurological disease which affects the muscular system. These stimuli have been recorded at the Hôpital Européen Georges Pompidou in Paris via an acoustic microphone only.

E. Validation

The reliability of the tracking of the cycle length time series via speech peak saliences has been tested by means of the modal speech signals, their numerical derivatives and integrals as well as the co-recorded throat microphone signals. The four time series that are so obtained for each speaker are expected to be quasi-identical given that they report the same glottal cyclicity via four different signals. The agreement between the four time series is evaluated by means of their inter-correlation.

In addition, the low-frequency spectra of the four time series have been obtained and inter-correlated. The reason is that in a later experiment, the tremor frequency is estimated based on the low-frequency spectrum ([0-15Hz] or [3-15Hz]) of the cycle duration time series.

F. Vocal cues

One acoustic cue is the abscissa of the center of gravity of the low-frequency spectrum of the cycle length series. Two frequency intervals ([0-15Hz] or [3-15Hz]) have been considered. Indeed, cardiac beat, breathing and bloodflow are expected to influence strongly the spectrum below $3Hz$. The second cue is the coefficient of variation of the

cycle length series, which characterizes the excursion of the cycle durations with respect to the average. These two cues are rough estimates of respectively the modulation frequency and the modulation depth of vocal tremor.

III. RESULTS

A. Experiment 1

The cycle length series have been estimated for the corpus of normophonic speakers. Visual inspection of the cycle duration time series (Fig. 4) shows that the speech cycles are correctly discovered. The event sequence and signal polarity that have been retained are generally different for the four time series. Even so, one observes a good agreement between the four time series reporting glottal cycle durations. In Fig. 4, one also observes a quasi-perfect agreement between the time series that have been obtained from the acoustic and throat microphone signals. This is however not true in general.

Tables I(a) and I(b) show the inter-correlation coefficients for the length time series and low-frequency spectra.

Table I : Inter-correlation coefficients for cycle length time series (a) and low-frequency spectra (b) obtained via the acoustic speech signal (1), its derivative (2) and integral (3) as well as the throat microphone signal (4) for each speaker i.

(a) Cycle length time series

i	F_0	Correlation coefficient					
		1-2	1-3	1-4	2-3	2-4	3-4
1	90	0.99	0.99	0.99	0.99	0.99	0.99
2	114	0.99	0.99	0.98	0.98	0.98	0.98
3	116	0.98	0.98	0.96	0.98	0.97	0.97
4	120	0.95	0.97	0.89	0.96	0.86	0.89
5	212	0.96	0.96	0.71	0.97	0.73	0.76
6	228	0.99	0.99	0.96	0.99	0.96	0.96
7	241	0.98	0.97	0.87	0.97	0.87	0.85
8	244	0.98	0.55	0.95	0.60	0.93	0.70

(b) Low-frequency spectra $(0 - 15Hz)$

i	F_0	Correlation coefficient					
		1-2	1-3	1-4	2-3	2-4	3-4
1	90	0.99	0.99	0.99	0.99	0.99	0.99
2	114	0.99	0.99	0.99	0.99	0.99	0.99
3	116	0.99	0.99	0.99	0.99	0.99	0.99
4	120	0.99	0.99	0.99	0.99	0.98	0.99
5	212	0.99	0.99	0.98	0.99	0.98	0.98
6	228	0.99	0.99	0.97	0.99	0.97	0.97
7	241	0.99	0.99	0.99	0.99	0.99	0.99
8	244	0.99	0.99	0.99	0.99	0.99	0.99

B. Experiment 2

The cycle duration extraction has been applied to the corpus of dysphonic speakers. In Table II one observes that the low-frequency cycle duration perturbations are higher for the ALS speakers. The increase of the tremor frequency is small, however.

Table II : Estimates of the modulation frequency (center of gravity) and modulation depth (coefficient of variation) of vocal tremor for normophonic and ALS speakers

(a) Normophonic speakers

	Center of gravity (Hz)		C.V.(%)
	$[0 - 15Hz]$	$[3 - 15Hz]$	
Minimum	3.88	6.67	0.53
First quartile	4.47	6.91	0.75
Median	5.03	7.27	0.81
Third quartile	5.59	7.50	0.93
Maximum	6.01	7.62	1.12

(b) ALS speakers

	Center of gravity (Hz)		C.V.(%)
	$[0 - 15Hz]$	$[3 - 15Hz]$	
Minimum	2.97	5.97	0.75
First quartile	4.24	7.24	1.19
Median	4.64	7.69	1.69
Third quartile	5.45	8.17	2.70
Maximum	6.77	9.10	4.52

IV. DISCUSSION AND CONCLUSION

A temporal method for the tracking of cycle lengths in voiced speech has been proposed. It is based on the speech sample saliences and does not request that the signal is locally periodic and the average cycle length known a priori. The good agreement between the four time series that have been obtained for each speaker suggests that salience analyses may be able to track reliably glottal cycles in the speech signal.

One additional condition, which is not developed in the text, is that from one cycle to the next, the maximum salience is not reaffiliated from the main to a previously subordinated peak or vice versa. The extraction of several candidate cycle duration time series and the selection of the least perturbed one is likely to discard any time series that has been affected by peak reaffiliation. This cannot be mathematically guaranteed however.

Salience analysis has been applied to tracking cycle duration in the voices of dysphonic speakers affected by a neurological disease. Results suggest that SLA speakers are characterized by higher vocal tremor depths and frequencies than normophonic speakers.

REFERENCES

[1] C.Mertens, F.Grenez, and J.Schoentgen, "Preliminary evaluation of speech sample salience analysis for speech cycle detection," in *Proceedings 3rd International Workshop on Advanced Voice Function Assessment, Madrid*, 2009, pp. 29–32.

[2] C.Mertens, F.Grenez, and J.Schoentgen, "Speech sample salience analysis for speech cycle detection," in *Proceedings 10th Annual Conference of the International Speech Communication Association INTERSPEECH, Brighton*, 2009.

TEMPORAL MEASURES OF THE INITIAL PHASE OF VOCAL FOLD OPENING ACROSS DIFFERENT PHONATION TYPES

P. J. Murphy

Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland.
Peter.Murphy@ul.ie

Abstract: **The present contribution introduces three temporal measures of vocal fold opening – as indicated by the time of decreasing contact of the vocal folds estimated from the electroglottogram signal. The sustained vowel [a:], produced when simulating the phonation types very pressed, pressed, neutral, strained (hyperfunctional) breathy and (hypofunctional) breathy, is analysed. The results indicate discrimination of phonation type along the adduction dimension for each of the measures of vocal fold opening duration.**

Keywords : **vocal fold opening, phonation type, voice quality, electroglottography**

I. INTRODUCTION

Quantitative characterization of phonation type is useful for categorization of voice quality [1] (including pathological voice) and for improving naturalness in speech synthesis. Furthermore, studies of (indicators of) phonation type in conjunction with vocal loading studies, which document resulting pathology and/or impressions of vocal fatigue, may help to set guidelines for the avoidance of inappropriate and potentially harmful use of the phonatory mechanism.

The present contribution introduces three measurements of vocal fold opening as estimated via the electroglottogram signal (section II) and tests the behaviour of these indices for different phonation types (section III).

II. METHODS

A Electroglottograph

The electroglottograph comprises two electrodes placed external to the larynx. During use a high frequency current passes between the electrodes and the output signal varies depending on the impedance of the path between the electrodes. As the vocal folds vibrate they move through levels of high impedance (open glottis) to low impedance (closed glottis). As the impedance decreases (conductance increases) with contact the electroglottogram (EGG) signal provides a measure of vocal fold contact [2] (top row Fig. 1). The electroglottogram provides complimentary information to the glottal flow waveform; the maximum in the

electroglottogram occurs when contact is maximum (glottal flow is minimum or zero) while the maximum in the glottal flow occurs during the open phase (when the EGG amplitude is minimum or zero). A possible advantage of using the EGG signal is that it is less affected by supra-glottal acoustic influences, which can produce source–filter interaction, making glottal flow determination and interpretation challenging.
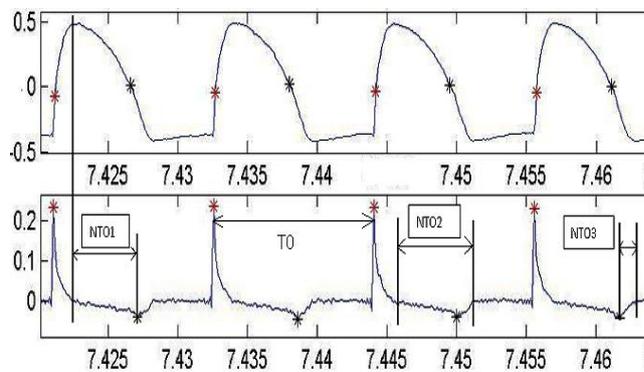
Fig. 1 Upper row - EGG signal (neutral phonation), breathy [x-axis indicates sample number (or time) and y-axis (arbitrary scale) indicates amplitude (amount of contact)]
Lower row – 1st derivative of EGG signal (DEGG), normalized time of vocal fold opening (NTO) 1, 2 and 3 are shown in the lower row [x-axis indicates sample number (or time) and y-axis (arbitrary scale) indicates rate of change of contact]

B Measures of Vocal Fold Opening Time

Fig. 1 and Table 1 indicate how the measures cycle duration (and hence fundamental frequency) and the vocal fold opening times (NTO1, 2 and 3) are estimated. Comparison of vocal fold vibration images with EGG signals (cf. [3]) illustrates that the minimum point in the DEGG signal corresponds with the point of glottal opening. NT01 is a measure of the time of decreasing contact of the vocal folds from the point of maximum contact to the point where the glottis opens. The final phase of opening corresponds to where the glottis is open but the folds are still in contact in certain places along the length and depth of their structure.

NT01 is estimated by determining the time between the positive peak of the EGG and the negative peak of the DEGG signals, per cycle and dividing by the cycle duration. NT02 also uses the EGG positive peak as its starting point and the first zero-crossing after the negative

peak in the DEGG signal is taken as the end of the contact phase. NT03 is estimated as the difference between these times (NTO3=NTO2-NTO1). NT01 is an indicator of vocal fold opening during the glottal closed phase while NT03 provides a measure of opening that takes place during the glottal open phase. NT02 provides a measure of vocal fold opening in its entirety as represented by vocal fold contact via the EGG signal (the EGG signal on its own does not provide detailed information on vocal fold characteristics during the advanced stages of glottal opening – during this region the glottal flow signal or the photoglottogram (PGG), which provides an approximation of the projected glottal area, can provide additional information).

Table 1 Opening time measures estimated from the electroglottogram (EGG) and the first derivative of the electroglottogram (DEGG) signals

Measure Symbol	Description of Measure	Measurement Method
T_0	glottal cycle duration	measured between positive peaks in the DEGG signal (points of glottal closure)
NTO1	normalised vocal fold opening time 1 – vocal fold opening duration during the closed phase	opening time 1 is measured from the peak in the EGG signal to the negative peak in the DEGG signal – dividing by the cycle duration provides the normalised index
NTO2	normalised vocal fold opening time 2 – vocal fold opening duration in its entirety (as estimated from the EGG signal)	opening time 2 is measured from the peak in the EGG signal to the next zero crossing of the DEGG signal – dividing by the cycle duration provides the normalised index
NTO3	normalised vocal fold opening time 3 – vocal fold opening duration (as estimated from the EGG signal) during the open phase	opening time 3 is the difference between NTO2 and NTO1

C Recording

The EGG and speech signals were recorded in a sound treated room in the Department of Speech Communication and Voice Research, University of Tampere, Tampere, Finland. The sustained vowel [a:] was phonated while simulating the phonation types very pressed, pressed, neutral, strained (hyperfunctional) breathy and (hypofunctional) breathy by a single female speaker experienced in portrayals of phonation type.

Thirty cycles of a steady portion of the vowel were selected for analysis.
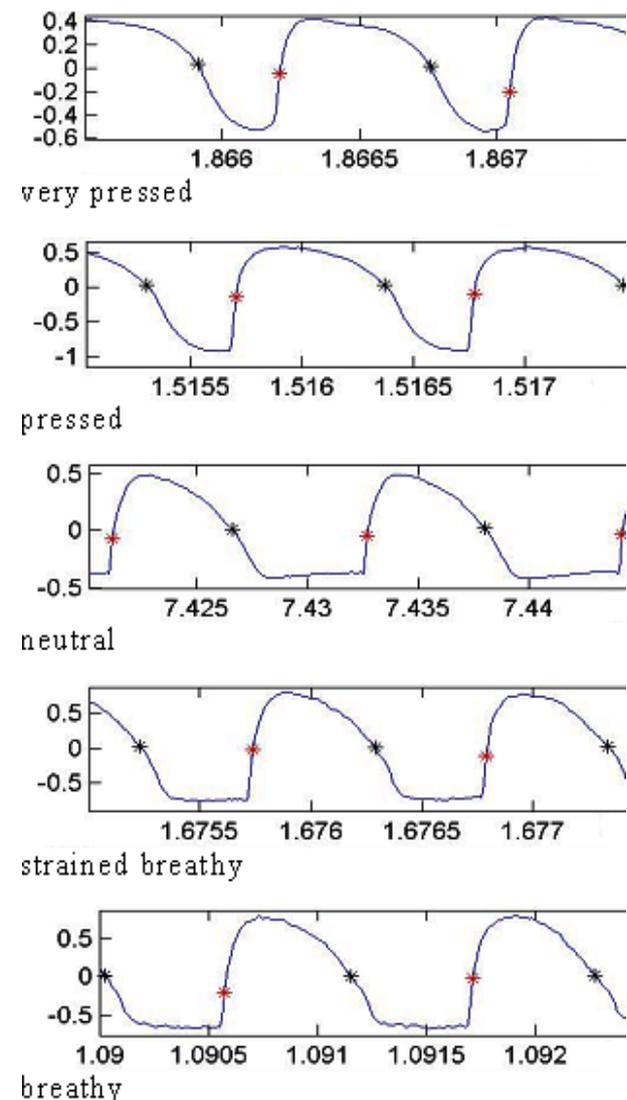
Fig. 2 Electroglottogram (EGG) displays produced for the vowel [a:] for five phonation types: 1st row (top) - very pressed, 2nd row - pressed, 3rd row - neutral, 4th row - strained (or hyperfunctional) breathy, 5th row - (hypofunctional) breathy [x-axis indicates sample number (or time) and y-axis (arbitrary scale) indicates amplitude (amount of contact)]

III. RESULTS

Fig. 3 shows fundamental frequency (f0) versus cycle number for the phonation types very pressed, pressed, neutral, strained (hyperfunctional) breathy and (hypofunctional) breathy. Fig. 4, 5 and 6 show NTO1, 2 and 3, respectively, versus cycle number for the five phonation types. Average values are provided in Table 2.
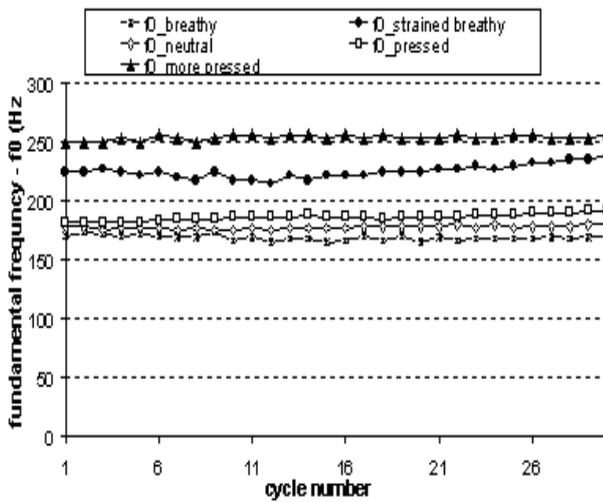
Fig. 3 Fundamental frequency (f0) for the five phonation types
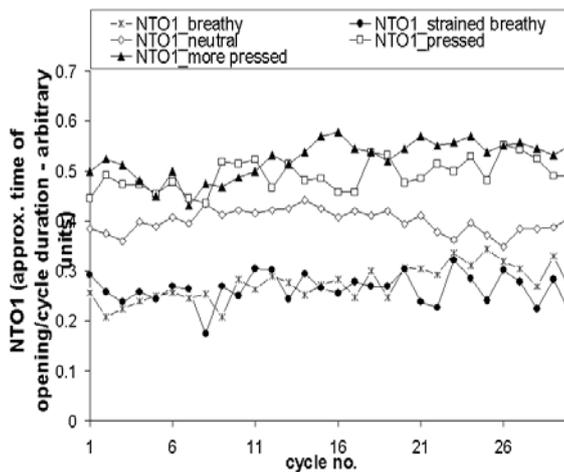
Fig. 4 Normalised time of opening 1 (NTO1) for the five phonation types
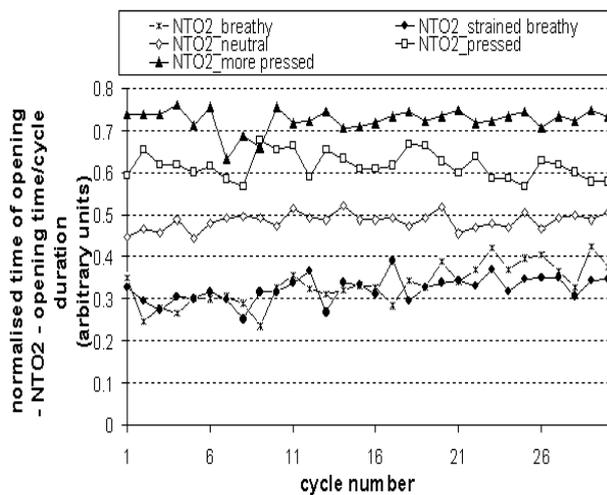
Fig. 5 Normalised time of opening 2 (NTO2) for the five phonation types
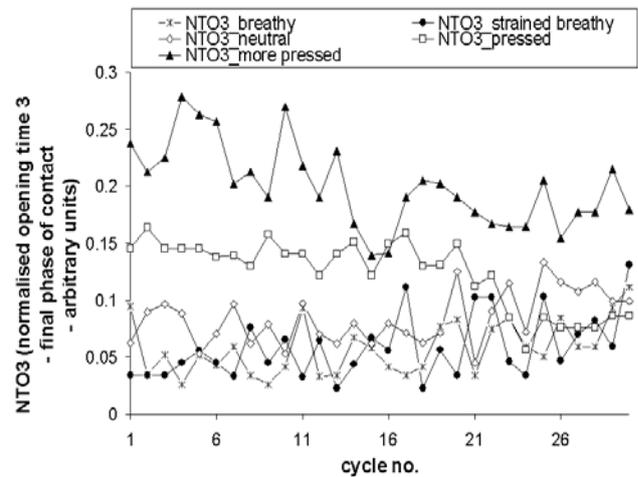
Fig. 6 Normalised time of opening 3 (NTO3) for the five phonation types

Table 2 Electroglottogram based fundamental frequency and glottal opening time measures averaged over 30 glottal cycles

Measure /Phonati -on Type	Mean Fundam -ental Frequen -cy (f0)	Mean Normali- sed Time of Glottal Opening 1 (NTO1)	Mean Normali- sed Time of Glottal Opening 2 (NTO2)	Mean Normali- sed Time of Glottal Opening 3 (NTO3)
Very Pressed	254	0.52	0.72	0.200
Pressed	187	0.49	0.62	0.123
Neutral	178	0.40	0.48	0.084
Strained (Hyperfu -nctional) Breathy	225	0.26	0.32	0.058
Hypofun- ctional Breathy	169	0.27	0.33	0.058

IV. Discussion

Normalised time of vocal fold opening shows discriminatory ability of phonation type along the adduction dimension. Pressed voice is associated with a higher than normal level of adduction, while breathy voice is associated with an abducted vocal fold configuration (negative adduction) [4]. The NTO1, 2 and measures differentiate the data along this adduction dimension; the averaged data are ordered as very pressed>pressed>neutral>breathy. The opening times are greater for the adducted configuration (compared to neutral) as the centre of mass of each fold is closer at closure. Conversely the opening times are less for the

abducted configuration as the centre of mass of each fold is further apart at closure (if a closed phase exists). Further information regarding vocal fold opening during the glottal open phase can be supplied using the inverse filtered flow signal or the PGG signal.

V. CONCLUSION

The analysis of the EGG signal for the vowel [a:] suggests that measures of vocal fold opening time are useful for the discrimination of phonation types. The three measures of opening; the initial opening phase (NTO1), the entire opening phase (NT02) and the final phase of opening (NTO3) discriminate the data along the adduction dimension with values ordered as follows; very pressed>pressed>neutral>breathy (hyperfunctional and hyopfunctional). Future work will examine these and other glottal measures with data from a number of speakers in combination with additional measurement modalities.

VI. ACKNOWLEDGEMENTS

REFERENCES

[1] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers," *Folia Phoniatr Logop* vol. **48**, pp. 240–254, 1996.
[2] A. Fourcin, E. Abberton, D. Miller, and D. Howells, "Laryngograph: Speech pattern element tools for therapy, training and assessment," *Eur. J. Disord. Commun*. vol. 30, pp. 101-115, 1995.
[3] 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.
[4] M. Rothenberg and J.J. Mashie, "Monitoring vocal fold abduction through vocal fold contact area," *J. Speech Hear Res. vol.* 31, pp. 338-351, 1988.

A NOVEL METHOD FOR THE EXTRACTION OF VOCAL TREMOR

Yannis Pantazis, Maria Koutsogiannaki and Yannis Stylianou

Computer Science Dept, University of Crete and Institute of Computer Science, FORTH, Crete, Greece

Abstract: **Vocal tremor is defined as slow modulation of fundamental frequency or its amplitude [1, 2]. Even though vocal tremor may be attributed to neurological diseases, it may also be a natural stochastic modulation of voice. Many studies try to measure these modulations assuming that they are stationary. Hence, their analysis is limited to small intervals loosing important information about vocal tremor. We propose a novel method for the estimation of the modulations which is able to adapt to nonstationary environments. The method is mainly based on an AM-FM signal decomposition algorithm which is able to estimate the instantaneous components of speech signals. Results confirm that the method successfully extract the modulations of large speech segments and robustly estimate the time-varying modulation frequency and the time-varying modulation level of vocal tremor.**
Keywords: Voice quality, Vocal tremor, AM-FM decomposition

I. INTRODUCTION

Typically, tremor in phonation is defined as modulations of the fundamental frequency and modulations of the amplitude due to the inability of humans to keep constant the tension of their vocal folds [3]. This phenomenon affects the glottal cycle in voiced speech making the fundamental frequency and the amplitude to vary stochastically. Vocal tremor is usually categorized into the physiological tremor which is a slow natural modulation of glottal cycle and the pathological tremor which is attributed to neurological diseases such as Parkinson or tremor of the limbs [4], [2]. Most importantly, while physiological tremor makes speech sound more natural and possibly more individual, pathological tremor may influence the quality of patients voice, hence, may influence the ability of patient's communication.

Moreover, while pathological tremor is characterized by stronger periodic patterns –a property that vibrato singing style has, too–, physiological tremor is more stochastic [4]. The analysis of physiological tremor is of great importance since vocal tremor in normophonic speakers may be an early sign of a neurological disease [5], [6]. Thus, it is useful to develop an estimation algorithm that is able to measure or extract the vocal tremor even for normal voices. In the literature, acoustic analysis of tremor is usually based on the accurate estimation of fundamental frequency and then the characterization of the variations of fundamental frequency [1], [2]. Modulation frequency and modulation level are prominent attributes that are extracted from the instantaneous fundamental frequency [1], [2].

However, there are some issues not addressed in previous studies. Indeed, many studies are intersted only for the 1st harmonic which is related with the fundamental frequency but not for the higher harmonics. But 1st harmonic may be modulated by first formant which may lead to biased results. A more serious limitation of the previous studies is that the analyzed sustained vowel has duration that is one to two seconds. The reason for using short duration is that the modulation frequencies as well the modulation levels should be constant in order to apply classical frequency estimation analysis. This is a real drawback since the analysis of larger segments of speech may show interesting properties on vocal tremor [7], [8].

The objective of this paper is to present and validate a novel method for the estimation of the vocal tremor on sustained vowels uttered by normophonic subjects. The proposed method assumes speech as a sum of time-varying sinusoids whose instantaneous amplitude and instantaneous frequency are estimated using a recently proposed AM-FM decomposition algorithm [9], [10]. The prime advantage of this algorithm is its ability to demodulate multicomponent signals (like speech) very accurately. Interestingly, any of the instantaneous components can be used for the analysis of vocal tremor and not only the 1st harmonic. Then, the second step of the algorithm is to subtract from the analyzed instantaneous component the very slow modulations ($< 2Hz$) in order to reveal the higher frequency modulations. This is achieved by filtering the instantaneous component using a Savinzky-Golay smoothing filter [11]. The final step is to estimate the modulation frequency and the modulation level which now are time-varying attributes because the modulations are primarily nonstationary. The estimation is performed using the same AM-FM decomposition algorithm applied for the extraction of instantaneous components. Results on sustained vowels uttered by normophonic speakers showed that the proposed method accurately estimate the instantaneous components of speech signals and then robustly extract the time-varying modulation frequency and modulation level.

The organization of the paper is as follows. Section II

presents the tremor analysis method while in Section III the results are shown. Finally, Section IV concludes the paper.

II. ESTIMATION OF VOCAL TREMOR

In this paper, we will consider that speech signals can be effectively modeled as a sum of time-varying sinusoidal components

$$s(t) = \sum_{k=1}^{K} a_k(t)cos(\phi_k(t)) \tag{1}$$

where K is the number of components, while $a_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude and instantaneous phase of the kth component, respectively. Moreover, instantaneous frequency, $f_k(t)$ is defined as the first derivative over time of the instantaneous phase:

$$f_k(t) = \frac{1}{2\pi}\frac{d\phi_k(t)}{dt} \tag{2}$$

In order to extract the characteristics of vocal tremor, the first crucial step is the estimation of the instantaneous components, $\{a_k(t), f_k(t)\}_{k=1}^{K}$ from the speech signal. The second step is to estimate and then remove the very slow modulation from the analyzed instantaneous component. The third and final step is the extraction of the modulation frequency and modulation level which are the vocal tremor characteristics we are interested. We will now describe each step in details.

A. Step 1: Estimation of Instantaneous Components

The estimation of the instantaneous components is achieved by an AM-FM decomposition algorithm recently proposed by Pantazis et al. [9], [10]. The AM-FM decomposition algorithm (aQHM in abbreviation) is an adaptive algorithm which is based on a time-varying sinusoidal model. The sinusoidal model is called quasi-harmonic model (QHM) and its parameters are estimated frame-by-frame through linear Least Square method. The prominent property of QHM is its ability to capture and then correct frequency estimation errors even if the analysis was intended to be realized using wrong frequencies. Thus, QHM is able to estimate the frequency mismatches and iteratively eliminate them.

The initialization of aQHM algorithm necessitates a rough estimate of the analysis frequencies for the first frame. During this study, the initial frequencies were assigned as integer multiples of an estimated fundamental frequency computed using the autocorrelation function of the first frame [12]. Moreover, time resolution of aQHM is determined by the hop-size of the algorithm while frequency resolution is determined by the window type and window length. We choose hop-size of $5ms$ and Hamming window as window function. The window's duration was chosen to be three times the period of the smaller frequency, i.e. three times the pitch period.

The main advantage of aQHM algorithm is its ability to adapt to the signals characteristics. Indeed, after the first pass of the signal with QHM, an estimate for the instantaneous frequencies and instantaneous amplitudes, $\{\hat{a}_k(t), \hat{f}_k(t)\}_{k=1}^{K}$, are obtained. Then, in the following passes, aQHM algorithm adapts the QHM basis functions using the estimated instantaneous frequencies. Thus, the bias due to the nonstationarity of the AM-FM signal is reduced and more accurate estimates for the instantaneous components are obtained. As an example, Fig. 1 shows the five first estimated instantaneous frequencies of a sustained vowel uttered by a male speaker using the aQHM algorithm.

B. Step 2: Removal of Very Slow Modulations

After choosing which instantaneous component will be analyzed, the second step of the analysis is to eliminate modulations which are less than $2Hz$. The removal of the trend is necessary in order to reveal the quasi-periodical modulations attributed to vocal tremor (compare Fig. 2a before elimination and Fig. 3a after elimination). However, before the removal of the trend, as a preprocessing step, we downsample the instantaneous component to have sampling frequency 1000 Hz. Indeed, since we are interested for modulations which are less than $20Hz$, the downsampled instantaneous component do not miss any important information.

The smoothing of the instantaneous component is performed using the Savinzky-Golay (S-G) filter [11], [13]. S-G smoothing filter essentially performs a local polynomial regression on a distribution of equally spaced points to determine the smoothed value for each point. The main advantage of this approach is that it tends to preserve features of the distribution such as relative maxima, minima and width, which are usually "flattened" by other adjacent averaging techniques like moving averages. The order of the local polynomial used in this study was 4 while the frame size was set to $1s$ (1000 samples). Fig. 2a shows the instantaneous component as well its smoothed version for a sustained vowel. Fig. 2b implies that S-G filter captures the frequencies that are less than $2Hz$. Then, the smoothed instantaneous component is subtracted from the unsmoothed in order to reveal the remaining modulations of the component. Note that using different parameters for the S-G filter the smoothed signal will capture more or less of the signal's frequencies.

C. Step 2: Extracting Vocal Tremor Characteristics

The final step is the modeling and estimation of the remaining modulations. As already stated, these modulations are nonstationary, hence, FFT-based approaches are not appropriate for this task. We suggest modelling the remaining nonstationary modulations as an amplitude mod-

ulated and frequency modulated signal. Mathematically, it is given by

$$x(t) = m(t)cos(\psi(t)) \tag{3}$$

where $x(t)$ are the remaining modulations of the instantaneous components, $m(t)$ is the instantaneous amplitude which with the appropriate scaling corresponds to the modulation level, and $\psi(t)$ corresponds to the instantaneous phase. Once again, instantaneous frequency is given by $\zeta(t) = \frac{1}{2\pi}\frac{d\psi(t)}{dt}$ and corresponds to the modulation frequency.

aQHM algorithm is applied a second time for the estimation of the instantaneous components, $m(t)$ and $\zeta(t)$. The initial frequency of the first frame was computed by the largest peak of the FFT of the first frame while hopsize was set to $1ms$. Hamming window was used and its duration was set to $0.6s$. Fig. 3a shows the reconstructed signal obtained from the aQHM algorithm, while Fig. 4 shows the estimated modulation frequency and estimated modulation level.

IIII. RESULTS

In this section, the output of the proposed method for vocal tremor analysis is presented for normophonic speakers. The method is validated on a database of normal voices developed in our recording lab. 11 male and 5 female healthy subjects whose age varies between 23 and 45 were participated. Sustained vowels $/a/, /e/, /i/, /o/$ and $/ou/$ have been recorded at $48kHz$ and then downsampled at $16kHz$. The duration of sustained vowels varies from $2s$ to $8s$ depending primarily on the speaker.

Illustratively, Fig. 1 shows the first five harmonics extracted from sustained vowel $/a/$ using aQHM algorithm (Step 1). The signal which is reconstructed from the instantaneous components has signal-to-reconstruction error of about $32dB$ which proves that the analysis is very accurate. For the total database, the average signal-to-reconstruction error was more than $30dB$. Fig. 1 shows also that the modulations of higher harmonics are more evident which explains the use of normalization for the estimation of modulation level.

Fig. 2 shows the instantaneous frequency of the 1st harmonic after removing its mean value and its filtered version using the S-G smoothing filter (Step 2). The smoothed instantaneous component contains information about the frequencies which are less than 2Hz. This component is then removed in order to reveal the modulations that are attributed to vocal tremor. Thus, the remaining component is analyzed using aQHM algorithm (Step 3). Fig. 3 indicates that the decomposition algorithm adapts to the nonstationary modulations of the signal. Extended tests on the database confirmed the ability of aQHM to adapt to the signal. The extracted time-varying modulation frequency as well the extracted time-varying modulation level
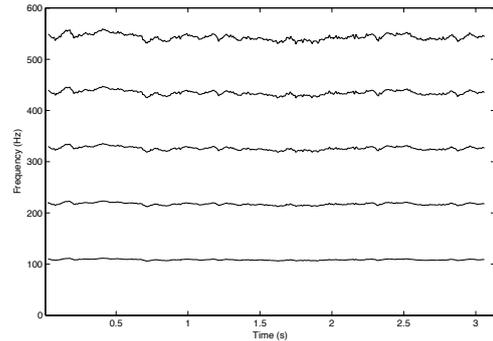
Figure 1: First five instantaneous frequencies of a normophonic male speaker uttered the sustained vowel /a/.

are shown in Fig. 4. Modulation frequency takes values in this example between $2Hz$ and $13Hz$.

Table I reports the averages of fundamental frequency, $\mu(f_0)$, of mean value, $\mu(mf)$, and standard deviation, $\sigma(mf)$, of modulation frequency and of mean value, $\mu(ml)$, and standard deviation, $\sigma(ml)$, of modulation level for male and female speakers uttering various vowels. Visual inspection shows that the standard deviation of modulation frequency is higher for the male speakers while the mean value of modulation frequency shows no tendency between the genders.
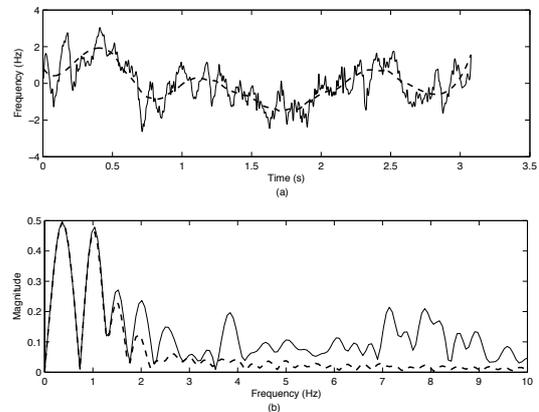
Figure 2: (a) First harmonic of Fig. 1 without its mean value (continuous line) and its smoothed version (dashed line) are shown. (b) Fourier transform of signals in (a). S-G smoothing filter captures the frequencies that are below $2Hz$.

IV. CONCLUSION & FUTURE WORK

A novel method for the acoustical analysis of vocal tremor was presented. It is based on a AM-FM demodulation algorithm which is used for the extraction of both instantaneous amplitudes and instantaneous frequencies
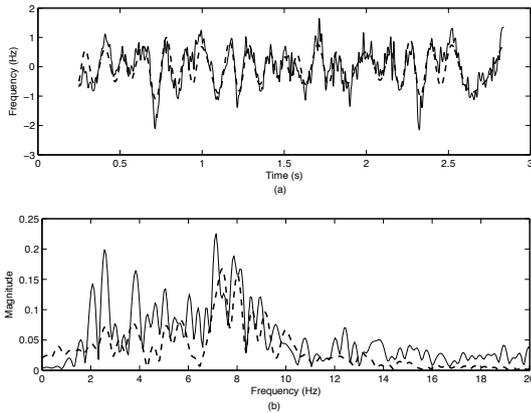
Figure 3: (a) Instantaneous component after subtracting its smoothed version (continuous line) and the reconstruction of the AM-FM decomposition algorithm (dashed line). (b) Fourier transforms of the components in (a).
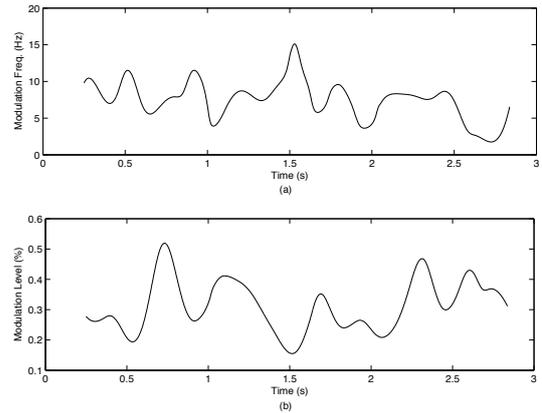
Figure 4: (a) Modulation frequency of the signal in Fig. 3. (b) Modulation level of the same signal. Note that neither modulation frequency nor modulation level have constant values during the phonation.

Table 1: Summary of modulation features for five vowels and both genders.

		$\mu(f_0)$ (Hz)	$\mu(mf)$ (Hz)	$\sigma(mf)$ (Hz)	$\mu(ml)$ (%)	$\sigma(ml)$ (%)
Male	/a/	113	4.4	1.4	0.25	0.11
	/e/	116	4.3	1.2	0.28	0.13
	/i/	119	4.1	1.3	0.25	0.11
	/o/	121	6.2	1.7	0.22	0.09
	/ou/	122	8.0	2.0	0.20	0.08
Female	/a/	233	6.6	0.9	0.36	0.14
	/e/	228	9.3	0.9	0.33	0.14
	/i/	239	3.1	0.8	0.29	0.12
	/o/	235	4.7	0.9	0.27	0.10
	/ou/	236	3.4	0.8	0.27	0.10

from the speech signal (Step 1) and the extraction of modulation frequency and modulation level from the analyzed instantaneous component (Step 3). Results indicate that the proposed method is cabable of handling large segments of sustained vowels where the assumption of modulations' stationarity is invalid and provide robust time-varying estimates for the modulation frequency and the modulation level.

Finally, while the proposed method was validated only on normophonic subjects it is of great importance to apply and test it in pathological subjects. Future work will be devoted on analyzing pathological vocal tremor and possibly on other applications such as the analysis of vibrato singing style where the objective is to achieve a particular amount of modulation frequency and/or modulation level.

REFERENCES

[1] W. Winholtz and L. Ramig. Vocal Tremor Analysis with the Vocal Demodulator. 35:562–573, 1992.

[2] J. Schoentgen. Modulation Frequency and Modulation Level owing to Vocal Microtremor. *J. Acoust. Soc. Am.*, pages 690–700, Aug 2002.

[3] I. R. Titze. Motor and Sensory Components of a Feedback Control Model of Fundamental Frequency. *Producing Speech: Contemporaty Issues*, pages 309–320, 1995.

[4] H.J. Freund. *Central Rhythmicities in Moter Control and its Pertubances*, pages 79–82. Springer, Berlin, 1987.

[5] C.A. Meeuwis and E.A. Baarsma. Essential (Voice) Tremor. *Clinical Otolaryngology*, 5, 1985.

[6] L.J. Findley and M.A. Gresty. *Head, Face and Voice Tremor*, pages 239–253. New York: Raven, 1988.

[7] J. Kreiman, B. Gabelman, and B.R. Gerratt. Perception of Vocal Tremor. *Journal of Speech, Language and Hearing Research*, 46:203–214, 2003.

[8] H. Ackermann and W. Zeigler. Acoustic Analysis of Vocal Instability in Cerebellar Dysfunctions. *Annals of Otology, Rhinology and Laryngology*, 103:98–104, 1994.

[9] Y. Pantazis, O. Rosec, and Y. Stylianou. AM-FM Estimation for Speech based on a Time-varying Sinusoidal Model. In *Interspeech*, Brighton, 2009.

[10] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AM-FM Signal Decomposition with Application to Speech Analysis. *IEEE Trans. on Audio, Speech and Language Processing*, submitted.

[11] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[12] T. F. Quatieri. *Speech Signal Processing*. Prentice Hall, Signal Processing Series, 2002.

[13] J. Steinier, Y. Termonia, and J. Deltour. Comments on smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, 44:1906–1909, 1972.

INDIRECT ESTIMATION OF FORMANT FREQUENCIES THROUGH MEAN SPECTRAL VARIANCE WITH APPLICATION TO AUTOMATIC GENDER RECOGNITION

U. K. Laine[1], O. J. Räsänen[1]

[1]Helsinki University of Technology, Department of Signal Processing and Acoustics, Espoo, Finland

Abstract: **A novel approach for estimation of speaker specific vocal tract properties is presented in this paper. Instead of using the well-known long-term average spectrum (LTAS) of speech, it is shown that the variance of the magnitude of the spectrum in each band is also suitable for estimation of formant frequencies. This representation, called mean spectral variance (MSV), is applied to an automatic gender classification task, where it is shown to achieve good classification accuracy in combination with the fundamental frequency of speech. The MSV is compared with LTAS and their similarities and differences are discussed.**

Keywords: **Formant estimation, gender classification, long-term feature averaging**

I. INTRODUCTION

Speaker dependent variability in vocal apparatus properties has a notable impact on the acoustic properties of speech signals. Cross-speaker variation in characteristic formant frequencies poses a difficult challenge for speech processing systems designed to work independently of speaker identity, while it also plays an important role in speaker identity detection [1] and gender classification (e.g., [2]).

One possible approach for analyzing speaker and gender specific properties of the vocal apparatus is through long-term averaging of the acoustic parameters [3]. The long-term average spectrum (LTAS) has been widely studied in speaker recognition, and although its performance falls behind state-of-the-art Gaussian-mixture models (GMM) using Mel-cepstral coefficients (MFCCs), the computational simplicity of LTAS is appealing for many applications [4-5]. In addition to LTAS, averaging of, e.g., autocorrelation-, LPC-, cepstral-, and reflection coefficients, have also been studied [6].

However, all these studies have concentrated on the averaging of feature vectors per se, but none to our knowledge have studied modeling of feature variance in isolation of the spectral mean. In this paper we show that instead of utilizing the long-term spectrum directly, the spectral variability of speech signals also reflects the speaker and gender specific average formant structure. For estimation of speaker specific acoustic parameters, we introduce a straightforward method for estimating average formant frequencies (AFF) indirectly from continuous speech. More specifically, we show that the AFFs can be easily obtained by computing the mean spectral variance (MSV) separately for each frequency band during voiced speech. The basic idea behind our method is simple; while each formant is moving mainly around its mean value these movements should cause the largest spectral variance to occur around the mean as well.

The MSV representation is compared to the well-known LTAS, and it is shown that the methods contain complementary information regarding speech signals. The general quality and usability of the MSV method is assessed in a classification task where MSV templates and pitch of the speaker are combined as cues to perform automatic gender detection.

II. METHODS

A. Computation of mean spectral variance (MSV)

The speech signal (f_s = 16 kHz) is first pre-emphasized with a standard 1st order FIR-filter. Voicing is estimated using standard cepstral analysis and only voiced frames are preserved for further analysis. The signal is then windowed using a 6 ms Hamming window with 2 ms window shifts. The small window length causes the absence of pitch periodicity in spectral representations and leads to regularly good matches between window position and the maximal excitation of vocal tract resonances during glottal closure. Spectral tilt and mean are removed from each frame by fitting a line to the spectrum and the frames are normalized into unit vectors. All spectral frames are collected into a spectrogram and the mean spectral variance for each frequency band is

computed over the entire set of frames to produce the MSV representation. The tilt and mean of the MSV are removed and then this vector is normalized to a unit vector. In addition to MSV, the long-term average spectrum (LTAS) is extracted from the speech material. The procedure for LTAS is identical to MSV except that the mean of the spectrum is taken over the spectrogram instead of the variance.

Figures 1 and 2 illustrate the LTAS and MSV representations computed over several speakers from the TIMIT corpus. The average, gender specific, formant structure is readily seen. The AFF estimates provided by both methods are relatively close to each other as predicted. Two general observations can be made; first, both genders most actively utilize the frequency band of 300-3400 Hz that was historically selected to be the band of analog telephone systems (see Fig. 1 bottom frame), and secondly, the shape of MSV between genders is very contrastive in the 1000-5000 Hz frequency band.

B. Automatic gender detection based on formant structure and pitch

There are notable structural differences in the vocal tracts for men and women, and therefore the average formant information can be utilized for automatic detection of speaker gender (e.g., [7]). In addition, vocal fold structure can be considered as at least partially independent of vocal tract length (cf., e.g., source-filter modeling), and it also serves as a reliable cue to gender identity. Therefore the mean pitch of a speaker is also utilized in the recognition process.

In the training of the recognizer, MSV vectors v_m and v_f are computed across several speakers from the TIMIT training set ($N = 560$ for both genders) in order to estimate the average male and female spectral structures with formant peaks. The common mean $v_c=(v_m+v_f)/2$ of the vectors is subtracted from both v_m and v_f in order to maximize contrast:

$$v'_g = v_g - v_c \qquad (1)$$

Finally, the obtained templates are normalized to unit vectors.

Only variation in the frequency band of 1000-5166 Hz is used for recognition, since it was found to lead to maximal performance. The use of this frequency band is also in line with the work of Mendoza et al. [5], who performed a statistical discriminant analysis of male and female voices and found that gender specific differences in LTAS are concentrated in the frequency region of 0.8 – 5 kHz.
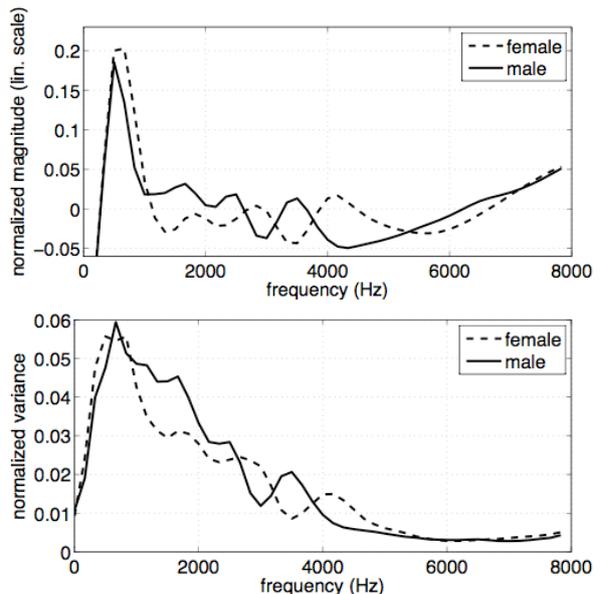
Figure 1: Average LTAS (top) and MSV (bottom) according to gender from the TIMIT training corpus.
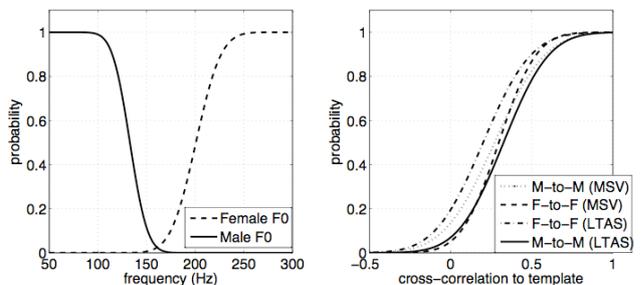
Figure 2: Gender specific cumulative probability distributions for pitch (left) and MSV & LTAS (right).

Once the template vectors for both genders are created, the training set is processed again and the distance d_g between MSV of the analyzed utterance and the templates is measured by cross-correlation. Distributions of d_g values from male utterances to the male template and female utterances to the female template are modeled as a cumulative normal distribution (fig. 2, right). Pitch is also modeled for both genders as two separate cumulative Gaussian distributions estimated from the training data (Fig. 2, left).

In the classification phase, MSV is computed from the input utterance according to section 2.A and vector v_c is again subtracted from the representation. The mean pitch of the utterance is also extracted. Ultimately, the probability for a gender is estimated using the trained probability distributions and by assuming the independence of probabilities:

$$P(gender) = P(f_0|gender) * P(d_g|gender) \qquad (2)$$

where f_0 is the mean pitch of the utterance and d_g is the cross-correlation between the gender specific MSV template v'_g and the MSV representation estimated from the utterance. When LTAS is used for comparison, the same training and classification procedure is used to obtain gender templates and respective cross-correlation distributions.

III. RESULTS

A. The templates

Fig. 3 shows the obtained limited-band templates used for gender classification for both MSV and LTAS. The structure of both features clearly differentiates between male and female speakers. Although the behavior at higher frequencies is quite similar for both MSV and LTAS, there are notable differences in the region between 1 and 2.5 kHz. One major difference is that the male LTAS contains two peaks at approximately 1300 Hz and 2300 Hz, whereas the male MSV has only one wide peak in between centered around 1800 Hz. Since the range of male F2 is usually between 900 Hz and 2300 Hz, and F3 receives values between 1700 Hz and 3000 Hz [7], this may suggest that MSV computed from sub-pitch periodic windows reacts more strongly to the movement of formants (describing their frequency range) whereas LTAS indicates mean formant locations. MSV peaks are slightly wider than LTAS peaks also at higher frequencies, thus supporting this assumption.

It is also well known that active articulation mainly affects the three lowest formants (especially the second), whereas higher formants are more stationary, reacting relatively passively to articulatory movements. This is also reflected in both the LTSA and MSV templates, where the shape of normalized mean and variance models approach each other at higher frequencies.
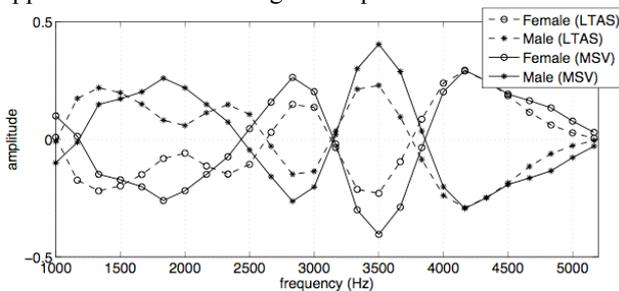
Figure 3: MSV and LTAS templates used in recognition.

B. Baseline classification results

When gender classification is evaluated with the TIMIT test set (56 males and 56 females, 10 utterances per speaker, 1120 utterances in total), a correct classification rate of 98.6 % is achieved (Table 1). This compares well with the approaches reported in the literature. For example, Zeng et al. [8] achieved a 98.2 % gender classification accuracy using a GMM based approach. Vergin et al. [2] achieved a classification rate of 85 % with a different corpus by using the average values of the two first formants as reference values for gender classification. Interestingly, they reported that no improvement was gained by including the higher formants, whereas the current approach leads to optimal results when the analyzed frequency region includes formants F2-F4 (1000 Hz – 5166 Hz) but not F1.

Table 1: Gender classification results for the full TIMIT test set (560+560 utterances).

gender	F0+MSV	F0+LTAS	LTAS	MSV	F0
male	99.3	98.8	82.9	85.7	98.6
female	97.9	97.3	87.0	84.3	95.7
mean	*98.60*	98.05	84.95	85.00	97.15

While MSV and LTAS both carry information regarding gender identity, their overall effect is small compared to F0, which alone leads to an over 97 % classification rate.

C. Feature combinations and noise

To gain a better insight of feature performance in different signal conditions, the gender classification task was performed separately for each possible combination of the three features (F0, MSV and LTAS) using a subset of 300 + 300 utterances (30 + 30 speakers) from the TIMIT test set. Three different noise conditions were used: the clean signal, and SNRs of 20 dB and 10 dB (Table 2).

The results indicate that MSV + F0 again yield the best recognition results (98.5 %), although the differences to LTAS + F0 and MSV + LTAS + F0 are not large. Although the recognition result at 10 dB SNR is still above 90 %, the noise robustness of this approach falls behind a GMM-model using F0 and RASTA-PLP features, where gender recognition rates of 97.9 % for an SNR = 20 dB and 97.5 % for an SNR = 10 dB have been reported [8]. The results obtained with solely LTAS are in line with previous gender recognition systems (e.g., [9], where the LTAS above 1 kHz was used for classification).

Table 2: Gender recognition results for different
feature sets in noise (TIMIT test, 300 + 300 utterances).

Features	Male	Female	Mean
Clean speech (SNR = ∞)			
MSV + LTAS + F0	99.1	97.0	98.05
LTAS + F0	98.7	96.3	97.50
MSV + F0	100.0	97.0	**98.50**
MSV + LTAS	89.0	88.3	88.65
MSV	89.7	86.7	88.20
LTAS	87.0	85.3	86.15
F0	99.0	93.3	96.15
White noise (SNR = 20 dB)			
MSV + LTAS + F0	98.0	97.3	**97.65**
LTAS + F0	98.0	97.0	97.50
MSV + F0	99.0	96.0	97.50
MSV + LTAS	91.7	83.3	87.50
MSV	90.0	76.7	83.35
LTAS	88.7	82.3	85.50
F0	97.3	94.0	95.65
White noise (SNR = 10 dB)			
MSV + LTAS + F0	86.0	97.7	**91.85**
LTAS + F0	86.0	97.3	91.65
MSV + F0	87.3	96.0	91.65
F0	87.7	95.3	91.50
MSV + LTAS	80.0	79.3	79.65
MSV	78.0	72.3	75.15
LTAS	77.7	83.3	80.50

A closer error analysis revealed that while MSV and LTAS have a similar overall performance on clean speech, they do not always make errors in the same utterances. In 76 cases of the total 600 utterances (clean speech), MSV and LTAS were giving contradictory information, i.e., one of the two was supporting the wrong gender hypothesis. However, the probabilistic framework used in the recognition compensates for this by assigning small probabilities to features that do not match either of the models. When the SNR drops to 10 dB, MSV performs significantly worse than LTAS, which is a reasonable result since white noise has a larger impact on the variance than the mean.

IV. CONCLUSIONS

A straightforward and efficient method for estimating the average formant frequencies (AFF) through mean spectral variance (MSV) from continuous speech was presented in this paper. As predicted, the MSV method provides comparable AFF estimates compared with those of long-term average spectrum (LTAS).

The usefulness of this approach was demonstrated in a gender classification task where speaker-specific MSV-information and pitch were combined in a straightforward manner as cues for gender identity. In addition, MSV was compared and combined with LTAS. The achieved gender classification rate compares well to other approaches reported in the literature (e.g., [2], [8]) and MSV performance was slightly higher than LTAS for clean speech. However, and as can be expected, MSV is not a particularly robust feature for long-term averaging in severe white noise. The obtained gender classification results are also in line with previous literature, showing that F0 alone is a very strong cue to gender identity in speech.

REFERENCES

[1] Faundez-Zanuy M. & Monte-Moreno E.: State-of-the-art in speaker recognition. *IEEE Aerospace and Electronic Systems Magazine,* Vol. 20, No. 5, pp. 7-12, 2005

[2] Vergin R., Farhat A. & O'Shaughnessy D.: Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification. *Proc. ICLSP'96,* pp. 1081-1084, 1996

[3] Markel J. D., Oshika B. T. & Gray A. H.: Long-Term Feature Averaging for Speaker Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing.* Vol. 25, No. 4, pp. 330-337, 1977

[4] Kinnunen T., Hautamäki V. & Fränti P.: On the use of long-term average spectrum in automatic speaker recognition. *International Symposium on Chinese Spoken Language Processing (ISCSLP'06)*, Singapore, pp. 559-567, 2006

[5] Mendoza E., Valencia N., Muñoz J. & Trujillo H.: Differences in Voice Quality Between Men and Women: Use of the Long-Term Average Spectrum (LTAS). *Journal of Voice*, Vol. 10, pp. 59-66, 1996

[6] Wu K. & Childers D. G.: Gender recognition from speech. Part I: Coarse analysis. *J. Acoustical Society of America*, Vol. 90, No. 4, pp. 1828-1840, 1991

[7] Hillenbrand J., Getty L. A., Clark M. J. & Wheeler K.: Acoustic characteristics of American English vowels. *Journal of Acoustical Society of America*, Vol. 97, No. 5, pp. 3099-3111, 1995

[8] Zeng Y.-M., Wu Z.-Y., Falk T. & Chan W.-Y.: Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. *Proceedings of 5th Int. Conference on Machine learning and Cybernetics*, pp. 3376-3379, Dalian, 2006

[9] Hertrich I. & Ziegelmayer G.: Sexual dimorphism in the long term speech spectrum. *Human Evolution,* Vol. 2, No. 3, 1987

A BOTTOM-UP PROCEDURE TO EXTRACT PERIODICITY STRUCTURE OF VOICED SOUNDS AND ITS APPLICATION TO REPRESENT AND RESTORATION OF PATHOLOGICAL VOICES

Hanae Itagaki[1], Masanori Morise[2], Ryuichi Nisimura[1], Toshio Irino[1] and Hideki Kawahara[1]

[1]Wakayama University, Wakayama, Japan
[2]Ritsumeikan University, Japan

Abstract: **A bottom up procedure for extracting repetitive structures in speech sounds has been developed on the basis of a temporally stable representation of periodic sounds (TANDEM) and adaptive spectral smoothing (STRAIGHT). The proposed method evaluates local periodic structures in the frequency domain to detect repetition in the time domain. A group of dedicated periodicity detectors are combined to construct the proposed procedure for a repetitive structure extractor called an excitation structure extractor (XSX). The proposed procedure is tested using a set of stylized test signals with artificial shimmer and jitter to investigate the applicability of such aperiodic signals. [The test results indicated that the proposed procedure outperformed in descriptive power of those complex excitation modes over existing F0 detectors. Finally, the proposed procedure is applied to analyze pathological voice examples to investigate the feasibility of voice quality restoration applications.**
Keywords: **periodicity extraction, fundamental frequency, TANDEM-STRAIGHT, XSX, aperiodicity, pathological voice**

I. INTRODUCTION

Fundamental frequency (F0) extraction is essential to analyze speech signal processing (e.g., speech synthesis, speech conversion, singing synthesis, and so on). A majority of the conventional methods[1,2] are specialized for extracting a specific periodic structure represented by a single value (F0). However, actual vocal cord vibration consists of irregularities, especially in boundaries of voiced segments. In addition, pathological voices show various types of aperiodic structures even in the middle of voiced segments. These irregularities make conventional F0 extraction methods ineffective for representing and analyzing biological aspects of vocal cord vibration.

A new power spectral representation[3] enabled a structural analysis of periodicity. This method does not rely on the assumption that the signal understudy has only one periodicity represented by F0. The proposed method is outlined in the next section. In this method, periodicity structure is represented as a collection of local maxima at each analysis frame[4]. Using this bottom-up procedure, existing local repetitions of signals are extracted without any prior information.

The proposed method also provides a procedure to remove irregularities in pathological voices. It enables patients' voices to be simulated after medial treatment.

II. PROPOSED METHODS

This section briefly outlines the proposed method. Details can be found in the other articles.

A temporally stable power spectral representation of periodic signals is yielded by averaging two power spectra. They are calculated using two time windows located a half pitch period apart. This simple procedure is called TANDEM.

A power spectral envelope that does not have any trace of periodicity is yielded by smoothing a TANDEM spectrum using an F0 adaptive smoothing function. The simplest smoothing function for this is a rectangular function and the width is adjusted to F0. This is called STRAIGHT spectrum.

The difference between TANDEM and STRAIGHT spectra is in the fine structure reflecting the signal periodicity. Consequently, dividing a TANDEM spectrum by a STRAIGHT spectrum leaves the fine structure. When the dominant contributing factor of the fine structure is periodicity due to F0, inverse Fourier transform shows a dominant peak at 1/F0. Figure 1 illustrates this concept.
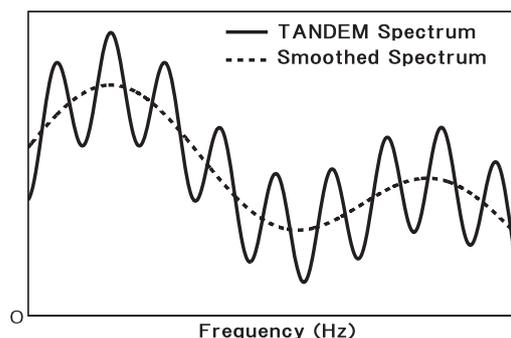
Fig. 1 TANDEM spectrum and STRAIGHT spectrum

However, there is a contradiction. TANDEM and STRAIGHT spectra are calculated using F0 information. However, F0 is the very thing we are to extract. It is impossible to implement TANDEM and STRAIGHT without F0 information in advance. One solution is to hypothesize a tentative F0. The designed periodicity detector responds best when the actual F0 coincides with the assumed F0. In other words, it is a dedicated periodicity detector for the assumed F0.

116

There are some details of the dedicated F0 detector design. A weighting function must be prepared to select reasonable numbers of base-band harmonic components. This is because spectral side bands smear out harmonic structures when the instantaneous frequency of F0 varies over time. This is inevitable, because the temporal variation of F0 is an important carrier of prosodic information in speech. A raised cosine function is used to implement this weighting.

The final stage is to integrate information extracted by a set of the dedicated detectors spanning the possible F0 range. Currently 40Hz to 800Hz is assumed to be the possible F0 range.

Information was integrated by averaging the outputs of dedicated F0 detectors by weighting each detector output using a raised cosine function centered around the assumed 1/F0. This structure enables bottom-up extraction of local periodicity. All locally periodic components are represented as local peaks of the integrated output of the final stage. Figure 2 illustrates conceptual architecture of the proposed periodicity detector, called an excitation structure extractor (XSX).
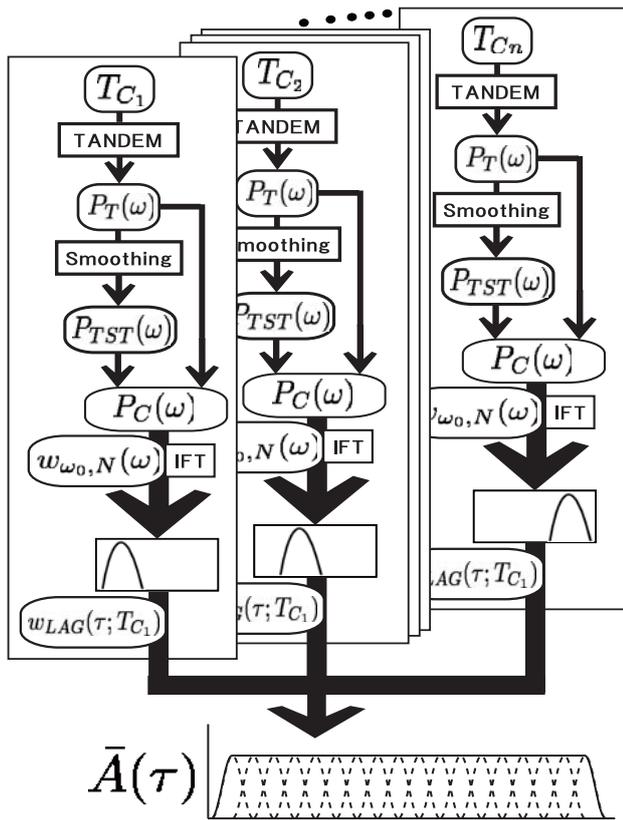
Fig. 2 Conceptual flow diagram of XSX

III. ANALYSIS EXAMPLES

This section shows examples of periodicity analysis using artificial test signals and actual pathological voices. Artificial signals were used to provide clues for interpreting typical representations found in the real analysis results.

A. Simulation results with test signals

The behavior of XSX was investigated using modulated pulse trains. The first test signal simulated a simmer. This test signal started as a periodic pulse train and incrementally reduced pulse amplitude every other pulse. This was an AM signal with a very fast modulation frequency. This AM signal is shown in Fig. 3. Figure 4 shows extraction results. The upper plot shows the integrated periodicity score for each maximum in each frame. The first five local maxima of each frame are shown. The score is normalized to have the value one when the signal is purely periodic in an analysis frame. The bottom plot shows the frequency of each peak. The size of the dots is determined on the basis of their periodicity score. Lager dots represent stronger periodicity.
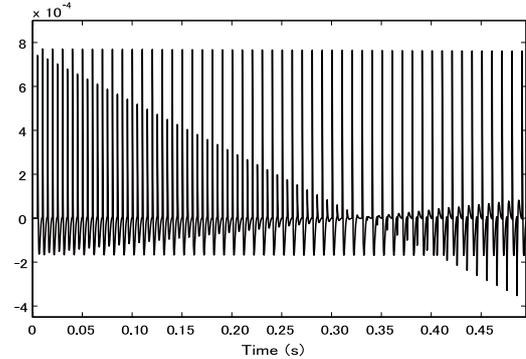
Fig. 3 Amplitude-modulated pulse trains used in this simulation
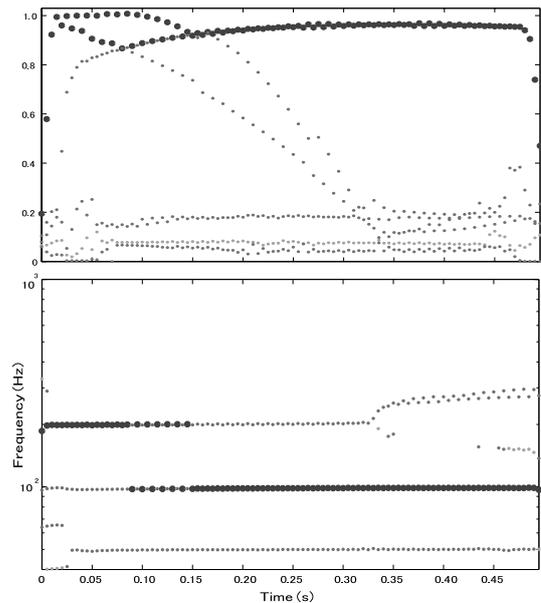
Fig. 4 Periodicity extraction results for AM signal using XSX

In the beginning, the score shows close to pure periodicity, and the corresponding frequency stays around 200Hz. The periodicity score decreases in accordance with increase in amplitude deviation of every other pulse. Another periodicity that corresponds to paired pulses increases. It corresponds to a trajectory around 100Hz. The plots around 0.33s indicate that the 200Hz periodicity disappears there. This suggests that the paired pulse dominates repetitive structures. All this illustrates that the XSX provides rich information for investigating signal periodicity structure.

The next test signal simulated jitter. This test signal was also started from periodic pulse train and shifted its periods every other cycle. The test wave is shown in Fig. 5. The analysis results are shown in Fig. 6. This manipulation is FM with very fast modulation frequency.
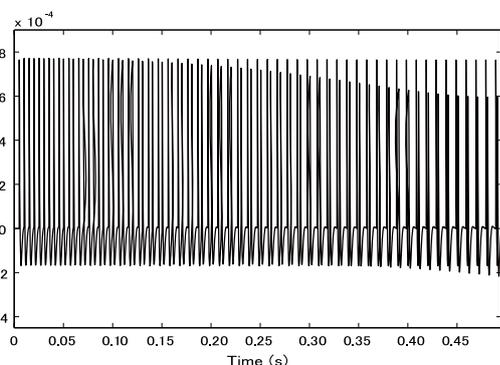
Fig. 5 Frequency-modulated pulse trains used in this simulation
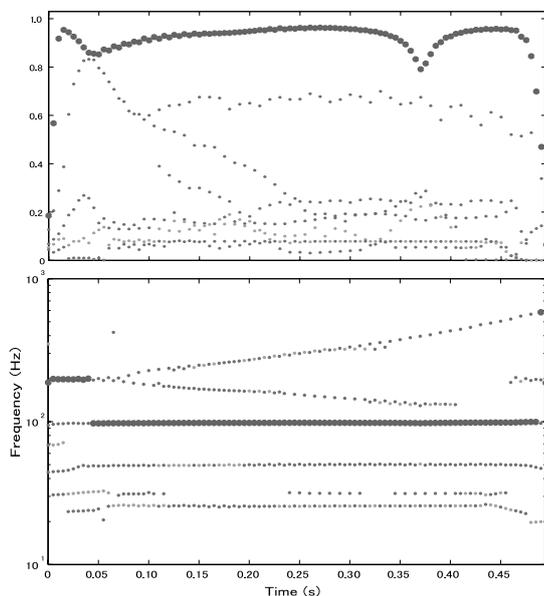
Fig. 6 Periodicity extraction results for FM signal using XSX

The periodicity score plotted in the upper panel of Fig. 6 also shows transition from smaller periodicity unit to the larger unit. The initial part has 200Hz periodicity. The lower panel of Fig. 6 shows this 200Hz periodicity split into two intermittent trajectories. The higher trajectory corresponds to a shorter period of the paired periods. The lower trajectory corresponds to a longer period. It also shows that the paired period, which corresponds to 100Hz periodicity, increasingly gains a higher periodicity score. This indicates that the initial sub-harmonic structure takes over the role of the fundamental frequency.

A series of simulations were conducted to establish an empirical functional relationship between the integrated score and probability of random fluctuation to yield the score. Conceptually straightforward architecture enables this probabilistic interpretation of analysis results.

B. Analysis results for pathological voices

The proposed procedure was applied to analyze pathological voices in a database [5]. Similar structures explained in the previous section were also found in analysis results of real voice examples. It will be helpful to compare these real results with the simulation results for diagnosis applications. These real voice examples are shown in Fig. 7 (a) and Fig. 8 (a) and their analysis results are shown in Fig. 7 (b) and Fig. 8 (b). Figure 7 shows the vowel /e/ and the analysis results of a female patient suffering from polypoid vocal cords saying it. Figure 8 shows the vowel /i/ and analysis results of a male patient suffering from larynx cancer saying it.
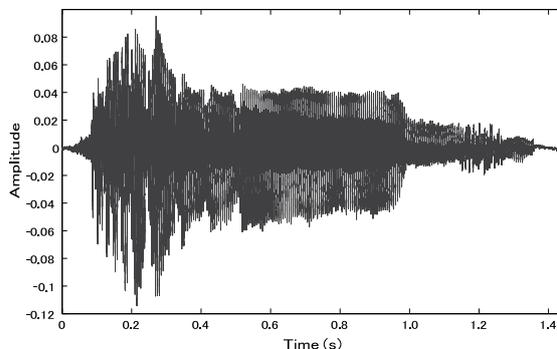
Fig. 7 (a) Pathological voice example (Polypoid Vocal Cords)
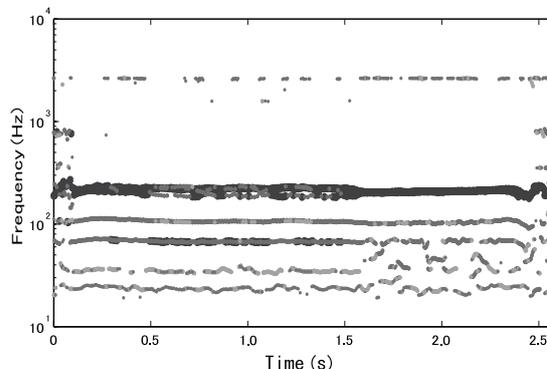
Fig. 7 (b) Periodicity structure extracted by XSX
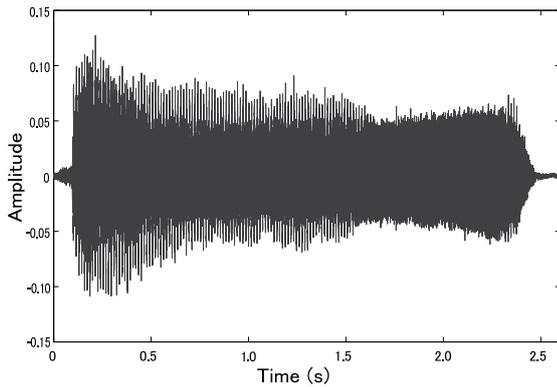
Fig. 8 (a) Pathological voice example
(Larynx Cancer)
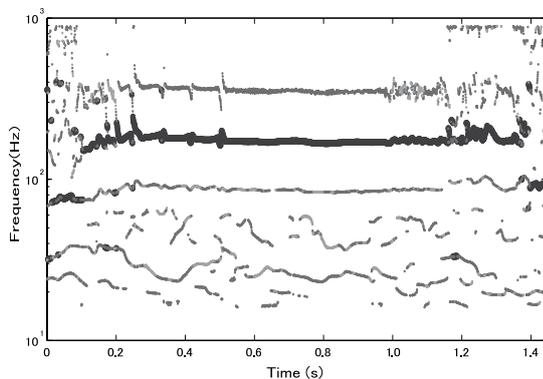
Fig. 8 (b) Periodicity structure extracted by XSX

IV. SIMULATING HEALING PROCESS

The proposed method provides new possibilities for enabling objective assessment of healing processes after medical treatment of pathological voices. Effects of local periodicity can be cancelled out by using periodicity adaptive smoothing procedures. AM and FM effects on F0 are cancelled out by averaging period lengths of every other period. Recovered F0 trajectory is calculated using the procedure mentioned above. Spectral effects are removed by applying an adaptive moving average using the period length that corresponds to the sub-harmonic frequency. The aperiodicity index in each time-frequency region can be replaced by values of normal voices. Applying these recovered parameters to STRAIGHT-based synthesis procedures [3,6] enables a hypothetical recovered voice to be produced after the healing process is completed. By using this recovered voice and the current voice before the medical treatment as exemplar samples for voice morphing, generated morphing results provide a set of reference voices to be compared with voices of the patients in their healing processes. This may provide an objective index to represent the degree of healing.

V. CONCLUSION

A new procedure to extract local periodicity structures has been developed on the basis of a new power spectral representation. The proposed procedure produces simple descriptions of periodic structure relying on no strong assumptions. Simulation results using test signals with known aperiodicity and results using actual examples of pathological voices illustrated the usefulness of the proposed method. Application to objective assessment of healing processes is promising but it also requires further investigations and collaboration.

REFERENCES

[1] Arturo Camacho and John G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music", J. Acoustical Society of America, 124, Issue 3, pp. 1638-1652(2008)
[2] Worfgang Hess, "Pitch Determination of Speech Signals", Springer-Verlag, Berlin(1983)
[3] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino and Hideki Banno, "TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation", Proc. ICASSP 2008, Las Vegas,pp.3933-3936(2008)
[4] Osamu Fujimura, Kiyoshi Honda, Hideki Kawahara, Yasuyuki Konparu and Masanori Morise, "Noh Voice Quality", J. Logopedics Phoniatrics Vocology, 04 June 2009
[5] The Japan Society of Logopedics and Phoniatrics, Interuna Publishers, Inc, ISBN:978-4-900637-21-4 (2005).
[6] Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveign´e. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds", Speech Communication, 27, 3, pp.187-207(1999).

ACOUSTIC/AERODYNAMIC ASSESSMENT OF NORMAL AND DYSPHONIC VOICE

G. Cantarella[1], G. N. Baracca[1], S. Forti[2], L. Pignataro[1]

[1] Otolaryngology Department and [2] Audiology Unit, University of Milan, Fondazione IRCCS Ospedale Policlinico, Mangiagalli e Regina Elena, Milan, Italy

giovanna.cantarella@policlinico.mi.it

Abstract: Voice production is a complex multidimensional phenomenon resulting from the combination of acoustic, aerodynamic and elastic forces. The evaluation of voice characteristics in clinical practice is often based only on perceptual and acoustic evaluation, but an exhaustive assessment should take into consideration also aerodynamic parameters.

In this study two groups of subjects, the first one composed by patients affected by organic dysphonia and the second one by controls with normal voice, underwent: simultaneous acoustic/aerodynamic voice assessment by means of EVA device (SQ-Lab, Aix-en-Provence, F); maximum phonation time measurement; GIRBAS perceptual evaluation. Statistical analysis allowed to search for correlations between the perceptual voice quality grading and the recorded acoustic/aerodynamic parameters.

Keywords : **dysphonia, acoustic assessment, aerodynamic evaluation**

I. INTRODUCTION

The evaluation of voice characteristics in clinical practice is often based only on perceptual and acoustic evaluation. Due to the diffusion of digital systems and software for voice recording and analysis, acoustic measurements are commonly obtained in patients affected by laryngeal disorders. Nevertheless there are no widely accepted and standardized methods; therefore assessing objectively vocal emission is an unsolved problem.

Up to now, only a few studies have been dedicated to the analysis of aerodynamic parameters. Aerodynamic methods have been described since several decades [1-7], but their diffusion has been limited due to the scarcity of specifically designed instruments.

Voice production is a complex multidimensional phenomenon resulting from the combination of acoustic, aerodynamic and elastic forces; it was stated by Hirano [2] that the glottis should be considered as an energy transducer which converts aerodynamic power into acoustic energy. Therefore an exhaustive assessment of vocal function should ideally take into consideration both aerodynamic and acoustic parameters.

The assessment of voice should be based on objective measurements in order to allow a comparison of the results across different voice clinics and different therapeutic protocols.

The goal of this study was to analyze acoustic and aerodynamic indexes in a group of patients affected by dysphonia and in a control group of subjects with normal voice, in order to search for the parameters which better correlate with the dysphonia severity and which allow to discriminate dysphonic vocal emissions from normal ones. We also looked for correlations between the subjective parameters obtained with GRBAS Scale and the objective indexes acquired by the acoustic/aerodynamic evaluation.

II. METHODS

The study includes 51 patients (34 women and 17 men) affected by benign organic dysphonia (22 patients with vocal fold polyps, 12 with cysts, 12 with Reinke's edema, 2 with nodules, 1 with sulcus glottidis). The control group is composed by 23 subjects with normal voice, homogeneous for age and gender.

Both the dysphonic and the normal subjects underwent videolaryngostroboscopy, to ascertain and obtain a documentation of the status of vocal folds.

The protocol for the multidimensional voice evaluation consisted of maximum phonation time measurement, the perceptual voice assessment by means of the GIRBAS Scale [8], and the acoustic/aerodynamic evaluation with EVA device (SQ-Lab, Aix-en-Provence, F).

More in details:

a) Maximum Phonation Time (MPT) measurement was obtained during emission of the vowel /a/ at a comfortable pitch and loudness; three consecutive trials were performed and we considered the best one.

b) Perceptual voice evaluation was by means of the GIRBAS scale [8], which includes the six parameters of the grade of dysphonia (G), instability (I), roughness (R), breathiness (B), asthenia (A), and strain (S). The voice samples were computer recorded using a dynamic microphone (AKG, model C 1000 S) at a constant distance of five centimetres from the patient's mouth during the production of a sustained /a/, the repetition of single words and sentences, and conversation. All of the voice

120

samples were subsequently evaluated by a jury of four experienced listeners (two voice therapists and two phoniatricians), and scored in the usual manner (0=normal; 1=slight disturbance; 2=moderate disturbance; 3=severe disturbance). The parameters G,R,B were then taken into consideration for statistics in this study.

c) Acoustic/aerodynamic evaluation. Voice recordings were performed by means of the EVA system, which allows simultaneous analysis of acoustic and aerodynamic indexes. A rubber mask, connected to a mouthpiece, is placed on the mouth of the patients strictly adherent to the skin in order to avoid any air leak. The mouthpiece contains a calibrated directional microphone and a grid pneumotachograph. The microphone and the aerodynamic sensor are coaxial so that voice sound and phonatory airflow may be recorded at the same time. The microphone is at the distance of two centimeters from the patient's mouth.

Two different tests were performed:

1. Voice range profile: the recording is made while the patient pronounces a sustained /a/ at comfortable pitch and intensity for at least 4-5 seconds. Afterwards four traces are displayed on the computer screen: 1- sound wave form, 2 - intensity, 3 - fundamental frequency, 4 - airflow. A one-second segment is selected for analysis in the most stable part of the wave form. Among the available ones, the following acoustic indexes were taken into consideration: Mean Fundamental Frequency (F_0), coefficient of variation of F_0 (CV F_0), Jitter %, Shimmer %, Harmonics to noise ratio (HNR), Intensity coefficient of variation (I CV).

The considered aerodynamic indexes were: mean oral airflow (OAF), expressed in cc per second; Oral airflow coefficient of variation; this index relativizes the OAF standard deviation to the mean airflow value per second; it is an indicator of airflow instability and, indirectly, of the capability to achieve and maintain glottic closure during phonation; Glottic leakage (= Mean OAF/Mean Intensity) is expressed in cc/s/dB; this index evaluates the amount of air utilized to produce one decibel in one second; it estimates the efficiency of the glottis in transforming aerodynamic power into acoustic energy.

2. Airway interrupted method for indirect estimation of subglottic pressure [9] (P, in hPa), during the emission of a sequence of "pa". For this test a

pressure sensor allowed to measure intraoral pressure and to derive subglottal pressure.

The values of laryngeal resistance (LR, calculated as the ratio P/OAF), the Glottal efficiency index (GEI, calculated as the ratio dB/hPa) and the laryngeal efficiency (LE, measured as the ratio dB/(hPa·dm^3/s) have been derived by the EVA software.

Statistical analysis

Data are presented as mean ± SD.

Intergroup comparison was performed with the Mann-Whitney test. Associations between parameters were determined by Pearson's correlation coefficients. Two-sided exact tests were used and p values of less than .05 were considered significant. All statistics were calculated using the Statistical Package for the Social Sciences 17.0 for Windows software package (SPSS Inc, Chicago, IL).

III. RESULTS

Table 1 reports results of statistical analysis concerning 13 considered acoustic/aerodynamic parameters of the voice and maximum phonation time (MPT). All variables but L.R., FO and oral airflow were significantly different in dysphonic subjects when compared to normal controls (Tab. 1).

Tab. 1

	Case	Control	p
G.E.I	7.33 ± 2.58	11.11 ± 3.90	<0.001
L.E.	35.47 ± 26.88	127.60±146.15	<0.001
L.R.	53.34 ± 31.79	77.55 ± 68.41	ns
P	12.14 ± 4.44	8.13 ± 2.77	<0.001
MPT	11.45 ± 5.59	18.08 ± 6.44	<0.001
F0 (Hz)	163.22±46.61	163.24±48.65	ns
CV F0	2.46 ± 3.26	0.74 ± 0.32	<0.001
I CV	1.26 ± 0.58	0.85 ± 0.30	0.002
Jitter %	2.30 ± 4.73	0.44 ± 0.24	<0.001
Shimmer %	1.02 ± 0.90	0.30 ± 0.15	<0.001
HNR	13.13 ± 8.69	20.33 ± 3.23	<0.001
Glottic leakage	3.00 ± 2.28	1.92 ± 0.70	0.017

	Case	Control	p
Oral airflow	225.05±179.90	150.95± 5.46	ns
OA CV	7.24 ± 8.40	3.55 ± 2.14	0.048

The correlation between the grade of dysphonia (measured by GIRBAS scale) and the six parameters reported in table 2 - in particular three aerodynamic parameters and three acoustic ones - resulted significant. The values of CV FO, Jitter and Shimmer increased when G increased, whereas G.E.I., L.E. and HNR decreased. Also the correlation between G and the other variables was evaluated, but no significant results were found (Tab. 2). Roughness was significantly correlated with the same variables and also with glottic leakage, oral airflow and OA CV. Breathiness, instead, was correlated with different variables: L.R., MPT and oral airflow; in particular values of airflow increased when B raised, while LR and MPT were inversely related to B value.

Tab. 2

	Grade of dysphonia		Roughness		Breathiness	
	correlation	p	correlation	p	correlation	p
G.E.I	- 0.347	0.024	-0.374	0.018	-0.173	ns
L.E.	- 0.364	0.019	-0.332	0.039	-0.316	ns
L.R.	-0.095	ns	0.029	ns	-0.322	0.046
MPT	-0.281	ns	0.05	ns	-0.405	0.007
CV F_0	0.537	<0.001	0.545	<0.001	0.137	ns
Jitter %	0.534	<0.001	0.554	<0.001	0.174	ns
Shimmer %	0.527	<0.001	0.590	<0.001	0.142	ns
HNR	- 0.489	<0.001	-0.559	<0.001	-0.081	ns
Glottic leakage	-0.032	ns	-0.331	0.026	0.281	ns
Oral airflow	0.00	ns	-0.319	0.031	0.293	0.048
OA CV	0.275	ns	0.333	0.026	-0.072	ns

IV. DISCUSSION

Our results confirm that both acoustic and aerodynamic parameters are useful in the assessment of dysphonia as they allow to differentiate an hoarse voice from a "normal" one. Nevertheless overlapping between data obtained by the analysis of dysphonic and normal voices was found; this result is in agreement with the current literature [10-12]. In particular three aerodynamic parameters and three acoustic ones were significantly related to the degree of dysphonia as perceptually measured by the GRBAS scale; nine parameters were significantly related to roughness changes and three with breathiness. These data highlight that perceptual voice evaluation is a reliable means for voice evaluation.

V. CONCLUSION

This study confirms the utility of both acoustic and aerodynamic indexes for the objective assessment of voice pathologies. Further work will analyze from the acoustic /aerodynamic point of view the outcome achieved by phonosurgery in patients affected by organic pathologies of the vocal folds.

REFERENCES

[1] H.K. Schutte. "Aerodynamics of phonation". Acta Oto-Rhino-Laryngol Belg, vol 40, pp 344-357, 1986

[2] M. Hirano "Objective evaluation of the human voice: clinical aspects". Folia Phoniatr;vol 41, pp 89-144, 1989.

[3] A. Giovanni, V. Molines, N. Nguyen, B. Teston, D. Robert, M. Cannoni, A. Pech. "Une méthode multiparamétrique d'évaluation vocale objective assistée par ordinateur . Ann Oto-laryng (Paris), vol 109, pp 200-206, 1992

[4] A. Giovanni, N. Estublier, D Robert, B. Teston, M Zanaret, M. Cannoni. "Evaluation vocale objective des dysphonies par la mesure simultanée de paramètres acoustiques et aérodynamiques à l'aide de l'appareillage EVA" . Ann Otolaryngol Chir Cervicofac; vol 112, pp 85-90, 1995

[5] A. Giovanni, D. Robert, N. Estublier, B. Teston, M. Zanaret, M. Cannoni. "Objective evaluation of dysphonia: preliminary results of a device allowing simultaneous acoustic and aerodynamic measurements ". Folia Phoniatr Logop; vol 48, pp 175-185 1996.

[6] P. Dejonckere, P. Bradley, P Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, V. Woisard. "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques". Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS), Eur Arch Otorhinolaryngol vol Feb. 258(2), pp 77-82, 2001.

[7] B. Teston. "L'évaluation objective des dysfonctionnements de la voix et de la parole, 2ème partie: les dysphonies" . Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, vol. 20, pp 169-232, 2001.

[8] M. Hirano. "Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice"; in: *Clinical examination of voice*. Hirano M (ed). Vienna, Springer, 1981, pp 81-84.

[9] Smitheran J, Hixon T. A clinical method for estimating laryngeal airway resistance during vowel production. J Speech Hear Disord 1981; 46:138-46

[10] H.K. Schutte. *The efficiency of voice production*. Groningen: Kemper, 1980.

[11] F. Ursino, F. Matteucci,, V. Picasso . "L'elettroglottografia e gli indici aerodinamici". Acta Phoniatrica Latina; vol. 24, pp 153-161, 2002.

[12] P. Woo, R.H. Colton, L. Shangold. "Phonatory airflow analysis in patients with laryngeal disease". Ann Otol Rhinol Laryngol, vol 96, pp 549-555, 1987.

Voice images

TOWARDS VIDEO LARYNGOSTROBOSCOPY-BASED AUTOMATED SCREENING FOR LARYNGEAL DISORDERS

A. Gelzinis[2], A. Verikas[1,2], M. Bacauskiene[2], E. Vaiciukynas[2], E. Kelertas[2], V. Uloza[3], A. Vegiene[3]

[1]Intelligent Systems Laboratory, Halmstad University, Halmstad, Sweden
[2]Department of Electrical and Control Equipment, Kaunas University of Technology, Kaunas, Lithuania
[3]Department of Otolaryngology, Kaunas University of Medicine, Kaunas, Lithuania

Abstract: **This paper is concerned with kernel-based techniques for automated categorization of laryngeal colour image sequences obtained by video laryngostroboscopy. Features used to characterize a laryngeal image are given by the kernel principal components computed using the N-vector of the 3-D colour histogram. The least squares support vector machine (LS-SVM) is designed for categorizing an image sequence (video) into the** *healthy*, *cancerous* **and** *noncancerous* **classes. The kernel function employed by the LS-SVM is defined over a pair of matrices, rather than over a pair of vectors. The classification accuracy of over 85% was obtained when testing the developed tools on data recorded during routine laryngeal videostroboscopy.**

Keywords: **Larynx pathology, Image sequence, Classification, Support vector machine**

I. INTRODUCTION

Video, laryngeal still images, voice signal, and patient's questionnaire data are considered as the main information sources to characterize human larynx. Nowadays, automated analysis of voice is increasingly used for detecting and screening laryngeal pathologies [1], [2], [3].

However, there were very few attempts to create systems for automated analysis of still colour laryngeal images. Ilgner et al. [4] proposed a CCD camera-based technique for automated categorization of manually marked suspect lesions into *healthy* and *diseased* classes. The categorization is based on textural features extracted from co-occurrence matrices [5], [6] computed from manually marked areas of vocal fold images. The classification accuracy of 81.4% was reported when testing the technique on a very small set of 35 images. A set of 785 colour laryngeal images obtained by direct microlaryngoscopy has been used in studies presented in [7], [8], [9]. The classification accuracy of over 95% was achieved when categorizing the images into one healthy and two pathological (nodular and diffuse) classes. When categorizing the same set of images into seven classes (one healthy and six pathological), the classification accuracy of over 80% was reported [10]. Image texture, distribution of colour, and geometry of edges of vocal folds are the types of features used for the categorization. It was found that colour is amongst the most discriminative types of features.

A. *Video laryngostroboscopy*

Video laryngostroboscopy is used extensively for inspecting vocal folds and in the clinical practice for diagnosing voice disorders [11]. Video laryngostroboscopy is a well-established technique for measuring the glottal gap or examining the glottic closure [12]. Videostroboscopy is one of the standard methods used to examine moving objects. Flashing light is used to illuminate an object in stroboscopy. When the flashes are synchronized with the vocal fold vibrations, a stationary view of the vocal folds is obtained.

However, the single-flash-timing video laryngostroboscopy has a limitation that it is effective only when vocal fold vibrations exhibit only one single fundamental frequency. Multiple tones (fundamental frequencies) may be recorded in the case of some diseases, such as polyps, nodules, and cysts [13]. In such cases, a clear view of the vibrating vocal folds can not be obtained with the single-flash-timing video laryngostroboscopy. A multiple-flash-timing technique of video laryngostroboscopy was proposed by Deguchi et al. [13] to deal with such cases.

In [14] image sequences recorded with the stroboscopy system have been used to measure the glottic angle and the angular velocities of vocal fold abduction and adduction. The authors point out that semi-automated edge tracking would be an important improvement of the technique.

It is worth mentioning that not only edge tracking but also other tasks usually carried out when analyzing video data need automated or semi-automated analysis. Decision making is one of such tasks. In clinical practice, decision making is quite often based on subjective evaluation of video data. Quantitative measures of motion, colour distribution and geometry of vocal folds can provide objective information and be useful in medical treatment planning and greatly facilitate tracing progress over time.

The long-term goal of this work is a decision support system to facilitate screening for laryngeal disorders. A voice signal, sequences of colour vocal fold images obtained from video laryngostroboscopy, and questionnaire data [15] are the information sources to be used in the analysis. This paper is concerned with automated categorization of image sequences obtained from laryngeal videostroboscopy into a healthy class and two classes of disorders, namely cancerous and noncancerous.

II. THE DATA

The task considered in this paper concerns automated categorization of colour image sequences obtained from

video laryngostroboscopy into three decision classes, namely a *healthy* class and two *pathological* classes—mass lesions of vocal folds. We distinguished two groups of mass lesions of vocal folds i.e. *noncancerous* lesions—*nodules*, *polyps*, *papillomata*, *keratosis*, and *cysts*—and *cancerous* lesions—*carcinoma*. The diagnosis was confirmed by histological examination of laryngeal specimens removed during endolaryngeal microsurgical intervention. To illustrate the three decision classes, Fig. 1 presents examples of vocal fold images obtained by the direct microlaryngoscopy.
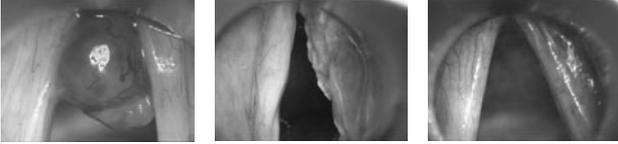
Fig. 1. Images from the *noncancerous* (left), *cancerous* (middle), and *healthy* (right) classes.

The data have been recorded at the Department of Otolaryngology, Kaunas University of Medicine, Lithuania. The image sequences were acquired during routine videostroboscopy, using the "EndoSTROB" device. The duration of one image sequence was equal to 8s. The resolution of 720×576 pixels was used to record the image sequences. Data from 87 patients were available. Among those, 63 patients belong to the noncancerous class, 18 to the cancerous class and 6 to the healthy class.

III. FEATURES

Various types of features characterizing colour, texture, and geometry of the biological structures seen in colour images of vocal folds can be extracted [8]. Features characterizing the distribution of image colour are used in this study. The approximately uniform $L^*a^*b^*$ colour space is employed to represent colours. We characterize the colour content of an image by the probability distribution of the colour represented by the 3–D colour histogram of $N = 4096$ $(16 \times 16 \times 16)$ bins and consider the histogram as an N-vector. Most of bins of the histograms were empty or almost empty. Therefore, to reduce the number of components of the N-vector, the histograms built from a set of training images were summed up and the N-vector components corresponding to the bins containing less than N_α hits in the summed histogram were left aside. Hereby, when using $N_\alpha = 50$ we were left with 918 bins—a ψ vector of measurements with 918 components.

Having a vector of measurements ψ, the feature vector $\mathbf{x}$ is computed in the following way. We assume that κ is a kernel [16] and Φ is a mapping of ψ onto the feature space F, such that $\kappa(\psi_i, \psi_j) = \langle \Phi(\psi_i), \Phi(\psi_j) \rangle$, where $\langle \cdot, \cdot \rangle$ stands for the inner product. Let $\widetilde{\Phi}(\psi_i)$

$$\widetilde{\Phi}(\psi_i) := \Phi(\psi_i) - \frac{1}{M} \sum_{i=1}^{M} \Phi(\psi_i) \qquad (1)$$

with M being the number of data points. The features x are then given by the kernel principal components computed as

projections of $\widetilde{\Phi}(\psi)$ onto the eigenvectors

$$\mathbf{v} = \sum_{i=1}^{M} \alpha_i \widetilde{\Phi}(\psi_i) \qquad (2)$$

of the covariance matrix $K_{ij} = \langle \widetilde{\Phi}(\psi_i), \widetilde{\Phi}(\psi_j) \rangle$, where the expansion coefficients α_i of the eigenvector are found from the eigenvalue problem

$$\lambda \boldsymbol{\alpha} = \mathbf{K} \boldsymbol{\alpha} \qquad (3)$$

where, the solutions $\boldsymbol{\alpha}$ are normalized by requiring $\lambda \langle \boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle = 1$. Thus, the feature x is given by

$$x = \langle \mathbf{v}, \widetilde{\Phi}(\psi) \rangle = \sum_{i=1}^{M} \alpha_i \langle \widetilde{\Phi}(\psi_i), \widetilde{\Phi}(\psi) \rangle \qquad (4)$$

The optimal number of components (features) used is determined experimentally.

IV. THE CLASSIFIER

We use a support vector machine (SVM) as a classifier in this work. Assuming that $\boldsymbol{\Upsilon}(\mathbf{x})$ is the non-linear mapping of $\mathbf{x}$ into the new space, the 1-norm soft margin SVM can be constructed by solving the following problem:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \sum_{i=1}^{M} \xi_i \qquad (5)$$

subject to

$$y_i(\langle \mathbf{w}, \boldsymbol{\Upsilon}(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \quad i = 1, ..., M \qquad (6)$$

where $\mathbf{w}$ is the weight vector, $y_i = \pm 1$ is the desired output, M is the number of training data, $\langle \rangle$ stands for the inner product, ξ_i are the slack variables, b is the threshold, and γ is the regularization constant controlling the trade-off between the margin and the slack variables. The discriminant function for a new data point $\mathbf{x}$ is given by:

$$f(\mathbf{x}) = \mathcal{H} \left[\sum_{i=1}^{M} \alpha_i^* y_j k(\mathbf{x}, \mathbf{x}_i) + b^* \right], \qquad (7)$$

where $k(\mathbf{x}, \mathbf{x}_i)$ stands for the kernel and the Heaviside function $\mathcal{H}[y(\mathbf{x})] = -1$, if $y(\mathbf{x}) \leq 0$ and $\mathcal{H}[y(\mathbf{x})] = 1$ otherwise. The optimal values α_i^*, b^* of the parameters α_i and b are found during training.

A. Least squares SVM

Suykens and Vandewalle [17] have introduced a least squares version of the SVM classifier (LS-SVM). We use this type of SVM in this work. Parameters of the LS-SVM are estimated by solving the following optimization problem:

$$\min_{\mathbf{w}, b, \mathbf{e}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \frac{1}{2} \sum_{i=1}^{M} e_i^2 \qquad (8)$$

subject to

$$y_i(\langle \mathbf{w}, \boldsymbol{\Upsilon}(\mathbf{x}_i) \rangle + b) = 1 - e_i, \quad i = 1, ..., M \qquad (9)$$

The main difference between the LS-SVM and SVM is the equality constraints (Eq.(9)) used in the LS-SVM instead of unequally constraints defined by Eq.(6). Due to the equality constraints, the optimal parameter values can be found by solving a set of linear equations, instead of quadratic programming applied in the case of SVM. The solution is given by [17]

$$\begin{bmatrix} 0 & -\mathbf{y}^T \\ \mathbf{y} & \mathbf{Z} + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix} \qquad (10)$$

where $Z_{ij} = y_i y_j \kappa(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{I}$ is the identity matrix, $\mathbf{1} = [1_1, ..., 1_M]$, $\mathbf{y} = [y_1, ..., y_M]$, and $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_M]$.

Since an SVM is a binary classifier while the task is to distinguish between three classes, the one-against-one scheme is used to make the categorization in this work.

B. Kernel function

For $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, one usually uses the linear: $\mathbf{x}_i^T\mathbf{x}_j$, Gaussian: $\exp\{-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma\}$ or polynomial: $(\mathbf{x}_i^T\mathbf{x}_j + 1)^d$ kernel. The kernel is defined over a pair of vectors.

In this work, classification is based on a set of vectors rather than on a single vector. A sequence of images is recorded from a patient. Each image is represented by a feature vector. Feature vectors are then collected into a matrix (each vector constitutes a matrix column) and used to make a decision. Therefore, a kernel function utilized by the LS-SVM classifier is defined over a pair of matrices $(\mathbf{A}, \mathbf{B})$ rather than over a pair of vectors. A positive definite kernel of such type has been recently proposed by Wolf and Shashua [18]. The authors use the principal angles between the two column spaces defined by the matrices $(\mathbf{A}, \mathbf{B})$ to assess the matching between the spaces and derive a positive definite kernel based on that concept. The "QR" factorization of the matrices $(\mathbf{A}, \mathbf{B})$ and the kernel Gram-Schmidt orthogonalization process are used to derive the kernel. Applying the "QR" factorization the matrices $(\mathbf{A}, \mathbf{B})$ can be written as $\mathbf{A} = \mathbf{Q}_A\mathbf{R}_A$ and $\mathbf{B} = \mathbf{Q}_B\mathbf{R}_B$, where $\mathbf{Q}$ is an orthonormal basis and $\mathbf{R}$ is an upper-diagonal matrix of size $M \times M$ of the Gram-Schmidt coefficients representing the columns of the original matrix in the new basis. The principal angles $\cos(\theta_i)$ are given by the singular values σ_i of the matrix $\mathbf{Q}_A^T\mathbf{Q}_B$, $\cos(\theta_i) = \sigma_i, i = 1, ..., M$. It was shown that

$$\kappa(\mathbf{A}, \mathbf{B}) = \det(\mathbf{Q}_A^T\mathbf{Q}_B)^2 = \prod_{i=1}^{M} \cos(\theta_i)^2 \qquad (11)$$

is a positive define kernel [18]. We use this kernel in our work. The algorithm for evaluating the kernel without explicit computation of $\mathbf{Q}_A$ and $\mathbf{Q}_B$ can be found in [18]. Only inner-products between the columns of $\mathbf{A}$ and the columns of $\mathbf{B}$ are used.

V. EXPERIMENTAL INVESTIGATIONS

A. Experimental setup

There were 200 image frames in one image sequence. However, only 20 image frames were used to estimate the kernel defined over a pair of matrices. Due to the small number of data available for the experiments, the leave-one-out test has been used to estimate the classification accuracy. The data used were normalized to zero mean and variance one. The polynomial kernel of degree $q = 2$ was used to extract the kernel principal components, while the Gaussian kernel was used to estimate the kernel defined over a pair of matrices. The experimental tests performed concern the influence of the LS-SVM regularization constant γ, the Gaussian kernel width parameter σ, and the number of the kernel principal components used on the test set data classification accuracy. The dependence of the classification accuracy on the percentage of the data variance accounted for by the number of the kernel principal components used was also studied.

B. Results

In Fig. 2, shown is the classification accuracy of the test set data as a function of the number of the kernel principal components used to characterize colour of one image frame, for given values of the regularization constant γ and the kernel width parameter σ. As can be seen, nine principal components is the best choice.
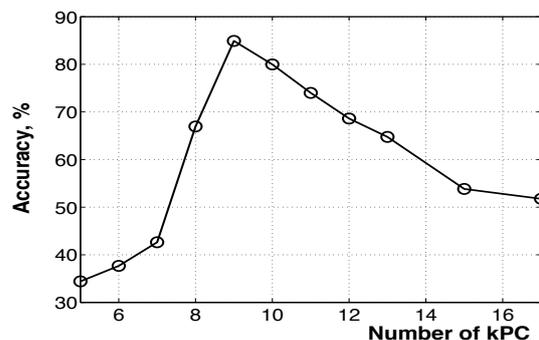
Fig. 2. The classification accuracy of the test set data as a function of the number of the kernel principal components, for given values of the regularization constant γ and the kernel width parameter σ.

The graph presented in Fig. 3 plots the test set data classification accuracy as a function of the percentage of the data variance accounted for by the number of the kernel principal directions used to represent colour. As can be seen from Fig. 3, the percentage of the data variance accounted for by the optimal number of the components is close to 90. Fig. 4 relates the test set data classification accuracy the regularization constant γ, and the number of the kernel principal components used to represent colour. As can be seen from Fig. 4, a large number of principal components significantly deteriorates the classification accuracy. Fig. 5 plots the test set data classification accuracy as a function of the regularization constant γ and the kernel width parameter σ.

VI. CONCLUSIONS

A technique for automated categorization of laryngeal colour image sequences obtained by video laryngostro-
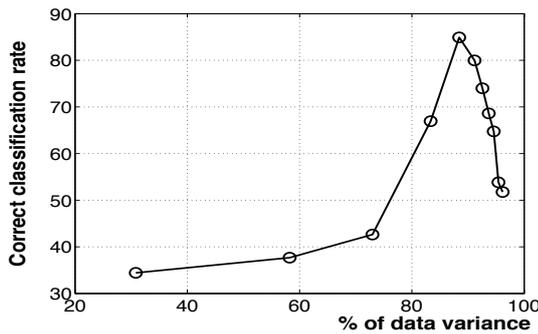
Fig. 3. The classification accuracy of the test set data as a function of the percentage of the data variance accounted for by the number of the kernel principal components used.
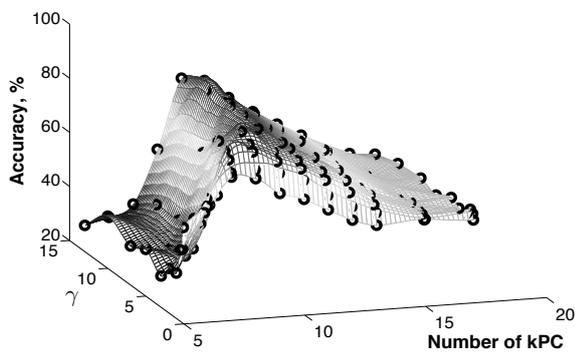
Fig. 4. The classification accuracy as a function of the regularization constant γ and the number of kernel principal components.
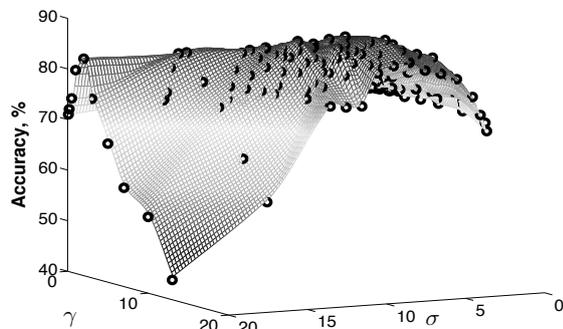
Fig. 5. The classification accuracy as a function of the regularization constant γ and the kernel width parameter σ.

boscopy was developed. The LS-SVM employed to categorize an image sequence into the *healthy*, *cancerous* and *noncancerous* classes exploits a kernel function defined over a pair of matrices, rather than over a pair of vectors. The classification accuracy of over 85% was obtained when testing the developed tools on data recorded during routine laryngeal videostroboscopy. One can expect increasing the accuracy even further by adding features of other types. A larger database needs to be collected for the comprehensive examination of the technique.

REFERENCES

[1] R. J. Moran, R. B. Reilly, P. de Chazal, P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis", *IEEE Trans Biomedical Engineering* vol. 53 (3), pp. 468–477, 2006.
[2] J. I. Godino-Llorente, P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors", *IEEE Trans Biomedical Engineering* vol. 51 (2), pp. 380–384, 2004.
[3] A. Gelzinis, A. Verikas, M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization", *Computer Methods and Programs in Biomedicine* vol. 91 (1), pp. 36–47, 2008.
[4] J. F. R. Ilgner, C. Palm, A. G. Schutz, K. Spitzer, M. Westhofen, T. M. Lehmann, "Colour texture analysis for quantitative laryngoscopy", *Acta Oto-Laryngologica* vol. 123 (6), pp. 730–734, 2003.
[5] R. M. Haralick, K. Shanmugam, I. Dinstein, "Textural features for image classification", *IEEE Trans System, Man and Cybernetics* vol. 3 (6), pp. 610–621, 1973.
[6] A. Gelzinis, A. Verikas, M. Bacauskiene, "Increasing the discrimination power of the co-occurrence matrix-based features", *Pattern Recognition* vol. 40 (9), pp. 2367–2372, 2007.
[7] A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, "Towards a computer-aided diagnosis system for vocal cord diseases", *Artificial Intelligence in Medicine* vol. 36 (1), pp. 71–84, 2006.
[8] A. Verikas, A. Gelzinis, D. Valincius, M. Bacauskiene, V. Uloza, "Multiple feature sets based categorization of laryngeal images", *Computer Methods and Programs in Biomedicine* vol. 85 (3), pp. 257–266, 2007.
[9] M. Bacauskiene, A. Verikas, A. Gelzinis, D. Valincius, "A feature selection technique for generation of classification committees and its application to categorization of laryngeal images", *Pattern Recognition* vol. 42 (5), pp. 645–654, 2009.
[10] A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, "Integrating global and local analysis of colour, texture and geometrical information for categorizing laryngeal images", *International Journal of Pattern Recognition and Artificial Intelligence* vol. 20 (8), pp. 1187–1205, 2006.
[11] P. S. Popolo, I. R. Titze, "Qualification of a quantitative laryngeal imaging system using videostroboscopy and videokymography", *Ann. Oto. Rhinol. Laryn* vol. 117 (6), pp. 404–412, 2008.
[12] H. Rihkanen, P. Reijonen, S. Lehikoinen-Soderlund, E. R. Lauri, "Videostroboscopic assessment of unilateral vocal fold paralysis after augmentation with autologous fascia", *European Archives of Oto-Rhino-Laryngology* vol. 261, pp. 177–183, 2004.
[13] S. Deguchi, Y. Ishimaru, S. Washio, "Preliminary evaluation of stroboscopy system using multiple light sources for observation of pathological vocal fold oscillatory pattern", *Ann. Oto. Rhinol. Laryn* vol. 116 (9), pp. 687–694, 2007.
[14] S. H. Dailey, J. Kobler, R. E. Hillman, K. Tangrom, E. Thananart, M. Mauri, S. M. Zeitels, "Endoscopic measurement of vocal fold movement during adduction and abduction", *Laryngoscope* vol. 115 (1), pp. 178–183, 2005.
[15] A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, M. Kaseta, "Using the patient's questionnaire data to screen laryngeal disorders", *Computers in Biology and Medicine* vol. 39 (2), pp. 148–155, 2009.
[16] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004.
[17] J. A. K. Suykens, J. Vandewalle, "Least squares support vector machine classifiers", *Neural Processing Letters* vol. 9 (3), pp. 293–300, 1999.
[18] L. Wolf, A. Shashua, "Learning over sets using kernel principal angles", *Journal of Machine Learning Research* vol. 4, pp. 913–931, 2003.

ADVANCED PREPROCESSING OF LARYNX IMAGES TO IMPROVE THE SEGMENTATION OF GLOTTAL AREA

V. Osma-Ruiz [a], J.M. Gutiérrez-Arriola [a], J.I. Godino-Llorente [a], N. Sáenz-Lechón [a], R. Fraile [a], J.D. Arias-Londoño [b]

[a] E.U.I.T. Telecomunicación-Universidad Politécnica de Madrid, Cra Valencia Km 7, 28031, Madrid.
[b] Grupo de Control y Procesamiento Digital de Señales. Universidad Nacional de Colombia. Manizales
Corresponding autor: V. Osma-Ruiz (vosma@ics.upm.es)

Abstract: **The present work describes an advanced method for image preprocessing to improve the automatic detection of the glottal space from laryngeal images obtained either with high speed or with conventional video cameras attached to a laryngoscope. Images are filtered using an anisotropic diffusion technique that combines smoothing properties with image enhancement qualities. The preprocessing technique improves the performance of the previous system based in watershed transform and merging. Results show that 38% of the mismatches in delineating the glottis are fixed or reduced. 111 larynx images have been segmented to obtain the glottal area, 11 of the 29 previous errors have been corrected.**
Keywords: **Segmentation, preprocessing, anisotropic diffusion, glottis.**

I. INTRODUCTION

Pathologies that may affect voice production are many and varied. However, all tend to have a common effect, that is, difficulties to achieve a correct vocal fold vibration during phonation. This usually comes with deficiencies in the closure of the glottis what implies further distortion. Analysis of these two effects, particularly fold vibration, is essential for otolaryngology physicians to diagnose laryngeal disorders.

One of the main problems faced by specialists is the high speed of vocal fold movement, which makes it impossible for the human eye to see the vibration with enough accuracy. Several systems have been developed over the last century to overcome this drawback: subjective methods as stroboscopy [1], [2] and high speed recordings [3], [4]; or objective techniques such as kymography [5], glottal area diagrams [6] and phonovibrography [7]. The latter are becoming increasingly more important because they empower the expert to quantify the movement in addition to visualize it.

All objective techniques mentioned above need an image processing oriented to the segmentation of the glottal area, either as part of its development, either to resolve various errors introduced during the recording, like movements suffered by the recording device and/or the patient.

This paper describes an advanced method for image preprocessing to improve the automatic detection of the glottal space from laryngeal images obtained with the previous system [9] that combines several relevant techniques in the field of digital image processing.

The remainder of this paper is organized as follows: in section II the tools used to achieve accurate detection of the glottal area are described including the system already designed and preprocessing with the anisotropic diffusion filter. In section III results obtained after combining the two methods are discussed and section IV highlights the main conclusions.

II. METHODS

A. Segmentation system

The method described in [9] allows to individualize the glottis in laryngeal images following the scheme presented in Fig. 1. The operation of each of the blocks is as follows:

*Watershed transform [10] of the gradient image: t*he first step is to convert the original image (RGB) into a grey scale image by means of a transformation to the YIQ model. The luminance component (Y) is chosen and its
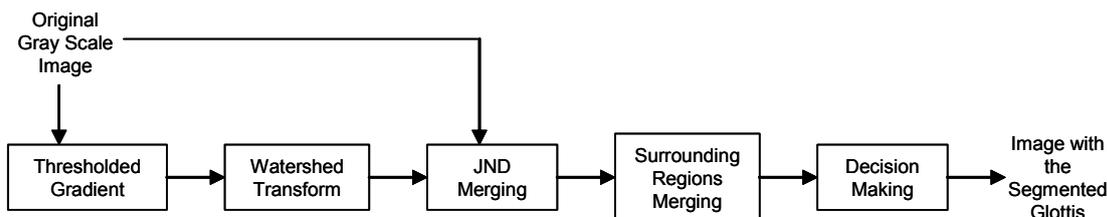
Fig. 1 - Scheme that represents the steps followed for the segmentation of the glottal space.

gradient is calculated. A threshold with a value of 2 is applied to the gradient image (i.e. those pixels of the gradient image with a grey level below 2 are assigned to 0), removing those edges that appeared due to the noise present in the image. After the thresholding, the watershed transform is applied to the resulting image, achieving the first region determination.

JND based merging: one of the drawbacks of the system is that Watershed transform is very sensitive to noise, causing the image to be divided into multiple regions where there are only a few (and the delimitation of only one is the goal of the system). The preprocessing mentioned above partially alleviates the problem but doesn't solve it. It is necessary to apply a subsequent merging to ensure the union of homogeneous regions. In this sense, the system presented in Fig. 1 introduces this block to merge the regions that are homogeneous to the human eye following on the basis of JND (Just Noticeable Difference) [11].

Surrounding regions merging: the third step consists of another merging process, now attempting to merge all the neighbours that surround a region with a lower grey level than all of them. Now the goal is to reduce the number of segmented objects by merging regions that can not correspond to the glottis (note that from a human observer's point of view, the glottis should always be a dark object surrounded by a lighter area).

Decision making: the last step is a classification process to detect the glottis among the rest of the objects present in the image. For this purpose, a linear predictor trained with the 7 invariant moments of the different objects is used.

111 images are processed to segment de glottal area. The presented system allows the automatic detection of the glottis in 75% of the analyzed images. In the remaining 25% glottis is detected varying one threshold. However, in 29 images small mismatches are presented in delineating glottal area boundaries, such as those presented in Fig. 2.

B. Preprocessing. Anisotropic Diffusion

A preprocessing more powerful than the presented in the system represented by Fig. 1 can reduce the number of regions resulting from the division provided by the Watershed transform, making the subsequent merging process easier and improving the results.

The image preprocessing will be then the combination of two processes: first the original gray scale image *I* will be smoothed without blurring the most significant edges with scale–spaced using anisotropic diffusion [13]; and

afterwards a threshold is applied to the gradient image to eliminate the remaining insignificant edges.
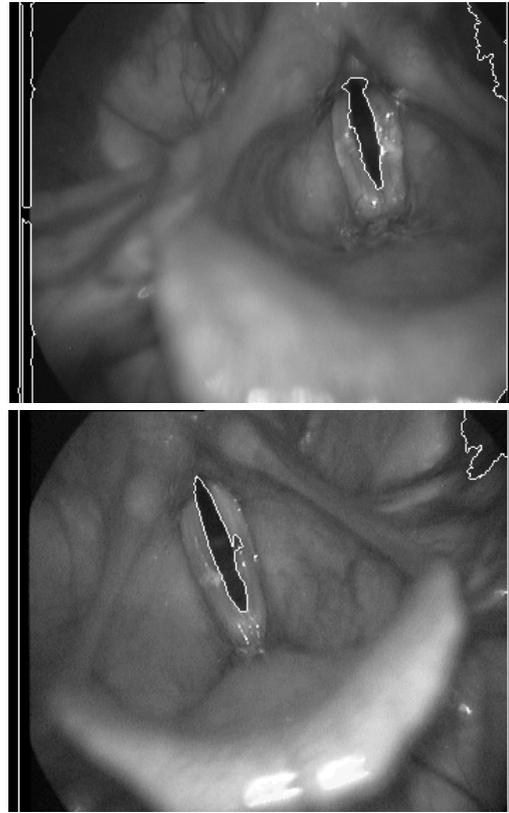
Fig. 2. Some mistakes reported by the system described in Fig 1.

The first process is implemented by equation 1 applied iteratively to each pixel of *I*.

$$I_{i,j}^{t+1} = I_{i,j}^t + \lambda \sum_{l=N,S,E,W} \left[c_l \cdot \nabla_l I \right]_{i,j}^t \qquad (1)$$

Where:
1. *i* and *j* are the row and column of the pixel.
2. *t* is th scale-space level, that is, the iteration number. Then $I_{i,j}^t$ is the gray value for pixel in row *i* and column *j* after *t* iterations.
3. ∇ is the intensity difference between the gray level of the pixel and that of its 4-conectivity neighbours (North, South, East and West)
4. *c* is the conduction coefficient and depends on the difference of the gray level in each direction. It must approximate zero for large differences where it is probable to locate an edge and almost one for small differences assuming the pixels are likely to belong to the same region. Equation 2 presents the function used in this work proposed in [5] for the North

direction. Conduction coefficients are calculated for each direction and each iteration.

$$c_{N_{i,j}}^{t} = e^{-\left(\frac{\left|I_{i-1,j}-I_{i,j}\right|}{K}\right)^2} \qquad (2)$$

5. λ controls diffusion speed and it should be under ¼ for the numerical scheme to be stable.

For each pixel (i,j) of the image, differences in the gray level are calculated in the four directions. Those differences that are high (representative of a border) will have a conduction coefficient equal to zero and will not affect the result of the iteration, while those that are small shall be added (or subtracted) to the value of the pixel in study. In this way, after the successive iterations, all of the pixels of the image tend to become closer to the similar neighbours maintaining the highest differences in gray level.

As example two images of the larynx are shown in Fig. 3: the image on the left is the result of a filtering with anisotropic diffusion (K=10, λ=0,2 and 50 iterations); the image on the right is the low pass filtered version of the same original image. It can be observed that anisotropic diffusion homogenizes the different tissues of the larynx without damaging most significant edges, while the standard filter creates a blur in the image.
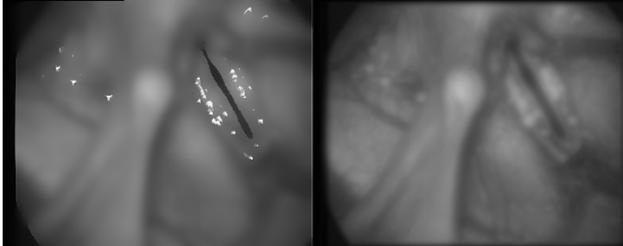
Fig. 3. Example of anisotropic diffusion. Left: image obtained from anisotropic diffusion after 50 iterations. Right: low pass filtered image.

III. RESULTS

The research was conducted in two steps:
1. Anisotropic diffusion as described in paragraph II-B is added to the system presented in Section II-A and results are analyzed. 12 images were randomly chosen: in 6 of them the glottis was correctly segmented with the original system and the other 6 presented some mismatches in glottal area delimitation. Diffusion parameters are varied to finally obtained 140 combinations:
   - K varies from 3 to 18 with a step of 5
   - λ varies from 0,05 to 0,25 with a step of 0,05
   - The number of iterations ranges from 5 to 65 with a step of 10

The four combinations that gave the best results were selected to carry out the next experiment. It is important to point out that as important as an improvement on a bad segmented image is that the correct segmentation remains unchanged.

2. The 4 combinations selected are used to segment all the 111 available images. Best results are obtained with the following values: K=8, λ=0,05 and 55 iterations. The error is removed in 14 out of 29 images and improved in 4. However a new mismatch appears in 7 of the 82 images that were correctly segmented with the original system. The overall result shows improvement in 11 images representing a percentage of 38% of errors fixed.

Results for the images shown in Fig. 2 are presented in Fig. 4. It can be observed that the glottis is detected without errors.
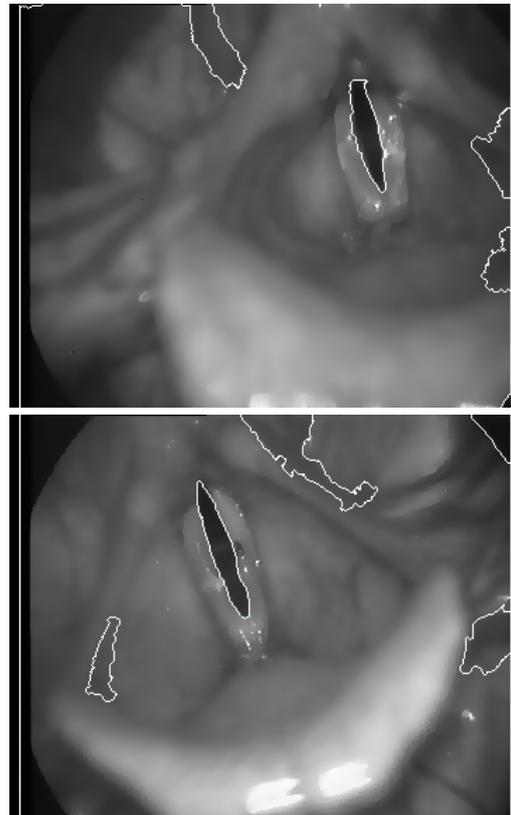
Fig. 4. These two images presented mismatches in the glottal area delimitation when processed with the original system as shown in Fig. 2. Error has been corrected with the anisotropic diffusion filter.

Nevertheless in Fig. 5 two new errors are shown. The glottal area was correctly segmented with the system described in section II-A but it presents same mismatch when anisotropic diffusion is included in the system.
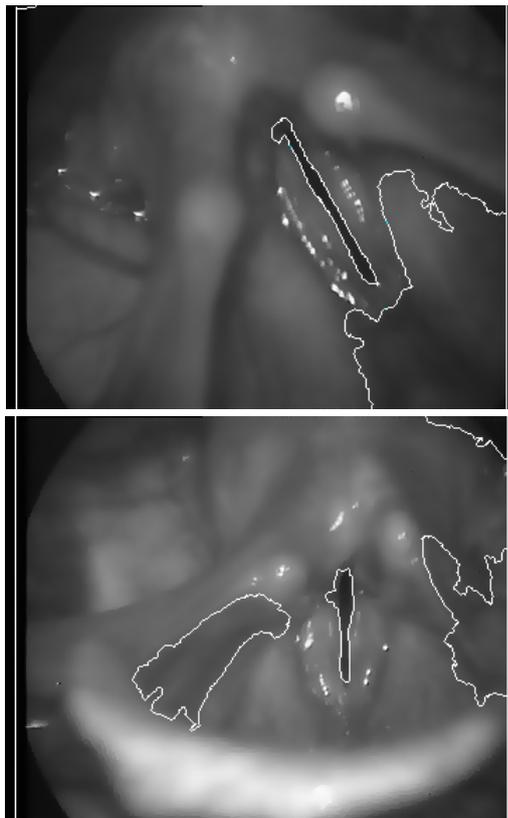
Fig. 5. Two images with an error when anisotropic diffusion is applied as a preprocessing technique.

IV. CONCLUSIONS

Anisotropic diffusion is an image preprocessing method that can increase the homogeneity of areas with similar gray levels, while maintaining, and even enhancing, the edges that separate areas with abrupt changes.

Anisotropic filtering has been used as a preprocessing step to improve the performance of a system that detect the glottal space from laryngeal images [9]. The main goal is to correct certain errors that appear in the delineation of the glottis.

With an anisotropic diffusion filter with parameters: $K=8$, $\lambda=0,55$ and 55 iterations, 111 laryngeal images have been processed, 29 of them with previous errors in the delineation of the glottis. Results after segmentation solve the problem in 14 cases and significantly improve another 4 images. On the other hand, 7 of the 82 remaining images have an error that was not previously present. Given these two aspects we can conclude that the preprocessing method presented in this paper achieved an improvement of 38% over the previous system.

REFERENCES

[1] Oertel,M.J., "Über eine neue 'laryngostroboskopische' untersuchungsmethode des kehlkopfes," *Zentralbl.f.d. Mediz. Wissenschaften Heft*, vol. 16, pp. 81-82, 1878.

[2] Rosen, C. A., "Stroboscopy as a research instrument: development of a perceptual evaluation tool," *Laryngoscope*, vol. 115, no. 3, pp. 423-428, 2005.

[3] Schwarz, R., Hoppe, U., Schuster, M., Wurzbacher, T., Eysholdt, U., and Lohscheller, J., "Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1099-1108, 2006.

[4] Zhang, Y., Bieging, E., Tsui, H., and Jiang, J. J., "Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging," *Journal of Voice*, 2009, In press.

[5] Wittenberg, T., Tigges, M., Mergell, P., and Eysholdt, U., "Functional imaging of vocal fold vibration: digital multislice high-speed kymography," *Journal of Voice*, vol. 14, no. 3, pp. 422-442, 2000.

[6] Yan, Y., Ahmad, K., Kunduk, M., and Bless, D., "Analysis of vocal-fold vibrations from high-speed laryngeal images using a Hilbert transform-based methodology," *Journal of Voice*, vol. 19, no. 2, pp. 161-175, 2005.

[7] Lohscheller, J., Eysholdt, U., Toy, H., and Dollinger, M., "Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2D-diagrams for visualizing and analyzing the underlying laryngeal dynamics," *IEEE Transactions on Medical Imaging*, vol. 27, no. 3, pp. 300-309, 2008.

[8] Manfredi, C., Bocchi, L., Bianchi, S., Migali, N., and Cantarella, G., "Objetive vocal fold vibration assessment from videokymographic images," *Biomedical signal processing and control*, vol. 1, no. 2, pp. 129-136, 2006.

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

[10] Osma-Ruiz, V. J., Godino-Llorente, J. I., Sáenz-Lechón, N., and Gómez-Vilda, P., "An improved watershed algorithm based on efficient computation of shortest paths," *Pattern Recognition*, vol. 40, no. 3, pp. 1078-1090, 2007.

[11] Shen, D. F. and Huang, M. T., "A watershed-based image segmentation using JND property," in *Proceedings of IEEE ICASSP 2003*, vol. 3, pp. 377-380, Apr.2003.

[12] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*, 2 ed., Wiley-Interscience, 2001.

[13] Perona, P. and Malik, J., "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629-639, 1990.

ARTICULATORY MODELLING OF THE VOCAL TRACT IN FEEDING FROM X-RAY IMAGES

A. Serrurier[1], A. Barney [1]

[1] Institute of Sound and Vibration Research, University of Southampton, UK

Abstract: **Two of the major functions of the human vocal tract are feeding and speaking. Ontogenetically and phylogenetically feeding tasks precede speaking tasks and it has been hypothesized that speaking movements constitute a subset of feeding movements. This study investigates whether the vowels /a/ /i/ /u/ can be articulated using feeding movements. Midsagittal tongue surfaces have been extracted from a Digital Videofluoroscopy film of liquid swallowing, and a 5-component articulatory model has been derived, explaining 96% of the tongue variance. Acoustic transfer functions have been estimated by means of an expansion model from midsagittal measurements to area function and an acoustic wave propagation model. The articulations optimally approaching the acoustic and articulatory characteristics of /a/ /i/ /u/ have been extracted from both the data and the model. The results show that the model can produce three /a/ /i/ /u/-like articulations whose points in the acoustic plane F1-F2 reach the /a/ /i/ /u/ ellipses of the literature, suggesting that speech articulations could indeed be producible from feeding movements. These results support the hypothesis that speech movements might have evolved from feeding movements.**
Keywords : **Speech, Feeding, Articulatory Modelling**

I. INTRODUCTION

Three main functions can be ascribed to the human vocal tract: breathing, eating and speaking. The tasks of eating and speaking are associated with involved articulatory movements requiring significant coordination. Since both ontogenetically and phylogenetically feeding tasks precede speaking tasks, MacNeilage [1] suggests in his significant paper on the frame/content theory of the evolution of speech production that speech cyclicity may have evolved from feeding cyclicity. Following this theory, Hiiemae & Palmer [2] suggest that, although controversial, it is reasonable to hypothesis that tongue movements in speech are derived from tongue movements in feeding. Our study explores this hypothesis on the basis of articulatory measurement and modelling to determine if the movements found in speech are a subset of those found in feeding. We present here an analysis based on an imaging study of the vocal tract (VT) during feeding.

An earlier study [3] used ElectroMagnetic Articulography (EMA) to record the motion of the jaw tongue and lips during speech and feeding tasks. The articulatory model developed from these data recorded on a single subject suggested that the movements of the tongue and jaw in speech might indeed form a subset of the tongue- and jaw-based feeding movements. For speech the underlying assumptions of models from EMA data have generally been well tested, but for feeding we must question whether they are equally valid. For example it has been demonstrated that three tongue points, as measured by EMA, are sufficient to recover the full 3D shape of the tongue for speech tasks (see e.g. [4]), but no such validation has been attempted for feeding. A similar question arises for other articulators. Further, EMA gives no data about the static parts of the VT boundary and therefore does not allow estimates of the acoustic response associated with the various geometrical configurations adopted in feeding.

To overcome these limitations it is necessary to use a data collection method that images the VT boundaries from glottis to lips but also has sufficient time resolution to capture the dynamics of the motion of the articulators. We have therefore chosen to use Digital Video Fluoroscopy (DVF) which permits the recording of X-ray films with a relatively low dose exposure. For the work presented here we will focus on the motion of the tongue only.

In the debate on speech evolution, emphasis has been placed on the three point vowels /a/, /i/ and /u/ known as the quantal vowels. These three vowels are considered to delimit acoustically the maximum space of a vowel system (e.g. [5]). We are therefore interested here in whether we can extract from the data, or from an articulatory model based on the data, articulations which are geometrically and/or acoustically close to the quantal vowels. As benchmarks we define the acoustic targets as typical F1-F2 values for /a/, /i/ and /u/ and the articulatory targets as the tongue shapes representing the typical /a/, /i/ and /u/ geometry [6].

II. METHOD

Data was collected at Hospitais da Universidade de Coimbra, Portugal using a single, native European Portuguese, female subject, age 40 years, with no known pathology of the upper respiratory tract and normal

swallowing function. Full local ethical approval was obtained prior to carrying out the study.

DVF images were recorded during feeding tasks of the upper VT from the lips to just below the glottis with a frame rate of 25 images/s. The feeding corpus [3] was designed in collaboration with a speech and language therapist guided by the UK National Descriptors for Texture Modification in Adults [7].

The food was divided into three categories covering a wide range of food textures from saliva to Hard Food. For this study, DVF images were obtained of Liquids (water, pourable custard), Solid Food (thick custard) and Hard Food (shortbread biscuit). A typical image from the DVF sequence is shown in Fig. 1a.

| a) Original DVF image | b) Manually segmented image |
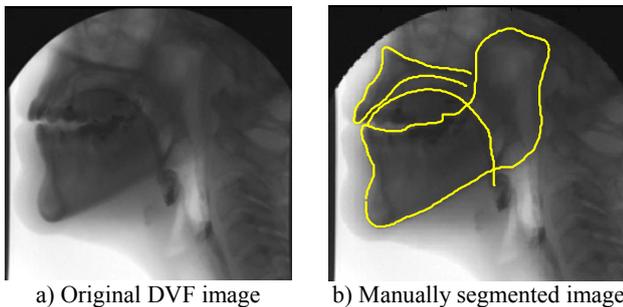
Fig. 1 *A typical frame from the DVF image sequence (a) and the manual segmentation of the hard palate the tongue and the jaw contours (b).*

Ideally, all images from all recordings would be analysed. Manual processing of images is, however, time consuming, so we have limited our pilot study to the recordings of pourable custard; a film sequence lasting about 14 sec. and containing two complete swallows. This choice is motivated by our EMA study [3] where we could extract most of the representative articulations from liquid swallowing. Custard was chosen here over water since it presents more contrasted images which facilitate analysis. Additionally, to make processing tractable only every second image has been analysed (i.e. 186 images), equivalent to a frame rate of 12.5 images per second.

Each image has been manually segmented to extract the contours of the tongue from epiglottis to tip. Despite some spatial blurring and some masking of the tongue surface by other anatomical features on certain images, the complete tongue surface has been carefully outlined (Fig. 1b) by the first author of this article, who has significant experience in manual VT segmenting. The contour of the hard palate has also been extracted for each image to act as a reference for image alignment.

Finally, the full set of VT articulators (i.e. epiglottis, larynx, back pharyngeal wall, velum, jaw, and lips) has been outlined for one specific articulation for use as a basis for acoustic modelling. The entire set of contours has then been calibrated in cm and aligned using the hard palate reference. Following common practice these contours are considered as the midsagittal outlines of the VT (see e.g. [8]).

Two methods were used to asses whether the feeding articulations gave good approximations to the targets, one based on the raw tongue profile data and the other based on an articulatory tongue model derived from the data.

The modelling process consists of extracting the statistically independent articulatory components of the tongue movement during the feeding task. The articulatory model has the advantage over the raw data of encompassing all task-derivable articulations theoretically producible by the tongue even though they may not be present in the recorded data. The tongue surface modelling was based on Principal Component Analysis (PCA) following the protocol described in [9]. Each tongue contour was re-interpolated using 200 regularly spaced points from tongue tip to epiglottis. A PCA has been applied to the $186\square200\square2$ X-Y coordinates of the tongue points. Table 1 shows that five components are enough to explain 96% of the variance of the tongue points with a reconstruction error of 0.12 cm.

Table 1 *Explained variance, cumulative explained variance and cumulative RMS reconstruction error for the tongue for each of the articulatory parameters*

Parameter	Var.	Cum.Var.	RMS
P1	62 %	62 %	0.35 cm
P2	17 %	79 %	0.26 cm
P3	9 %	88 %	0.2 cm
P4	5 %	93 %	0.15 cm
P5	3 %	96 %	0.12 cm

For both the articulatory and the acoustic cases we have defined a distance measure for the error between prediction and target. The articulatory distance is the Root Mean Square (RMS) distance between the target sets of points and a measured set describing the tongue surface. The acoustic distance is the relative RMS distance between a target set of formants and those derived from an estimated acoustic transfer function.

The process to obtain from the data and from the model the closest articulations to our articulatory and acoustic targets was as follows:

(1) An optimisation was performed on the articulatory model parameters to obtain the tongue contours that minimized the articulatory distance to three manual approximates of the articulatory target contours for the quantal vowels (Figs. 2a to 2c).

(2) The three tongue contours were inserted into a fixed, midsagittal VT contour (Fig. 2d). Note that currently the epiglottis and the larynx are fixed, leading sometimes to some anatomical abnormalities in the lower pharyngeal region (see e.g. Fig. 2a).

(3) The VT contours with the three tongue articulations were intersected from the glottis to the lips by a fixed grid (Fig. 2d).

(4) For each grid line we measured the distance between the points of intersection with the VT contours, considered as the diameter of the VT tube for the purpose of area function generation. Wherever grid lines were not perpendicular to the VT tube central axis, the measured distances were corrected using the method described by [9].

(5) Area functions were computed from each of the three sets of midsagittal distances according to the alpha-beta model [10] using values appropriate for women (Fig. 3). Since articulations represent vowels, areas calculated as less that 0.1 cm2 have been set equal to 0.1 cm2.

(6) Planar acoustic wave propagation was simulated with these area functions as described by [6]. The area function for /u/ was artificially lengthened by a tube 2 cm long and 0.2 cm2 in area to simulate the expected lip protrusion.

(7) F1-F2 for each of the acoustic transfer functions were extracted. The acoustic distances between these F1-F2 and the target F1-F2 were computed.

(8) An optimisation loop on the model parameters including steps (2) to (7) was performed to obtain the three tongue contours minimizing the acoustic distance to the quantal vowel targets. These articulations and associated area functions are considered as the closest model articulations to the acoustic targets (solid lines Figs. 2 & 3).

(9) The articulations from the data with minimal articulatory distance to the articulations obtained in step (8) were finally extracted (dashed lines Figs. 2 & 3).

III. RESULTS

Three articulations extracted from the DVF film of thin custard swallowing, have been selected as the closest articulatory and acoustic articulations to the quantal vowels. Similarly, three articulations have been produced from the articulatory model as the closest articulatory and acoustic estimates theoretically producible by the subject.

We observe (Figs. 2 & 3) that the tongue shape and position match well overall to the pattern expected for these articulations: low and backwards for /a/; up and forwards for /i/; and up and backwards for /u/ although two constrictions are observed for the midsagittal contour of the model /u/ (solid line Fig. 2c).

Predicted formants are shown in the F1-F2 plane in Fig. 4 with dispersion ellipses adapted from [11]. For the articulations extracted from the data, none of the F1-F2 points are inside the ellipses. The low F1 observed for /i/ and /u/ may relate to the smaller than expected constrictions in the area functions (in fact complete occlusion of the tract (see Figs. 2b & 2c); their null measurements have been artificially set to 0.1 cm2 in

the area functions (Fig. 3). The F1-F2 computed for the three model articulations fall just inside the ellipses, (on the border for /a/ and /i/).

a) /a/   b) /i/

c) /u/   d) Intersection with the grid
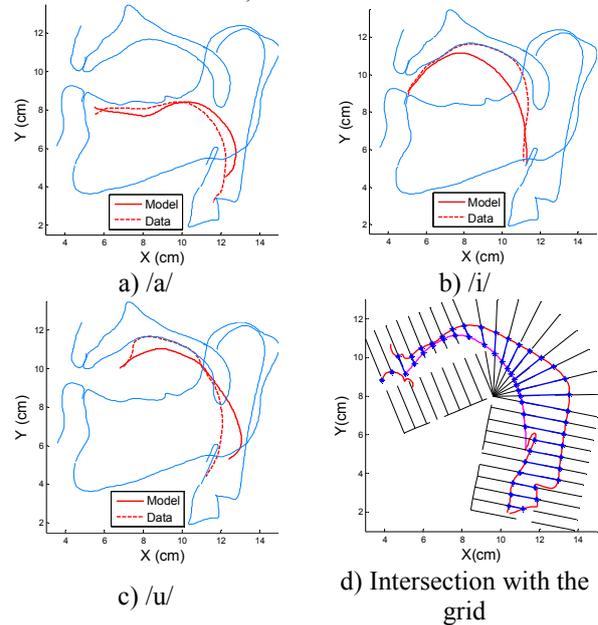
Fig. 2 *Midsagittal contours of the tongue (a, b, and c) from the model (solid line) and from the data (dashed line) having best fit to articulatory targets /a/, /i/ and /u/, plotted on fixed midsagittal VT contours to give context. In d, intersection of the VT articulated for a /i/ with the grid.*
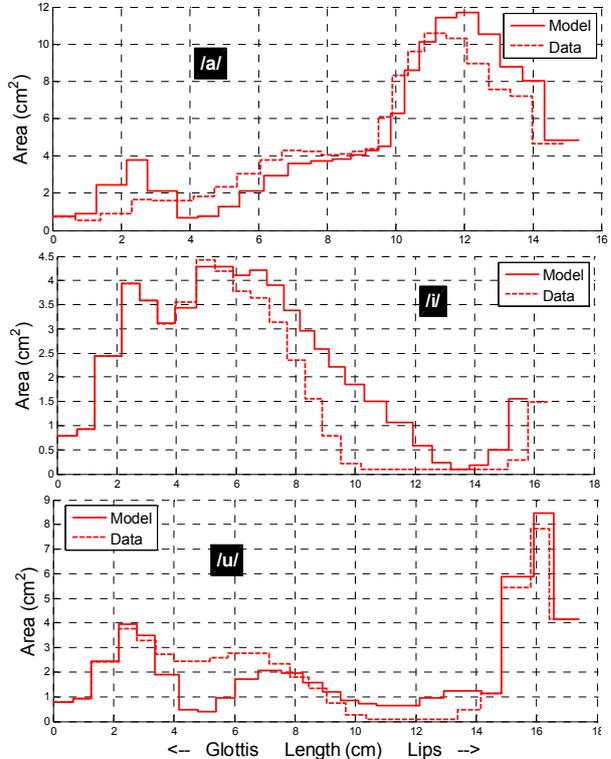
Fig. 3 *Area functions derived from the articulations plotted in Fig. 2.*
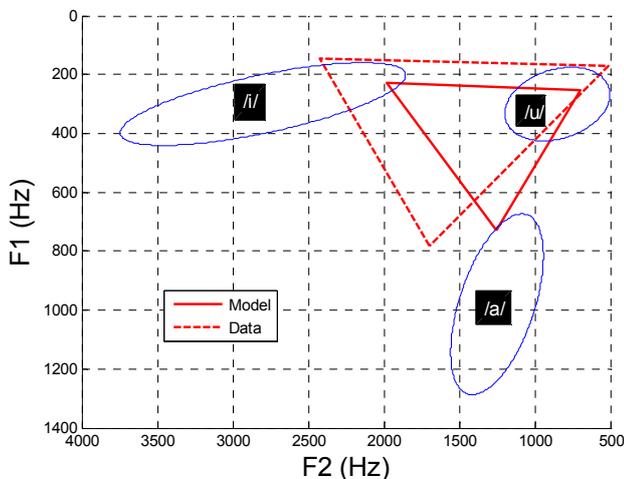
Fig. 4 *Position in the F1-F2 plane of the articulations of Figs. 2 & 3. The ellipses for the vowels /a/, /i/ and /u/ are adapted from [11].*

IV. DISCUSSION AND CONCLUSION

Our results show that while feeding we can articulate tongue shapes that match the typical midsagittal geometry of the quantal vowels with predicted F1-F2 close to the corresponding ellipses reported by [11].

Using an articulatory model, we have shown that the feeding movements representative of liquid swallowing allow articulation of tongue shapes which follow the typical midsagittal patterns found for the quantal vowels with predicted F1-F2 falling within the corresponding ellipses reported by [11]. In other words, it seems possible to articulate the point vowels /a/, /i/ and /u/ from feeding movements.

This article complements our previous study based on EMA recordings of a French male subject [3] by investigating another recording technique and a new subject (Portuguese, female). Our results support the general conclusion of [3] that speech articulations can largely be found within the set of feeding movements. More generally, our findings support the hypothesis that speech movements might evolve from feeding movements with the caveat that we have not considered the question of the control.

The results obtained from this pilot study appear promising for the future. We have shown the benefit of using DVF films which allow derivation of acoustic propagation simulations in the VT. An extended study should however include a more representative task of feeding extending analysis beyond the two swallows of liquid considered here to the complete data set. We have moreover seen that limiting our analysis to the tongue contours may lead to geometrical artefact in the predicted VT. Future work should also include the other articulators which impact on area function computation (e.g. larynx, velum, lips).

REFERENCES

[1] P. F. MacNeilage, "The frame/content theory of evolution of speech production". *Behav. Brain Sci.*, vol. 21, pp 499-546, 1998.

[2] K. M. Hiiemae and J. B. Palmer, "Tongue Movements in feeding and speech", *Crit. Rev. Oral Biol. M.*, vol. 14(6), pp. 413-429, 2003.

[3] A. Serrurier, A. Barney, P. Badin, L.-J. Boë and C. Savariaux, "Comparative articulatory modelling of the tongue in speech and feeding." *Proc. 8th ISSP*, Strasbourg, France, pp 325-328, 2008.

[4] Y. Tarabalka, P. Badin, F. Elisei and G. Bailly, "Can you "read tongue movements"? Evaluation of the contribution of tongue display to speech understanding." *Proc. ASSISTH'2007*, France, 2007.

[5] J. Liljencrants and B. Lindblom, "Numerical simulation of vowel quality systems: the role of perceptual contrast", *Language*, vol. 48, pp. 839-862, 1972.

[6] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton & co., 1960.

[7] National Descriptors for Texture Modification in Adults. British Dietetic Association and the Royal College of Speech and Language Therapists. http://tinyurl.com/mf9lsc, last accessed September 3rd 2009.

[8] D. Beautemps, P. Badin and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modelling", *J. Acoust. Soc. Am.,* vol. 109(5), pp. 2165-2180, 2001.

[9] A. Serrurier and P. Badin, "A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data." *J. Acoust. Soc. Am.,* vol. 123(4), pp. 2335-2355, 2008.

[10] A. Soquet, V. Lecuit, T. Metens and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI", *Speech Commun.*, vol. 36(3-4), pp. 169-180, 2002.

[11] G.E. Peterson and H.L. Barney, "Control methods used in a study of the vowels", *J. Acoust. Soc. Am.*, vol. 24, pp. 175-184, 1952.

SEGMENTATION OF LARYNGEAL HIGH-SPEED VIDEOENDOSCOPY IN TEMPORAL DOMAIN USING PAIRED ACTIVE CONTOURS

Habib J. Moukalled[1,2], Dimitar D. Deliyski[1,2], Raphael R. Schwarz[1,3], Song Wang[2]

[1] Department of Communication Sciences and Disorders, University of South Carolina, Columbia, SC, USA
[2] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA
[3] Siemens AG, Erlangen, Germany

This paper introduces a method for segmention of the vocal-fold edges in temporal domain from laryngeal high-speed videoendoscopy (HSV). The method employs a pair of active contours (snakes), which deform within a series of kymographic images derived from the HSV data. By following a set of deformation rules, this pair of active contours converges to the desired boundaries of the glottis. The proposed method was tested on a dataset of 98 HSV samples, of which 96 were successfully segmented. The new method substantially outperforms existing methods. However, more precise analysis revealed that of the 96 successfully segmented HSV samples, 18 exhibited a fine error up to ±1 pixel, and 78 samples exhibited errors exceeding a pixel. The large majority of the gross errors (76%) were due to problems near the posterior and anterior commissures, which warrants further investigation for improving the accuracy and reliability of the method.

Keywords: **high-speed videoendoscopy, active contour segmentation, snakes, glottis, digital kymography**

I. INTRODUCTION

Laryngeal high-speed videoendoscopy (HSV) contains unprecedented amount of information about the vibration of the vocal folds that is potentially clinically useful. However, navigating through the enormous amount of HSV data is difficult and impractical. In order for HSV to gain widespread clinical use, there is a need for image-processing algorithms for automatic extraction of the relevant vocal-fold vibratory features. That is the long-term purpose of this project.

This problem has been investigated in recent years. Yan *et al.* developed an algorithm to segment the glottis from HSV data by globally thresholding pixel intensities on a per-frame basis [1]. Lohscheller *et al.* developed an algorithm that takes advantage of HSV's 3D structure by performing a modified 3D seeded region growing for segmentation of the glottis and post-processing for reconstruction of the vocal-fold boundaries [2]. However, such local image thresholding or region growing algorithms are usually sensitive to image homogeneity and noise.

Active contours, or snakes, are deformable models that can dynamically converge towards the desired image features [3]. The deformation of a snake follows certain specified rules on the whole contour, which may make it more robust to image noise. A closed-loop snake has been used to analyze the PE-segment within HSV data [4]. A pair of open-curve snakes has been applied to the right and left vocal folds to segment the glottis from videokymography [5]. A HSV movie can be represented in temporal domain as a digital kymography (DKG) playback [6]. Therefore, an attractive approach for HSV-segmentation can be achieved by segmenting the glottis from all spatial-temporal kymographic images of HSV.

In this study, we employed a pair of open-curve snakes (right and left) on DKG images to segment the glottis, for which deformation rules enforce the temporal resolution of HSV. Fig. 1 illustrates two open-curve snakes, right and left, attracted to pixels with large gradient magnitude (aligned with the glottal boundaries), which is derived from DKG. In Fig. 1 (not drawn to scale) the white squares are the vertices, termed snaxels, which make up the right and left snakes. The white lines connecting the snaxels are spline segments. And the space between the vertical white lines denotes time.
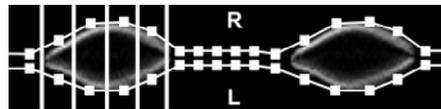
Fig. 1: Snakes are attracted to the pixels with large magnitude of gradient within a kymographic image.

The proposed method exhibits the following merits over previous methods: (a) the snake convergence is facilitated due to the absence of complex geometries in kymographic images; (b) the deformation of the snakes can be optimized by using time-delayed discrete dynamic programming; (c) the temporal resolution of HSV helps constrain snake deformation since DKG images exhibit continuity along the time axis; (d) the method is robust to the disappearing glottis during the closing phase; (e) the initialization procedure is simple and scalable; and (f) the method segments the right and left vocal-fold edges concurrently, while maintaining separate left and right segmentation results.

II. METHOD

A. Snake -energy Minimization.

Energy Minimizing Splines. A snake is a spline deformed in the spatial domain of a digital image in order

to minimize an energy functional comprised of internal forces derived from the snake's shape, and external forces derived from image features [3]. A snake is parameterized by the vector $v(s) = [x(s),y(s)]$, where $s \in [0,1]$, and seeks to minimize the following energy functional [3,7]:

$$E = \int_0^1 E_{\text{int}}(v(s)) + E_{image}(v(s))ds \quad (1)$$

The internal force E_{int} acting on snake $v(s)$ is a soft constraint used to make the snake's shape smooth and is given by:

$$E_{\text{int}}(v(s)) = \frac{1}{2}[\alpha(s) | v'(s) |^2 + \beta(s) | v''(s) |^2] \quad (2)$$

Where $v'(s)$ and $v''(s)$ are the first and second derivatives, respectively; α and β are two weights used to adjust the snakes elasticity and rigidity, respectively, which in turn influences the snake's shape. The image forces E_{image} acting on the snake $v(s)$, is a force that counter-balances the internal force E_{int}, and makes the snake align with desirable image features. For example E_{image} can be:

$$E_{image}(v(s)) = -| \nabla I(x,y) |^2, \quad (3)$$

where ∇I is the image gradient. By combining Eqs. (1) and (2) we obtain:

$$E = \int_0^1 \frac{1}{2}[\alpha(s)|v'(s)|^2 + \beta(s)|v''(s)|^2] + E_{image}(v(s))ds. \quad (4)$$

Using the calculus of variations, Eq. (4) has a numerical solution that can be obtained in $O(n)$ time [3]. By using specialized external force fields the convergence of snakes using the variational calculus framework can be significantly improved.

Snake Deformation Rules: In order to enhance convergence of the paired temporal snakes, three snake-deformation rules are applied: (1) no closed loops are permitted in the right and left snakes (i.e. snaxels of right and left snakes are defined by the time axis of kymographic images and can only move up or down within a kymographic slice during deformation); (2) in the absence of glottal-edge information right and left snakes are attracted to each other (i.e. regions in the kymograms where the vocal folds are in contact); and (3) right and left snakes are not allowed to pass each other in the deformation.

Time-delayed Discrete Dynamic Programming. The variational calculus framework for snake-energy minimization uses higher-order derivatives in order to approximate an energy minimizing spline from discrete data. Hard constraints are typically non-differentiable; as a consequence numerical instability occurs. In order to

overcome the instability of variational approaches, snake energy is minimized using discrete dynamic programming [7].

The discretization of the internal energy term of a snake given in Eq. (2), yields:

$$E_{\text{int}}(v_i) = \frac{1}{2}[\alpha_i | v_i - v_{i-1} |^2 + \beta_i | v_{i+1} - 2v_i + v_i |^2], \quad (5)$$

where v_i corresponds to the ith snaxel. By discretizing Eq. (4) we obtain:

$$E_{total} = \sum_{i=1}^n \left(E_{\text{int}}(v_i) + E_{image}(v_i) \right), \quad (6)$$

which can be viewed as a discrete multistage decision-making process, or better yet, a dynamic-programming problem [7].

Before dynamic programming can be applied, we must make the observation of a correspondence between minimizing the total energy measure of a snake and the problem of minimizing a function of the form [7]:

$$E_{total}(v_1,v_2,...,v_n) =$$
$$E_1(v_1,v_2,v_3) + E_2(v_2,v_3,v_4) + ... + E_{n-2}(v_{n-2},v_{n-1},v_n), \quad (7)$$

where each v is a state variable that can take m possible values. In the general case,

$$E_{i-1}(v_{i-1},v_i,v_{i+1}) = E_{\text{int}}(v_{i-1},v_i,v_{i+1}) + E_{image}(v_i). \quad (8)$$

Now, the dynamic programming solution involves generating a sequence of functions of one variable, $\{S_i\}_{i=1}^{n-1}$ (the optimal value function), where for obtaining each S_i a minimization is performed over a single variable. For example, given the energy function shown in Eq. (8), with $n = 4$, we have:

$$S_1(v_2,v_1) = \min_{v_0}[S_0(v_1,v_0) + E_1(v_1,v_2,v_3)],$$
$$S_2(v_3,v_2) = \min_{v_1}[S_1(v_2,v_1) + E_2(v_2,v_3,v_4)],$$
$$S_3(v_4,v_3) = \min_{v_2}[S_2(v_3,v_2) + E_3(v_3,v_4,v_5)],$$
$$\min_{v_0,...,v_4} E(v_0,v_1,v_2,v_3,v_4) = \min_{v_3}[S_3(v_4,v_3) + E_4(v_4,v_5,v_6)]. \quad (9)$$

And in the general case [7],

$$S_i(v_{i+1},v_i) =$$
$$\min_{v_{i-1}} [S_{i-1}(v_i,v_{i-1}) + E_i(v_i,v_{i+1},v_{i+2})]. \quad (10)$$

The discrete dynamic-programming solution for snake-energy minimization has a $O(nm^2)$ memory requirement and $O(nm^3)$ theoretical complexity, where n is the total number of stages (number of snaxels) and m is the total number of decisions at a given stage (neighborhood size).

Fig. 2 gives insight to the dynamic programming solution for snake-energy minimization as a pair of

temporal snakes deform within a glottal opening in a DKG image. The gray tiles of Fig. 2 represent the magnitude of the gradient, black tiles correspond to the snaxels of the right snake, black tiles with a dot correspond to snaxels from the left snake, and black tiles with a square correspond to snaxels where the right and left snake are overlapping. During the snake-deformation procedure snaxel movement is limited to a column-wise neighborhood, which prohibits the occurrence of closed loops (self intersections) in the right and left snakes and significantly reduces the search space needed for snake-energy minimization.
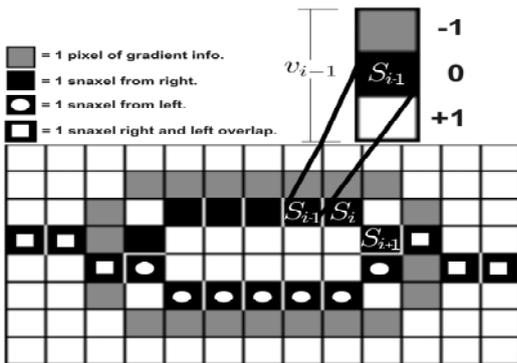
Fig. 2: Snake energy is minimized by finding the optimal state variable v_{i-1} along the direction of the column.

B. Experimental Design.

DKG Snake Toolbox. In order to test the new method, a custom software, DKG Snake Toolbox, was built. It allows a user to scroll through the HSV and DKG frames, which are dynamically linked. After the user manually marks the anterior and posterior commissures, an initial DKG image is selected at the 50% anterior-posterior level. Then, a snake-initialization line is placed in the middle of the glottis, spanning through the time axis of the kymogram. The right and left snakes are deformed in order to segment the glottis. After the result is verified, the remaining DKG images are automatically segmented.

Preprocessing the HSVs. The laryngeal tissues being observed are covered with a superficial layer, the lamina propria, which is highly reflective due to hydration and/or mucus presence. In general, light reflections represent a significant problem for snakes, because they introduce spurious noise into the gradient maps governing the snake deformations. Through the duration of a HSV recording, glottal openings exhibit distinctly dark intensities. Pixels having intensity values higher than the median intensity value of the entire recording more than likely correspond to light reflections, which can be easily suppressed.

Once light reflections have been suppressed, specialized gradient maps are computed. The gradient in the spatial domain is calculated for every frame of the HSV. Since the snaxels of the right and left snakes are

restricted to column-wise neighborhoods of movement, calculation of the gradient is performed using only the rows of a given frame in order to accent the horizontal edge information. A custom gradient map with gradients normal to the right vocal-fold edge is computed for the right snake, and a gradient map with gradients normal to the left vocal-fold edge is computed for the left snake.

Contour Embedding. In order to keep the right and left snakes attracted to each other in regions of the kymogram where the vocal folds are in contact, a new parameter, snake intensity (*snakeInt*), is devised. After each iteration of the dynamic programming, the right and left snakes are embedded in the opposing snake's edge map as a salient edge with intensity values between 0 and 255. This effectively bounds the right snake between the right vocal-fold edge and the left snake, and the left snake is bounded by the right vocal-fold edge and the right snake. This can prevent the right and left snakes from moving across one another during deformation.

Human Data. Fourteen vocally-normal speakers (7 men and 7 women between 22 and 29 years of age) have been recorded using with a Phantom V.7.1 (Vision Research, Inc., Wayne, NJ) monochromatic camera (16,000 fps, 320x320 pixels, 12-bit depth) connected to a 70° rigid laryngeal endoscope and a 300-W xenon light source. Each speaker produced the vowel /i/ in seven phonatory conditions, varying in register, pitch and loudness. Thousand-frame tokens of sustained phonation have been extracted from each recording to yield a total of 98 HSV samples.

III. RESULTS AND DISCUSSION

Snake Parameter Adjustment. In early works on snakes, the parameters α and β were shown to be sensitive parameters used to weight the snake model's continuity and rigidity, respectively. In the time-delayed discrete dynamic programming algorithm, α and β are not as sensitive as their classical counterparts [7]. For all results obtained in this paper, we have set $\alpha = 10$ and $\beta = 3$. The only parameters that have been adjusted were the *snakeInt* and the column-wise neighborhood size (*colSize*). *colSize* and *snakeInt* are adjusted twice per recording, once in order to initialize the right and left snakes, and one additional time for the automated segmentation stage.

Figs. 3 and 4 show the values of *colSize* and *snakeInt* used for the initialization and segmentation stages for female and male subjects, respectively. Fig. 5 provides an example of (A) the initial positions of the right and left snakes in the toolbox, (B) the deformation results for the initial kymogram, and (C) phases of the opening cycle with the deformation results (the white contours along the glottis) presented in the spatial domain of the HSV for a female subject.
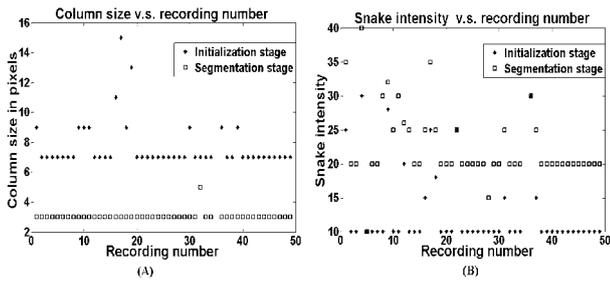
Fig. 3: *colSize* and *snakeInt* for 49 female subjects.
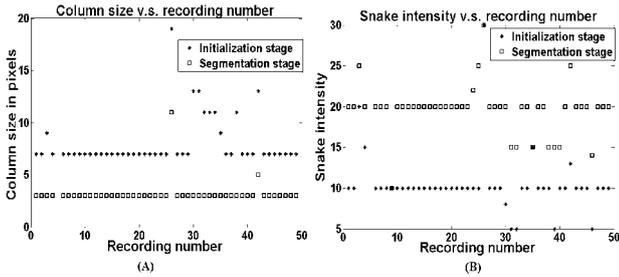
Fig. 4: *colSize* and *snakeInt* for 49 male subjects.
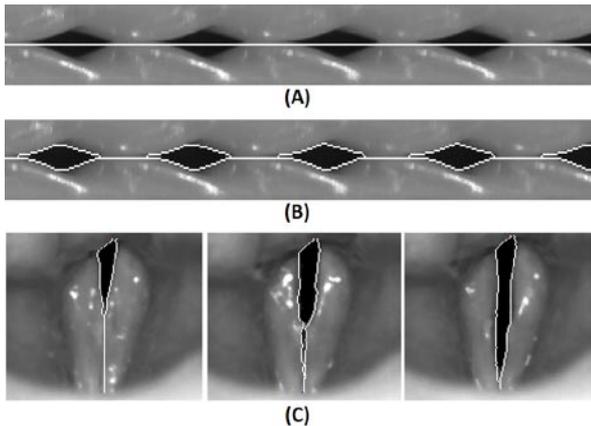
Fig. 5: (A) Initialization of the right and left snakes, (B) deformation results of right and left snakes for the initial kymogram, (C) three phases of the opening cycle with right and left snakes presented in spatial domain.

Validity and Reliability. From the 98 samples in the dataset, 96 samples were successfully segmented using the temporal paired snakes, and 2 presented difficulties due to poor lighting. That is an overall reliability of 98%, which is a highly-encouraging result. In all HSV samples, most DKG images were analyzed without gross errors, i.e. divergence of the snake from the correct edge by more than one pixel, usually due to attraction to the wrong nearby edge. Of the 96 successfully-segmented samples, 78 exhibited at least one DKG image with at least one gross error, 59 of which (76%) were due to a failure of the right and left snakes to attract to each other near the commissures, mainly the posterior commissure. Those instances can be easily corrected by introducing an

adaptively sized column-wise neighborhood and appropriate pre-processing when automating the method.

Accuracy. In all HSV samples, most DKG images exhibited sub-pixel accuracy of segmentation. Of the 18 samples free of gross errors, 1 had no single DKG image with a snake differing from the target edge, and 17 exhibited at least one DKG image containing an instance of an error up to ±1 pixel.

IV. CONCLUSION

The proposed paired temporal snake algorithm exploits the HSV temporal resolution for obtaining a segmentation of the glottis by following a set of snake-deformation rules. The snake deformation strategy employs a dynamic programming algorithm, in which the optimization of the snake-energy function decreases monotonically with respect to the asymptotic rate of growth of the algorithm, and thus the global convergence is guaranteed. The development of the algorithm is still in progress, to be extended to a fully-automatic method for segmentation of the glottis from HSV. This algorithm is reliable and fast, yet highly scalable in terms of the degrees of parallelism that can be exploited in the future.

REFERENCES

[1] Yan Y, Chen X, Bless D. Automatic tracing of vocal-fold motion from high-speed digital images. *IEEE Transactions on Biomedical Engineering*, 53(7):1394-1400, 2006.

[2] Lohscheller J, Toy H, Rosanowaski F, Eysholdt U, Dollinger M. Clinically evaluated procedure for reconstruction of vocal-fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, 11(1):400-413, 2007.

[3] Kass M, Witkin A, Terzopoulos D. Active contour models. *International Journal of Computer Vision*, 1(1):321-331, 1987.

[4] Lohscheller J, Dollinger M, Schuster M, Schwarz R, Eysholdt U, Hoppe U. Quantitative investigation of the vibration pattern of the substitute voice generator. *IEEE Transactions on Biomedical Engineering*, 51(8):1394-1400, 2004.

[5] Manfredi C, Bocchi L, Bianchi S, Migali N, Cantarella G. Objective vocal fold vibration assessment from videokymographic images. *Biomedical Signal Processing and Control*, 1:129–136, 2006.

[6] Deliyski D, Petrushev P, Bonilha H, Gerlach T, Martin-Harris B, Hillman R. Clinical Implementation of Laryngeal High-Speed Videoendoscopy: Challenges and Evolution. *Folia Phoniatrica et Logopaedica*, 60(1):33-44, 2008.

[7] Amini A, Weymouth T, Jain R. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):855–867, 1990.

OBJECTIVE COMPARISON OF THE ELECTROGLOTTOGRAM TO SYNCHRONOUS HIGH-SPEED IMAGES OF VOCAL-FOLD CONTACT DURING VIBRATION

Maria E. Golla[1], Dimitar D. Deliyski[1,2], Robert F. Orlikoff[3], Habib J. Moukalled[1,2]

[1] Department of Communication Sciences and Disorders, University of South Carolina, Columbia, SC, USA
[2] Department of Computer Sciences and Engineering, University of South Carolina, Columbia, SC, USA
[3] Department of Speech Pathology and Audiology, West Virginia University, Morgantown, WV, USA

Abstract: **This study investigated vocal-fold contact characteristics through electroglottography (EGG) and related them to vibratory behavior as seen through high-speed videoendoscopy (HSV). When the EGG cycle was broken down into phases, the contacting phase represented an increasing percentage of the whole cycle as the EGG signal moved through three registers (pulse, modal, and falsetto). Conversely, the decontacting phase corresponded to a decreasing percentage of the EGG cycle as it moved through the same registers. Furthermore, comparisons of the HSV images and the EGG signal indicated close relationships between specific EGG features and the onset of contact of the vocal folds, maximal contact between the vocal folds, and maximal loss of contact between mucus bridges.**

Keywords: **Voice; Electroglottography; High-Speed Videoendoscopy; Vocal-Fold Vibration**

I. INTRODUCTION

Electroglottography (EGG), a valuable tool for both voice researchers and clinicians, is sensitive to changes in vocal-fold contact area during phonation. Clinical observation and the application of various physical and mathematical models have been used to identify important EGG signal landmarks and relate changes in signal morphology to specific aspects of laryngeal physiology. The continued refinement and applicability of high-speed videoendoscopy (HSV) allows for the synchronization of the EGG signal with endoscopic images of the vocal folds.

The purpose of this study is to investigate variations of specific EGG features and relate them to HSV-observed changes in vibratory behavior. To this end, the following **research questions** are addressed: **(1)** Are the objective measures of fundamental frequency (*Fo*) consistent with the elicited samples across three registers (pulse, modal, falsetto)? **(2)** To what degree do five established EGG landmark features (Fig. 1) [1] vary as related to objective measures? **(3)** What are the relationships between the EGG markers and the physiology of the vocal fold movement as visible through HSV and digital kymography (DKG)?

II. METHODS

Human Data: Fourteen vocally-normal speakers (7 men and 7 women, between 22 and 29 years of age) were recorded using precisely-synchronized (≤ 11 μs) HSV (16,000 fps) with EGG (96,000 Hz) as they produced one or more trials of the vowel /i/ sustained in three different registers: pulse, modal, and falsetto [2]. After the data was collected, each HSV trial recording was reviewed by 2 experts who selected three 1,000-frame segments extracted from the whole recording, producing 3 smaller samples. One pulse register trial was excluded due to significant supraglottic compression which precluded visualization of the true vocal folds. The dataset included 72 modal register samples, 42 pulse register samples, and 45 falsetto samples. All 159 samples were used in answering the first two research questions; however, for the third research question, the data set was narrowed to allow for adequate analysis of the large amount of data. Only the middle of each three samples was analyzed for each trial, producing 24 modal register samples, 14 pulse register samples, and 15 falsetto register samples, a total of 53 samples.

Analysis: Using custom-designed software with a specialized graphic user interface, the EGG signals were visually aligned with DKGs taken at 5 equally-spaced locations along the anterior-posterior axis of the vocal folds (Fig. 2). Based on contemporary EGG models [1,4], 5 EGG landmark features were identified
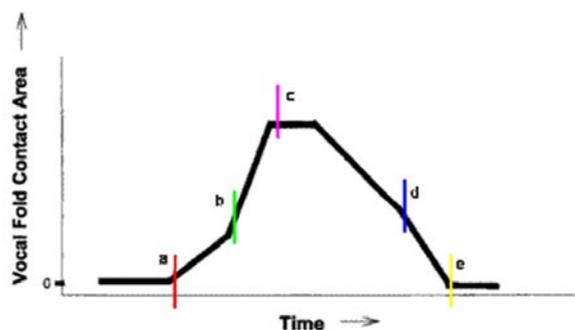
Fig. 1: Model Waveform of the EGG. a) Red marker; b) Green marker (estimated); c) Purple marker; d) Blue marker; e) Yellow marker.
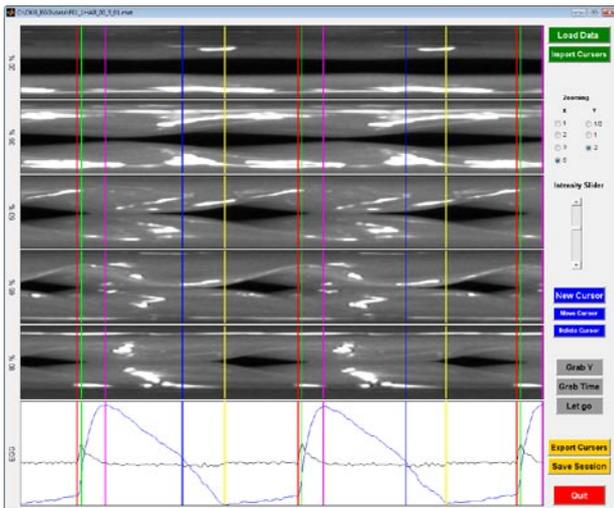
Fig. 2: EGG and DKG visually aligned for color-coded tagging.

and coded with a unique colored marker: (1) intra-cycle onset of contact during the increasing-contact (contacting) phase (red marker); (2) maximum of the derivative/velocity during the contacting phase (green marker); (3) intra-cycle EGG maximum or maximum contact (purple marker), (4) EGG "knee" formed during the decreasing-contact (decontacting) phase (blue marker); and (5) intra-cycle offset of contact during the decontacting phase (yellow marker). Using the custom software, each EGG sample was manually tagged for the 5 landmark features using this color coding system (Fig. 1) and a consensus of the markings was established.

Once the EGG signals were tagged, the time stamps of all markers for each of the 5 color sets were imported into a custom Matlab script. First, the time stamps were converted into vectors of period measurements corresponding to each different feature (color) in every EGG sample. Based on period information, mean frequencies and first-order perturbation functions were computed for each feature and sample to determine the most stable EGG feature.

The 5 EGG feature markers were then exported to another custom-designed software with a specialized graphic user interface, which allowed concurrent visualization and playback of HSV and DKG, with the colored EGG feature markers overlaid in both the HSV and DKG (Fig. 3). User-controlled interface allowed playback of either: HSV frames dynamically-linked to a time stamp on the DKG display or DKG frames dynamically-linked to the corresponding anterior-posterior line. Using each frame as the base measuring unit, each of the 5 EGG feature markers were measured relative to the following 4 HSV landmark features: (1) first contact of the vocal folds, (2) maximum contact of the vocal folds, (3) complete loss of contact between the vocal folds, and (4) complete loss of contact of any mucus bridges.

III. RESULTS

Consistency of Fo with registers.

Analysis of the acoustic signal determined that the *Fo* for the modal and falsetto registers fell within normal limits [2] for the elicited register. Moreover, the modal registers and falsetto registers did not overlap within or across sexes (Table 1). The *Fo* could not be calculated for the pulse register phonations based on the acoustic signal due to the aperiodicity of the samples; therefore, the mean frequency for each sample was computed based on the EGG signal. Several of the pulse register samples had higher frequencies than expected [2]. Closer visual inspection of the DKG for these samples revealed multiple pulses for an individual glottal cycle. Since the EGG signal tracks the change in contact between the vocal folds, it would naturally be sensitive to these multiple pulses. After excluding the markers corresponding to the repeated pulses within a glottal cycle as determined from DKG, the frequency of the glottal cycle fell within normal limits for the pulse register [3,5].
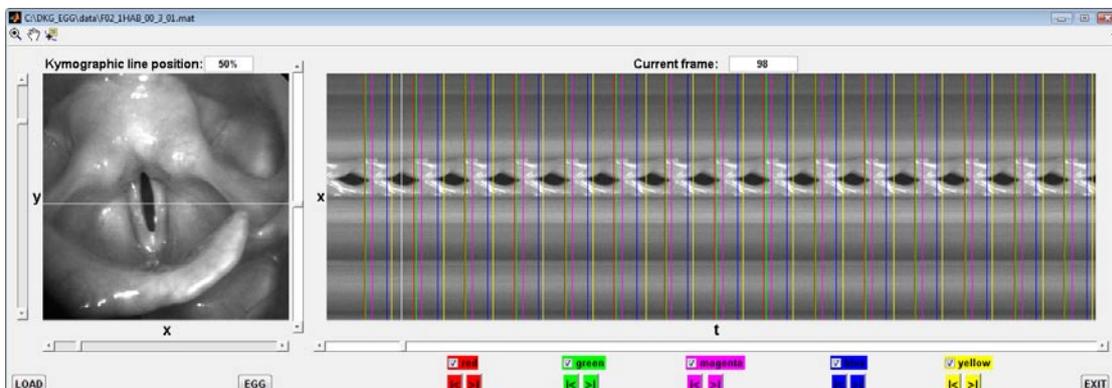
Fig. 3: HSV and DKG visually aligned with EGG markers for concurrent playback.

Table 1: *Fo* (Hz) of habitual and falsetto registers.

Register	Range	Male Range	Female Range
Modal	90.02 Hz – 269.05 Hz	90.02 Hz – 164.45 Hz	152.29 Hz – 269.05 Hz
Falsetto	300.5 Hz – 1041.71	300.05 – 496.04 Hz	411.59 Hz – 1041.71 Hz

One trial (and its subsequent three samples) had a higher frequency of EGG cycles than expected for the pulse register which could not be explained by double or multiple pulses. Listener judgment found this trial consistent with pulse register phonation, and visually the open phase of the cycle represents less than 25% of the entire cycle. It is reasonable to assume that despite the high frequency, this trial met the characteristics of pulse register, and thus, the sample should be considered in the remaining portions of the study. It is likely that this sample contained elements of both pulse and modal register as described by Hollien, Girard, and Coleman [3].

Measuring variation of five EGG landmark features.

Across registers perturbation values reveal significant trends, specifically the large degree of perturbation within the pulse register compared to the modal and falsetto registers. Table 2 records the range of perturbation for each register. It is likely that the higher perturbation levels within the pulse register were related to the overall aperiodicity expected for the register and the presence of the double and multiple pulse phenomena within the register, given that the secondary or tertiary pulses are significantly shorter than the initial pulse of the vibratory cycle. It was not surprising that the modal register demonstrated the least amount of perturbation due to its periodicity (relative to the pulse register) and consistent contact (relative to the falsetto register).

Significant variability was noted within samples and across registers. Table 3 reports the extent of the variability of perturbation for each of the 5 EGG landmark features. For all markers variability was noted within the 1,000-frame samples, as demonstrated by not only the high mean of perturbation values, but also the differences between the mean and median values (indicating the presence of outliers within the data). The most variant marker was the yellow marker, whereas

Table 2: Breakdown of relative perturbation range (%) for each register.

Phonatory Behavior	Mean (%)	Median (%)
Pulse	17.91 – 19.95	8.65 – 11.17
Modal	0.78 – 3.38	0.50 – 3.04
Falsetto	5.98 – 7.19	2.75 – 7.13

Table 3: Mean (median) values of relative perturbation (%) for each EGG landmark feature.

Phonatory Behavior	Red Marker	Green Marker	Purple Marker	Blue Marker	Yellow Marker
Pulse	19.54 (8.65)	19.31 (8.97)	19.95 (11.17)	17.91 (9.71)	19.64 (10.61)
Modal	.82 (.65)	.78 (.50)	1.74 (1.49)	2.77 (2.57)	3.38 (3.04)
Falsetto	5.98 (4.38)	6.81 (2.75)	6.30 (3.73)	7.19 (7.13)	6.49 (5.67)
Combined	5.94 (1.12)	6.12 (.83)	6.61 (2.34)	7.02 (3.41)	7.47 (3/91)

the marker with the least overall perturbation was the green marker. Based on these results, the green marker was considered to be the most stable feature. Thus, a glottal cycle was defined as the distance between consecutive green markers.

Since the green marker was established as the most consistent feature, the intra-cycle ratios of the glottal cycle were calculated in percent relative to the green marker. By doing this, the EGG landmark features naturally break the EGG signal into 5 phases:

Phase 1: Red-Green—time between the red and green markers (onset of contact and the point of maximum velocity during the closing phase).
Phase 2: Green-Purple—time between the green and purple markers (point of maximum velocity during the closing phase and the point of maximum contact between the vocal folds).
Phase 3: Purple-Blue—time between the purple and blue markers (point of maximum contact between the vocal folds to the EGG "Knee").
Phase 4: Blue-Yellow—time between the blue and yellow markers (EGG "Knee" and the offset of contact between the vocal folds).
Phase 5: Yellow-Red—time between the yellow and red markers (offset of contact between the vocal folds and the onset of contact between the vocal folds.)

Table 4 summarizes the percentage to which each of these phases comprise the entire glottal cycle. The results indicate that, although Phase 1 comprises the smallest percentage of the entire glottal cycle for every register, there are visible trends between registers. Specifically, the pulse register has the shortest Phase 1 (relative to the overall glottal cycle), followed by the modal register, whereas Phase 1 comprises slightly more of the overall glottal cycle in the falsetto register. This trend continues for Phase 2, so that the time between the onset of contact and the point at which maximum contact is achieved becomes a greater part of the overall glottal cycle as the subject's phonation moves through the registers.

Phases 3 and 4 could be grouped together to represent the time in which the vocal folds are losing contact. When viewed this way, clear trends relative to the register are again visible. The combination of Phases 3 and 4 represent approximately 61% of the entire glottal

144

Table 4: Means of the intra-cycle ratios (%) for the 5 EGG phases relative to the green marker.

Cycle Phase	All Registers	Pulse Register	Modal Register	Falsetto Register
Phase 1: Red-Green	3.85%	1.69%	2.97%	6.68%
Phase 2: Green-Purple	9.49%	4.98%	7.41%	15.74%
Phase 3: Purple-Blue	33.37%	46.51%	34.14%	23.76%
Phase 4: Blue-Yellow	13.97%	14.87%	12.75%	15.39%
Phase 5: Yellow-Red	39.32%	31.94%	42.73%	38.43%

cycle within the pulse register, whereas the two phases make up approximately 46% of the cycle for the modal register, and approximately 37% for falsetto. These findings are consistent with our understanding of the physiology of the vocal folds and the degree of contact expected for each of the registers [2-4].

Interestingly, Phase 5 does not follow the expected trend for the registers. It would be reasonable to assume that if the yellow marker represents the offset of contact, and the red marker represents the onset of contact of the next cycle, then there should be maximum loss of contact between the vocal folds during Phase 5. It would also be reasonable to assume that since falsetto is thought to have the least amount of contact for the entire cycle then Phase 5 should represent the largest percentage of the entire glottal cycle. However, Phase 5 represents a greater portion of the cycle in modal register than in falsetto.

Relationship between EGG markers and HSV

Results indicate there is a strong relationship between the red and green markers (which generally fell within 100 μs of each other) and the onset of contact between the vocal folds. There also appears to be a strong relationship between the purple marker and maximum contact of the vocal folds. The blue marker was calculated relative to both the maximum contact and the offset of contact between the vocal folds, and results indicate it is more closely related to the offset of contact of the vocal folds than the maximum contact between the vocal folds. Generally the blue marker was placed before the loss of contact of the vocal folds; however, occasionally the blue marker was placed at the loss of contact when a mucus bridge was present. The yellow marker is strongly related to the offset of contact between the vocal folds or the offset of contact between mucus bridges if present.

IV. DISCUSSION

Results of this study indicate that EGG signal does directly relate to the changing contact between the vocal folds. When broken down into phases, the contact phases (Phases 1-2: Red-Purple Marker) constitute the smallest percentage of the cycle in pulse register, a slightly larger percentage of the cycle in modal register, and an even greater percentage of the cycle in falsetto. Conversely, the loss of contact phases (Phases 3-4: Purple-Yellow Marker) constitutes the smallest percentage of the falsetto register cycle, a larger percentage of the modal register cycle, and the largest percentage of the pulse register cycle. These findings are consistent with current literature on the physiology of the vocal-fold vibration in various registers [2-5].

Comparison of the EGG signal and HSV recordings reveal that the EGG markers do closely align with the onset of contact, the point of maximum contact, and the offset of contact between mucus bridges. Also, the blue marker was found to be more closely related to loss of contact between the vocal folds–sometimes appearing at the point of loss of contact. Additionally, mucus bridges play a significant role in the morphology of the EGG signal at the offset of contact.

V. CONCLUSION

This study is unique in terms of the method's accuracy and the direct linkage of an indirect measure of vocal-fold contact through EGG, to visual physiologic measures of vocal-fold contact through HSV. The results cross-validate several EGG features and pose new questions about others, especially the EGG knee appearing during the opening phase.

REFERENCES

[1] Baken RJ, Orlikoff RF. *Clinical Measurement of Speech and Voice.* 2nd ed. San Diego: Singular Publishing Group; 2000.

[2] Hollien H. On vocal registers. *J Phonetics* 1974; 2:125-143.

[3] Hollien H, Girard GT, Coleman RF. Vocal fold vibratory patterns of pulse register phonation. *Folia Phoniatr* 1977;29(3):200-205.

[4] Titze IR. Interpretation of the electroglottographic signal. *J Voice* 1990;4(1):1-9.

[5] Whitehead RL, Metz DE, Whitehead BH. Vibratory patterns of the vocal folds during pulse register phonation. *J Accoust Soc Am* 1984;75(4):1293-1297.

QUANTITATIVE ANALYSIS OF DIPLOPHONIC VOCAL FOLD VIBRATARY PATTERN FROM HIGH-SPEED DIGITAL IMAGING OF GLOTTIS

Y. Yan[1,2], K. Izdebski[2,3], E. Damrose[2], D. Bless[4]

[1]School of Engineering, Santa Clara University, Santa Clara, California, USA
[2]Department of Otolaryngology, Stanford University, Stanford, California, USA
[3]Pacific Voice and Speech Foundation, California, USA
[2]Department of Surgery, University of Wisconsin-Madison, Wisconsin, USA

Abstract: This paper investigates vocal fold (VF) vibratory properties using quantitative analysis of high-speed digital imaging (HSDI) based on Nyquist-plot method derived from voicing during production of aberrant voice quality (VQ), clinically referred to as diplophonia, and defines the mechanism responsible for diplophonia and show how treatment (Tx) effects this VQ and VF behavior. In particular, pre- and post-Tx HSDI recordings of a female patient with muscular tension dysphonia (MTD) were analyzed using new quantitative analysis system for HSDI that involves tracing of VF edge and generation of glottal waveform and VF displacement, allowing us to define quantitative measures of vibratory symmetry and synchronization of VF vibrations, with subsequent analyses of glottal waveforms using Nyquist formula, to reveal vibratory pattern and characteristics of the vocal folds during this aberrant sound production, and later during normative phonation post Tx.

This is first ever HSDI and Nyquist-plot based analyses of aberrant voice known as diplophonia derived here from vocalization of a MTD case. The results reveal definitive and specific character of VF vibration responsible for this VQ.

Keywords : **High-speed digital imaging, vocal-fold vibration, diplophonia, Nyqist plot**

I. INTRODUCTION

Control of VF vibrations can be variably affected by organic and functional causation resulting in vibratory irregularity and deviant/aberrant acoustic product [1]. Such deviant vocal outputs have been analyzed in the past using various techniques including acoustics, aerodynamics and visualization, leading to improved understanding of how vocal fold physiology relates to the actual sound production and how to treat the underlying pathology. Usage of acoustics combined with visualization has been shown to provide improved information on the underlying pathology and underscores the value of multidisciplinary approach and clinical power [2], hence improving characterization of VF vibrations associated with some voice disorders, specifically for those that generate similar perceptual effects, confusing unequivocal clinical diagnosis that often is made by the ear alone.

To improve the analysis and to provide up-to-date explanation, newest visualization technique to study VF dynamics based on HSDI, also termed high speed videoendoscopy (HSV), combined with application of Nyquist-plot analysis [3], [4] were used here to study aberrant vocalization known clinically as diplophonia [1]. The HSDI system records images of the vocal folds at an acquisition rate of 2000 frames per second, fast enough to capture the vibration of the vocal folds in more details. To overcome the cumbersome subjective evaluation of analyzing such massive data, quantitative methods for HSDI-based and acoustic analyses have recently been developed [3-5], and were used here. These analyses generate comprehensive patterns of VF vibrations and are capable of defining the characteristics of the VF vibration in terms of robust, quantitative measures that enhance understanding of the mechanism of phonation not only in normative voice, but specifically in voice pathologies, which can not be handled easily by traditional stroboscopic illumination.

Phonatory VQ referred to in broad terms as "diplophonia, multiphonia or biphonation" is assumed to represent simultaneous production of at least two distinct tones during phonation. This pattern may be induced by an imbalance in bilateral tension or mass of the VF or by events within each VF, as evidenced in clinical cases representing mass, paralysis, neurological driven or functional dysphonias including non-true VF driven voice pathology. The mechanism underlying diplophonia has been studied by different researches using diverse experimental and theoretical approaches including biomechanical modeling, analysis of acoustic recordings and direct imaging of voice production. These studies suggest that diplophonia represents 1) an abnormal

glottic cycle with double or even multiple phased opening and closing; 2) asymmetric vibrations of the vocal folds, in which the left and right folds vibrate at a different frequency, or 3) out-of-phase vibratory pattern; 4) within cord variability or/and 5) combinations of the above. In this study we aim to provide a comprehensive, quantitative analysis of glottal area waveform (GAW) derived from HSDI recordings of diplophonic phonation from a clinical case representing idiopathic diplophonia, of MTD type.

II. METHODOS

Data were acquired from HSDI recordings of a female patient with idiopathic diplophonia of MTD type, prior to and following voice therapy. In this way the subject was serving as her own control. Post data processing involved automated image segmentation, detection of VF edge and generation of GAW and VF displacements described by us previously [3-5]. Of specific interest here was the usage of Nyquist plot based analysis of the HSDI-derived waveforms and associated perturbation measures [3], [4]. This technology is used to generate characteristic patterns of the VF vibration and to describe quantitatively the VF vibration in the MTD patient's voicing before and after treatment.

New measures of the VF vibration that include symmetry/ homogeneity, synchronization, and sustainability are defined. These measures along with the jitter metrics formed the basis of a robust quantitative analysis of the VF vibrations in pre- and post-Tx MTD voices. Further, the analysis provided an automatic, robust calculation of the glottic closure characteristics including the open quotient (OQ), speed quotient (SQ) and glottal closure index. Nyquist plot based waveform analyses and associated perturbation measures of the MTD voice were shown to generate not only at-a-glance patterns of the spatial- and temporal characteristic but also quantitative measures of the VF vibrations for the MTD voice.

III. RESULTS

HSDI-based quantitative analyses of the MTD voice demonstrated that the anterior and posterior of the VF undergo non-identical vibrations with distinct patterns of opening and closing during the diplophonic phase (Fig. 1). In particular, the anterior and medial portions of the VF exhibited more obvious bi-cyclic vibratory pattern compared to the posterior portion of the VF (data not shown). In contrast, vibrations within the left-right folds are almost symmetrical throughout the anterior-posterior (A-P) locations during the diplophonic phase (data not shown).
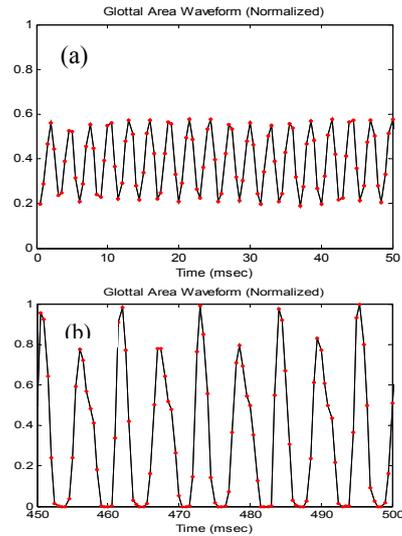
Figure 1 Normalized GAW derived from HSDI recordings of the MTD patient, showing a transition from (a), single-cyclic phase (phase I) to (b), diplophonic phase (phase II)

These findings suggest that it is the non-homogeneity in the A-P VF vibrations, but not the asymmetry, that is linked to the diplophonia. Interestingly, both asymmetry and non-homogeneity were evaluated to exist in the VF vibration prior to transitioning to the diplophonic phase, and is characterized by glottal incompetence and a breathy and rough VQ. Our analyses also revealed an improvement in both symmetry and homogeneity of the VF vibration after the phase transition (during diplophonic phase) (data not shown).
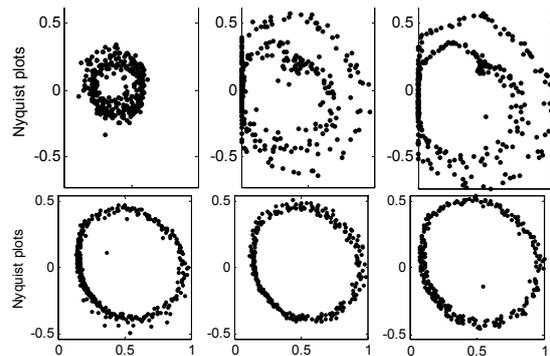
Figure 2 GAW Nyquist plots representing three consecutive time periods of HSDI recordings (0~300ms; 300~600ms and 600~900ms) before (upper row) and after (lower row) Tx

Results from an identical analysis of HSDI recordings from the same patient after voice Tx revealed a significant improvement in the synchronization of vibrations within the VF that led to almost normophonic GAW and Nyquist pattern (Fig. 2).

Further, the jitter measures were calculated and the results showed that post-Tx voice was significantly improved as evaluated by a jitter of 1.9%, as compared to 8.8% for pre-Tx voice during phase I, or 2.2% and 2.5% during phase II (diplophonic phase), representing perturbation measures for the two respective simultaneous vibrations.

IV. Discussion

Quantitative evaluation of symmetry /homogeneity of VF vibrations in the pre- and post-Tx MTD voice were performed and the results showed an overall improvement in the VF vibration after Tx. In particular, the anterior (A), medial (M) and posterior (P) portions of the vocal flods vibrate at the same frequency (Fo~194 Hz) and the A-M-P vibrations are better synchronized, as evidenced by high correlation coefficient between the A-P vibrations (0.6171), compared to pre-Tx results (-0.13), indicating complete out-of-phase vibrations in the A-P vocal folds.

V. Conclusion

To conclude, asynchrony in VF vibration at different portions of the A-P axis of the vocal folds is associated with the diplophonia perceived in this MTD patient. Analyses of the HSDI recordings from the same patient after Tx demonstrated significant improvement in vibratory synchronization along the VF length, consistent with the subject having a normal perceptual rating after Tx as rated by the GRBS (G-grade, R-roughness, B-breathiness and S-strain) scale.

References

[1] K. Izdebski, "Clinical Voice Assessment: The Role and Value of the Phonatory Function Studies," in *Current Diagnosis & Treatment, Otolaryngology Head and Neck Surgery*, A. Lalwani Ed. Lange Publications, 2008, pp. 416-429.
[2] K. Pedersen, and U. Yousaf, "Advanced voice assessment: videostroboscopy and objective voice measurement," *European Archives of Oto-Rhino-Laryngology and Head and Neck*, vol. 264, Supplement S50 HL 119, 2007.
[3] Y. Yan, K. Ahmad, M. Kunduk, and D. Bless, "Analysis of vocal-fold vibrations from high-speed laryngeal images using a Hilbert transform based methodology," *J. of Voice*, vol. 19, pp.161-175, 2005.
[4] Y. Yan, X. Chen, and D. Bless, "Automatic Tracing of Vocal-fold Motion from High Speed Digital Images," *IEEE Trans. Biomedical Engineering*, vol 53(7), pp. 1394-1400, 2006.
[5] Y. Yan, E. Damrose, and D. Bless, "Functional Analysis of Voice Using Simultaneous High-speed Imaging and Acoustic Recordings," *J. of Voice*, vol. 21, pp. 604-616, 2007.

EXPERIMENTAL SYSTEM FOR NEUROLOGICAL CASE STUDIES

Jana Klečková, Petr Maule, Jiří Polívka [1], Vladimír Rohan [2]

[1]Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Pilsen, Czech Republic
[2] Department of Neurology, University Hospital, Pilsen, Czech Republic

Abstract: **Diagnostics and treatment of neurological disorders is based on continuous evaluation of the amount of clinical data and their various characteristics. Rising of the quantity of information with different clinical meaning which needs to be assessed is connected with the development of new diagnostic and medical methods. To understand recovery processes in the brain, some researchers are using new graphic diagnostic methods include for instance perfusion computed tomography (CTP), CT angiography or diffusion weighted magnetic resonance MR DWI to better understand the human brain regions i.a. involved in speaking and understanding language. The goal is achieved in various projects funded by the Czech Research Foundation (project number 106/09/0740) and Czech Ministry of Education (the project number 2C06009).**
Keywords: **neurology, aphasia, computed tomography, magnetic resonance**

I. INTRODUCTION

A special problem are impairments of speech which may be congenital (e.g. the cleft lip and palate) or acquired by disease (e.g. cancer of the larynx). Impairments are, among others, treated with speech training by speech therapists. They score the speech quality subjectively according to various criteria. The idea is that the word accuracy (WA) of an automatic speech recognizer should be highly correlated with the human rating. Using speech samples from laryngectomees it is shown that the machine rating is about as good as the rating of human experts and can also be done via telephone. This opens the possibility of an objective and standardized rating of speech quality. Serious problem is some of the aphasia in the wake of cerebral vascular diseases (Thrombotic (Ischemic) Stroke, Hemorrhagic Stroke, and Cerebral Aneurysm). Although now considered less frequent, thrombosis is still estimated to be the etiology of about two-thirds of all ischemic strokes [4]. Occlusion of a cerebral artery produces immediate cessation of blood flow and subsequent death of brain tissue (ischemic infarction) in the territory supplied by the involved vessel. The resulting neurologic defect (syndrome) reflects the vascular territory involved. Table 1 present the most common aphasic syndrome associated with particular cerebrovascular territories.

The contributed paper deals with proposal and usage of modern database neuroinformatics information technologies in research, educational and clinical application.

II. METHODOS

The Content-based visual information retrieval (CBVIR) has been one of the most growing research area over the last few years. The reason is steadily increasing amount of multimedia, especially visual data in wide range of today professional activities. The extensive multimedia databases, often with the internet underline, can contain many thousands of specialized images. This status ask for the new methods, which help browsing large multimedia databases, find the right case not only by the simple text-based queries or matching exact selected field.

Table 1 Association between Aphasic Syndromes and Selected Cerebrovascular Territories

Cerebrovascular Territory	Aphasic Syndrom
Anterior Cerebral Artery Occlusion	Extrasylvian motor aphaisa
Posterior Cerebral Artery Occlusion	Occupital alexia
Middle Cerebral Artery Occlusion	
Total	Global aphasia
Orbitofrotal branch	Broca aphasia
Rolandic branch	Broca aphasia, cortical dysarthreia
Anterior parietal branch	Conduction aphasia
Posterior parietal branch	Wernicke aphasia, extrasylvian sensory aphasia
Angular branch	Anomia, extrasylvian sensory aphasia
Posterior temporal branch	Wernicke aphasia
Anterior temporal branch	Anomia

The CBVIR technics offer intelligent similarity data search, which can be used if the comparison among many cases are needed. In medical applications the recent diagnostic and therapy procedures usually involve work with the latest technical equipments and imaging devices.

The new graphic diagnostic methods include for instance perfusion computed tomography (CTP), CT angiography or diffusion weighted magnetic resonance MR DWI [2]. This approach produces huge amounts of medical images for each patient and study case, generally in the international standard Digital Imaging and Communications in Medicine (DICOM) [1]. Images produced by these technics are usually stored in medical information systems. The CBVIR doorway should be helpful in the future work with this great amount of stored medical images for the clinical staff as well as the researchers and scientists.
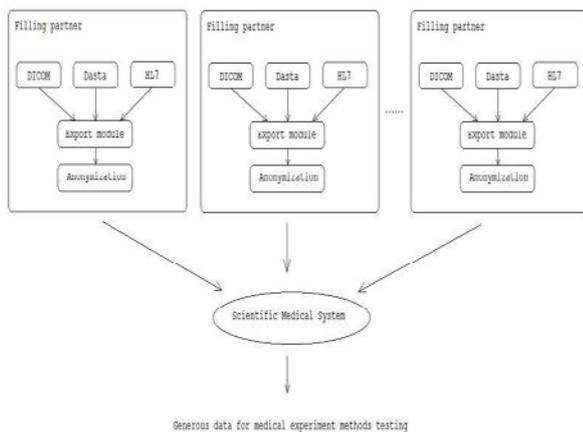
Figure 1: Filling system schema

III. RESULTS

The common CBVIR system engine architecture contains some basic functional modules – storage and access methods, visual feature extraction, distance measures and similarity calculations and user interface and interaction methods. Each incoming case (image or array of images) is analysed, the distinguished visual features are extracted and are compared with the features of stored images. Then the retrieved images are presented. The visual features are classified into primitive features (color, grey level, shape or texture), logical features like an object identity and abstract features such as significance of scenes pictured. In the medical applications, the color and grey level features are often neglected, especially due to the lack of the contrast reference point in the radiology images. More frequently the texture and shape features are applied. The technics for texture identification use for instance Canny

operators, invariant moments, scale-space filtering, Gabor filters, wavelets and Markov texture characteristics or Fourier descriptors for shape characterization. Also the segmentation of the incoming image into smaller parts is investigated.

IV. DISCUSSION

In our contemporary research we tried to develop the open web-based multimedia database system of complex neurological information for each real medical case. This system will first store the medical images in DICOM format and more information about the current study (Fig.1). One of the objectives is to design and set up CBVIR engine for our experimental database. The beginning basic proposal of the CBVIR in context of our database system is presented, the visual features are chosen. The retrieval engine will use comparison of histogram vectors and the optimal method for similarity calculations must be selected. The basic visual features of the one case´s images will be combined with the other information from the database, such as from the textual section or form section. Then the aggregate histogram vector will be created, stored in the database and can be used in the further similarity study retrieval. The presented concept of the CBVIR system will be used in future work, mainly in design process of our experimental database system.

V. CONCLUSION

The contribution to the problem solution of consequences or mortality in the area of cerebral vascular diseases is expected. All data will be anonymous but mutual relations will be preserved.

The project will ease work of individuals and groups that are interested in medical data processing.

REFERENCES

[1] ACR/NEMA 2008. The DICOM standard. http://dicom.nema.org
[2] H. Muller and N. Michoux and D. Bando nand A. Geissbuhler, 2003, "*A review of content-based image retrieval systems in medical applications – clinical benefits*", International Journal of Medical Informatics.
[3] J.P. Eakins and M.E. Graham, 2002, "*Content-based image retrieval*", Tech. Rep. JTAP-039 JISC.
[4] D.Benson, A. Ardila, 1996, "*Aphasia: a clinical pespective*", New York, Oxford University Press, ISBN 0-19-508934-0.

Devices

ACCELLERATION SENSOR MEASUREMENTS OF SUBGLOTTAL SOUND PRESSURE FOR MODAL AND BREATHY PHONATION QUALITY

Wolfgang Wokurek[1], Manfred Pützer[2]

[1]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

[2]Institut für Phonetik, Universität des Saarlandes, Saarbrücken, Deutschland

Abstract: We present a non-invasive attempt to indirectly measure the subglottal sound pressure. This quantity opens an additional acoustical path to observe the voiced sound source. The subglottal sound pressure contours of two phonation qualities, the modal phonation quality and the breathy phonation quality, are compared. The electroglottographic signal was recorded simultaneously as a well known reference basis for physiological details of voice production.

Keywords: accelleration sensor, subglottal sound pressure, phonation quality

I. Introduction

The subglottal sound pressure is of interest in order to study fine details of speech production, because the subglottal resonances may interact with voiced sound production, diphthongs in particular [1], [2], [3]. So far, every attempt to place a pressure transducer in the subglottal cavity resulted in an invasive method. That sort of methods was ruled out in advance for our study.

Placing a microphone as transducer at the fossa jugularis (suprasternal notch) was repoted to yield good agreement of the transducer signal with direct subglottal pressure signal [4]. We decided for this study not to measure at the fossa jugularis to avoid a possible resonance influence from the chest that is known by audio engineers to be in the 1 kHz region. But the fossa jugularis is a place that is less covered by the cables of the electroglottogram electrodes and is likely to be investigated in the future.

The availability of micro electro mechanical systems (MEMS) to measure the accelleration, a quantity that is proportional to the force moving the sensor mass, led us to construct an external sensor to track the subglottal sound pressure at the skin of the neck. The sensor is gently pressed at the neck of the speaker in front of the cricothyroid ligament, located near the lower end of the larynx. The acceleration signals are recorded and chances may not be too bad that the tissue passes the subglottal pressure to the sensor. Due to the mass an the compressibility of the tissue only a (low pass) filtered version of the subglottal pressure may arrive at the sensor. Moreover we do not have a true reference signal of the subglottal pressure. But we do have the electroglottographic (EGG) signal as a phonatory reference. And when the EGG indicates a degcreasing tissue contact of the vocal folds, we attribute the subglottal pressure to be the cause.

This acceleration sensor method was previously applied to measure the resonance parameters of the subglottal cavity [3]. Each of our studies is used to review and possibly improve the sensor and the eevaluation procedure.

II. Method

A. Sensor

A three axis acceleration sensor is pressed gently against the skin of the neck and the sensor signals are recorded. The precise position at the neck is crucial. We identified the skin over the cricothyroid ligament as a potentially very good position to access the subglottal pressure. The cricothyroid ligament is a soft elastic tissue between the cricoid and the thyroid cartilage in the lower part of the larynx. It may be localized by touching the larynx and sensing for a small soft gap in the elsewhere hard larynx structure.

Figure 1 shows the sensor in a test environment. The sensor currently is a cube of 1 cm edge length containing two ADXL202E two axis micro electro mechanical acceleration sensors. At one side of the cube a plastic nose of 5 mm height made out of hot glue is attached to improve the contact to the cricothyroid ligament. This nose is important to avoid loss
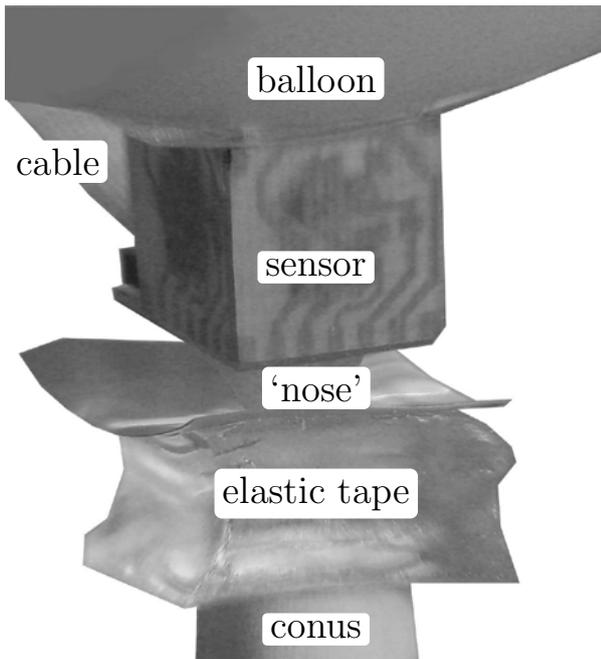
Figure 1: Acceleration sensor in test environment: pressed on compressible elastic tape

of contact during speech when the larynx moves up and down. The opposite side of the cube is glued to a balloon inflated with air to a diameter of about 6 cm. The upper frequency limit of the accelleration transducer is between 3 and 4 kHz.

The suspension by a hand held ballon yields more stable results than previous attempts to attach the sensor by a glue tape or by pressing the sensor to the neck by the second finger. The elastic suspension of the sensor together with the elastic and tissue (having its own mass) is a resonatory system by its own. And reasonable measurements of the subglottal pressure can not be expected in the frequency range of this resonance. The test environment shown in Figure 1 indicates a wide maximum between 100 and 200 Hz (similar to Figures 4 and 5). By a softer and less inflated ball the peak can be shiftet to the 100 Hz region, but so far not further down and we currently interpret this fact as a limit caused by the elastic tissue and the mass of the sensor.

B.   Signal processing

Each axis signals is lowpass filtered by an analog active filter to achieve the maximum bandwidth of 3.5 kHz specified by the data sheet of the sensor. As

in a previous study [6] a 4th order Butterworth characteristic is implemented by two active Sallen-Key lowpass filter circuits. An additional 10 dB amplifier and a line buffer prepare each channel for cable transmission and for the level required by the soundcard.

During the recording session the sensor is slowly turned relative to the movement direction of the cricothyroid ligament by unconsciously changing the position of the hand holding the balloon and by vertical movements of the larynx. Hence no single axis signal shows the subglottal pressure contour. A principle component analysis of the vector signal uncovers the direction of the strongest oscillation, and the projection of the acceleration vector signal on this direction is considered as the subglottal pressure signal. It is labeled as main pressure component, or MPC signal.

The two sensors are orthogonally mounted at two sides of a cube and one axis of each points into the same direction, to and from the neck, where the main oscillation is expected. A simplifying assumtion of this study is that the cube is only moved parallel and not turned by the vibrating skin of the neck. In this case the signals from the common axis should be basically the same. At the end of this study we discovered, that some of the recordings show different waveforms in that direction, contradicting the simplifying assumtion. Hence, an advanced kinematic model should be added in further surveys, in order to transform the sensor acceleration measurements to the acceleration of the skin of the neck.

C.   Speech material

The accelleration sensor signals and the EGG signal were recorded simultaneously for a single male speaker. Sustained vowels, nasals and diphthongs were uttered with two phonation qualities: modal and breathy. These phonation qualities are produced with different tensions of the vocal folds in the larynx. Breathy phonation has shorter and less complete closure phases and longer open phases compared with modal phonation. The cavity resonance oscillations are stronger damped when the vocal folds are open, due to the larger wall surface of the total coupled cavity [5]. Furthermore the center frequencies of the coupled cavity are slightly decreased.

III.   RESULTS

The excitation signal is compared to the electroglottographic signal for vowels uttered with modal and breathy phonation quality. In Fig. 2 (modal phona-
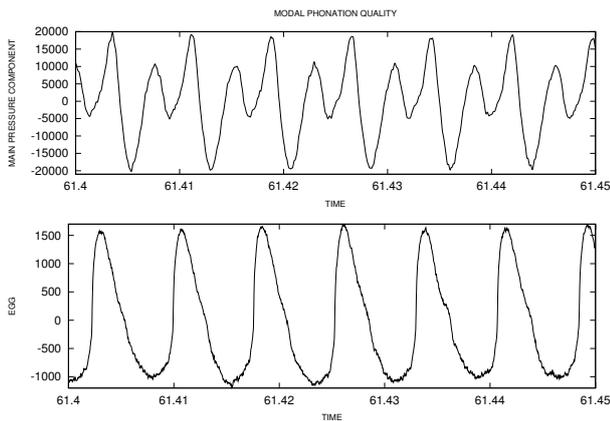
Figure 2: Modal phonation quality.

tion quality), the beginning of the closing phase of each pitch cycle is displayed as a steep ascent of the EGG contour. The ascent ends in the contact phase. The locally maximal contact is marked by the upper peak. With a short delay, the first cycle of the MPC signal starts. The opening phase increases the damping and slightly lowers the frequency.
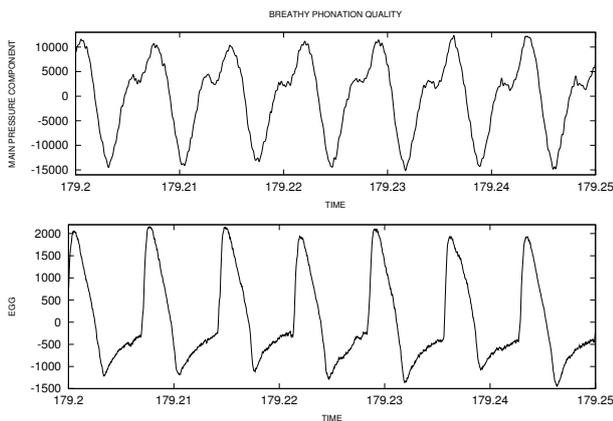
Figure 3: Breathy phonation quality.

In Fig. 3, breathy phonation quality is characterized by shorter and less complete closure phases and longer open phases. The longer open phase lets the oscillation amplitude of the main pressure component (MPC) signal descend much more compared to modal phonation. The shorter and less complete closure phase reduces the excitation and the amplitude of the first cycle of the MPC signal.

Both Figures 2 and 3 clearly show oscillations, but in spite of the elastically suspended sensor discussed in section II./A., a short term spectrum may increase insight. The MPC contours are quasi periodic. In order to ignore the associated harmonic structure of the spectrum, the analysis window is limited to the

fundamental period of each signal. A kaiser window shape is selected to have control on the contrast between frequency resolution and spectral leakage. The window parameter $\alpha = 4$ suited both phonation qualities.
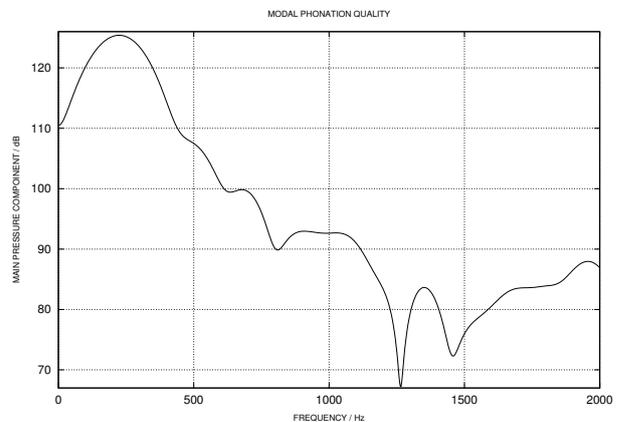
Figure 4: Spectrum of the main pressure component (MPC) signal for modal phonation quality.

The modal recording has a lower fundamental frequency and an 8 ms window is the longest that is not dominated by a harmonic structure. Fig. 4 shows the magnitude spectrum from 0 to 2 kHz. This frequency range includes the regions of the first and second subglottal resonance of [500 Hz-700 Hz] and [1300 Hz-1500 Hz] respectively. It clearly shows a dominant component slightly above 200 Hz. Very likely it corresponds to the oscillation that is visible in the time domain in Fig. 2. There are peaks in the range of the first and second subglottal resonance, but we have no means to identify them.
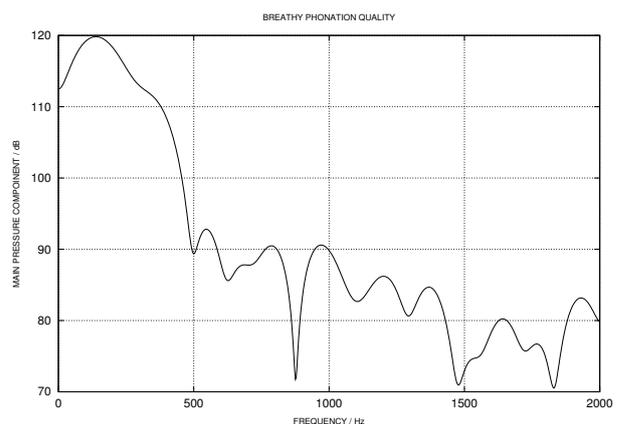
Figure 5: Spectrum of the main pressure component (MPC) signal for breathy phonation quality.

The speech sample from the breathy voice quality recording has a higher fundamental frequency and re-

156

quires a shorter analysis window of 7 ms duration. Now the strongest component is centered slightly below 200 Hz. Again this is the excited and damped oscillation with a period of about 5 ms that is visible in the MPC part of Fig. 3. The subglottal resonance may be visible in Fig. 5 but is not identified.

IV. Discussion

One intention of this study was to obtain a detailed recording of the subglottal pressure up to 2 kHz by pressing an accelleration sensor to the skin of the neck. After the interpretation of modal and brathy vowel recordings in time and short term frequency domain we see oscillations that may be caused by a resonant structure consisting of the balloon suspension, the sensor and the skin and subcutan tissue. This structure has a center frequency around 200 Hz and a bandwidth in the 100 Hz range. Attempts to lower the resonance frequency of the suspension sensor system, did not substantually decrease the center frequency of the system including skin and subcutan tissue.

Otherwise, this unintended oscillation is driven by the main 'force' of interest, the voice source. This cycle of glottal movements and contacts introduces boundary conditions that influence the subglottal pressure contour that drives our oscillation as well as the time variant damping of that oscillation. Quantities related to the temporal open quotient and the damping that is related to the degree of opening between the vocal folds may be extracted by advanced signal processing.

From the point of causality a conservative summary might be the following: the sequence of opening and closing of the vocal folds is visible to a certain extent in the EGG waveform and our new MPC waveform does not contradict.

Finally, since the turning movements of the sensor do not proof to be neglectible, a kinematic model of the sensor is required to transform the accelleration sensor signals to the accelleration of the sensor nose at the neck.

V. Conclusion

The present study demonstrates that a subglottal sound pressure signal (the MPC signal) reveals the phonation physiology of modal and breathy phonation quality, similar to the electroglottographic signal. During different phases (closing phase, closed phase, opening phase, open phase) of the glottal cycle, the intensity of subglottal pressure changes due to a different contact status of the vocal folds. The physiologic differences cause corresponding changes in the amplitude, frequency, and damping of the oscillations in the MPC signal. These observations encourage a further look at other phonation qualities (e.g. hoarseness quality) for a better understanding of the representation of the healthy and pathological phonatory cycle in the MPC signal.

References

[1] X. Chi and M. Sonderegger. Subglottal coupling and its influence on vowel formants. *The Journal of the Acoustical Society of America*, 122(3):1735–1745, 2007.

[2] S. M. Lulich. *The Role of Lower Airway Resonances in Defining Vowel Feature Contrasts*. PhD thesis, MIT, 2006.

[3] A. Madsack, S. M. Lulich, W. Wokurek, and G. Dogil. Subglottal resonances and vowel formant variability: A case study of Swabian diphthongs. LabPhon11, Wellington, Juli 2008.

[4] K. Neumann, V. Gall, H. K. Schutte, and D. G. Miller. A new method to record subglottal pressure waves: potential applications. *J Voice*, 17(2):140–59, 2003.

[5] K. N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1998.

[6] W. Wokurek and A. Madsack. Messungen subglottaler Resonanzen mit Beschleunigungssensoren. In *34. Jahrestagung für Akustik*. DAGA, März 2008.

BASIC REQUIREMENTS ON MICROPHONES FOR VOICE RECORDINGS

J. G. Svec[1], H. Sramkova[1], S. Granqvist[2]

[1] Laboratory of Biophysics, Dept. Experimental Physics, Palacky University Olomouc,
tr. Svobody 26, CZ 771 46 Olomouc, the Czech Republic, e-mail: svecjan@vol.cz
[2] Department of Speech, Music and Hearing, School of Computer Science and Communication,
Royal Institute of Technology, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden, e-mail: svante@speech.kth.se

Abstract: **There is a need for specifications of microphone characteristics which shall be fulfilled in order to make the microphone acceptable for voice measurements. In this preliminary study we address the most basic parameters – the frequency response, the frequency range, the dynamic range, and the directional characteristics of the microphones. We argue that the frequency response of the microphones shall be flat (i.e., less than 2 dB variation) within the frequency range between the lowest expected fundamental frequency of voice and the highest component of the voice spectrum of interest. The equivalent noise level of the microphones is recommended to be at least 15 dB lower than the sound level of the softest phonations produced. The upper limit of the dynamic range of the microphone shall be the same or higher as the sound level of the loudest phonations. In case of directional microphones, their placement shall be at the distance that corresponds to maximally flat response of the microphone in order to avoid the proximity effect. If this distance is not known, a directional microphone is considered unsuitable for SPL and spectral measurements of voice.**

Keywords : **Microphones, measurement, voice recording, recommendations**

I. INTRODUCTION

In voice and speech research, the purpose of the microphone is to convert the sound pressure signal to an electric signal with the same characteristics. However, most microphones are not developed for this purpose but for recording of music, performance, public address systems, broadcasting etc. [1;2]. Consequently, many of the microphones are not suited for accurate measurements of voice and speech. Despite of the fact that voice and speech measurements are carried routinely for clinical and research purposes, the subject of microphone selection has not received sufficient attention in the voice and speech literature. While there have been attempts to provide recommendations for the choice of microphones [3-5], so far there has been an insufficient explanation of the principles on which the recommendations should be based. As a result of this, the carried measurements often lack sufficient accuracy.

The purpose of this paper is to provide guidelines for selecting a microphone, which is suitable for measurement of voice and speech. Recommendations are formulated that can be used for selecting the proper microphone for voice and speech research. This paper is an extract from a more elaborate paper (in preparation), which will provide more detailed information.

II. METHODOS

To assure accurate recording of voice and speech, we consider three fundamental characteristics of sound: a) fundamental frequency, b) timbre, i.e., the sound spectrum, and c) the sound pressure level (SPL). These three characteristics shall ideally be identical in the captured sound to the sound emitted by the speaker or singer. While the first characteristic is typically well captured by microphones, the accuracy of capturing the sound spectrum and SPL depends on the frequency response and dynamic range of the microphone. An additional factor is the noise in the room, which can also influence the accuracy of the measurement. Also the proximity effect of directional microphones is considered since it changes the microphone frequency response when the mouth-to-microphone distance is changed.

III. RESULTS

The following requirements can be formulated on the frequency range, the frequency response and the dynamic range of microphones intended for voice measurement purposes:
- the equivalent noise level (i.e., the low dynamic limit) of the microphone shall be at least 15 dB below the SPL of the softest produced voice. Also the acoustic noise in the room shall be at least 15 dB below the SPL of the softest produced voice.

158

- the upper dynamic limit of the microphone shall be at least as high as the loudest produced voice level
- the low frequency limit (-2 dB) of the microphone shall be lower than the lowest produced fundamental frequency of the voice
- the upper frequency limit (-2 dB) of the microphone shall at least as high as the highest spectral frequency of interest. (The limit of 8 kHz is above the spectral maximum of most known sounds encountered in human speech and corresponds to the limit of laboratory standard microphones type LS1 according to ANSI [6]).
- the frequency response of the microphone between the low and upper frequency limit shall be flat within 2 dB
- microphones mounted to the side of the head are allowed to have a gain of up to 3-5 dB in the frequency region above 5 kHz, due to reduced radiation of the high-frequencies to the side
- directional microphones should be used for SPL and spectral measurements only at the distance, at which their frequency response is flat, in order to avoid proximity effect. That distance should be found in microphone specifications. If the distance is not known, the microphone is not considered suitable for the SPL and spectral measurements of voice and speech.

IV. DISCUSSION

The requirements imply that different phonation tasks put different demands on microphones. For instance, measurements of voice produced at more or less comfortable loudness (with SPLs around 60-80 dB re 20 µPa at 30 cm microphone distance) are less demanding than measurements of voice across the whole dynamic range (with SPLs@30 cm ranging between 40 and 130 dB re 20 µPa). For measurements of comfortable phonations microphones with relatively small dynamic range, i.e., with equivalent noise levels smaller than 40 dB and upper dynamic level of 90 dB may be sufficient. But for measurements over the whole dynamic range of voice, microphones with an equivalent noise level smaller than 25 dB and upper dynamic level above 130 dB are needed when they are positioned at the distance of 30 cm from the mouth. Head-mounted microphones are positioned at much closer distances to the mouth thus are exposed to voice SPLs about 15 dB higher than the microphones positioned at 30 cm. Therefore the head-mounted microphones are expected to have an equivalent noise level below 40 dB and the upper level limit of at least 145 in order to be able to capture voice over the whole dynamic range. Also, the requirements imply that measurements of fundamental frequency perturbations have different demands than measurements of voice SPL or voice spectrum.

A 2008 Internet survey of microphone characteristics revealed that many of the commonly offered microphones do not fulfill specifications required for measurements over the whole dynamic range of voice [7]. The survey results for the omnidirectional microphones are shown in Table 1 and those for the directional microphones are given in Table 2. In many cases the information on some of the microphone characteristics was not provided, making such microphones questionable for voice measurements. In some head-mounted microphones the upper dynamic limit was around 130 dB, which is too low to capture the loudest voice at the distance of 5 cm. The frequency response of many microphones was not sufficiently flat and exhibited a "presence peak", i.e., level gain of up to 7 dB at frequencies around 3-10 kHz. In directional microphones, the reference distance for the flattest response was often not provided. This indicates that the task of selecting a microphone shall not be taken lightly. It is our hope that an improved knowledge on microphones and their characteristics will allow more accurate measurements of voice and speech in future.

V. CONCLUSION

While the specified requirements can be considered preliminary, they provide a basis for improvement of accuracy of voice measurements. Before using a microphone for measurement purposes it is important to study the microphone specifications. Most of the manufacturers offer these specifications at their websites. When the specifications are not known, the microphone should not be considered suitable for measurement purposes.

REFERENCES

[1] The ABC's of AKG: Microphone basics & fundamentals of usage: Nashville, TN, AKG Acoustics, U.S., 2003.
[2] Howard DM, Murphy D: Voice Science, Acoustics, and Recording. Plural Publishing, 2007.
[3] Baken RJ, Orlikoff RF: Clinical measurement of speech and voice. ed 2, San Diego, CA, Singular Publishing Group, 2000.
[4] Spielman J, Starr AC, Popolo PS, Hunter EJ: Recommendations for the creation of a voice acoustics laboratory. NCVS Online Technical Memo 2007;No.7, v.1.4.: 1-8.
[5] Titze IR: Workshop on acoustic voice analysis. Summary statement. National Center for Voice and Speech, 1995.

[6] ANSI S1.15-1997/Part 1 (R2006) American National Standard Measurement Microphones. Part 1: Specifications for Laboratory Standard Microphones, 2006.

[7] Sramkova H: Technicke pozadavky pro akustickou registraci hlasu a reci. [Technical requirements for acoustic registration of voice and speech] (B.Sc. Thesis, in Czech). Olomouc, Czech Republic, Dept. Experimental Physics, Faculty of Science, Palacky University, 2008.

Brand and model	Type	d_m cm	L_{noise} dBA	L_{max} dB	F_{min} Hz	F_{max} kHz	ΔL dB	L_{pp} dB	F_{pp} kHz	F_{ppmax} kHz	Price CZK
S :HSP 2	HM	5	28	150	20	20	2	4	3	13	13140
S :HS 2	HM	5	26	142	20	20	N	N	N	N	11700
AKG :HC 577	HM	5	26	133	20	20	3	2	10	15	12687
Shure :WBH53T	HM	5	35	142	20	20	N	N	N	N	8991
S :MKE 2-4 GoldC	LPL	5	26	N	10	20	3	6	4	12	7893
S :MKE Platinum 4C	LPL	5	26	140	20	20	1	4	7	12.5	7812
S :MKE 2EW GOLD	LPL	5	N	142	20	20	2	6	4	12.5	5670
Sony :MCM-C10	LPL	5	N	N	50	15	N	N	N	N	N
S :MKH 800-P48	CL	30	10	136	30	20	0	7	10	30	73055
S :MKH 20-P48	CL	30	10	134	12	20	0	0	Flat	Flat	26910
AKG :C 414 LTD	CL	30	20	140	20	20	2	7	5	12	27200
AKG :C 12 VR	CL	30	22	138	30	20	3	5	2.1	7	107400
AKG :C 4000 B	CL	30	8	145/155	20	20	3	5	1.5	11	15171
AKG :CK 62-ULS	CL	30	13	140	20	20	0	2	5.2	5.2	5511
AKG :CK 92	CL	30	17	132/142	20	20	1	2	4	10	5051
AKG :Perception 420	CL	30	16	135/155	20	20	2	6	5	10	7800
B&K :4958	CL	30	28	140	10	20	1	2	4	10.5	N
B&K :4188	CL	30	15.8	146	8	12.5	N	N	N	N	N
B&K :4950	CL	30	15	142	8	16	0	3	5	10	N
B&K :4942	CL	30	14.6	146	6.3	16	0.5	1.5	5	10.3	N
B&K :4145	CL	30	10	146	3	18	1	2	0.7	10	N
Olympus:ME30W	CL	30	N	N	20	20	N	N	N	N	N
Shure :SM63L	CL	30	N	N	80	20	10	4	2	3	4941

Table 1: Characteristics of some *omnidirectional* microphones obtained from an Internet search in 2008 (adapted from [6]). The colored cells mark values which were not found or which would not fulfill the recommendations for the overall human voice range considered here (minimum F0=50 Hz, maximum spectral frequency of interest 8 kHz, minimum SPL of voice 40/55 dB(A) at the distance of 30/5 cm, maximum SPL of voice 130/145 dB at the distance 30/5 cm).

S = Sennheiser microphones
B&K = Brüel&Kjaer microphones
Type: HM=head-mounted, CL=classical, LPL=lapel
Dir=Directionality (O=omnidirectional, D=directional)
dm = selected distance for voice measurement (can be changed in omnidirectional microphones)
L_{noise}- equivalent noise level
L_{max}- maximum recordable level
F_{min} – low frequency limit
F_{max} – high frequency limit
ΔL– level variation (difference between the maximum and minimum sensitivity) in the frequency response between 70 - 5000 Hz
L_{pp}– maximum gain at the high-frequency presence peak
F_{pp}– starting frequency of the presence peak (+2dB)
F_{ppmax}– frequency of the maximum of the presence peak
N – value not found
The price is in Czech crowns, the exchange rate was about 25 CZK / EUR.

Brand and model	Type	d_m cm	L_{noise} dBA	L_{max} dB	F_{min} Hz	F_{max} kHz	ΔL dB	L_{pp} dB	F_{pp} kHz	F_{ppmax} kHz	Price CZK
S : HSP 4	HM	1	37	150	40	20	2	9	5	10	13140
AKG : C 520/C	HM	N	31	130	60	20	12	6	3	8.5	N
Shure : SM12A	HM	N	N	N	50	50	N	N	N	N	6291
S : ME 104	LPL	N	30	N	40	20	10	6	3	10	4920
S : ME 4-N	LPL	1	N	120	40	20	12	3	8	12	1171
Sony:ECMCS10	LPL	N	N	N	100	16	N	N	N	N	1785
S :MKH 800 TWIN	CL	N	12	134	30	50	1	5	5.5	42.5	N
S :MKH 800-P48	CL	1	10	136	30	20	1	5	20.5	30.5	73055
S : MKH 40-P48	CL	N	12	134	40	20	0	0	Flat	Flat	N
S : MD 441 U	CL	N	N	135	20	20	N	N	N	N	15217
S : MD 421 II	CL	N	N	N	30	17	8	8	1.1	5	21400
S : e 914	CL	N	N	N	30	17	10	10	1.2	5	10700
AKG :C 414 LTD	CL	N	19	137	20	20	N	N	N	N	11240
AKG : C 12 VR	CL	N	6	152	20	20	4	5	1	13	27200
AKG : C 451 B	CL	N	22	128	30	20	8	10	1.5	8	107400
AKG : C 4000 B	CL	N	18	135	20	20	4	4	3	12	10111
AKG : C 3000 B	CL	N	8	145/155	20	20	7	6	3	6.5	15171
AKG : C 1000 S	CL	N	14	140/150	20	20	6	8	1	6.5	7133
AKG : Solid Tube	CL	1	21	137	50	20	12	2	3	10	N
AKG :CK 61ULS	CL	N	20	130/145	20	20	2	0	Flat	Flat	25840
AKG : C 391 B	CL	N	13	140	20	20	0	2	10	10	5511
AKG :Perception 420	CL	N	17	132/142	20	20	2	2	2	10	9651
Shure : PG58	CL	N	16	135/155	20	20	4	4	1.6	11	7800
Shure : SM81-LC	CL	N	N	N	60	15	7	3	1.5	3.5	1521
Audix : SCX25	CL	N	16	146	20	20	2	2	1	4.5	10790

Table 2: Characteristics of some *directional* microphones obtained from an Internet search in 2008 (adapted from [6]). dm = distance at which the frequency response was measured (in directional microphones the frequency response changes when the distance of the microphone from the sound source is changed)
The rest of the abbreviations and symbols are the same as in Table 1.

COMPARISON OF EXCITATION SIGNALS FOR AN ELECTRONIC LARYNX

Christian Jochum, Peter Reiner and Martin Hagmüller

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

Abstract: **This paper deals with the sound quality of electro-larynx devices, which is one method of communication for people who have lost their larynx. Current commercially available devices are characterized by an unnatural, mechanical sound. Assuming the availability of a linear transducer several alternative excitation signals are compared to the sound of a state-of-the-art electro-larynx. The signals considered are both physical models and waveform models. In a listening test 10 sentences, recorded by two healthy electro-larynx speakers using the different excitation signals were evaluated by 20 listeners. Results suggest that a more natural speech sound may be possible without sacrificing intelligibility.**

Keywords: **Electro-Larynx, Excitation Models, Listening Test, Linear Transducer**

I. INTRODUCTION

In case of laryngeal cancer at an advanced stage the last possibility to stop further advancement of the cancer is to remove the entire larynx. This results in the loss of the usual voice production mechanism, based on vibration of the vocal folds. The patients have then to rely on a substitute voice production mechanism. One of the methods available is the electro-larynx (EL), a hand held device, which produces a buzz-like sound and is held against the neck for talking. The vocal tract is excited by this sound and can be used to shape the speech sound. This paper will present different models which were used for an alternative excitation signal and are evaluated with subjective listening tests.

II. BACKGROUND AND RELATED WORK

All current EL devices use a nonlinear transducer concept that was introduced with the first ELs by Gilbert Wright and, later, the Aurex Corp. in the early 40s [7]. An armature is vibrating at a set fundamental frequency (f_0) in a plunger coil and is thereby hammering against a hard plate, just like in old doorbells. This design concept has several shortcomings. First, there are generally significant deficits in low-frequency energy levels (below approx. 500 Hz) [9] and there is a very high level of ambient or self-noise which makes the speaker harder to understand. Further, there is only little variation in the harmonic structure of such generated speech. This is one of the main reasons why the voice is perceived as robotic. Most important, the shape of the source signal is determined by the mechanical characteristics of this concept and thus very hard to control. In [6], a linear transducer design

was proposed, which is similar to current loudspeaker design. Such a design introduces the possibility of using a different excitation signal than currently used for EL devices.

III. DESCRIPTION THE EXCITATION MODELS

Both physical and waveform models have been used as excitation models. The former model the physics of the voice production with the vocal cords, while the later do not care about the mechanism of voice production, but only model the resulting waveform.

We considered two different one-mass model implementations. One approach, which also models the mucosal wave, assumes a single mass-spring system at the entrance of the glottis, and a transmission line that provides the phase delay between the lower and upper part of the vocal folds [2]. This model is called D_1-model in the rest of the paper. A second approach extends the D_1 model by learning a nonlinear mapping from a recorded excitation signal [2], which is further called the D_2-model. In both models, jitter was implemented by varying the stiffness factor k of the spring in the mass-spring model in every cycle (Fig. 1).
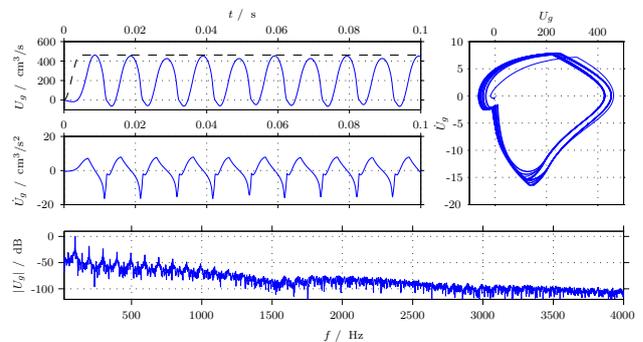
Fig. 1: Simulation using the D_2-model and a jitter of $\approx \pm 1.22$ Hz. The upper left panel shows the glottal flow $U_g(t)$ and—dashed—the lung pressure $P_l(t)$ while the middle panel shows the corresponding derivative $\dot{U}_g(t)$. The right panel shows the phase space and the bottom shows the frequency spectrum of U_g.

As on of the waveform models, we used the R++-model which is a computationally efficient derivative of the Rosenberg-B-model [12]. The voice source waveform is described with the time constants that were introduced by the LF-model [3]. Jitter was obtained by adding a uniformly distributed white noise component to the fundamental period T_0 (Fig. 2).

Hanquinet et al. [4] introduced a phonatory excitation model particularly suitable for the synthesis of disordered
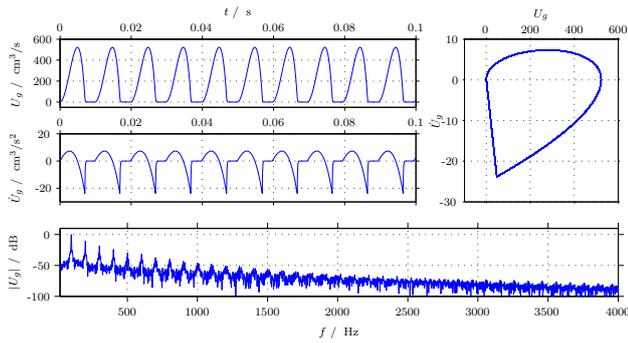
Fig. 2: Simulation using the R++-model with $E_e = -5.96 \text{ m}^3/\text{s}^2$, $T_0 = 10$ ms, $t_a = 109$ μs, $t_e = 7$ ms, and $t_p = 4.94$ ms. The upper left panel shows the glottal flow U_g while the middle panel shows the corresponding derivative $\dot{U}_g(t)$. Right: Phase space. Bottom: Frequency spectrum of U_g.

speech It is further called the HGS model. It utilizes a shaping function to transform a trigonometric driving function into a desired waveform whereby the amplitude and the fundamental frequency of the driving function are used to control the instantaneous frequency and the spectral richness of the output signal independent from each other. While the driving function is represented by a cosine function, the shaping function is defined as an equivalent polynomial formulation of the Fourier series,

$$U_g[n] \approx \frac{1}{2}a_0 + \sum_{k=1}^{\tilde{M}} a_k A^k \cos(k\Theta_n) + b_k A^k \sin(k\Theta_n) \tag{1}$$

which is truncated after $\tilde{M}$ harmonics. a_k and b_k are the Fourier coefficients. A^k is used to modify the Fourier coefficients to influence the spectral richness of the synthetic source signal. Jitter and micro-tremor was added using a stochastic model for jitter [11] (Fig. 3)
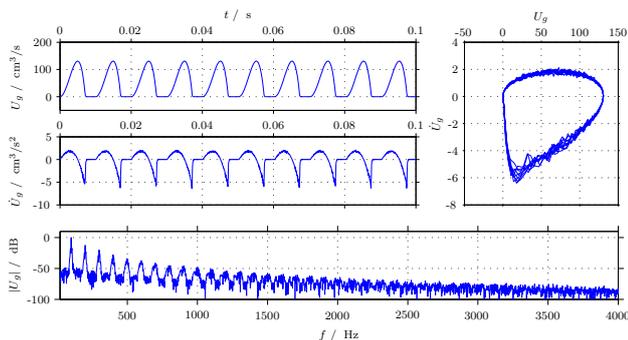
Fig. 3: Simulation of the HGS-model with $f_0 = 100$ Hz, $M = 40$, $b = 0.05$, $B = 4$ Hz, and $\Psi = 6$ Hz. The upper left panel shows the glottal flow U_g while the middle panel shows the corresponding derivative $\dot{U}_g(t)$. The right panel shows the phase space and the bottom plot shows the frequency spectrum of U_g.

Nonlinear oscillators complete the set of source signal models. The Van der Pol (VdP) equation is one of the simplest auto-oscillating systems with linear restoring force and nonlinear damping governed by a nonlinear

second order ordinary differential equation [1].

$$\ddot{x} - \varepsilon\left(1 - x^2\right)\dot{x} + \omega_0^2 x = 0 \tag{2}$$

This model has nothing to do with modeling the human voice, but has been included because of the aim to find a quasi-stationary oscillation and its simple implementation with an analog circuit. Unfortunately, the no useful quasi-stationary oscillation pattern has been found.

Finally, the non-linear oscillator-plus-noise (O+N) model [10] is an autonomous nonlinear deterministic system with a system equation of $y[n+1] = f\left(\mathbf{y}[n]\right)$ where $\mathbf{y}[n]$ represents a time-delay embedding of the oscillators output signal. The nonlinear predictor $f(\cdot)$, which is realized in terms of nonlinear radial basis functions (RBF) using Gaussian kernels, has its parameters learned from a recorded speech signal. The amplitude modulation of the noise signal is achieved by an additional RBF network such that $a[n+1] = f_n\left(\mathbf{y}[n]\right)$. This models the noise signal according to the actual oscillator state and, therefore, it can be added pitch synchronous to the output of the oscillator (Fig. 4).
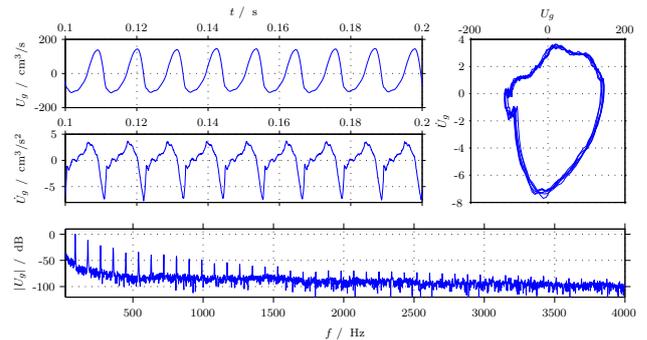
Fig. 4: Simulation of the O+N-model with $f_0 = 91$ Hz and without jitter. The upper left panel shows the glottal flow U_g while the middle panel shows the corresponding derivative $\dot{U}_g(t)$. The right panel shows the phase space and the bottom-most shows the frequency spectrum of U_g.

For a detailed description of the models, see the corresponding references.

IV. SUBJECTIVE LISTENING TEST

To generate an electro-larynx sound the excitation signal was fed into a mini-shaker Brüel & Kjær 4810, which was used as a linear transducer EL. The source signal was generated with a laptop computer at a sampling frequency of 22kHz, with an external digital to analog converter (Edirol UA-25). The signal was fed into a power amplifier, which drove the shaker (Fig. 5). The shaker transfer function has been measured and the neck transfer-function was estimated. The calculated excitation signal was filter with the inverse of both transfer functions to have a better approximation of the desired excitation signal in the vocal tract. The shaker was put in a sound attenuating box to reduce the directly radiated noise.
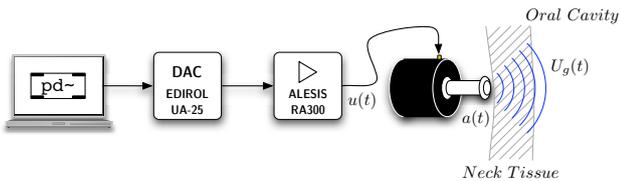
Fig. 5: Block diagram of the experimental system.

In addition, a commercially available Electro-Larynx (Servox Digital) was also used as a reference. Two healthy, German speaking subjects were recorded in a recording studio. Both were experienced in speaking with an EL device. They spoke the inventory of the Oldenburger Satztest [13], which approximates the phoneme frequency in German language with each of the described excitation signals including the EL device. The speech signal was recorded with a high-quality omni-directional head-set microphone (AKG HC 577), sampled with 44.1kHz and recorded on the laptop computer.

In speech pathology, the RBH-Scale [8] and the GRBAS-Scale [5] define commonly used standard methods and are both based on the judgement of subjectively perceived hoarseness. While both of these tests would provide us with a good estimate regarding the healthiness of a voice, they are not particularly useful to access the naturalness of our artificial source models.

Based on the known shortcomings of today's ELs, we defined five quality indicators as follows:

Spectrum, Intonation: *How would you judge the naturalness of the frequency spectrum (speech sound and intonation)?*

Noise: *How disturbing or irritating is the direct sound of the artificial larynx (noise, background sound)?*

Listening Effort: *How much effort does it take to listen to and to understand the speaker?*

Overall Quality: *How would you judge the overall quality?*

The 4 female and 16 male listeners were all native German speakers. The mean age was 26.8 years varying from 21 to 30 years. They were using high quality headphones (Beyerdynamic DT 770 PRO) which were attached to the same D/A converter as used above. First, they were offered 7 pairs to get used to the material. The listeners were allowed to adjust the volume to a comfortable level. Then, they were asked to evaluate seven source signals with 10 sentences each. Eight null pairs were included to check the listener reliability. This gives 88 pairs of EL speech for every listener using the above questions on a category comparison scale (CCR) (Tab. I). A short break was required in the middle of the listening session.

V. RESULTS

One male listener was excluded because several of the null-pairs were not correctly identified, which leaves 19 listeners for further evaluation. The pairwise comparison

Score	The quality of the second token compared to the quality of the first one is:
3:	Much Better
2:	Better
1:	Slightly Better
0:	About the Same
-1:	Slightly Worse
-2:	Worse
-3:	Much Worse

TABLE I: Subjective opinion scale for CCR testing.

of 7 different excitation signals gives 21 comparison mean opinion score (CMOS) results for every evaluation quality. To increase the readability, the results are presented in an order of preference, which is calculated by averaging the scores of the CCR test for each method. The order of preferences shows a ranking and the distance between the excitation signals, but the scale does not correspond to the CCR test. Fig. 6 shows the mean order of preference values $\bar{X}$ with the 95% confidence interval CI_{95} and the standard deviation s.

Even though the standard deviation of the results is high, some significant results may be derived. First, the commercial electro-larynx device (Servox) is on the lower end of the rating for all qualities, but the listening effort, meaning even though it does not sound nice, it may be the most intelligible sound source. At the end of the day, intelligibility is what really counts, so it makes sense, that an electro-larynx device is optimized for this quality. Consistently low ratings have been given to the D_1-model and the VdP-Model, both for the quality and the listening effort. Both do not incorporate jitter or use only a very simple jitter model.

VI. CONCLUSION

We have compared seven excitation signals and evaluated them using a perceptual listening test. The model which is preferred most in terms of overall quality, is the HGS-model, which is also well rated concerning the listening effort. One common property of the better rated model was the incorporation of a jitter model, that makes the speech signal sound more natural.

It has to be noted that the speech recordings were produced by subjects with normal anatomy of the vocal tract. Further studies would have to include laryngectomized speakers. Concerning the intelligibility, dedicated intelligibility test, such as a modified rhyme test would be necessary to better evaluate the intelligibility. Further work would involve generating a more natural fundamental frequency contour, as the flat pitch contour is another main reason for the mechanical sound of an electro-larynx device.

We can conclude, that in case a linear transducer EL device will be available in the future, we may be able improve the perceived quality of EL speech, without having to reduce intelligibility.

164
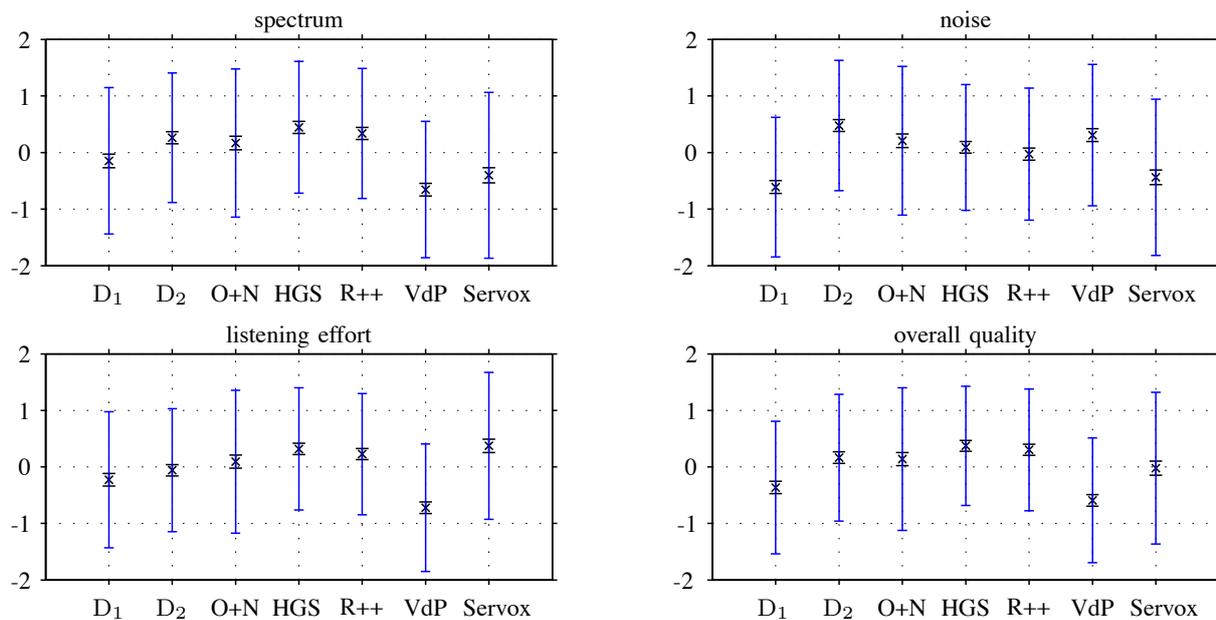
Fig. 6: The order of preference calculated from the CCR of 19 listening subjects. All four panels show the mean scores $\bar{X}$ together with 95% confidence interval CI_{95} and the standard deviation s.

REFERENCES

[1] Balthasar Van der Pol and J. Van der Mark. Frequency demultiplication. *Nature*, 120(3019):363–364, September 10, 1927.

[2] Carlo Drioli. A flow waveform adaptive mechanical glottal model. Technical Report TMH-QPSR, KTH, Vol. 43:69-79, KTH, Department of Speech Music and Hearing, 2002.

[3] G. Fant, J. Liljencrants, and Q.-G. Lin. A four parameter model of glottal flow. Technical Report STL-QPSR Nos. 2-3, Royal Institute of Technology, Stockholm, Sweden, 1985.

[4] Julien Hanquinet, Francis Grenez, and Jean Schoentgen. Synthesis of disordered voices. *Nonlinear Analyses and Algorithms for Speech Processing*, pages 231–241, 2006.

[5] M. Hirano. *Clinical Examination of Voice*. Springer, New York, 1981.

[6] K.M. Houston, R.E. Hillman, J.B. Kobler, and G.S. Meltzner. Development of sound source components for a new electrolarynx speech prosthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2347–2350, 1999.

[7] Hanjun Liu and Manwa L. Ng. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*, 34(3):327–332, September 2007.

[8] T. Nawka, L. C. Anders, and J. Wendler. Die Auditive Beurteilung Heiserer Stimmen nach dem RBH-System. *Sprache - Stimme - Gehör*, 18:130–133, 1994.

[9] Yingyong Qi and Bernd Weinberg. Low-frequency energy deficit in electrolaryngeal speech. *Journal of Speech and Hearing Research*, 34(6):1250–1256, 1991.

[10] Erhard Rank and Gernot Kubin. An oscillator-plus-noise model for speech synthesis. *Speech Communication*, 48(7):775–801, July 2006.

[11] Jean Schoentgen. Stochastic models of jitter. *The Journal of the Acoustical Society of America*, 109(4):1631–1650, April 2001.

[12] Raymond Veldhuis. A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *The Journal of the Acoustical Society of America*, 103(1):566–571, January 1998.

[13] Kirsten Wagner, Volker Khnel, and Birger Kollmeier. Entwicklung und Evaluation eines Satztests in deutscher Sprache-Teil I: Desing des Oldenburger Satztests. *Zeitschrift fr Audiologie*, 38(1):4–15, 1999. (in German).

DIA: A TOOL FOR OBJECTIVE INTELLIGIBILITY ASSESSMENT OF PATHOLOGICAL SPEECH

C. Middag[1], J.P. Martens[1], G. Van Nuffelen[2] and M. De Bodt[2]

[1]Department of Electronics and Information Systems, Ghent University, 9000 Ghent, Belgium
[2]Antwerp University Hospital, University of Antwerp, 2650 Edegem, Belgium

Abstract: **Intelligibility is generally accepted to be a very relevant measure in the assessment of pathological speech. In clinical practice, intelligibility is measured using one of the many existing perceptual tests. These tests usually have the drawback that they employ unnatural speech material (e.g. nonsense words) and that they cannot fully exclude errors due to the listener's bias. This raises the need for an objective and automated tool to measure intelligibility. Here, we present the Dutch Intelligibility Assessment (DIA), an objective tool that aids the speech therapist in evaluating the intelligibility of persons with pathological speech. This tool will soon be made publicly available.**

Keywords : **objective intelligibility assessment, pathological speech, speech therapy.**

I. INTRODUCTION

Communication is getting increasingly important in our society. People with communication disorders often suffer other social discomfort as well. Therefore, follow-up of these patients in order to improve their pronunciation is becoming increasingly important. An important measure in the assessment of communication efficiency is intelligibility. This is often determined in a perceptual way, which is subjective in nature. Consequently, there is a growing need for an automated, and thus objective method for measuring intelligibility.

Previous software packages have been developed to measure intelligibility of English patients suffering from dysarthria [1] and [2] describes a system doing the same for German laryngectomees and children with cleft lip and/or palate. The former is based on a measure called goodness of fit of the alignment between the uttered speech and the target speech, while the latter uses word accuracy rate, after doing speech recognition on the uttered speech.

We present the Dutch Intelligibility Assessment (DIA), a tool to assist speech therapists when dealing with patients suffering from pathological speech, which is based on a novel methodology, described in [3]. The method underlying the DIA tool extracts phonemic, phonological and context-dependent phonological features from automatic speech alignment on the basis of acoustic models that were trained on normal speech.

Based on those features, intelligibility is predicted using a compact model that can be trained on pathological speech samples. The experimental evaluation of the system shows standard errors between perceived and computed intelligibilities lower than 8%. This is a sufficiently strong basis for the development of an automated version of the Dutch Intelligibility Assessment.

II. THE PERCEPTUAL DIA TEST

The –initially subjective- test we have automated is the Dutch Intelligibility Assessment (DIA) test [4]. This test consists of 50 consonant-vowel-consonant (CVC) words, mostly nonsense but well pronounceable words. These 50 words are divided into three subtests: one testing the Dutch consonants in the initial position, one in the final position and the last one testing the vowels and diphtongs in the middle position in the word. To avoid guessing by the listener, there are 25 variants of each subtest, of which one is chosen at random for each execution of the test. The perceptual intelligibility score is then calculated as the percentage of tested phonemes which are correctly identified. This test is proven to be highly reliable (an interrater correlation of 0.91 and an intrarater correlation of 0.93 [4,5]).

III. THE COMPUTERIZED DIA TOOL

Within the framework of the SPACE[1] project, this perceptual DIA test has been automated, as described in [6]. While the perceptual test only uses the 50 tested phonemes, the computerized version takes every phoneme of the 50 words into account. All uttered speech is lined up against the target words using forced alignment of two automatic speech recognizers (ASRs). This results in three feature sets: phonemic features, phonological features and context-dependent phonological features. The phonemic features describe how well on average the Dutch phonemes are recognized by the used ASRs, while the phonological features describe how well a phonological feature can be realized by the speaker. The context-dependent phonological features point to transitions between two articulatory positions. These feature sets are then used in a simple regression model to predict the intelligibility of the speaker.

---

[1] http://www.esat.kuleuven.be/psi/spraak/projects/SPACE/

Figure 1 Screenshot of the recording environment of the tool. The patient is presented with a sequence of words, which are displayed one after the other. These words are automatically recorded by the tool for later analysis.

Different models have been designed: a general model, as well as pathology-specific models for people with hearing impairment, dysarthria and laryngectomy. We recently also added a model for children with cleft lip and palate. As shown in [3,6], the correlations between the computed intelligibility scores and the perceptual scores are about as high as the interrater reliability, which means the automated version can compete with the human judging. Moreover, the DIA tool could be a more objective and less time-consuming way for the speech therapist to administer the test.

IV. TOOL DESCRIPTION

Our purpose was to design a user-friendly, easily available tool which does not require a complex setup to administer the test. To use the DIA tool, the user only needs a PC or laptop with a web browser, a head set and sound card, and an up-to-date Java runtime environment. The tool works in a client/server environment and can be used both in online or offline mode.

Once a user has an account, patients can be added and edited. As we respect the privacy of the patients, every user can only view recordings of its own patients. When a patient is added, the user can start the test. We advice to do a microphone test first, to be assured that the recording quality is well enough and the microphone is in the right position (e.g. not too close to the mouth). When starting the test, a sequence of words is presented to the patient (Fig. 1). Each of these is recorded as a separate .wav file, which is stored for subsequent analysis.

When the recording is finished, the speech therapist can analyze the recordings by listening to every word and filling in the missing phoneme (Fig. 2). This results in a perceptual score and a report displaying the nature of the errors, e.g. wrong place/manner of articulation, as described in [4,5]. Every recording can be judged by several listeners, which can easily be added in the recording menu.

In a final step, the user can also run an automatic analysis. This step results in an objective intelligibility

score, as well as a number of statistics of the analysis (Fig. 3). These statistics display the speech profile of the current patient, compared to normal speakers, as well as a number of well-defined pathologies.

V. TOOL VALIDATION

To validate the tool, a master student recorded 33 laryngectomees, 19 hearing impaired, and 9 dysarthric patients. The recording settings were not always ideal and sometimes a lot of background noise could be noticed. Every patient performed the test, which was recorded using our DIA tool.

Figure 2 Perceptual analysis of the recordings. When clicking on the button, the corresponding .wav file is played, and the listener can fill in the missing part.

Automatic score

94%

Figure 3 A part of the final report, showing the results of the automatic analysis. Here, an overall score for the objective intelligibility is calculated (upper left), as well as a number of statistics, showing a.o. how the current patient performs with respect to a normal speaker (label "testpreker") as well as a number of well-defined speech pathologies.

Apart from the objective score calculation, the subjective evaluation of the speech intelligibility was performed by two professional listeners.

The interrater agreement between the two listeners was measured using the Pearson correlation coefficient between their scores and reached values as high as 94%. The Pearson correlation between the mean of the listener's scores and the objective scores reached 90%, which is almost as good as the interrater agreement.

VI. FUTURE WORK

These results are very promising and reveal that an objective evaluation of pathological speech can indeed be useful in the clinical practice. In a next step, we will investigate the possibility of replacing the nonsense words by more natural speech such as existing words or even phrases. We are also working towards a more profound articulatory assessment, which can then lead to the determination of an appropriate therapy for every patient.

REFERENCES

[1] J. N. Carmichael, "Introducing Objective Acoustic Metrics for the Frenchay Dysarthria Assessment Procedure", PhD Thesis, 2007, University of Sheffield.

[2] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster and E. Nöth, "PEAKS – A system for the automatic evaluation of voice and speech disorders", in Speech Communication vol. 51, 2009, pp. 425-437

[3] C. Middag, J.P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features," in EURASIP Journal on Advances in Signal Processing, vol. 2009, Article ID 629030, 9 pages, 2009. doi:10.1155/2009/629030

[4] M. De Bodt, C. Guns, and G. V. Nuffelen, NSVO: Nederlandstalig SpraakVerstaanbaarheidsOnderzoek, Vlaamse Vereniging voor Logopedisten, Herentals, Belgium, 2006.

[5] G. Van Nuffelen, M. De Bodt, C. Guns, F. Wuyts, and P. Van de Heyning, "Reliability and clinical relevance of segmental analysis based on intelligibility assessment," Folia Phoniatrica et Logopaedica, vol. 60, no. 5, pp. 264–268, 2008.

[6] C. Middag, G. Van Nuffelen, J. P. Martens, and M. De Bodt, "Objective intelligibility assessment of pathological speakers," in Proceedings of the International Conference on Spoken Language Processing (Interspeech '08), pp. 1745–1748, Brisbane, Australia, September 2008.

Special Session on Singing voice

Chairperson and Introduction:
J. Sundberg, Sweden

SINGING VOICE AND PREGNANCY
PRELIMINARY RESULTS FROM A CASE STUDY

Filipa M. Lã
Department of Communication and Arts, University of Aveiro, Portugal
Johan Sundberg
Department of Speech, Music and Hearing, School of Computer Science and Communication, KTH, Stockholm, Sweden

During pregnancy significant changes in bodily tissues occur. For example, the cervix undergoes deep structural/biomechanical alterations due to an increase in concentration of progesterone. Previous studies have found a significant correlation between the changes that both cervical and vocal fold smears undergo during the menstrual cycle, demonstrating a relevant hormonal influence on laryngeal tissues. It can be hypothesised that such tissue changes that may occur during pregnancy affect conditions for phonation with respect to e.g. vocal fold motility. To test this hypothesis recordings of audio, electrolaryngograph, oral pressure and air flow signals were made during pregnancy, at birth and after pregnancy of a semi-professional classically trained soprano. The tasks involved repetitions of the syllable [pae] while performing a diminuendo at various pitches, thus allowing determination of the lowest pressures producing vocal fold vibration and vocal fold contact, i.e. the phonation and contact threshold pressures. Oral pressure during the occlusion for the consonant [p] was accepted as a measure of subglottal pressure. Concentrations of sex female steroid hormones were measured during pregnancy, at birth and post-partum. Results showed a steep decrease of concentrations of progesterone and oestrogens from pregnancy to post-partum conditions. Likewise, phonation and collision thresholds decreased markedly at birth and post-partum, shifts that are in accordance with expectations based on the effects of sex steroid hormones on tissue viscosity and water retention. The results thus demonstrate an effect of pregnancy on the voice.

INTRODUCTION

Previous research has suggested that the larynx is subject to hormonal influence [1]. Physiological and acoustical changes of the female voice, such as those occurring during the menstrual cycle and at climacterium, have been reported to be associated with significant variations of sex steroid hormonal concentrations (i.e. oestrogens, progesterone and testosterone) [2-11]. Significant physiological and structural similarities have been claimed to exist between cervical and vocal fold mucosas. Changes in the mucosa of the cervix observed at the three phases of the menstrual cycle were also observed in the mucosa of the vocal folds due to variations in the female sex steroid hormonal concentrations [1, 2, 10].

Like the menstrual cycle, pregnancy is also associated with significant hormonal changes. Elevated concentrations of both oestrogens and progesterone, with a dominant role of high concentrations of progesterone can be observed, especially during the third trimester of pregnancy [13, 14]. According to Shiff and Burn [15], high concentrations of oestrogens increase the viscosity of bodily tissues. This is caused by a shift in the solid-gel equilibrium of the interstitial fluid towards a more solid state, with a consequent water retention and oedema [15]. Additionally, high progesterone content has been reported to increase the viscosity of vocal fold glandular secretions [1, 2]. Thus, pregnancy can be expected to affect voice production in terms of changed vocal fold motility.

Few studies have been reported concerning the effects of pregnancy in the voice. van Gelder (1974) reported vocal symptoms in pregnant singers, such as small submucous haemorrhages, redness and swelling of laryngeal tissues, which he referred to as "*laryngopathia gravidarium*" [16]. He pointed out that these symptoms were similar to those observed during the menstrual cycle in some opera singers, a condition that he named "*laryngopathia menstrualis*" [16]. A more recent study of the speaking voice of pregnant and non pregnant women assessed the incidence of vocal symptoms (e.g. hoarseness, vocal fatigue, and aphonia), and compared maximum phonation time (MPT) and voice turbulence index (VTI) between pregnant and post-partum conditions [17]. No significant differences were observed in the incidence of vocal symptoms between the groups. However, vocal fatigue seemed more prevalent in the pregnant women. This group also presented significant decrease in maximum phonation time (MPT). When comparing pregnancy with post-partum conditions (12-24 hours after birth), the authors further found a significant increase in MPT and a decrease in voice turbulence index (VTI) for the post-partum condition [17].

With respect to the singing voice, studies of effects of pregnancy were mainly based on singers' perceptions. Mostly, positive effects have been reported, e.g. improved voice quality [18].

Previous research has suggested that both phonation and collision threshold pressures (PTP and CTP,

respectively) reflect vocal fold motility [19, 20]. The current investigation analyses biomechanical properties of the vocal fold vibration in terms of these thresholds pressures in a pregnant singer.

METHOD

The subject was a healthy, non-smoker classically trained soprano, aged 28. Following a longitudinal study design, she was recorded every week, from week 28 of pregnancy until week 8 after birth. This yielded a total of 21 recordings: (i) 12 during the last weeks of pregnancy (the *Prae* group of recordings); (ii) one at 48 hours after birth (the *At* recording); and (iii) 8 during the weeks following birth (the *Post* group of recordings). For technical reasons, one of the recordings in the *Prae* group had to be discarded.

Furthermore, three blood samples were collected for each of the above recording groups: (i) one in week 29th of pregnancy; (ii) one 48 hours after birth; and (iii) one 8 weeks after birth.

A combination of the Digital Laryngograph Microprocessor and the Glottal Enterprises MS-110 computer interface was specially designed for the purposes of this study. It allows simultaneous recording of two AC signals and two DC signals. The audio and electroglottograph signals were recorded by the Laryngograph component. It also imported the flow and the oral pressure signals from the Glottal Enterprises unit, which were collected by means of a Rothenberg flow mask and a pressure transducer, respectively. The latter was attached to a thin plastic tube inserted into the flow mask, such that its end was located inside the subject's lip opening at the corner of the mouth.

All these four signals were digitized and sent over a USB contact into a PC provided with the Speech Studio software; thus, audio, EGG, subglottal pressure and airflow signals were obtained as separate tracks of wav computer files.

Vocal tasks included six performances of a set of repetitions of the syllable [pae] sung as diminuendos at pitches A3, E4, B4 and F5. This allowed determination of the lowest pressures producing vocal fold vibration and contact vocal fold, i.e. PTP and CTP [20]. Oral pressure during the occlusion for the consonant [p] was accepted as an estimate of subglottal pressure (Psub).

The wav files were analyzed by means of the SoundSwell software. PTP was calculated as the mean of the lowest pressure that caused phonation and the highest pressure that did not produce phonation as evidenced by the flow signal. Loss of vocal fold contact decreases the amplitude of the EGG signal considerably; therefore, CTP was calculated as the mean of the lowest pressure that caused vocal fold contact and the highest pressure that failed to produce vocal fold contact, according to the EGG signal amplitude. The threshold values obtained

were averaged across the six versions produced on each of the four pitches.

RESULTS

Fig. 1 shows the concentrations of progesterone and oestradiol for the *Prae*, *At* and *Post* conditions. Results show the highest values for the *Prae* condition for both hormones.

Figs. 2 and 3 compare the PTP and CTP averages with PTP values calculated according to Titze's equation [19]. As can be seen, the values obtained from the current study show a dependence on fundamental frequency (F0) similar to that predicted by Titze's equation. The scatter of the data points in the graphs reflects a considerable variability of the thresholds during the recorded weeks. Additionally, it can be observed that PTP shows a closer approximation to the Titze reference for the *Post* than for the *Prae* conditions.
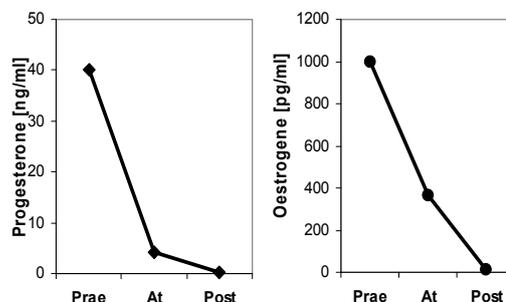
Fig.1: Subject's progesterone and oestradiol concentrations for *Prae*, *At* and *Post* conditions.
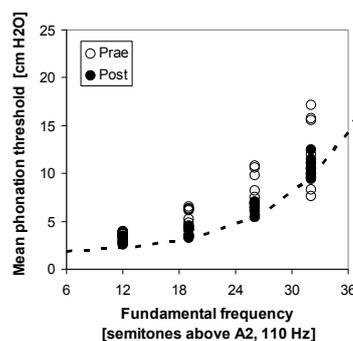
Fig. 2: Mean PTP for all recordings during *Prae* and *Post* conditions, as function of F0 compared with Titze's equation (dashed curve).

On average, CTP exceeded PTP by 40 to 50%, corresponding to 1 and 5 cm H_2O. Mostly both thresholds showed similar changes from week to week, as illustrated in Fig. 4.
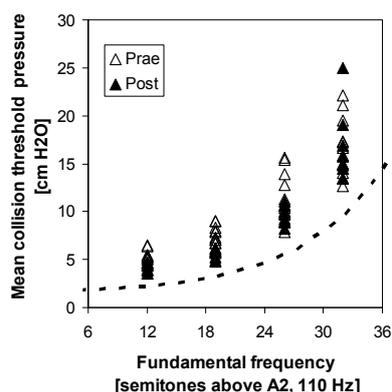
Fig. 3: Mean CTP for all recordings during *Prae* and *Post* conditions, as function of F0 compared with Titze's equation (dashed curve).
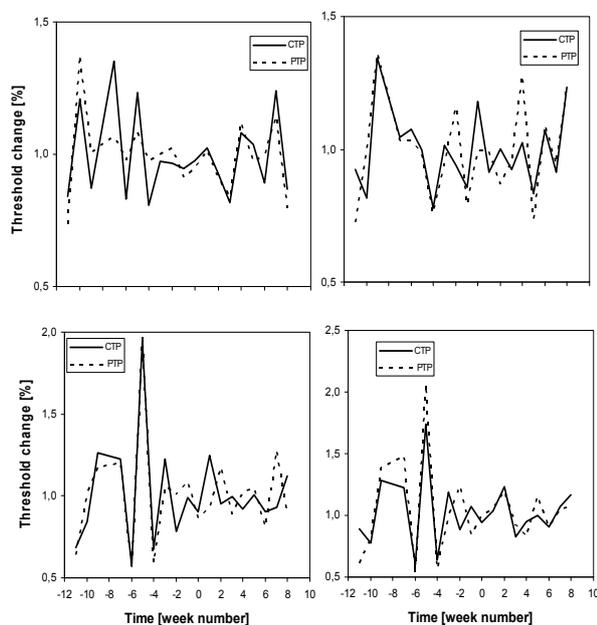
Fig. 4: Relative week-to-week variations of CTP and PTP. Weeks number -12 to -1 correspond to the *Prae* condition; week number 0 represents the *At* birth condition; and weeks number 1 to 8 correspond to *Post* birth condition.

Fig. 5 shows the threshold values grouped *Prae*, *At* and *Post* conditions for each of the four pitches. Like the results obtained for hormonal concentrations, the highest values for both thresholds occurred for the *Prae* condition, for all pitches.

DISCUSSION

The main question raised in this pilot investigation was whether sex steroid hormonal variations during

pregnancy affect vocal folds motility. The results showed a clear effect of pregnancy on voice production, complementing previous findings [17].
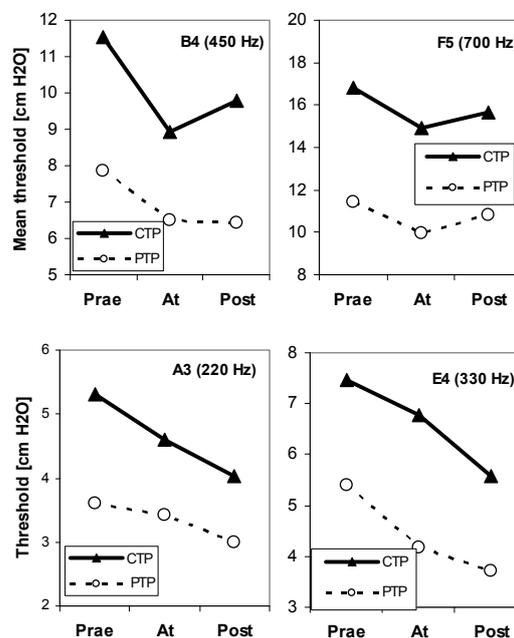
Fig. 5: Mean CTP and PTP for pitches A3, E4, B4 and F5 for the *Prae*, *At* and *Post* conditions.

It has been assumed that CTP and PTP reflect vocal fold motility [19, 20]. Therefore, these parameters were used as measurements to compare vocal fold motility between the last trimester of pregnancy, at birth and post-partum. The PTP values obtained basically follow Titze's equation [19] (see Fig. 1). This supports the assumption that the PTP data obtained yielded reliable information. Moreover, the CTP showed a week-to-week variation that, by and large, was similar to that of the PTP (see Fig. 2). This leads to the assumption that CTP, like PTP, reflects vocal fold motility.

Concerning female sex steroid hormones, results were in accordance with the expected pattern of highest concentrations of progesterone and oestrogens during pregnancy, followed by a sharp decrease for the *At* and *Post* conditions [13, 14]. Likewise, the PTP and CTP showed highest values for the *Prae* condition, followed by a sharp decrease to the *At* and *Post* conditions. This is consistent with the assumption that vocal fold motility is affected by variations in female sex steroid hormones.

Substantial support for this assumption can be found in the field of endocrinology. Concentrations of oestrogens affect solid-gel equilibrium of interstitial body tissues; elevated concentrations swift this equilibrium towards a more solid state, causing an increased viscosity and oedema through water retention [15]. Moreover, elevated concentrations of progesterone have been related to an

174

increase in the viscosity of the vocal mucosal glandular secretions [1, 2]. Thus, one might expect that increased tissue viscosity, water retention, and viscosity of glandular mucosa secretions contribute to a decrease of vocal fold motility reflected in raised PTP and CTP.

CONCLUSIONS

This investigation has shown, for a single subject, considerable changes of PTP and CTP during pregnancy, implying effects in vocal fold motility. PTP and CTP shifted from high values for the *Prae* condition (i.e. prior to birth), when elevated progesterone and oestrogen concentrations were observed, to lower values *At* and *Post* conditions, when concentrations of these hormones were markedly reduced. The results thus demonstrate, for this singer, how hormonal variations during pregnancy affected the vocal fold motility.

REFERENCES
[1] J. Abitbol, P. Abitbol, and B. Abitbol, "Sex hormones and the female voice." *J Voice*, vol. 13, pp. 424-446, 1999.
[2] J. Abitbol, J. Brux, G. Millot, M. F. Masson, O. L. Mimoun, H. Pau, and B. Abitbol, "Does a hormonal vocal cord cycle exist in women? Study of vocal premenstrual syndrome in voice performers by videoscopy-glottography and citology on 38 women." *J Voice,* vol. 3, pp.157-162, 1989.
[3] F. M. Lã, and J. W Davidson, "Investigating the Relationship between Sexual Hormones and Female Western Classical Singing". *J Res Stud Mus Edu*, vol. 24, pp. 75-87, 2005.
[4] F. M. Lã, J. W. Davidson, W. Ledger, D. M. Howard, and G. Jones, "A case-study on the effects of the menstrual cycle and the use of a combined oral contraceptive pill on the performance of a western classical singer: an objective and subjective overview", *Musicae Sciencea. Special Issue: Performance Matters,* pp. 85-107, 2007.
[5] H. Isenberg, W. S. Brown, and H. B. Rothman, "Effects of menstruation on the singing voice; Part II: Further developments in research". In *Transcriptions of the twelfth symposium for the care of the professional voice*, Part I. V. Laurence, Ed. New York: The Voice Foundation. 1983, pp. 117-123.
[6] M. J. Boulet, and B. J. Oddens, "Female voice changes around and after the menopause – an initial investigation". *Maturitas*, vol. 23, pp. 15-21, 1996.

[7] O. Amir, T. Biron-Shental, and E. Shabtai, "Birth Control Pills and Nonprofessional Voice: Acoustic Analyses", *J Speech Lang Hear Res,* vol. 49, pp. 1114–1126, 2006.
[8] P. H. Damstè, "Voice change in adult women caused by virilizing agents", *J Speech Hear Disord*, vol. 32, pp. 126-132, 1967.
[9] F. M. Lã, W. Ledger, J. W. Davidson, D. M. Howard and G. Jones, "The effects of a third generation combined oral contraceptive pill on the classical singing voice", *J Voice*, vol. 21, pp. 754-761, 2007.
[10] J. Abitbol and Abitbol, P. "Endocrine disorders of the larynx", in *Diseases of the Larynx*. A. Ferlito, Ed. London: ARNOLD, 2000, pp. 311-333.
[11] S. W.Chae, G. Choi, H. J. Kang, J. O. Choi, and S. M. Jin, "Clinical analysis of voice change as a parameter of premenstrual syndrome", *J Voice*, vol. 15, pp. 278-283, 2001.
[12] K. Myers, S. Socrate, D. Tzeranis, M. House, "Changes in the biochemical constituents and morphologic appearance of the human cervical stroma during pregnancy" *Eur J Obstet Gynecol Reprod Biol*, vol. 144S, pp. S82-S89, 2009.
[13] G. Larciprete, H. Valensise, B. Vasapollo, F. Altomare, R. Sorge, B. Casalino, A. De Lorenzo and D. Arduini, "Body composition during normal pregnancy: reference ranges", *Acta Diabetol*, vol. 40, pp. S225–S232, 2003.
[14] R. E. Garfield, G. Saade, C. Buhimschi, I. Buhimschi, L. Shi, S-Q. Shi and K. Chwalisz, "Control and assessment of the uterus and cervix during pregnancy and labour", *Hum Reprod Update*, vol. 4 (5), pp. 673-695, 1998.
[15] M. Shiff and H. Burn, "The effect of intravenous estrogens on ground substance", *Arch Otolaryngol,* vol. 73: pp. 43-51, 1961
[16] L. van Gelder, "Psychosomatic aspects of endocrine disorders of the voice", *J Commun Disord,* vol. 7, pp. 257-262, 1974,
[17] A. L. Hamdan, L. Mahfoud, A. Sibai, and M. Seoud, "Effect of Pregnancy on the Speaking Voice", *J Voice*, [in Press], 2007.
[18] A. L. Abramson, B. M. Steinberg, W. J., Gould, E. Bianco, R. Kennedy and R. Stock, "Estrogen receptors in the human larynx: clinical study of the singing voice". In *Transcripts of the Thirteenth Symposium: Care of the Professional Voice*, Part II. V. Laurence, Ed. New York: The Voice Foundation, 1984, pp. 409-413.
[19] I. R. Titze, "Phonation threshold pressure: A missing link in glottal aerodynamics", *J Acoust Soc Am*, vol. 91, pp. 2926-35, 1992.
[20] L. Enflo and J. Sundberg, "Vocal fold collision threshold pressure: an alternative to phonation threshold pressure?" *Logoped Phoniatr Vocol* [accepted for publication], 2009.

CASE STUDY OF VOICE QUALITY DIFFERENCES IN A SOPRANO SINGING IN DIFFERENT EARLY MUSIC PERFORMANCE STYLES

David M Howard[1], Jude Brereton[1], Helena Daffern[2]

(1) Audio Lab, Department of Electronics, University of York, Heslington, York, YO10 5DD, UK
(2) Singer and researcher, London and York, UK

Abstract:
This paper considers the characteristics of three differing styles of singing early music, as characterized by Richard Bethell [1] of the National Early Music Association, UK. In particular, the sung outputs from a postgraduate soprano who was practiced in singing all three styles are analysed along with the output from an electrolaryngograph which provides data on cycle-by-cycle fundamental variation as well as vocal fold contact area. The results are compared and contrasted with those from a group of early music and opera singers analysed previously.
Keywords: **Singing, voice analysis, voice acoustics, electrolaryngography, closed quotient, opera, early music.**

I. INTRODUCTION

Sung performances by modern-day singers of music composed between approximately 1600 to 1900 (referred to herein as *early music*) can vary considerably in terms of technique and vocal output. Details of the exact techniques that would have been used by the singers of that period are scarce, but known major differences between then and now include tuning systems (non-equal temperament then and equal temperament now), pitch reference (typically higher today note-for-note based on A4 (440Hz) rather than A4 (415Hz), but this does depend on whether the musical key of the piece has been changed in modern editions), the size of audiences (modest then and much larger today), the timbre of accompanying instruments (today's instruments have developed considerably in terms of their timbral output, tuning stability and overall acoustic output power) and overall size of performance spaces (today's spaces are much larger requiring a singing technique that achieves greater acoustic output power).

Singing fashion has changed over the years and the performance of early music has been subjected to these variations. Since the 1960's revival of early music in the UK, many singers have become interested in performing early music and Potter [2, p3] notes that *"One of the consequences of the stylistic fragmentation of classical music has been the proliferation of singing styles associated with early music"*.

There continues to be much debate about appropriate singing styles and techniques for the performance of early music today. Recently Richard Bethell of the National Early Music Association, UK, [1] described three commonly used performance styles, the third being as yet less well established, that should be more widely considered for the performance of this repertoire.

A) *Operatic*: Institutionally/ academically trained singer's formant voice, with fairly wide continuous vibrato, lower larynx development (producing a rich and plummy sound) and capable of high volume.

B) *Early Music Mainstream*. When compared to the operatic voice, higher larynx position (producing a sound midway between categories A and C), narrower amplitude (but more or less continuous) vibrato, and generally lower volume.

C) *Clear Smooth Sweet Chaste*. Fairly soft, straight tone, without vibrato except as an ornament. Little or no lower larynx development, producing a sound close to the speaking voice

An initial analysis is conducted of vibrato samples from each style. An initial analysis is conducted of vibrato samples from each style and larynx closed quotient (the percentage of time for which the vocal folds remain in contact in each cycle) data is presented, for which differences have been shown for adult singers with training and experience [3].

These results are compared to those obtained for modern-day professional singers of early music and professional opera singers [4] to highlight similarities and differences.

II. METHOD

The experiment was carried out in a performance space in the Music Department at the University of York, UK. A young professional soprano sang 'Lascia ch'io pianga' from *Rinaldo* by Handel accompanied by a harpsichord in the three different styles identified above.

The singer was placed further away from the harpsichord than she would have been for a performance in order to keep the singer to harpsichord output ratio low on the audio recording, which was made with two closely positioned (~30cm off-axis) omnidirectional microphones (Sennheiser MKH20 and DPA 4060). In addition, the output from an electrolaryngograph was simultaneously recorded on an Audio Devices 744 4-channel digital recorder at 44.1 kHz sampling rate and 24 bit resolution. The harpsichord level was kept constant with the lid open

with the short stick for all three performances so as not to influence the timbral or intensity output of the singer with the accompaniment (Daffern et al., 2006).

An analysis was made of the note C5 from bar 14 of the aria, sung to the final syllable of the word *libertà* in the three different styles.
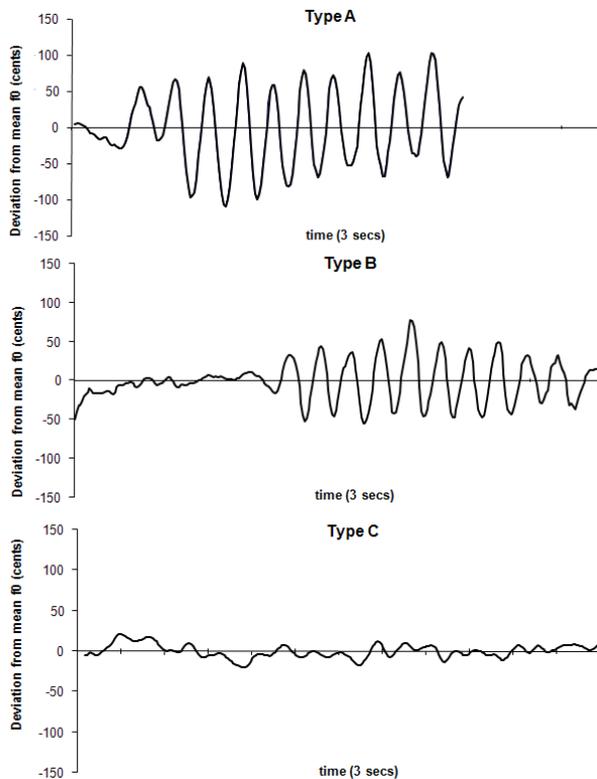
Figure 1: Fundamental frequency contours for the final syllable of the word *libertà* (C5) from bar 14 of 'Lascia ch'io pianga' from *Rinaldo* by Handel sung by a young professional soprano in the three styles A, B and C (see text).

III. RESULTS

Fig. 1 shows the fundamental frequency contours of the analysed samples in styles A, B and C respectively. There are clear differences between the three styles. Style A (operatic) shows vibrato more or less continuously throughout the tone, whereas in style B (early music) a periodic vibrato is not apparent until nearly half-way through the tone. Style C shows no discernible periodic vibrato, although there is some natural fluctuation in the sung fundamental frequency.

There are further differences in the vibrato, when present, in the tones in styles A and B. Although the rate of the vibrato is the same, at 6 oscillations per second for each style, the average peak-to-peak extent of vibrato in Style A (123.8 cents) was larger than that found in Style B (87.4 cents).
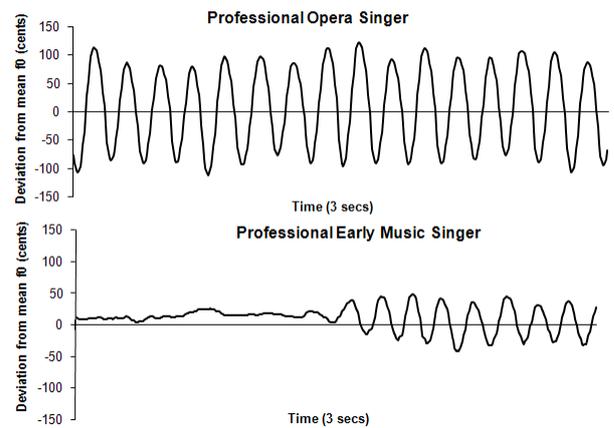
Figure 2: Fundamental frequency contours for individual notes sung by a professional opera singer and a professional Early Music singer (Data from [4]).

The differences in the vibrato found here between styles A and B mirror those found by Daffern [4] when comparing opera and early music singers. Example data from Daffern's study are shown in Fig. 2 for a a professional opera singer (adult female) and a professional early music singer (adult female). A comparison of the vibrato results of the style A sample here with the sample from a professional opera singer shown in Fig. 2, for which average peak-to-peak vibrato extent is 191.5 cents, indicates that the vibrato extent used by the current singer is not as extreme as that normally found in professional opera singers. Daffern also found for her opera singer group that they produced a more consistent vibrato from the very onset of the tone, whereas there is some delay in the onset of vibrato in style A produced by the singer in this study (compare plots in Figs. 1 and 2).

The early music singers in Daffern's study typically produced vibrato as a stylistic component in the context of the music, producing appropriate notes as straight tones with a late introduction of vibrato with an average peak-to-peak vibrato extent of 69.8 cents. This is illustrated in the graphs below.

The average peak to peak extent of the vibrato tones produced by the early music singers in Daffern's study was also generally lower than observed for the opera singers, which is a characteristic also observed in the results of the current study.

Larynx closed quotient (CQ) is measured with an electrolaryngograph [5] and it shows the percentage of each cycle for which the vocal folds are in contact. It is important to note that when the folds are *in contact*, it does not necessarily mean that they are *closed*, since they can be partly open. The output from the electrolaryngograph cannot show the difference between partly and fully closed.
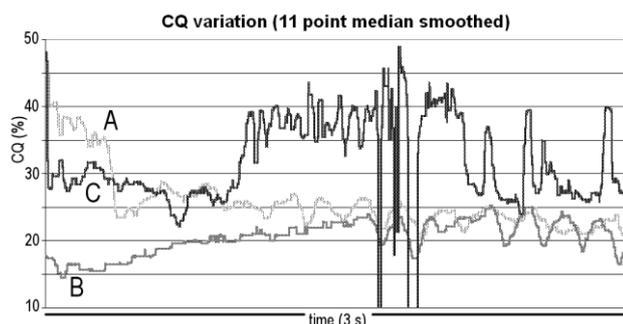
Figure 3: Eleven point median smoothed larynx closed quotient (CQ) plots against time for the final syllable of the word *libertà* (C5) from bar 14 of 'Lascia ch'io pianga' from *Rinaldo* by Handel sung by a young professional soprano in the three styles A, B and C (see text).
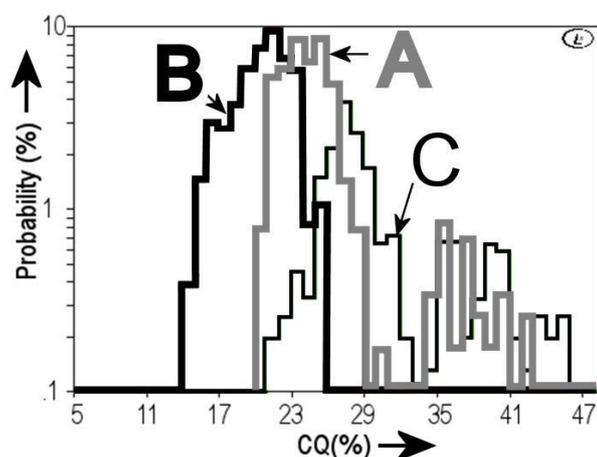
Figure 4: Larynx closed quotient (CQ) histograms for the final syllable of the word libertà (C5) from bar 14 of 'Lascia ch'io pianga' from Rinaldo by Handel sung by a young professional soprano in the three styles A, B and C (see text).

Fig 3 shows the CQ variation with time for the sung notes shown in Fig. 1; these values have been 11 point median smoothed due to the break-up that occurs around the midpoint of the type C plot. It can be seen that during the vibrato portions of the output during the type A and type B productions (compare the plot with Fig. 1), the CQ values are closely matched. However, the type A performance starts with relatively high CQ values during the early part of the note and then drops to around 23% as it meets the type B CQ output. The type C plot starts around 30% and then rises to nearly 40% before dropping again towards the end of the note.

To provide a different view of CQ values, the overall range of CQ values used during this note is shown in Fig 4 in the form of a second order histogram, which serves to remove non smooth values [5], for each of the types A, B and C. It is clear that the CQ used for types B occupies

its own range but that the CQ distributions used for types A and C are bimodal. The nature of this bimodality can be seen in the time plots (Fig. 3) for the type A output which starts high and then drops at the point where the vibrato starts (see Fig. 1). The type C version starts and ends around 30% and has a portion around the centre which is closer to 40%.
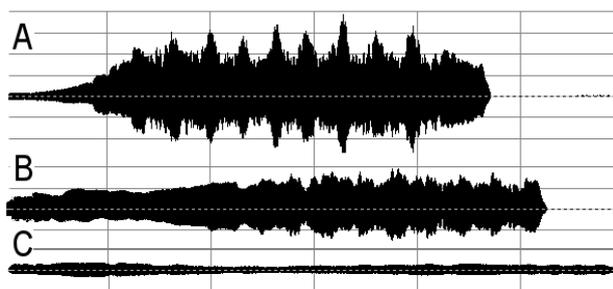
Figure 5: Sound pressure waveforms for the final syllable of the word *libertà* (C5) from bar 14 of 'Lascia ch'io pianga' from *Rinaldo* by Handel sung by a young professional soprano in the three styles A, B and C (see text).

Fig. 5 shows the acoustic pressure waveforms for the three types. These waveforms are plotted with the correct amplitudes relative to each other so that comparisons can be made. It can be seen that the output for type A has the greatest amplitude, followed by type B and then type C. The type C output becomes quieter during its mid portion, and it turns out that this is where the vocal fold contact variation is both quite low in amplitude and close to being sinusoidal (n.b. because of the algorithm used to calculate CQ, a sinusoidal electrolaryngograph output waveform will have a value around 40%).
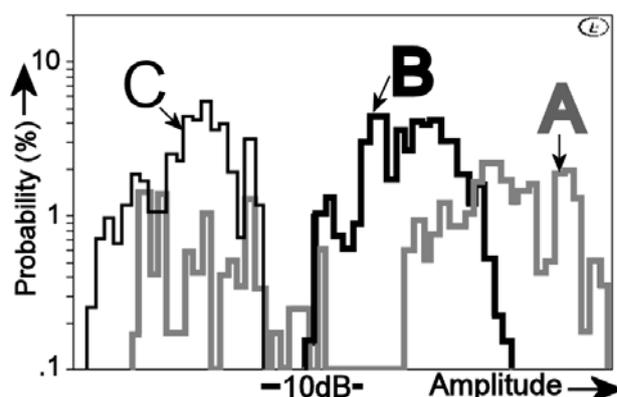
Figure 6: Sound pressure (uncalibrated hence no reference level and only dB given) output for the final syllable of the word *libertà* (C5) from bar 14 of 'Lascia ch'io pianga' from *Rinaldo* by Handel sung by a young professional soprano in the three styles A, B and C (see text).

Fig. 6 shows histograms of the output sound pressure amplitude for each type and it can be seen that type C is the quietest and that type A is the loudest, with around 17 dB difference between their modal values. The output for type B sits in between with its mode around 10 dB higher than that for type C and 7 dB lower than that for type A.. However, it should also be noted that the type A note is clearly bimodal; this occurs between the early part (soft) and the later part (loud) of the note (see Fig. 5).

IV. DISCUSSION

The differences observed between the three styles, as produced by the current singer, appear to reflect the observations made by Bethel [1] and the previous research conducted by Daffern [4]. Whilst Daffern found the vibrato rate varied within the two groups she compared, the extent to which the vibrato changed between styles A and B was not as great as Daffern found for the two groups overall. This is especially true of the peak-to-peak extent and the use of vibrato throughout tones as discussed above. This could be due to the training and age of the singer in this study, allowing for a freedom of vocal technique to execute the different styles but without the intensive opera training or regular specialised performance of one style, does not produce the characteristics to the same extent. This would also support the findings of Daffern which suggest that the early music singers and opera singers have vocal techniques which thrive in their own environment.

Larynx CQ values do vary between the three types, indicating that CQ is available to the singer for modification when singing in different styles This has been noted previously [6] for a professional tenor singing in three different styles: *opera*, *Elizabethan* and *conventional early music* for which the *opera* CQ values were higher than the *conventional early music* which in turn were higher than the *Elizabethan* style. These findings support those from this singer if the Elizabethan style can be approximately equated to the type C style herein.

V. CONCLUSION

The singer produced three differing singing styles which reflected the characteristics observed by Bethell [1], however styles A and B did not produce differences in vibrato as drastic as those found by Daffern [4] for professional singers specializing in those styles. In addition, the comparative distributions of CQ values between the three styles for this soprano confirm those found for a professional tenor by Howard [6].

Singing is a precious form of communication with many underlying facets. Understanding some of the subtleties of voice production strategies employed when singing in different styles will lead to a greater understanding and knowledge of the range of possible human vocal outputs and how best to achieve them in practice, whether pedagogically or clinically.

Knowledge of such differences has the potential to influence both performance practice and vocal coaching for both singing and speech. It has been shown that it lends itself well to implementation in real-time visual feedback systems for voice training such as WinSingad [7], SingandSee [8] and VoceVista [9]. In the future, there is the potential for enhancing such systems with additional displays, thus moving such work closer to being more "complete" for the professional voice user.

VI. Acknowledgements

REFERENCES

[1] Bethel, R. (2009), Preferred Vocal Emission for Handel's Arias: a case study, In: *Abstracts of NEMA International Conference*, 7 – 10 July, York, UK http://www.rma.ac.uk/news-and-events-html/Abstracts_for_NEMA_conference_at_York.pdf (accessed 16/09/09).

[2] Potter, J. Introduction: singing at the turn of the century, In: *The Cambridge companion to singing*, Potter, J. (Ed.), Cambridge: Cambridge University Press, 1-8, 2000.

[3] Howard, D.M. Variation of electrolaryngographically derived closed quotient for trained and untrained adult singers, *Journal of Voice*, **9**, (2), 163-172, 1995.

[4] Daffern, H. *Distinguishing characteristics of vocal techniques in the specialist performance of early music,* University of York: PhD Thesis, 2008.

[5] Howard, D.M. Electroglottography - electrolaryngography, In: *The Larynx*, 3rd Ed., Fried, M.P. and Ferlito, A. (Eds.), San Diego: Plural Press, 227-243, 2009.

[6] Howard, D.M. Quantifiable aspects of different singing styles – a case study, *Voice*, **1**, (1), 47-62, 1992.

[7] Howard, D.M., Welch, G.F., Brereton, J., Himonides, E., DeCosta, M., Williams, J. and Howard, A.W. WinSingad: A real-time display for the singing studio, *Logopedics Phoniatrics Vocology*, **29**, (3), 135-144, 2004.

[8] Nair, G. *Voice - Tradition and technology*, San Diego: Singular Publishing Company, 1999.

[9] Thorpe, C.W., Callaghan, J., and van Doorn, J. Visual feedback of acoustic voice features for the teaching of singing, *Australian Voice*, 5, 32-39, 1999.

VOCAL EFFORT IN SINGERS OF A NATIONAL LYRIC ORCHESTRA

R. Sisto[1], A.Pieroni[1], D.Annesi[1], P. Nataletti[1], F.Sanjust[1], C.Manfredi [2], M.Venzi[2]

[1]Department of Occupational Hygiene, ISPESL, Monteporzio Catone (ROME), Italy
[2]Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy

Abstract: **Scientific data in literature show that the singers of classical lyric orchestras are exposed to high risk of damage to the vocal apparatus due to the intense effort they have to face during the artistic performances.**

Vocal effort in a group of singers of a classical orchestra of a National lyric theatre is considered here. A specific protocol of measures has been defined with the aim of evaluating the quality of vocal emissions before and after the artistic performance during the rehearsal of a grand opera. Voice quality was parametrised in terms of average pitch value, quality ratio, vibrato frequency and extension.

A statistically significant difference was found between the quality ratio and the standard deviation of the fundamental frequency F0 and of the vibrato extension in the exercises executed before and after the vocal performance. These results confirm the hypothesis that such parameters are related to the laryngeal effort.

Keywords : **lyric singers, vocal effort, vocal quality**

I. INTRODUCTION

New protectionist laws impose to evaluate all different risk factors in the workplaces. All categories of workers are included and, among them, workers employed in recreational activities and shows should be considered. In this context, lyric national theatres are supplying to the assessment of the different risk factors to which their employees could be exposed.

Among such risk factors, noise and vocal effort are surely the most prominent. In addition, noise and vocal effort are often related to each other and both contribute to the injuries of the auditory and vocal apparatuses [1-3].

It is well know that professional singers are exposed to high vocal effort due to their performances. The stress to which the vocal apparatus is daily exposed can produce long-term effects ranging from the voice quality degradation to severe laryngeal pathologies. Previous scientific studies focused on the voice fatigue in singers and actors on the analogy of what was known about other workers categories exposed to vocal effort such as teachers [2,3]. The first studies lead on the teachers' vocal effort focused on the fundamental frequency (F0) analysis, on the phonation duration and on the emitted average sound pressure level at a certain distance during the working day [4].

Later, other parameters were specifically studied for singers to assess the vocal effort such as F0 variation, background noise, speech transmission index, signal to noise ratio, etc. [3-8]. These parameters were related to psychophysical evaluation subjectively reported by the subjects themselves [9,10].

The methodology of the vocal effort evaluation is based on the use of vocal dosimeters capable of registering the vocal emission during the whole working day [11,12].

The aim of this work was to individuate objective vocal parameters capable of an early detection of the voice quality degradation induced by the effort of the artistic performance. The main objective is to cast a non-invasive test to check the status of the vocal apparatus in workers exposed to vocal effort due to their working activity. Another main objective of the study is to understand the mechanism of the damage process with the aim of elaborating a prevention strategy.

II. METHODOS

Measurements were performed in the Teatro Regio in Turin during an experimental campaign finalized to the physical risks exposure evaluation in workplaces. Seven volunteer female lyric singers were enrolled into the present study: three Soprano, two Mezzo-Soprano, two Contralto. The singers were asked to execute some vocal exercises before and after the artistic performance during the rehearsal of a grand opera with the aim of comparing the voice quality before and after the vocal effort of a standard working day. Sound signals were recorded with a microphone and a sound analyzers Symphonie (01dB) in the rehearsal hall. Data were analyzed by means of the BioVoice software tool [13], that allows the extraction of vocal parameters also in singers. The emission quality was parametrized in terms of average fundamental frequency (F0) value, quality ratio, vibrato frequency and extension [13-16]. The following protocol was adapted from [10] :

Exercise n.1: Emit sustained \a\, \i\, \u\ vowels for 2 or 3 seconds with mild loudness and comfortable pitch (main tone of emission). Repeat 10 times without pauses, corresponding to a total time of about 30 s for each vowel.

Exercise n.2: Repeat exercise n.1 for the vowel \i\ with a very low sound intensity and moderately high pitch.

Exercise n.3: Emit vowel \i\ varying the main emission tone from low to high pitch at a low sound intensity

Exercise n.4: Emit 10 times 5 short-duration \i\ at low sound intensity and moderately acute pitch.

Exercise n.5: Repeat the first strophes of the song "Happy Birthday" at low sound intensity and acute pitch and fill out a questionnaire in which the difficulty in producing low intensity sounds had to be reported with scores from 1 (low difficulty) to 10 (high difficulty).

Exercise n.6: Count aloud from one to three (repeated three times). Fill out a questionnaire (same scores as in exercise 5) reporting the difficulty in producing high level sounds and the laryngeal perceived discomfort. The subjects were also asked to specify if the discomfort was perceived into the larynx, outside the larynx, or in both districts.

In this paper, we present preliminary results relative to the analysis of the vowel \a\ emission as in Exercise 1. Specifically, the first and the tenth emissions before and after the vocal performance during a chorus proof were analyzed. More results will be presented elsewhere.

The BioVoice tool was applied to objectively quantify voice quality. According to [15] the analyzed parameters are: F0 (pitch), vibrato rate (Vrate), vibrato extension (Vext), and the first five formants. Vrate and Vext represent respectively the number of oscillations per second and the oscillation amplitude of the pitch's modulation in time. The standard deviation (Std) of all parameters was also measured.

Moreover, the Singing Power Ratio (SPR) [15] was defined and measured. SPR is related to the energy content of the vocal formants, whose amplitude and frequency correspond to the resonant peaks of the power spectral density (PSD). In particular, in the singing voice, the SPR is defined as the ratio between the area under the curve of the PSD relative to the cluster of the first two formants ($Area_{1,2}$) and that of cluster of the third, fourth and fifth formants ($Area_{3,4,5}$):

$$SPR = \frac{Area_{1,2}}{Area_{3,4,5}} \qquad (1)$$

The better the singer voice quality, the more closely the SPR should approach the unit value. In fact, in this case, the singer voice can be clearly distinguishable from the background orchestra.

A major difficulty in the SPR measure has been finding a reference "threshold frequency" that cuts the PSD integral into $Area_{1,2}$ and $Area_{3,4,5}$. Both a "static threshold" S0, set at 2500 Hz (i.e. midpoint between 2000 and 3000 Hz, approximately representing the second and the third formant respectively), and two "dynamic thresholds", $F_{ref}1$ and $F_{ref}2$, have been defined and tested. $F_{ref}1$ corresponds to the local minimum of the PSD in the range 2 - 3 kHz, while $F_{ref}2$ to the mean frequency value of the second and the third formant. Both dynamic thresholds gave approximately the same results, while S0 gave worse results. In this paper, $F_{ref}2$ has been applied and is named here $F_{ref}inf$. Finally, an upper threshold

$F_{ref}sup$ has been introduced, corresponding to the first frequency minimum found after the 5th formant.

III. RESULTS AND DISCUSSION

Some figures, relative to a soprano singer, are reported here, as they are illustrative of a common behavior found in all cases. Fig. 1 shows the evolution in time of F0 that appears more unstable and irregular after the vocal performance as a consequence of the vocal effort.

In Fig. 2 the time evolution of vibrato is shown: the frequency modulation in time loses its sinusoidal behaviour after the vocal effort due to the performance. Along with the vibrato distortion, also the vocal intonation deteriorates and its time behaviour appears unstable.
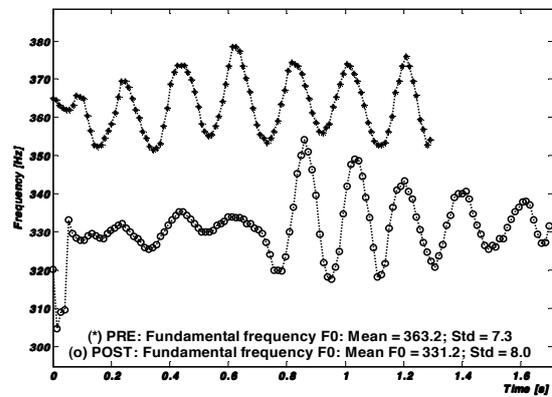
Figure 1- Pre-post performance F0 tracking

Moreover, the vocal effort causes a deterioraration in the SPR, which shows an increasing trend with the phonation fatigue. Fig. 3 shows the PSD before (grey) and after (black) the vocal performance, with SPR=3.9 and 15.3 respectively. In Fig. 3, dots correspond to the PSD maxima and stars to $F_{ref}inf$, $F_{ref}sup$ as obtained with BioVoice.

Finally, Figs. 4 and 5 show the signal spectrogram respectively before and after the vocal performance, pointing out a more regular behaviour of both harmonics and formants before the performance.

Though we analysed few cases, a statistical analysis was performed to find out possible significant differences between data before and after the vocal effort. Data were analyzed by means of a standard Student's t-test (significance criterion $p<0.05$) for paired samples to find statistically significant differences between voice quality parameters before and after the vocal performance. In particular, the first and the tenth vowel \a\ emissions before the performance were compared respectively to the first and the tenth emissions after the vocal effort. A mean emission was defined as the average between the first and the tenth emission. The mean emission characteristics before the vocal performance were

compared to the characteristics of the mean emission after the vocal effort. As data distributions were found not normal, the non parametric Wilcoxon rank test was also applied. Results are shown in Table I.
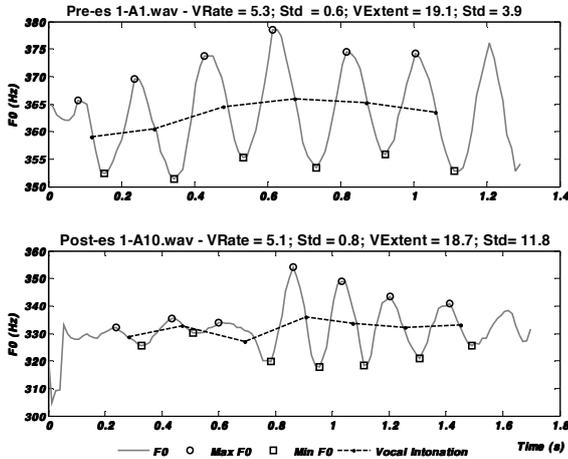
Figure 2: comparison between the vibrato before (first emission) and after (tenth emission) the vocal effort for a singer. Dots and squares correspond to estimated maximum and minimum F0 values, respectively.
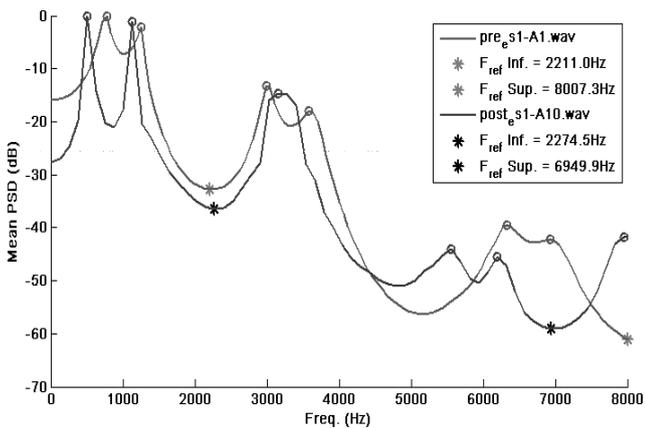
Figure 3: comparison between the PSD before (grey line) and after (black line) the vocal effort.

From the Table, the Std of F0,Vrate and Vext appears sensitive to the exposure to phonation fatigue, as all parameters show an increasing trend. Although the difference between F0 mean values before and after the vocal effort does not give statistical significance, F0 shows an increasing trend due to the exposure.

The parameter SPR seems to be one of the most sensitive to the exposure to the vocal effort. The differences between the SPR before and after the performance are in fact always statistically significant if the first, the tenth or the average of the two last emissions are considered. The statistical distribution of the SPR (mean value) before and after the performance is shown in Fig. 5.
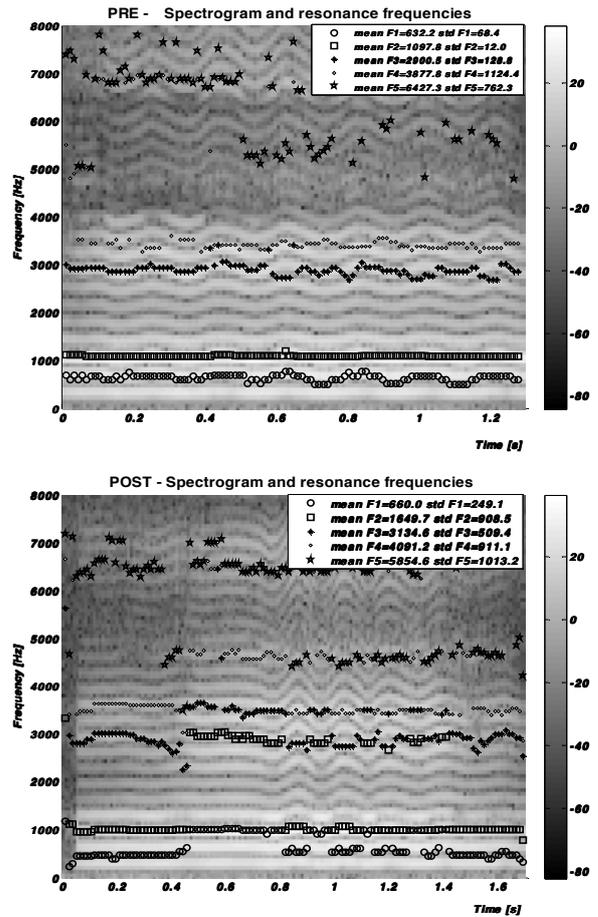
Figure 4: Spectrogram before (top) and after (bottom) the vocal performance.

As expected the parameter SPR, representing voice quality, deteriorates (increases) after a laryngeal sustained effort. Finally, notice that the parameters show a statistically significant difference not only between the exercises executed before and after the artistic proof but also between the first and the tenth vocal emission.

IV. CONCLUSION

Some of the voice parameters studied in this work before and after the artistic performance during the rehearsal of a grand opera seem to be sensitive to the vocal effort of a typical working day. In particular, statistically significant differences were found between the Std of F0, Vrate and Vext, before and after the artistic performance. Another sensitive parameter is SPR, specifically implemented in the BioVoice tool to define the quality of sung voice. Future work will be devoted to enlarge the data set for a better statistical analysis. Our results, if confirmed, could in fact be useful to define an effective protocol for monitoring long-term adverse effects of the vocal effort in exposed populations.

Table I: Student's t-test and Wilcoxon non-parametric rank test comparison between voice parameters measured before and after the vocal effort.

T-TEST	pre -post 1st emission	pre -post 10th emission	pre-post mean
F0	n.s.	n.s.	n.s.
Std F0	n.s.	0.037	0.055
SPR	n.s.	0.00044	0.00234
Vrate	n.s.	n.s.	n.s.
Std Vrate	0.01103	0.00669	0.00306
Vext	n.s.	n.s.	n.s.
Std Vext	0.02761	n.s.	n.s.
WILCOXON	pre -post 1st emission	pre -post 10th emission	pre-post mean
F0	n.s.	n.s.	n.s.
Std F0	n.s.	0.01563	0.01563
SPR	0.03125	0.01563	0.01563
Vrate	n.s.	n.s.	n.s.
Std Vrate	0.01563	0.03552	0.01991
Vext	n.s.	n.s.	n.s.
Std Vext	0.01563	0.07813	0.03125
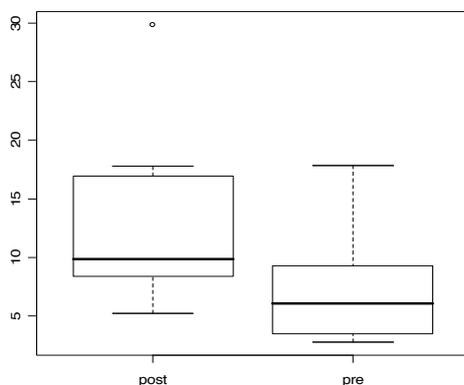
Figure 5: boxplot showing the mean SPR before (pre) and after (post) the vocal effort.

REFERENCES

[1] Vilkman E., "Occupational safety and health aspects of voice and speech professions." *Folia Phoniatr. Logop.* 56:220 –53, 2004.

[2] Titze IR., McCabe DJ., "Chant therapy for treating vocal fatigue among public school teachers: a preliminary study." *Am. J. Speech. Lang. Pathol.* 2002;11:356–69.

[3] Sala E., Airo E., Olkinuora P., et al. "Vocal loading among day care center teachers." *Logop. Phon. Vocol.* 27:21– 8, 2002.

[4] Sodersten M., Granqvist S., Hammarberg B. et al., "Vocal behavior and vocal loading factors for preschool teachers at work studied with binaural DAT recordings." *J. Voice* 16(3):356 –71, 202.

[5] Stemple JC., Stanley J., Lee L., "Objective measures of voice production in normal subjects following prolonged voice use." *J. Voice* 9(2):127–33, 1995.

[6] Sangiorgi T., Manfredi C., Bruscaglioni P., "Objective Analysis of the Singing Voice as a Training Aid", *Logop. Phon. Vocol.*, 30: 136-146, 2005.

[7] Manfredi C., Bocchi L., Calisti M., Vanello N., Cantarella G., Mazzei L., Sardi M., "Training of The Singing Voice: a Multimodal Feature Extraction Approach", *PEVOC8 Conf.*, 24-28 August, 2009, Dresden, DE.

[8] Szabo A., Hammarberg B., Hakansson A. et al., "A voice accumulator device: evaluation based on studio and field recordings." *Logop. Phon. Vocol* 26:102–17, 2001.

[9] Kitch JA., Oates J., "The perceptual features of vocal fatigue as self reported by a group of actors and singers." *J. Voice* 8(3):207–14, 1994.

[10] Sapir S., Mathers-Schmidt B., Larson GW., "Singers' and non-singers vocal health, vocal behaviors, and attitudes toward voice and singing: indirect findings from a questionnaire", *Eur.J.Disord.Comm.* 31:193–209, 1996.

[11] Carroll T., Nix J., Hunter E., Emerich K., Titze I., Abaza M., "Objective measurement of vocal fatigue in classical singers: A vocal dosimetry pilot study," *Otolaryng.Head &Neck Surg. ,* 135, 595-602, 2006.

[12] Titze IR., Svec JG., Popolo PS., "Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues." *J.Speech Lang.Hear.Res.* 46:922–35, 2003.

[13] Manfredi C., Bocchi L., Cantarella G.,"A Multipurpose User-Friendly Tool for Voice Analysis: Application to Pathological Adult Voices", *Biom. Signal Proc.& Control*, vol.4, p. 212–220, 2009.

[14] Shipp T, Leanderson R., Sundberg J., "Some acoustic characteristic of vocal vibrato." *J. Res. in Singing*, 4:18-25, 1980.

[15] Sundberg J., *The Science of Singing Voice*. DeKalb, Illinois: North. Illinois Univ. Press, 1987.

[16] Horward D.M., Welch G.F.,Brereton J., Himonides E., DeCosta M., Williams J., Horward AW. WinSingad, "A real-time display for the singing studio." *Logop. Phoniatr. Vocol.* 2004; 3: 135-144.

Obstructive Sleep Apnoea

AUTOMATIC DETECTION OF OBSTRUCTIVE SLEEP APNEA SYNDROME BASED ON SNORE SIGNALS

Barbara Calabrese[1], Franco Pucci[2], Miriam Sturniolo[3], Pierangelo Veltri[1], Antonio Gambardella[2,3], Mario Cannataro[1]

[1]Bioinformatics Laboratory, University Magna Graecia of Catanzaro, Italy
[2]Institute of Neurology, University Magna Graecia of Catanzaro, Italy
[3]Institute of Neurological Sciences- National Research Council (ISN-CNR), Italy

Abstract: **Obstructive sleep apnea syndrome (OSAS) is a human disease affecting the human breathing of a patient while sleeping. To be studied, a patient has to be screened while sleeping, thus diagnosis is often hard and costly. Polysomnography is the standard method for obstructive sleep apnea diagnosis. However it does not permit a mass screening of patients because it has high cost and requires long term monitoring. Different efforts are reported in literature for finding new diagnostic methods implemented on portable devices. This paper presents a preliminary study for the development of a portable system based on snore signals acquisition and spectral analysis for OSAS identification.**
Keywords : **Home monitoring, OSAS, Snore analysis.**

I. INTRODUCTION

Sleep apnea is a common disorder that affects both children and adults. An obstructive sleep apnea syndrome (OSAS) is defined as a complete cessation of airflow for more than 10 seconds which requires a significant respiratory effort to restart normal respiration. It requires immediate intervention to prevent it from becoming life-threatening [1].

This disease affects a significant percentage of the adult population which varies according to several studies by 15% to 35% in men and from 5% to 20% in women. The most obvious complications arising from OSAS are diminished quality of life brought on by chronic sleep deprivation and cardiovascular problems.

Currently, the 'gold' standard method for diagnosing OSAS is polysomnography (PSG) [1]. This diagnostic exam requires that the patients spend a full-night in hospital. Thus it is time consuming and high costly, because usually it is possible to monitor one patient for night for instrument. Furthermore it is labour intensive because the clinicians need to collect and analyze a large number of data (e.g. different and large signals, such as EEG, ECG, oxymetry, EMG, thoracic-abdominal movements).

Efforts are being directed to the identification of alternative methods for OSAS diagnosis to permit clinicians to detect automatically and objectively OSAS events saving time and work. Snore signals have been investigated as an alternate diagnostic tool for the detection of obstructive sleep apnea [1].

This work reports current approaches in OSAS diagnosis based on the analysis of snore signals and outlines a possible approach for developing a system for the automatic acquisition, analysis and classification of snore signals.

Although there exist some approaches to detect snore signals, used to discriminate OSAS patients from snorers, we are not aware of portable devices able to automatically detect apnea events and to discriminate their different sub-types (e.g. central, peripheral and mixed apnea).

The development of such a portable system would allow the diagnosis of OSAS events without using PSG instruments. Thus, our proposed system could be used for home-monitoring of suspected patients who turn to doctors accusing specific symptoms.

To detect apnea events, the proposed system collects only snoring signals, analyzing and classifying them to discriminate between simple snorers and OSAS patients. So the clinician must not examine a full night acquisition but only the portions of signal characterized by apnea events.

The rest of the paper is organized as follows. Section II describes current approaches for OSAS diagnosis focusing on the signal processing techniques. Section III presents a novel approach for the analysis of snore signals. Section IV outlines a possible procedure to analyse data. Finally, section V concludes the paper and outlines future work.

II. METHODS

The polysomnography [2] is a functional exam that permits the monitoring of different biological activities. Numerous physiological sensors are attached to the patient to record night-time breathing, brain activity and physical activity. Although the PSG is the standard approach for OSAS diagnosis, it requires technical expertise and is labour-intensive and time-consuming. Timely access is a problem for many patients, the majority of whom continue

to have undiagnosed sleep apnea. Thus, alternative approaches to diagnosis, such as portable monitoring, have been proposed as a substitute for polysomnography in the diagnostic assessment of patients with suspected sleep apnea.

Different portable PSGs are used in clinical practice as a first level of screening of OSAS [3]. Although portable PSG allows the home monitoring of patients, it is an invasive technique and the patient remains connected to a lot of sensors. Moreover, as standard PSG, also portable PSG produces a lot of data, which is inefficient to analyze if one relies on manual processing.

Different efforts are underway to find better methods for diagnosing or screening of OSAS. Current alternative methods to PSG are: overnight oximetry, which measures a patient's oxygen saturations throughout the night, ECG or snore monitoring. Overnight oximetry is not considered completely adequate as a screening test, since the oxygen levels in the blood of many patients with OSAS do not provide the information needed to understand their condition. Thus, there is a growing interest in developing portable snore-based devices for OSAS monitoring.

In [5] the development of a portable device for home monitoring of snoring is described. It performs detection and selection of the snores, while discarding any other events that are present in the sound recording, as cough, voice, and other artefacts. The device performs temporal analysis of signals. It detects snore events by evaluating signal amplitude and detects possible apnea events by measuring the delay between snores.

Another portable device for snore detection is described in [6]. The device itself also serves as a Web server. Doctors and caregivers can access real-time and historical data via a Microsoft Internet Explorer browser or a remote application program for telemonitoring of snoring and OSAS symptoms.

Both systems are able to detect only snore events through time analysis and they do not reach high success rate and sensitivity. They do not exploit frequency-based and time-frequency-based analysis.

For the detection of OSAS events, the analysis of snoring signals has been performed in time or in frequency domain [7, 8]. In the time domain the evaluated parameters are duration of snores, mean value/standard deviation of pitch and max/average intensity sound. In the frequency domain the parameters of interest are fundamental frequency, formants, median frequency, central frequency and max frequency. The spectral parameters are extracted from the power spectrum that is evaluated by parametric (AR model) or non parametric methods (FFT, Welch periodogram) [8].

For the discrimination between simple snorers and OSAS patients, it has been reported some variability between frequency parameters from simple snorers and OSAS patients [9]. This variability is evident not only in the segments after apnea event but also in all the snores of OSAS patients. It has been, also, reported variability relative to formants. Formants estimated for snoring signals coming from simple snorers show lower variance than those coming from OSAS patients [10].

From a biomechanical point of view, snoring sounds are caused by many factors: the strength of respiratory-related airflow, vibrations on the soft palate, the shape of upper airway, and the airway obstruction due to tongue subsidence. Moreover vibration parts are not held by some cartilage or bones. Thus in [11] the authors suggest to consider that snoring sounds are nonlinear acoustic vibrations caused by various factors. That makes it difficult to solve the unique eigenfrequency of snoring sounds by a traditional linear frequency analysis, generally adopted in [7, 8]. If such nonlinear properties can be extracted by some other methods and their relation to some degree of OSAS syndrome (e.g. Apnea/Hypopnea Index) is demonstrated, it would be possible to establish a new screening method which replaces the costly PSG.

III. RESULTS

Taking into account the results and approaches available in literature, we propose a novel signal processing workflow to analyze snore signals and outline the design of a portable device for snore analysis and OSAS diagnosis.

The development of a snore-based OSAS detector (Fig. 1) requires a good design of the acquisition stage because the snoring signal acquisition is affected by several problems. Different types of noise can contaminate the signals, such as background acoustical noise or electromagnetic interferences [11]. Although the use of unidirectional microphone can improve signal acquisition, however, noise reduction is needed to eliminate interferences. Therefore to build a reliable system, a robust pre-processing stage before signal analysis to improve signal to noise ratio and to allow a more accurate extraction of features is needed.

Fig. 1: Architecture of the system

To identify the occurrence of obstructive sleep apnea, as discussed previously, different features can be extracted from time and frequency domain. We have chosen to identify unambiguously apnea events trough frequency analysis of post-apneic snore events.

In the following we report some first experimental results related to the analysis of a vocal signal acquired, in the polysomnographic laboratory of the Institute of Neurology at University Magna Graecia of Catanzaro, from a patient affected by a moderate sleep apnea.

The vocal signal has been acquired with a digital audio recorder (Micro Track II Professional Audio Recorder) able to register mono and stereo signals at different acquisition rates. In this experiment, a signal long about 1 hour at 44 KHz has been recorded.

In this initial stage of the analysis, we have separated in a manual way the snoring events from the respiratory ones In this phase the separation between snoring and respiratory events has been performed with the help of doctors.

After this separation we have performed FFT analysis and power spectra evaluation on the selected portion of original signals. Figure 2 and Figure 3 reports, respectively, the power spectrum resulting from the analysis of snoring signals acquired from a patient affected by a moderate sleep apnea. In particular, the first plot is the power spectrum of a (generic) snore, whereas the second one represents the power spectrum of a post-apneic snore. The two spectra show significant differences because the post-apneic spectrum presents a larger number of frequency components at higher frequencies than the first one.
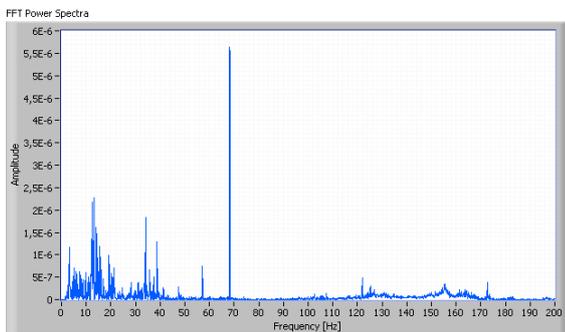
Fig. 2: Power spectrum of a generic snore
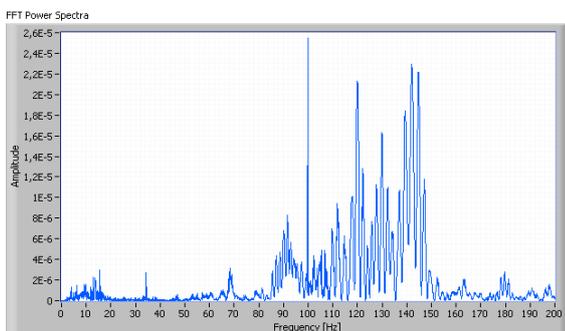
Fig. 3: Power spectrum of a post-apneic snore

IV. DISCUSSION

The differences between the power spectra of regular and post-apnea snores suggests us a possible method to differentiate snores from post-apnea snores by comparing in both spectra the number of frequency components above a certain power threshold.

The procedure to analyse the recorded signal has to extract snores, verify if they are post-apnea snores, then extract the characteristics of apnea events happening before such snores. The characteristics of such apnea events (e.g. number of events, duration, etc.) can be evaluated by the doctors to help the diagnosis of OSAS.

The signal analysis can be implemented by using the following procedure described in pseudo code:

```
PROCEDURE snoreAnalysis (VoiceSignal S)
BEGIN
    //preprocess S to increase signal-to-noise ratio;
    S.Preprocess ();

    //S_i is a snore identified in S
    Snore S_i;

    //apneaEvents is a list of apnea events in S
    apneaEventsList [] apneaEvents;

    WHILE NOT S.end_of_signal DO {
        // extract next snore S_i from the signal S;
        S_i = S.nextSnore ();
        IF (S_i.getSnoreType == post_apnea)
        THEN {
            // S_i is a post-apnea snore type
            //extract from S the apnea event A_i related to S_i
            A_i = get_apnea_event(S, S_i);
            // append A_i to a list of apnea events
            apneaEvents.Append();
        }
        ELSE {//skip S_i}
    };
    //analyse the Apnea events
    AnalyzeApneaEvents (apneaEvents);
END.
```

V. CONCLUSION

The paper presented a first approach for the automatic detection and characterization of snore signals related to the Obstructive Sleep Apnea Syndrome.

The proposed system is currently under development and a first prototype will be tested in the polysomnographic laboratory of the Institute of Neurology at University Magna Graecia of Catanzaro.

188

A first goal of the system is to extract from the signal only post-apnea snore events and then apnea events. The reduced signal will be validated by the doctors in a manual way avoiding the examination of all signal registrations as happens in present setting.

A set of signals registered in the Institute of Neurology will be used to further investigate and eventually validate the discriminatory characteristics of the power spectra of apnea and post-apnea snores with respect to respiratory and snoring events. After validation, a next step will regard the automatic classification of sleep apnea.

REFERENCES

[1] American Academy of Sleep Medicine, www.aasmnet.org/.
[2] Ferini Strambi L. et al., Linee Guida di Procedura Diagnostica nella Sindrome delle Apnee Ostruttive nel Sonno dell'Adulto, www.sonnomed.it.
[3] Flemons Ward W. et al., Home diagnosis of sleep apnea: a systematic review of the literature, Chest, col. 124, pp. 1543-1579, 2003.
[4] Stoohs R., Guilleminault C., MESAM 4: an ambulatory device for the detection of patients at risk for obstructive sleep apnea syndrome (OSAS), Chest, vol. 101, pp. 1221-1227, 1992.
[5] Y. L. Hsu, M. C. Chen, C. M. Cheng andC. H. Wu, Development of a portable device for home monitoring of snoring, IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2420 – 2424, 2005.
[6] C. M. Cheng, Y. L. Hsu, C. M. Young and C. H. Wu, Development of a portable device for telemonitoring of snoring and obstructive sleep apnea syndrome symptoms, Telemedicine Journal and E-Health , vol. 14, no. 1, pp. 55-68, 2008.
[7] J. Sola-Soler, R. Jane, J.A. Fiz and J. Morera, Pitch analysis in snoring signals from simple snorers and patients with obstructive sleep apnea, Proceedings of the Second Joint EMBS/BMES Conference, vol. 2, pp. 1527 – 1528, 2002.
[8] J. Sola-Soler, R. Jane, J.A. Fiz and J. Morera, Spectral envelope analysis in snoring signals from simple snorers and patients with Obstructive Sleep Apnea, Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol.3, pp. 2527 - 2530, 2003.
[9] J. Sola-Soler, R. Jane, J.A. Fiz and J. Morera, Variability of snore parameters in time and frequency domains in snoring subjects with and without Obstructive Sleep Apnea, 27th Annual International Conference of the Engineering in Medicine and Biology Society, pp. 2583 – 2586, 2006.
[10] Jane, J. Sola-Soler, J.A. Fiz, Morera, J., Automatic detection of snoring signals: validation with simple snorers and OSAS patients, Proceedings of the 22nd Annual International Conference of the IEEE Volume 4, 23-28 July 2000 Page(s):3129 – 3131.
[11] T. Mikami, Detecting Nonlinear Properties of Snoring Sounds for Sleep Apnea Diagnosis, The 2nd International Conference on Bioinformatics and Biomedical Engineering, pp.1173 – 1176, 2008.

AUTOMATIC DETECTION OF POST-APNOEIC SNORE EVENTS FROM HOME AND CLINICAL FULL NIGHT SLEEP RECORDINGS

M. Calisti [a], L. Bocchi [a], C. Manfredi [a], I. Romagnoli [b], F. Gigliotti [b], G. Donzelli [c]

[a] Department of Electronics and Telecommunications, Università degli Studi di Firenze, V.S.Marta 3, 50139 Firenze, Italy
[b] Fondazione Don C. Gnocchi, Via Imprunetana 124, 50023, Firenze, Italy
[c] Departments of Paediatrics, Università degli Studi di Firenze,V.le Pieraccini 24, 50139 Firenze Italy

Abstract: **Snoring is the hallmark of the Obstructive Sleep Apnoea Syndrome and several studies explore possible correlations between them. In this work an improved methodology with respect to [4] is proposed, based on a proper energy threshold applied on audio recordings for sound/silence detection, and on a feature vector of 14 elements (13 Mel Frequency Cepstral Coefficient plus the number of zero crossings) for sound classification. This feature vector is obtained from a 62-elements one by applying a genetic algorithm, fitted to obtain the best classification of the training/validation sets.**
The feature vector is analyzed by means of a radial basis neural network to perform snore events identification. Finally, formant frequencies and time analysis are also investigated to split up post-apnoeic snores and normal ones.
Audio data from 26 patients of different age and sex are used to test the methodology: 6 patients (3 male and 3 female) were used to train the nets (1800 snores) and 4 patients to validate the classification (600 snores). On the whole dataset of patients, a sensitivity between 69% and 84% is obtained in the detection of post-apnoeic snores.

Keywords: **Snore, neural network, Mel frequency cepstral coefficients, genetic algorithm, obstructive sleep apnoea.**

I. INTRODUCTION

Obstructive Sleep Apnoea (OSA) is a pathological condition where the upper airways collapse, reducing or cutting the flow to the mouth/nose. The diagnosis of Obstructive Sleep Apnoea Syndrome (OSAS) is commonly made by means of Polysomnographic (PSG) examination. PSG is mainly performed in a clinical environment (sleep laboratories), but could also be performed in home environment. However, PSG examination is bothering for patient, unsuited for mass screening purposes and expensive. Hence, new, simpler and non-invasive methods are investigated to detect OSAS. At present, according to the Italian guidelines, OSA is detected from full-night sleep analysis (uninterrupted recordings lasting from 6 to 10 hours) by means of PSG. Such a huge amount of data implies several technical problems concerning acquisition, storage and processing of data. Hence, efforts are made in the scientific community to define reliable OSA identification techniques from the audio signal only. At present, processing is made over the whole signal that is commonly classified into three classes: snore, breath, silence [1] or five classes: snore, breath, silence, duvet noise, other noise [2]. Other works consider just temporal features [3].
In this work we propose an automatic detection of snore events, that extends the results obtained in [4], followed by an evaluation of the number of Apnoeas or Hypopnoea events (AHI Index) with the methodology proposed in [5]. Our approach allows to split-up snores from other sounds, without predefining other sound classes, thus reducing the total length of the signal to be processed. The method is developed under Matlab2007a®. Full-night audio data (26 patients) come both from clinical and home recordings.

II. METHOD

The flow chart proposed in [4] is revised here, with the aim of performing a faster analysis and a more careful sound/silence segmentation. Firstly, we evaluate the histogram of the audio signal energy to perform an Otsu thresholding [6]. This method has the advantage that it does not require data pre-filtering, as a good energy separation between sound and silence is expected from our recordings [1], even in home environment. This assumption has been verified with a careful setup of the process, both as far as the device and the environmental setup are concerned. Specifically, a unidirectional microphone has been used to perform recordings connected to an external sound card to reduce noise artefacts of the laptop sound card. Patients were separated from bed partner and/or from pets, television and other predictable sources of noise. Moreover, the first 30 minutes of each recording were cut off, to avoid noise due to patient's movements, speaking with the clinician and similar ones. After the selection of sound events, a proper classification is proposed based on features extracted from 60 Mel Frequency Cepstral

Coefficients (MFCC) plus short term energy (STE) and the number of zero crossing (NZC), where:

$$STE = \log\left(\frac{\sum_{i=1}^{n} s(i)^2}{n}\right) + k \qquad (1)$$

$$NZC = \frac{\sum_{i=1}^{n-1}\left|sign\big(s(i+1)\big) - sign(s(i))\right|}{2} \qquad (2)$$

Where $n=441$ is the number of elements in each window, sign() is the sign function, s is the signal and k is a small constant value to avoid log(0). Mean and standard deviation of the MFCCs are obtained as in [4].

As the choice of the number of coefficients is often arbitrary or derived from boundary conditions, we performed a careful search of the most representative MFCCs by means of a genetic algorithm (GA), where each gene of a phenotype represents a MFCC. The population of feature vectors was processed by the neural network until we obtained the best fitting according to proper classification. Furthermore, after low-pass filtering the audio signal (2 kHz cut-off), the number of zero crossings has been used as a selection feature for snore/non-snore events.

Several methods are proposed in literature to separate OSA events from non-OSA ones. Here we adapted the one proposed in [5] with the aim of identifying snore episodes after apnoea ones. This allows obtaining an AHI index related to apnoeic events only. A detailed flow chart of the process for the best feature vector selection is reported in Fig.1.

Short term energy, number of zero crossing and MFCCs extraction from the signal are performed according to [4]. Mean (m) and standard deviation (std) of the MFCCs for all frames of an event are also evaluated.

To detect starting and ending points of the event, the Otsu methodology [6] was iteratively applied to obtain two thresholds, the upper one and the lower one. The histogram was settled up to 2000 levels. After a first upper threshold detection t_u, a second Otsu thresholding was performed from level zero to level t_u, to obtain a lower threshold t_l. When the STE of the signal overpasses t_u, a starting point is detected, when the STE of the signal falls down t_l, the ending point of the event is found. As in [4] this procedure allows to obtain two sets representing the starting and the ending points of the events. Filtering only these events instead of the whole signal greatly speeds up the signal processing.

Once we have obtained all the events from the recording, we listened and classified the various frames as snoring or non-snoring frames to prepare a training set. We classified about 600 events from 6 patients (3 male and 3 female) without regarding the prevalence of the pathology, for a total of about 1800 snoring frames and 1500 non-snoring frames.

At present, most of the approaches try to classify snoring and "other events", e.g. mainly breath. However different noise events are included in "other events" that are difficult to classify. Hence a different approach is presented here, where we train the net with feature vectors representing only snore.
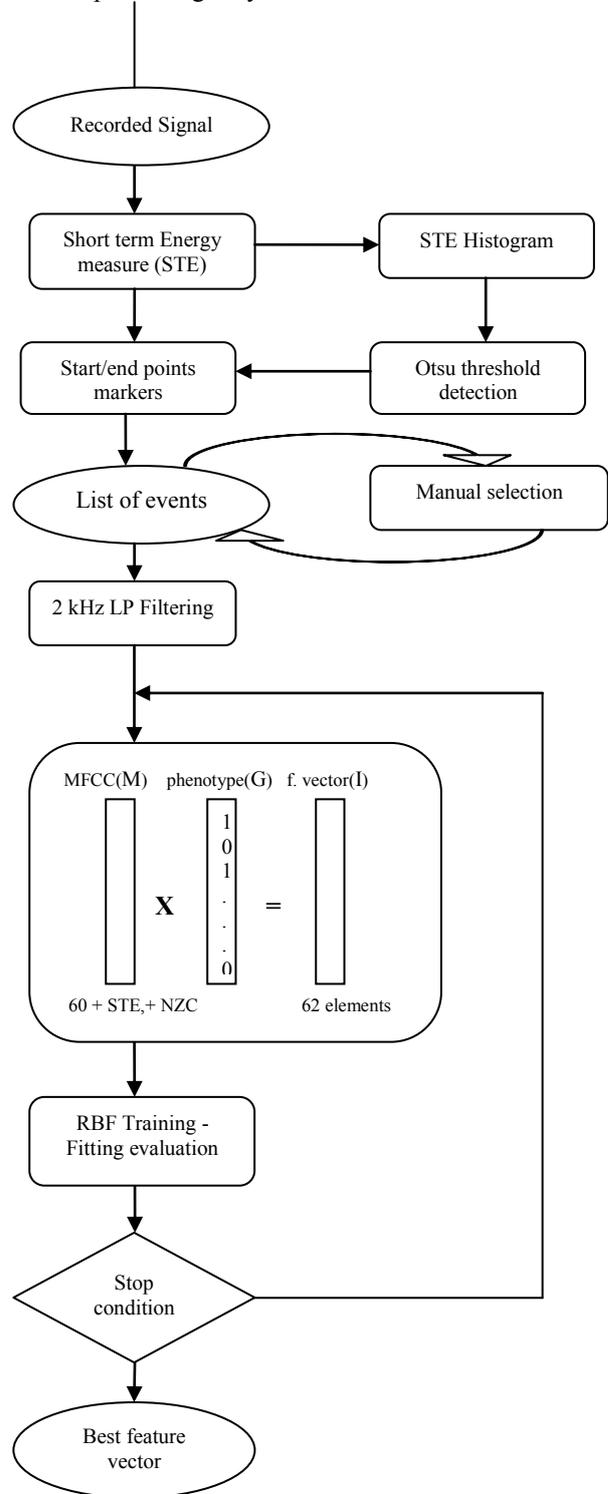
Figure 1. Flow chart of the sound detection and snore identification

Hence, if a unclassified feature vector (named here *void vector*) is presented to the net, according to the similarity of this new feature vector to the ones presented to the net during the training step, snore frames can be separated from non-snore ones.

In our opinion, this approach is more general as it is not proper to assume that cough, bed noise, breath and similar sounds belong to the same class, as was proposed in our previous work [4], though quite good results were obtained.

The improvement proposed here is based on a Radial Basis Neural Network (RBNN) that provides, as output of the hidden layer, a vector D representing the distance between the input vector (our feature vector of 62 elements) and the input weight matrix. For the radial basis neuron, the output is equal to 1 when the distance between the weight vector and the input vector is 0. Hence, the maximum value of D points out if the feature vector represents a snore or not.

All the 1800 frames representing snores are used as training set. Then, presenting several input vectors to the net, and taking into account the maximum of the output vector, 600 frames not already presented to the net are used to evaluate the network. The number of correct classifications over the validation set of 600 frames represents the value of our fitting function:

$$fit = -\frac{TP+TN}{600} \qquad (3)$$

Where TP=true positive (a snore correctly recognized as a snore), TN=true negative (a non-snore correctly recognized as a non-snore). The minus sign is due to the fact that the most common genetic algorithm tools aim to minimize the fitting function.

To perform the GA, the input vector I to the net was obtained as the product between the vector M of 62 elements (30 mean values and 30 std of the MFCCs plus STE and NZC) and a binary feature vector called phenotype G that represent the elements of the vector M that will belong to the input vector I or not. The various individuals of the population for the GA are different Gs with different combinations. The stop condition was set at 30 minutes of elaboration. The whole process is shown in Fig.1.

After GA optimization, the resulting best feature vector was used to train an Optimized Radial Basis Neural Network (ORBN) and to test the net on the whole database of patients. Here "optimized net" means a net trained with the optimized input set.

The snore event recognized as snore is than processed to identify the post-apnoeic snore event, according to [5]. Also, a temporal feature is taken into account, based on the assumption that at least 10s of silence should exist before the snore to satisfy the apnoea definition (air flow absence lasting 10s at least [7]).

III. RESULTS

Experiments were carried on under the same conditions as in [4], and with the same equipment. Mainly three blocks of the chain in Fig.1 affect our results: the sound detection from the whole recording; the snores recognition from the sound; the OSA-snores recognition from the snores.

The first step, mainly related to the reduction in time of the whole recording, gives good results. As an example, results for 4 subjects are shown in Table 1.

Table1.Examples of reduction in time of the recordings.

Patient	Time of whole recording (min)	Time of whole events (min)
Subject 1	572	37
Subject 2	446	47
Subject 3	592	70
Subject 4	476	31

The accuracy of this step, evaluated as the number of sounds detected over the total number of sounds, is about 96,65% (ranging from 93% to 99%). This accuracy was computed by listening to about 1 hour of recording for 10 patients.

The second step was validated by listening to 50 events extracted from 6 different patients. As in [4], an event is classified as snore if there is at least one frame recognized as snore in the whole event. The sensitivity, measured as TP/(TP+FN) varied from 87,1% to 97,82% with a good improvement with respect to [4]. Here FN (false negative), represents a snore wrongly recognized as non-snore.

The best phenotype was obtained running five times the GA, with the stop condition of 30min running, but in all cases the problem was optimized after 10 generations. From the five best phenotypes obtained, only the elements common to all of them were used, thus discarding 6 elements. Thus, the best phenotype is composed by 14 elements from the 62 of the original one, as shown in Table 2.

Notice that the OSA evaluation was carried on offline after the automatic snore extraction. Only the snores that follow a silence longer than 10s were analyzed.

Finally, we extended what suggested in [5], considering as apnoeic snores only the snores occurring after an apnoea event. In this way, we notice a little increasing of the post apneic snore formant frequencies.

Taking into account the difference on formant frequencies and the temporal consideration regarding a 10s silence before sound, we obtained a sensitivity varying from 85% to 87%.

Table 2. Best phenotype obtained from GA

Element of I	Element of M	Meaning
1	M(1)	Mean of 1° MFCC
2	M(4)	Std of 2° MFCC
3	M(5)	Mean of 3° MFCC
4	M(7)	Mean of 4° MFCC
5	M(14)	Std of 7° MFCC
6	M(15)	Mean of 8° MFCC
7	M(23)	Mean of 12° MFCC
8	M(24)	Std of 12° MFCC
9	M(30)	Std of 15° MFCC
10	M(37)	Mean of 19° MFCC
11	M(42)	Std of 21° MFCC
12	M(54)	Std of 27 ° MFCC
13	M(56)	Std of 28° MFCC
14	M(61)	Number of Zero Crossing

IV. DISCUSSION

The proposed sound/silence detection algorithm mainly fails with low intensity snores, as such events have not enough energy to be classified as sound signals by the Otsu methodology. However, as post apnoeic snore events are more intense than non-post apnoeic ones, this limitation could be acceptable. Moreover, the Otsu thresholding fails if very few snore events are present in the recording. Specifically in 2 cases out of the 26 analyzed, manual thresholding was required, as the patient snored few times as compared to the length of the whole recording. In this case, thresholds were not coherent with the sound. This happened for one laboratory recording where some devices added a continuous noise during the night and for one home recording, where the patient snored few times over the whole recording (about 6 minutes out of 7 hours of recording). However, as Table 1 points out, the reduction in time could be relevant. Hence further analysis is required to overcome these limitations and possibly define a time threshold that points out if the recording is acceptable or not.

The sensitivity of the ORBN was really good, achieving in some case the 98% of recognition. The large variety of different kind of snores does not allow for a perfect recognition, but these first results seem quite good also as compared to existing literature.

At the end of the whole chain, the post apnoeic snore recognition varies from 69% to 84%, using the approach in [5]. Actually, sound detection and sound classification are hold on in automatic way, while the post apnoeic snore is analyzed offline, with a methodology not yet implemented in the algorithm.

V. CONCLUSION

We provide a full automatic highly sensitive system for snore identification during sleep that takes into account aspects of the problem not considered in other approaches. The search of the most meaningful features that identify the snore from other sounds could be further explored to provide a link between snoring arousal and other sound features.

A post apnoeic classification provides a first attempt to validate the system from data recordings for syndrome evaluation. However, we point out two weaknesses: first, the non automatic performance of the post apnoeic identification step and second, the used approach that does not perfectly fit our needs, but that was chosen for its easy applicability.

Finally, larger testing is needed to further validate our approach and compare its capability against the traditional home polysomnography approach.

REFERENCES

1. A. S. Karunajeewa, U. R. Abeyratne, C. Hukins, "Silence-breathing-snore classification from snore-related sounds", Physiol. Meas. 29, p.227–243, 2008.
2. W. D. Duckitt, S. K. Tuomi, T. R. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data", Physiol. Meas. 27, p.1047–1056, 2006.
3. Y.-L. Hsu, M.-C. Chen, C.-M. Cheng, C.-H. Wu, "Development of a portable device for home monitoring of snoring", IEEE Int. Conf. Systems, Man and Cybernetics, vol.3, p.2420-24, 2005.
4. M. Calisti, L. Bocchi, C. Manfredi, I. Romagnoli, F. Gigliotti, G. Donzelli, "Automatic detection of snore episodes from fill night sound recordings: home and clinical application", AVFA09, Madrid, Spain, 2009.
5. A.K. Ng, T.S. Koh, E. Baey, T.H. Lee, U.R. Abeyratne, K. Puvanendran, "Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnoea?", Sleep Medicine, 2007.
6. N. Otsu, "A threshold selection method from gray–level histograms", IEEE Trans. Sys., Man., Cyber. Vol. 9(1), p.62–66, 1979.
7. M. Herzog, A. Schmidt, T. Bremert, B.Herzog, W. Hosemann, H.Kraftan, "Analysed snoring sounds correlate to obstructive sleep disorder breathing", Eur. Arch. Otorhinolaryngol. Vol. 265, p.105–113, 2008.
8. J. Sola-Soler, R. Jaine, J. A. Fiz, J. Morera, "Automatic classification of subjects with and without Sleep Apnoea through snoring analysis", Proc. 29th Int. Conf. IEEE EMBS, Lyon, France, 2007.

Mechanical models

ANALYSIS OF DEFORMATION CHARACTERISTICS OF EXCISED HUMAN VOCAL FOLDS BY OPTICAL STEREO-TRIANGULATION

B. Hüttner[1], A. Sutor[2], G. Luegmair[1], C. Bohr[1], U. Eysholdt[1], M. Döllinger[1]

[1]University Hospital for ENT medicine, University Hospital Erlangen, Erlangen, Germany
[2]Department of Sensor Technology, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Abstract: **For clinical treatment of voice disorders understanding of biomechanics of the voice producing parts in the human larynx is essential. An experimental setup is suggested to determine the deformations of the human vocal folds by inducing defined forces. In a static tensile test forces are applied to the fold of an excised human hemi-larynx. The resulting surface deformations of the tissue are detected using optical stereo-triangulation. For this purpose the positions of attached location markers are recorded by two cameras and reconstructed to three-dimensional points. The deformations of the vocal folds are derived from the displacements of the location markers. The correlation of the magnitude of induced forces and the elongation of tissue were analyzed and are presented.**

Keywords: **stereo-triangulation, vocal fold material parameters, hemi-larynx, vocal fold elasticity**

I. INTRODUCTION

Speech is the most important factor in human communication. Without the ability of speech the interaction in everyday life is seriously handicapped. The basis for speech is the primary voice-signal, which is produced in the larynx. There, on closing the vocal folds, a sub-glottal pressure is built up by the air flow from the lungs. The accumulated air is released in form of periodical bursts by a wave upon the vocal fold surface, which stimulates the folds for oscillation.

For clinical treatment of voice disorders the understanding of biomechanics of the voice producing parts in the human larynx is essential. Therefore we present an experimental setup to measure the correlation of applied forces and vocal fold deformations. The data is used to calculate material parameters as a starting point for creating artificial vocal folds that exhibit lifelike dynamics and deformation properties. These artificial vocal folds will be operated in a wind tunnel to analyze causalities in voice physiology.

II. METHODS

A. Experimental setup

In the experiments, human larynges, which were excised from cadavers, are used. By a sagittal cut one side of the larynx is removed. The resulting hemi-larynx allows free visibility to the remaining vocal fold (Fig. 1). The trachea is shortened to about 2 cm and is used to mount the hemi-larynx over a steal-tube. During the experiments the hemi-larynx is fixed in such a manner that pulling on the vocal fold influences only the fold, whereas the remaining parts of the hemi-larynx stay in place. 30 surgical micro sutures are sewn into the epithelium of the exposed vocal-fold and arranged in a 6 x 5 regular mesh. They serve as location markers. The
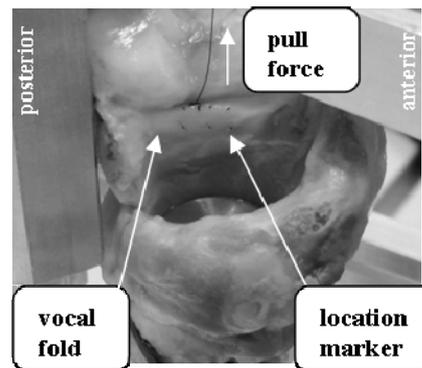
Figure 1: Fixed human hemi-larynx as example for the setup. On the vocal fold the sewn location markers are visible. The pull-forces are induced by medical strands and weights.

sutures of the first row are either sewn into the epithelium or into the muscle. They additionally serve as working points for the induced forces. The forces are generated by weights and are induced over medical strands and pulleys into the tissue. The forces are generated by weights of 10g, 20g, 50g and 100g and act vertically upwards. The pull-forces cause a deformation of the vocal fold which results in a displacement of the sewn location markers. The change in location is detected using the contact-free procedure of optical stereo-triangulation [1]. This technique allows the reconstruction of the three-dimensional (3D) marker locations by two two-dimensional (2D) images captured at different perspectives. The two images are generated by stationary digital cameras (1392 x 1040 pixel, 8-bit b/w) with a base-length of 1 m. The object-distance to the hemi-

larynx is 1.50 m and both cameras have a focal length of 300 mm. The cameras and the hemi-larynx are arranged in an isosceles triangle whereas the vocal fold presents the apex. The basis of the triangle is parallel to the vocal fold. The comparison of the 3D marker locations in different load-scenarios permits a precise metric determination of the vocal fold deformations depending on the induced forces. Before any force is applied to a vocal fold, the positions of the location markers are detected for reference.

B. Camera calibration & reconstruction

Stereo-triangulation allows the reconstruction of a point in the 3D Euclidian space on the basis of two 2D images captured at different perspectives. The reconstruction is divided into three steps: The calibration of the cameras, the detection of the location markers in the images and their reconstruction into 3D world-points. The calibration of the cameras has to be performed once as long as the mechanical setup of the cameras is not changed. For the mapping of a 3D world-point $\mathbf{m_w}$ from arbitrary Euclidian coordinates into a point $\mathbf{m}$ in a 2D pixel-coordinate system the pinhole camera model is used. The mapping is done by performing three transformations, namely the *extrinsic transformation*, the *perspective transformation* and the *intrinsic transformation* [2].

1. *extrinsic transformation*: A 3D arbitrary world-point $\mathbf{m_w}$ is transformed into a 3D point $\mathbf{m_c}$ in a camera-fixed coordinate system by a rotation $\mathbf{R}$ and a translation $\mathbf{t}$:

$$\mathbf{m}_c = \mathbf{R} \cdot \mathbf{m}_w + \mathbf{t} . \qquad (1)$$

2. *perspective transformation*: The point $\mathbf{m_c}$ is mapped from three dimensions into two dimensions:

$$\mathbf{m}_c = \begin{bmatrix} \mathbf{x}_c & \mathbf{y}_c & \mathbf{z}_c \end{bmatrix}^T \quad \mapsto \quad \mathbf{m}_p = \begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix}^T . \qquad (2)$$

3. *intrinsic transformation*: Because digital cameras can only display discrete points, the 2D metric coordinates of the point $\mathbf{m_p}$ have to be transformed into pixels. Furthermore, the origin of the coordinate system has to be adjusted to the top left corner of the image. This is done by the intrinsic transformation.

The result is the mapping for the pinhole camera:

$$\lambda \cdot \mathbf{m} = \mathbf{A} \cdot \mathbf{m}_c = \mathbf{A} \cdot (\mathbf{R} \cdot \mathbf{m}_w + \mathbf{t}), \qquad (3)$$

whereas $\mathbf{m}$ is the 2D image-point and $\mathbf{m_w}$ the 3D world-point. $\mathbf{A}$ is the matrix containing the intrinsic parameters of the camera (focal width, image center, scaling factor for metric- into pixel-coordinates). $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation of the extrinsic transformation. λ is a scaling-factor. It is necessary because all points $\mathbf{m}^i_w$ in 3D space that lie on a line through the camera's projection center are projected onto the same image point $\mathbf{m}$. Non-linear correction terms are considered because of non-ideal lens-systems [3]. The parameters of these corrections are summarized in the distortion vector $\mathbf{k}$.

The pinhole model is calibrated according to [4]. This method exploits the analogy between the pinhole camera and the *direct linear transformation (DLT)*. The DLT describes the transformation between points on a plane P_1 to another plane P_2 via homography $\mathbf{H}$. For camera calibration purposes a 2D chessboard is employed. The corners of the squares are used as calibration points. With the fact that all points lie in a 2D plane in 3D space, the orientation of the world coordinate system can be set in such a manner, that one coordinate of the calibration points equals zero. Thus we can describe the mapping of the i 3D world-points in the calibration-plane into the i 2D image-points by the DLT:

$$\lambda \mathbf{m}_i = \mathbf{H} \mathbf{m}^i_w, \qquad (4)$$

whereas $\mathbf{H}$ is the above mentioned homography-matrix. This equation is solved by least-square solvers. As the coordinates of the calibration points are known with respect to the coordinate-center of the chessboard, it is not necessary to know their exact 3D coordinates in space. So far non-linear effects of distortion have not been considered. To embed them in the mapping, a non-linear optimization is run, using the following cost-function [1]:

$$\min_{\mathbf{A},\mathbf{k},\mathbf{R}_i,\mathbf{t}_i} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \mathbf{m}_{i,j} - g(\mathbf{A},\mathbf{k},\mathbf{R}_i,\mathbf{t}_i,\mathbf{m}^i_w) \right\|_2^2 . \qquad (5)$$

During calibration $n=10$ images, each containing $m=100$ calibration-points (for each camera) are used. g is the mapping of the i-th 3D world-point $\mathbf{m}^i_w$. $\mathbf{m}_{i,j}$ are the detected points in pixel-coordinates. Equation (5) is a non-linear optimization problem. We apply the Levenberg-Marquardt algorithm to obtain the optimal parameters $\mathbf{A}$, $\mathbf{R}$, $\mathbf{t}$, and $\mathbf{k}$.

Before reconstruction, the 2D marker coordinates have to be extracted from the images. This is done by manually predefining a region of interest and applying a center-of-mass refinement of the pixel intensities.

Knowing the camera-parameters of equation (5) and the positions of both cameras to each other, the detected points can now be reconstructed. For this purpose the line of sight is reconstructed using the detected image-point $\mathbf{m}_i$ and the projection-center $\mathbf{c}_i$ ($i = 1, 2$) of camera 1 and 2. In the ideal case, the reconstructed 3D point is the intersection point of the both lines. Because of occurring errors in real experiments the reconstructed point has to be approximated as the center of the shortest distance between the two lines.

III. RESULTS

A. Application

Using the proposed method, the vocal fold of a 65 year old male was reconstructed and analyzed. The sutures were sewn into the epithelium. Fig. 2 shows exemplarily
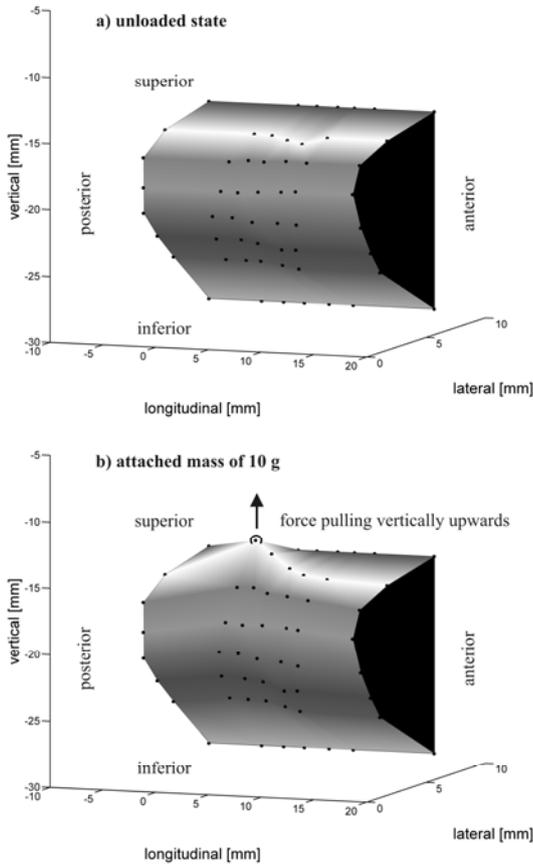
Figure 2: Model of a reconstructed vocal fold surface, a) in unloaded state, b) in loaded sate with an attached mass of 10 g. The 30 black dots in the middle represent the sewn location markers, the remaining ones serve as attached border. The black circle (b) indicates the working point of the applied force.

the reconstructed vocal fold surface. In Fig. 2a) the vocal fold is in an unloaded state, in 2b) it is deformed by an attached mass of 10 g (~ 0.1 N). The force is pulling vertically upward on the first marker (posterior) in the topmost row. The working point of the force is indicated by the black circle. The 30 inner black dots represent the 3D reconstructed location markers. The boundary dots are extrapolated with help of the coordinates of the location markers in the rows and columns respectively. They do not change their positions during the pulling and therefore act as attached border for the reconstructed vocal fold. The surface is created by generated triangles that connect the reconstructed marker positions and the boundary points, respectively.

Fig. 3 shows the deformation characteristics of this vocal fold. Plotted are the deformations of the five location markers in the topmost row from posterior to anterior. The symbols represent the deformations of the location marker being pulled. The curves were fitted to the data according to [5] by:

$$f(x) = a \cdot (e^{-b \cdot x} - 1) . \tag{6}$$

In this example a maximal deformation of the epithelium of approximately 6 mm was measured while applying a mass of 100 g to the vocal fold.
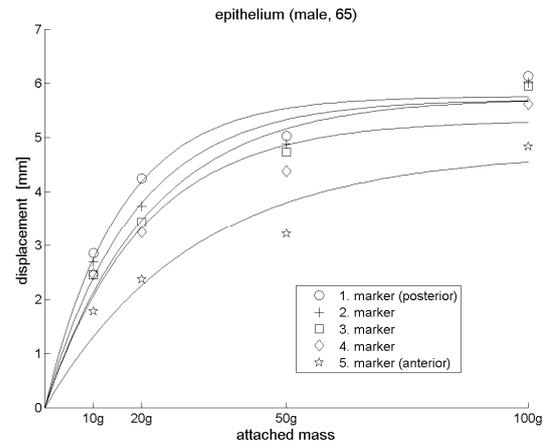
Figure 3: Deformation characteristics of the epithelium of a human vocal fold. The data shows a non-linear trend for weights higher 20g.

B. Camera calibration & reconstruction

For error estimation we have to distinguish between errors made during camera calibration and errors occurring during the process of reconstruction. In the calibration step of the cameras a non-linear optimization problem has to be minimized to obtain extrinsic and intrinsic parameters of the cameras (eq. 5). For this approximation the *rear projection error* (RPE) is minimized. The error quantifies the difference between the mapping of a known 3D calibration point in the Euclidian space with the pinhole camera model and its detected coordinates in the image. Table 1 presents the RPEs of the two cameras after calibration. In the ideal case the RPE is zero. The RPE was analyzed regarding the isotropic and the anisotropic error. All appearing errors are in the sub-pixel range. The errors of camera 1 and 2 are in the same magnitude, even though the errors of camera 2 are somewhat higher.

Table 1: Rear projection errors occurring in the calibration step due to the minimization of the non-linear optimization problems:

error	camera 1	camera 2
longitudinal [px]	0.48	0.49
vertical [px]	0.35	0.42
total [px]	0.65	0.72

The isotropic error of both cameras was in the magnitude of approximately 0.7 pixels. With the intercept theorem

$$\frac{I}{O} = \frac{i}{o} , \tag{7}$$

whereas I and O is the size of the image and object, respectively, i and o is the image distance and object distance, respectively, and under consideration of a pixel size of 4.65 μm we can compute that this corresponds to an error of approximately 54 μm.

Because of the RPE an image-point **m** is not projected into Euclidian space along a line, but into a cone around this line. This imprecision of the projection requires an approximation of the world-point $\mathbf{m}_w$ during the reconstruction procedure as mentioned above. The error that occurs during the approximation-process of the 3D point $\mathbf{m}_w$ is quantified by the reconstruction of the images of 15 x 10 calibration points lying in a 2D horizontal test-plane. The points had a distance of 100 μm to each other. In the ideal case all reconstructed points lie in a plane again. After the reconstruction process, the distances of the 3D points were calculated. The results are presented in Tab. 2. There you can see the mean distances between the reconstructed points and the concerning standard deviations. $\mathbf{x}_1$ and $\mathbf{x}_2$ are the basis-vectors of the 2D plane. $\mathbf{x}_1$ is perpendicular, $\mathbf{x}_2$ approximately parallel to the image-planes of the used cameras. The isotropic mean of 93.9 μm represents an error of 6%. The error in $\mathbf{x}_2$-direction (~ 20%) is more than twice the error in $\mathbf{x}_1$-direction (~ 8%).

Table 2: Mean distances of the reconstructed calibration points.

	Distance
$\mathbf{x}_1$	108.2 ± 13.4 μm
$\mathbf{x}_2$	80.0 ± 9.3 μm
$\mathbf{x}_1, \mathbf{x}_2$	93.9 ± 15.9 μm

IV. DISCUSSION

A method to measure the 3D deformations of human vocal folds was presented. The deformation characteristics of the analyzed vocal folds show a viscoelastic behavior. In the shown example (Fig. 2) the epithelium of the fold behaves nearly linear for low forces (up to 20 g = 0.2 N). For larger forces the deformations fade to saturation. This trend can be observed for all the five analyzed location markers. Consequently, the trend is independent of the longitudinal position of the suture on the vocal fold. This observation corresponds to results of [5] who analyzed the longitudinal deformations of canine vocal folds. The data in [5] was also fitted using equation (6). Although the trend is equal for all sutures, the elasticity of the epithelium is decreasing from posterior to anterior what can be seen on the shrinking amplitude of deformation in Fig. 2.

The anisotropic error in the reconstruction process is pointing to a systematic error during the calibration step of the cameras. For the mapping of the 3D calibration-points into the 2D image points the following ideal conditions are assumed: The test plane is ideally planar, the distances between the calibration-points are exactly 100 μm and there is no error during the detection of the mapped calibration points. All three assumptions cannot be fulfilled in real experimental conditions. Zhang did analyze the effects of such errors in his work [6]. He

found, that deformations of the calibration pattern (cylindrical or spherical) are more serious than Gaussian noise during the detection of the mapped 2D calibration points. Relying on [6], the occurring errors are results of a non-ideal calibration pattern and not of an imprecise detection of the image points.

V. CONCLUSION

In static pulling-experiments of human vocal folds the tissue-deformations were extracted and analyzed. With help of the presented method it could be shown that the epithelium obeys to a viscoelastic behavior: For low applied forces a linear deformation-process was observed, for strong pull forces the deformations were non-linear. The human vocal folds exhibit the same trend of deformation as canine vocal folds [5]. The obtained deformation characteristics of the human tissue will be used to compute the material parameters of artificial vocal folds [7]. Those will be used to analyze the dynamic behavior of vocal folds in an in vitro model.

VI. ACKNOWLEDGEMENTS

VII. LITERATURE

[1] O. Schreer: *Stereoanalyse und Bildsynthese*, Springer Berlin Heidelberg New York, 2005.
[2] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses", *IEEE Journal of Robotics and Automation*, 3(4), pp. 323-344, 1987.
[3] D. Brown, "A Close range camera calibration", *Photogramm Eng*, 37(8), pp. 855-866, 1971.
[4] Z. Zhang: "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11), pp. 1330-1334, 2000.
[5] F. Alipour, I. R. Titze, "Elastic models of vocal fold tissues", *J. Acoust. Soc. Am.*, 90(3), pp. 1326-1331, 1991.
[6] Z. Zhang, "A flexible new technique for camera calibration. Technical report", *Microsoft Research*, 1998.
[7] B. Schmidt, M. Stingl, G. Leugering, M. Döllinger, R. Lerch, M. Kaltenbacher, "Optimization material properties and geometry of a physical multi-layered vocal fold model ", *Proceedings of NAG/DAGA*, 306, 2009.

A LF-PULSE FROM A SIMPLE GLOTTAL FLOW MODEL

A. Aalto[1], P. Alku[2] and J. Malinen[1]

[1]Dept. of Mathematics and Systems Analysis, Helsinki Univ. Tech., Espoo, Finland

[2]Dept. of Signal Processing and Acoustics, Helsinki Univ. Tech., Espoo, Finland

Abstract: **We discuss a novel low-order mass-spring model of human vocal folds with incompressible 1D flow. Our model consists of three subsystems: a flow model, a nonsymmetric mass-spring model for the vocal folds, and a resonator representing the vocal tract (VT).**
Keywords: **Glottis model, Bernoulli flow, flow induced vibrations**

I. INTRODUCTION

This study addresses a mechanical glottis simulator to act as the source for the wave equation model of vowel production, see [4]. We require a validated glottal flow/pressure signal simulator that is computationally less demanding than Navier–Stokes and/or elasticity equations. We also use this model for studying the vocal tract feedback effect in [1] where we observe some phenomena reported in [10] and [11]. In this paper we validate our glottis model against the LF-signal model proposed in [3]. This is carried out by fitting the simulated pulse to LF-pulses measured from three different types of phonation.

Our model consists of three subsystems: a 1D flow model; a nonsymmetric, low-order mass-spring model for the vocal folds; and a resonator representing the vocal tract (VT), based on the Webster's equation. These subsystems are modelled using physically sound mathematical approximations. Firstly, we use the Bernoulli law and the mass conservation law for a static flow. Secondly, the flow through the glottis is assumed to be incompressible. Obviously, the flow is not truly static because of the moving vocal folds, and the Webster's equation is based on the very assumption that air is compressible. In addition to these inconsistencies between subsystems, also the glottis geometry is extremely simplified as, e.g., in the seminal work [7].

II. MATHEMATICAL MODEL

A. Flow

We assume an incompressible 1D air flow through the glottal opening whose velocity v_o satisfies

$$\dot{v}_o(t) = \frac{1}{C_{iner} h H_1} \left(p_{sub} - \frac{C_g}{\Delta W_1(t)^3} v_o(t) \right), \quad (1)$$

motivated by the Hagen–Poiseuille law; here p_{sub} is the subglottal pressure (subtracted by the ambient air pressure), and h is the width of the rectangular channel. The parameter C_{iner} regulates the flow inertia and C_g regulates the pressure loss in the glottis. The glottal opening is given by $\Delta W_1 = g + w_{21} - w_{11}$ at the narrow end (i.e., towards the supraglottal cavity). The opening at the wide end (towards the trachea) is $\Delta W_2 = H_0 + w_{22} - w_{12}$; see Fig. 1 for these and other used symbols.

In the glottis, the flow velocity $V(x,t)$ is assumed to satisfy the static mass conservation law for incompressible flow

$$H(x,t)V(x,t) = H_1 v_o(t)$$

where $H(x,t)$ is the height of the channel in the glottis. In our simple geometry it is

$$H(x,t) = \Delta W_2(t) + \frac{x}{L}(\Delta W_1(t) - \Delta W_2(t)), \quad x \in [0, L].$$

Now the pressure $p(x,t)$ in the glottis is given by the two equations above and the (static) Bernoulli law

$$p(x,t) + \frac{1}{2}\rho V(x,t)^2 = p_{sub}.$$

Since both vocal folds have two degrees of freedom, this pressure can be reduced to a force pair $(F_{A,1}, F_{A,2})^T$ where $F_{A,1}$ effects at the narrow end of the glottis ($x = L$) and $F_{A,2}$ at the wide end ($x = 0$). This reduction is done by using the total force and moment balance equations

$$F_{A,1} + F_{A,2} = h \int_0^L (p(x,t) - p_{sub}) \, dx,$$

$$L \cdot F_{A,1} = h \int_0^L x(p(x,t) - p_{sub}) \, dx - p_c \cdot h \frac{H_1}{2} \frac{H_0 - H_1}{2}.$$

The moment is evaluated with respect to point $(x, y) = (0, 0)$ for the lower fold and $(x, y) = (0, H_0)$ for the upper fold.

Evaluation of these integrals yields

$$
\begin{cases}
F_{A,1} = \frac{1}{2}\rho v_o^2 hL \left(-\frac{H_1^2}{\Delta W_1(\Delta W_2 - \Delta W_1)} \cdots \right. \tag{2} \\
\qquad \left. + \frac{H_1^2}{(\Delta W_1 - \Delta W_2)^2} \ln\left(\frac{\Delta W_2}{\Delta W_1}\right) \right) - \frac{H_1(H_0 - H_1/2)}{4L} h p_c, \\
F_{A,2} = \frac{1}{2}\rho v_o^2 hL \left(\frac{H_1^2}{\Delta W_2(\Delta W_2 - \Delta W_1)} \cdots \right. \\
\qquad \left. - \frac{H_1^2}{(\Delta W_1 - \Delta W_2)^2} \ln\left(\frac{\Delta W_2}{\Delta W_1}\right) \right) + \frac{H_1(H_0 - H_1/2)}{4L} h p_c.
\end{cases}
$$

B. Vocal folds

The vocal fold model consists of two wedge-shaped vibrating elements that have two degrees of freedom each (see Fig. 1). The distributed mass of these elements can be reduced into three mass points which are located so that m_{j1} is at $x = L$, m_{j2} at $x = 0$, and m_{j3} at $x = L/2$. The elastic support of the vocal folds is approximated by two springs at points $x = aL$ and $x = bL$. Thus the equations of motion for the vocal folds are

$$
\begin{cases}
M_1 \ddot{W}_1(t) + B_1 \dot{W}_1(t) + K_1 W_1(t) = -F(t), \\
M_2 \ddot{W}_2(t) + B_2 \dot{W}_2(t) + K_2 W_2(t) = F(t)
\end{cases} \tag{3}
$$

where $W_j = (w_{j1}, w_{j2})^T$ are the displacements of the right and left endpoints of the jth fold, $j = 1, 2$. Here M_j, B_j, and K_j are the mass, damping, and stiffness matrices, respectively

$$
\begin{aligned}
M_j &= P \begin{bmatrix} m_{j1} + \frac{m_{j3}}{4} & \frac{m_{j3}}{4} \\ \frac{m_{j3}}{4} & m_{j2} + \frac{m_{j3}}{4} \end{bmatrix}, \\
B_j &= \begin{bmatrix} b_{j1} & 0 \\ 0 & b_{j2} \end{bmatrix}, \\
K_j &= \frac{1}{P} \begin{bmatrix} a^2 k_{j1} + b^2 k_{j2} & ab(k_{j1} + k_{j2}) \\ ab(k_{j1} + k_{j2}) & b^2 k_{j1} + a^2 k_{j2} \end{bmatrix}.
\end{aligned} \tag{4}
$$

The entries of these matrices are computed by means of Lagrangian mechanics. The damping matrices B_j are diagonal since the dampers are located at the endpoints of the vocal folds. The springs are located symmetrically around the midpoint $x = L/2$, so that $a = (L/2 + l)/L$ and $b = (L/2 - l)/L$. The parameter P is used for tuning the oscillation frequency.

During the glottal open phase, the load terms of (3) are given by $F = (F_{A,1}, F_{A,2})^T$ as given in Eq. (2). During the glottal closed phase ($\Delta W_1 < 0$), there are no aerodynamic forces except the counter pressure from the VT. Instead, there is a nonlinear spring force given by the Hertz impact model for the collision of the vocal folds (see [5]):

$$
F_H = \begin{bmatrix} k_H |\Delta W_1|^{3/2} - \frac{H_0 - H_1/2}{2L} \frac{H_1}{2} h \cdot p_c \\ \frac{H_0 - H_1/2}{2L} \frac{H_1}{2} h \cdot p_c \end{bmatrix}.
$$
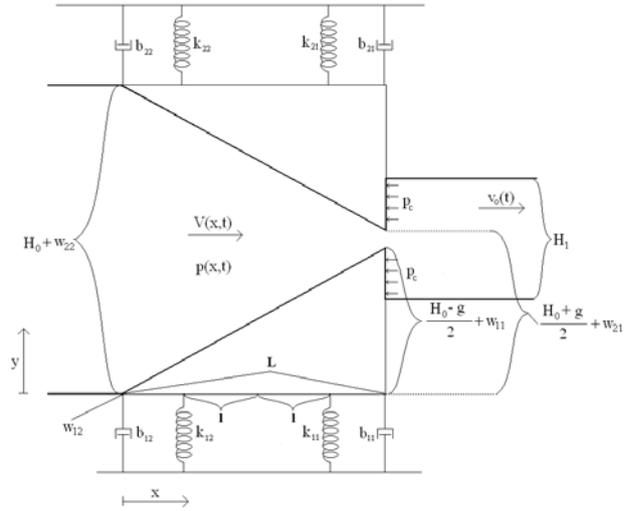
Figure 1: The geometry of the glottis model and the symbols used

C. Vocal tract

We use Webster's horn model resonator as an acoustic load. The Webster's horn equation is

$$
\Psi_{tt}(s,t) - \frac{c^2}{A(s)} \frac{\partial}{\partial s}\left(A(s) \frac{\partial \Psi(s,t)}{\partial s} \right) = 0
$$

where c is the sound velocity and $\Psi(s,t)$ is a velocity potential. Note that $p = \rho \Psi_t$ in the VT. The parameter $s \in [0, L_{VT}]$ is the distance from the narrow end of the glottis measured along the VT centreline and L_{VT} is the length of the VT. The area function $A(\cdot)$ is the cross-sectional area of the VT, perpendicular to the VT centreline. It is taken from the geometry of [Section 4, 4] corresponding [ø:].

The resonator is controlled by the flow velocity v_o from Eq. (1) through the boundary condition at the glottis end

$$
\Psi_s(0,t) = -v_o(t).
$$

The boundary condition at lips is a frequency-independent acoustic resistance of the form

$$
\Psi_t(L_{VT}, t) + \theta c \Psi_s(L_{VT}, t) = 0
$$

where θ is the normalised acoustic resistance (see [Chapter 7, 8]) regarded as a tuning parameter. This boundary condition represents flow resistance $p = \theta \rho c v$ where v is the flow velocity through the mouth.

The resonator exerts a counter pressure $p_c(t) = \rho \Psi_t(0,t)$ to the vocal folds equations (3) through Eqs. (2), thereby forming a mechanical feedback loop between the vocal folds and the vocal tract.

III. Numerical simulation

The equations of motion (3) are solved with the fourth order Runge–Kutta (RK) method, and the flow equation (1) is solved with the implicit Euler method. The VT is discretised by the Finite Element Method using piecewise linear elements and the physical energy norm of the Webster's equation. The Crank–Nicolson method is applied for temporal discretisation. Especially the FEM-solver performs faster with a constant time step, because the state update equations are the same, and we spare one matrix inversion on every step. Thus we keep time steps ΔT constant except when the glottis either closes or opens.

The load force in Eq. (3) is discontinuous (in fact, singular) when $\Delta W_1 \to 0^+$. The singularity is removed by replacing the aerodynamic force by zero when $\Delta W_1 < \epsilon$. Since v_o is small when ΔW_1 is small, the solution is not sensitive to the choice of $\epsilon > 0$. The discontinuity is dealt with by locating the time of closure/opening by interpolation and restarting computation from there, see [pp. 12-14, 1]. For this we use the second degree interpolating polynomial so that numerical error is of order $\mathcal{O}(\Delta T^3)$. Because the number of the exceptional steps (at times of opening or closure) does not depend on ΔT, the total error is also of order $\mathcal{O}(\Delta T^3)$. The total error of the RK-method is of order $\mathcal{O}(\Delta T^4)$ but considering its smaller computational burden, this interpolation method is an appropriate way to treat the discontinuous load in Eq. (3).

IV. Parameter estimation and model validation

The model parameters are listed in Table 2. When estimating model parameters, we use three LF-pulses, obtained by fitting the LF-waveforms to glottal flow derivatives, estimated from natural speech using an automatic inverse filtering method [2]. Speech data consisted of vowels [aː] produced in breathy, normal, and pressed phonation by a male speaker.

The damping matrix in Eq. (3) is adjusted so that

Figure 2: Three LF-pulses corresponding to breathy, normal, and pressed phonation and the corresponding simulated pulses after parameter tuning

the vibration is stable and sustained oscillation occurs. The rest of the parameters in Eqs. (2) and (3) are from literature (see Table 2). Note that only the relative values of parameters p_{sub}, C_{iner}, and C_g in Eq. (1) matter, and increasing the subglottal pressure p_{sub} only increases the height of the pulse.

If the changes in $\Delta W_1(t)$ are neglected, we have three parameters characterizing the flow pulse: the pulse length (parameter P), the height (parameter p_{sub}), and the inclination (parameter C_{iner}). The inertance C_{iner} and the parameter P in Eqs. (1) and (4)

Table 1: The inertance C_{iner} in Eq. (1) for different LF-pulses. The Mean error 1 is the error from the beginning of the pulse to the peak value and Mean error 2 from the peak value to the closure.

	Breathy	Normal	Pressed
C_{iner} $\left(\frac{kg}{m^4 s}\right)$	524.8	530.9	630.8
Mean error 1 (%)	2.93	1.45	2.09
Mean error 2 (%)	4.14	3.07	2.14

are tuned so that the pulse shape matches optimally a measured LF-pulse. Since the measured pulse height has no scale, the parameters p_{sub} and C_g are tuned so that their magnitudes are realistic and the total flow (area) of the pulse matches that of the LF-pulse.

This parameter tuning is performed so that the instants of maximal flow in the measured LF-pulse and the simulated pulse coincide. The length of the pulse (that is, parameter P) and the parameter C_{iner} are then varied in order to minimise the squared error of the two pulses. The measured and the simulated pulses after parameter tuning are shown in Fig. 2. The optimal values of C_{iner} are shown in Table 1.

V. Conclusions

We presented a flow mechanical glottis model to be used as a real-time source for a VT simulator. The model is validated against three LF-pulses that have been estimated by parameterising glottal flow derivatives, obtained by inverse filtering natural utterances. We conclude that LF-pulses corresponding to different types of phonation can be faithfully constructed.

It is interesting to observe how the inertance term C_{iner} in Eq. (1) increases monotonically when the phonation type changes from breathy to normal, and then to pressed. When producing pressed voices, speakers use a vibration mode which is characterised by a small abduction quotient (see [pp. 260, 9]) which, in turn, results in a flow pulse with a shorter relative length of the closing phase. The present study indicates that this phenomenon is reflected as the increase of the parameter value C_{iner}.

It is worth noting that here we have only optimised

the flow parameters of Eq. (1) according to measured data. Depending on the type of phonation, a speaker controls not only the subglottal pressure p_{sub} but also the mechanical properties of the vocal folds, corresponding to Eq. (3). For example, the glottis barely closes during breathy phonation, whereas the glottal vibration patterns are similar in all of our simulations.

References

[1] Aalto, A. (**2009**). A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load, Master's thesis, TKK, Helsinki. Available at `http://math.tkk.fi/research/sysnum/`.

[2] Alku, P. (**1992**). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication **11**, 109–118.

[3] Fant, G., Liljencrants, J., and Lin, Q. (**1986**). "A four-parameter model of glottal flow," Tech. rep., QPRS: Dept. for Speech, Music and Hearing, Stockholm.

[4] Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (**2007**). "Vowel formants from the wave equation," Journal of the Acoustical Society of America Express Letters **122**, EL1–EL7.

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

[6] Horáček, J. and Švec, J. (**2002**). "Aeroelastic model of vocal-fold-shaped vibrating element for studying the phonation threshold," Journal of Fluids and Structures **16**, 931–955.

[7] Ishizaka, K. and Flanagan, J. L. (**1972**). "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell System Technical Journal **51**, 1233–1268.

[8] Morse, P. and Ingard, K. (**1968**). Theoretical Acoustics, McGraw–Hill.

[9] Titze, I. (**1994**). Principles of voice production, Prentice Hall.

[10] Titze, I. (**2008**). "Nonlinear source-filter coupling in phonation: Theory," Journal of the Acoustical Society of America **123**, 2733–2749.

[11] Titze, I., Riede, T., and Popolo, P. (**2008**). "Nonlinear source-filter coupling in phonation: Vocal exercises," Journal of the Acoustical Society of America **123**, 1902–1915.

Table 2: The model parameters

	Source of the value
h, H_0, L, l	From [6]
H_1	Through condition $hH_1 = A(0)$
p_{sub}, C_g	Only the relative magnitudes of p_{sub}, C_{iner} and C_g are relevant.
g	Set so that glottal area matches typical observations
m_{ij}, k_{ij}	From mechanical properties of the model presented in [6]
b_{ij}	Tuned so that sustained, stable oscillation occurs
k_H	From [5]
$A(\cdot)$, L_{VT}	From [4]
θ	Tuning parameter regulating energy dissipation in the VT
P, C_{iner}	Optimised (see Sec. IV)

MATHEMATICAL MODELLING OF AIRFLOW IN THE GLOTTAL REGION AND ITS COMPARISON WITH EXPERIMENTAL DATA

J. Horáček, S. Gráf

Institute of Thermomechanics of the Academy of Sciences
Dolejškova 5, Prague 8, Czech Republic

Abstract: **Finite element method (FEM) was used for numerical simulation of the airflow field in a simplified model of the human vocal tract for vowel /a:/ with prescribed periodic oscillations of the vocal folds. The viscous fluid is modeled by 2D compressible Navier-Stokes equations in Arbitrary Lagrangian-Eulerian (ALE) formulation and considering the turbulence. The computed flow field pattern is compared with the original experimental results obtained by Particle Image Velocimetry (PIV) method.**
Keywords: **FEM, compressible Navier-Stokes equations, ALE method, k-ε turbulence model, PIV**

I. INTRODUCTION

The source of the human voice originates in an energy transfer of the air flowing from the lungs to the energy of the self-oscillating vocal folds and to the acoustic energy of the pressure fluctuations in the glottis that propagate subsequently through the acoustic resonance cavities of the human vocal tract to the mouth. The principles of the physical mechanism of such energy transfer into the acoustic pressure disturbances are not yet properly known and thus the experimental as well as theoretical studies on the flow field in the glottal region are currently encountered in papers on voice production modelling (see e.g. [1,2]). Here, the airflow velocity field in a simplified physical model of the vocal tract for vowel /a:/ with oscillating vocal folds is studied by numerical simulations and compared with the original measurement performed by the PIV method [3,4]. The computational domain is identical with the model of the airways used in the previous experimental study.

II. METHODS

The mathematical model of the viscous flow was considered in the form of unsteady compressible Navier-Stokes equations taking into account the turbulence and large vibration amplitude of the vocal folds. The numerical simulation of the unsteady 2D air flow-field in the glottis was performed by the finite element method using the ANSYS software with the FLOTRAN CFD code for the fluid flow modelling. The changes of the computational fluid domain during the prescribed vibrations of the vocal folds were respected by the ALE method. The standard k-ε turbulence model was used.

The fundamental vibration frequency, subglottal pressure and amplitude of the vocal folds oscillations were prescribed according to the parameters setting used in the experiment [4].

The computational domain was meshed by 15600 quadrilateral-shaped 2D FLUID141 elements. The transient analysis was performed with the time step $\Delta t = 10^{-5}$ s for 10 oscillation periods of the vocal folds.
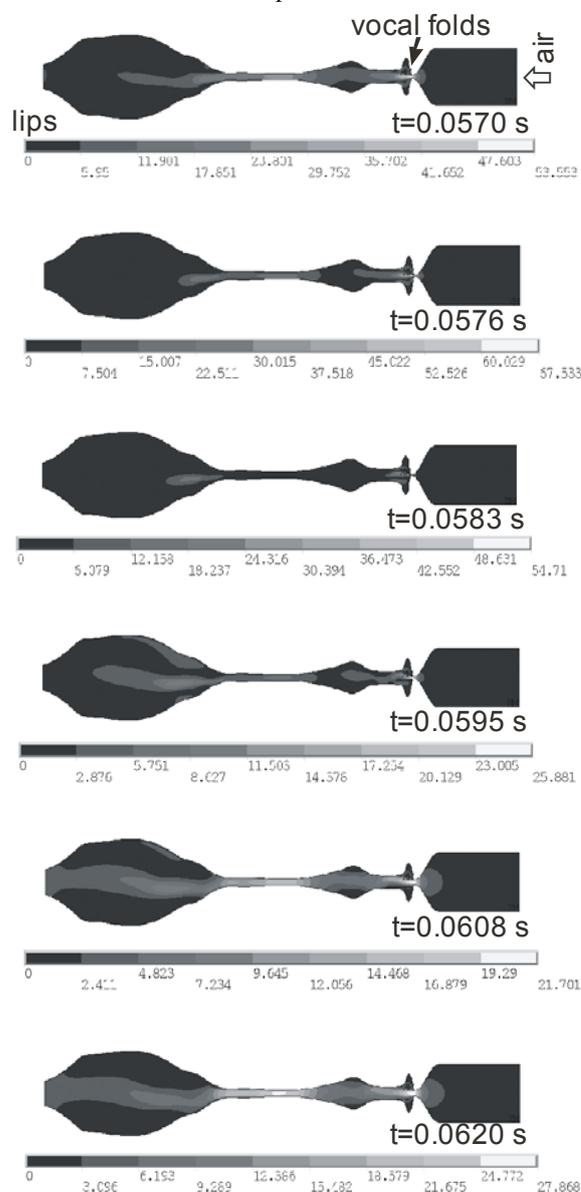
Fig. 1 Numerical simulation of the airflow– example.
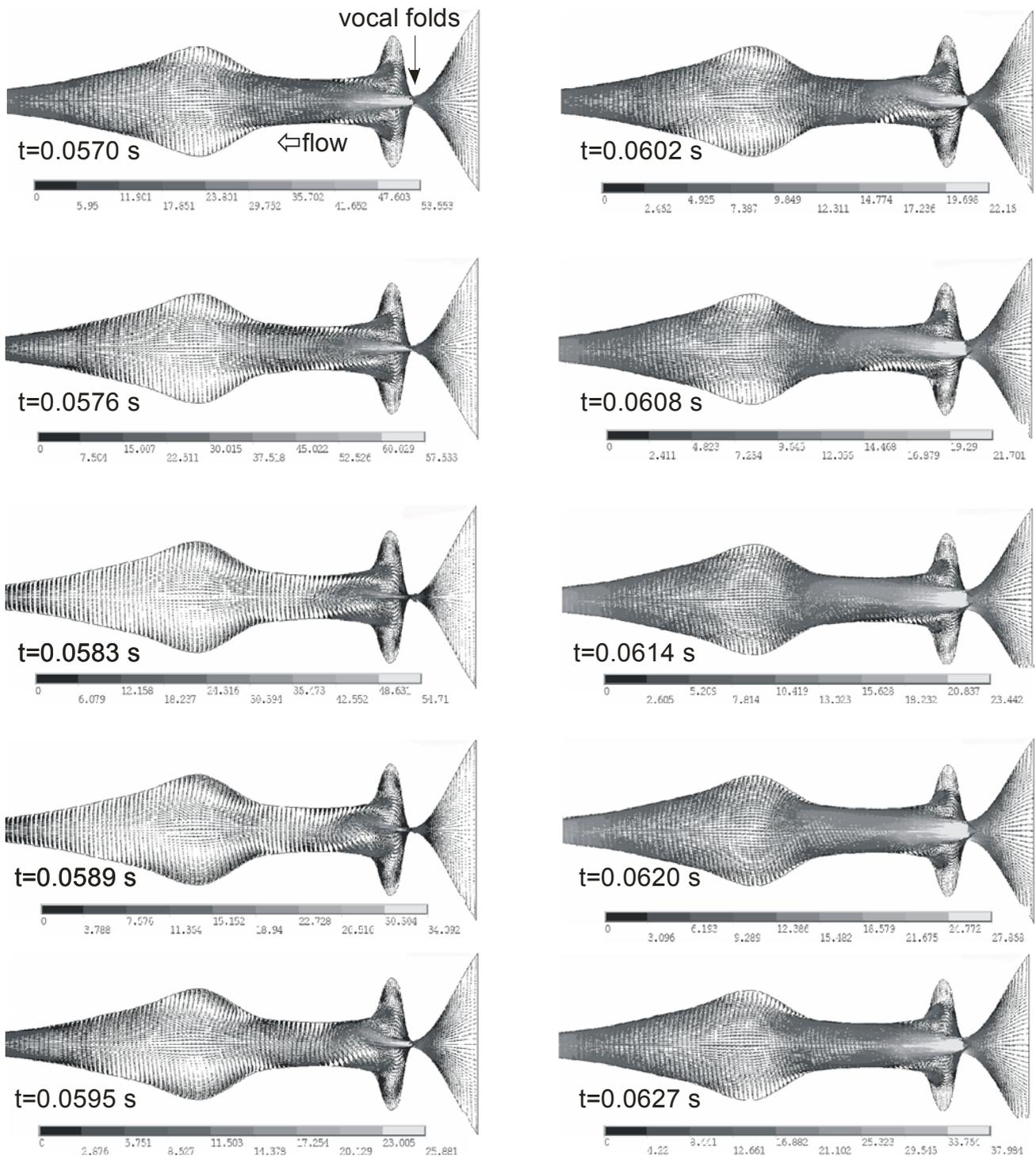
Fig. 2 Numerical simulation of the airflow field velocity in the glottal region during one oscillation period.

III. RESULTS

In the experiments, the airflow was coming from a simplified model of the trachea entering the model of the laryngeal part of the vocal tract with the self-oscillating artificial vocal folds and ending by the mouth cavity. The glottal region also included the ventricular folds model [3,4].

We present here the results computed for the velocity flow-field corresponding to the measurement for the average value of the subglottal pressure 900 Pa, air flow rate 0.25 l/s, fundamental vibration frequency of the vocal folds 158 Hz and the maximum vibration amplitude

of the glottis opening 2.5 mm. A complete closure of the glottis, as existed in the experiment, was not possible to model theoretically, and a very small minimum glottis opening (0.2 mm) was assumed. instead of the glottis closure. The air temperature 293.15 K, density 1.225 kg/m^3 and dynamic viscosity 15x10^{-6} Pa.s were considered in the computations.

The computed airflow velocity field is shown in Fig. 1 at 6 selected time instants during the oscillation period including the cases of maximum (67.5 m/s) and minimum (21.7 m/s) flow velocity in the glottis. The flow field pattern computed at 10 time instants over the tenth oscillation period is presented in the glottis region in Fig. 2. and the computational FE mesh is shown in detail in Fig. 3. Fig. 4 presents the time domain simulations of the glottal gap together with the axial v_x and lateral v_y components of the airflow velocities and the pressure drop Δp on the channel axis before entering the glottal region and in the narrowest cross-section of the orifice at two selected points numbered by 1 and 2 in Fig. 3.

The streamlines evaluated from the PIV measurement in the glottis near the vocal folds during one oscillation cycle are shown in Fig. 5. The vibrating vocal folds recorded by a high-speed camera are shown on the right snapshot margin.
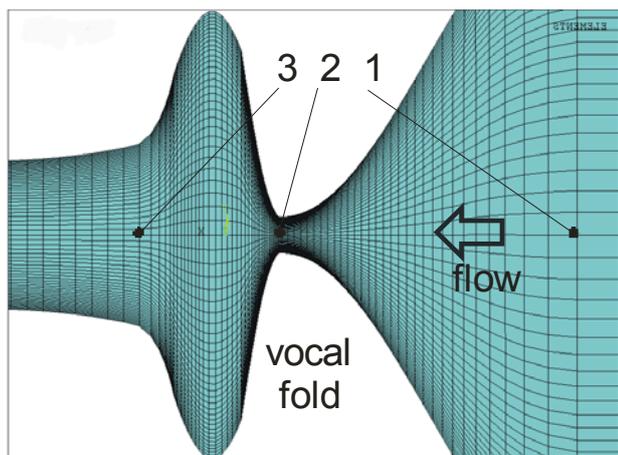
Fig. 3 Detail of the FE mesh in the glottal orifice.

IV. DISCUSSION

The computed results show that even if the channel geometry and the vocal folds motion are perfectly axisymmetric the flow field velocity pattern is changing over one oscillation period from a more symmetric (see, e.g., the flow field for t=0.057 s in Fig. 2) to evidently asymmetric flow (see, e.g., the flow for t=0.0595 s). The jet is flapping from a center position to the upper wall in the laryngeal region (see Fig. 2). This behavior resembles to the effect known from stationary flows called Coanda effect. The main air stream is changing the pattern by inflowing the wider part of the channel (laryngeal cavity),

where the large scale eddies are dominant, and then after inflowing to the narrower part of the channel (model of the epilarynx) the flow becomes uniform. Such qualitatively similar behaviour of the flow is possible to see in the experimentally obtained flow patterns (see Fig. 5).

Small-scale vortices computed in the model of the ventricular folds are produced as a result of the flow separation on the vibrating surface of the vocal folds model (see Fig. 2). The asymmetry of the flow (Coanda effect) is also possible to see in the model of the mouth cavity (Fig. 1).

Fig. 4 shows that the numerical solution becomes periodic after first 2 – 3 periods of the glottis gap oscillation. The pressure at the entrance to the glottis (in the node denoted by 1 in Fig. 3) is oscillating around the prescribed input pressure drop Δp =900 Pa. The oscillations of the lateral flow velocity v_y support the hypothesis on existence of the Coanda effect, this flow velocity in the narrowest cross-section of the glottal
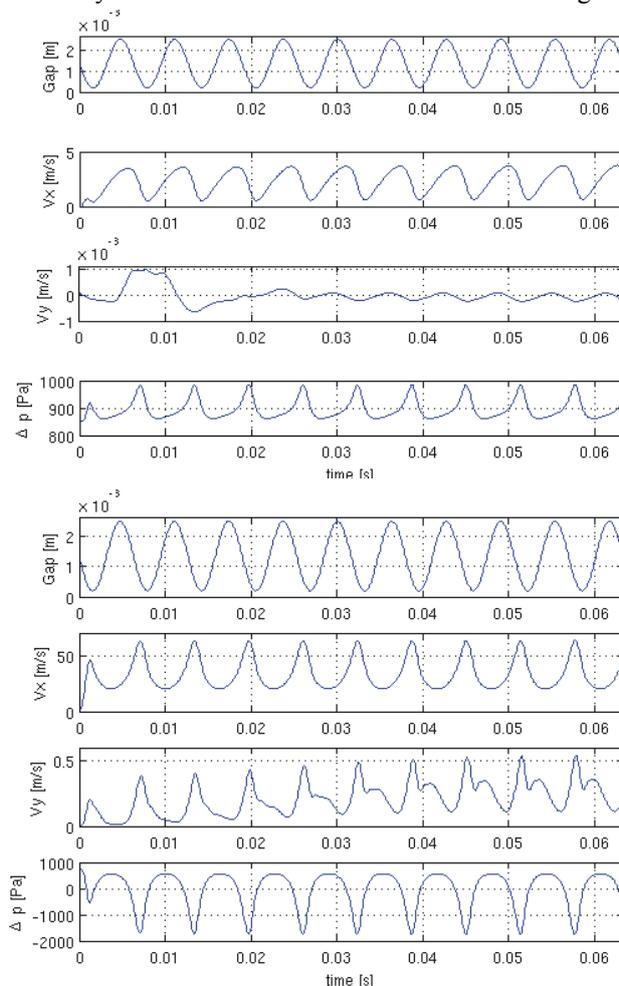
Fig. 4 Numerical simulation of 10 gap oscillation periods, axial and lateral flow velocity components and pressure drop at the points 1 (upper panel) and 2 (lower panel).

206
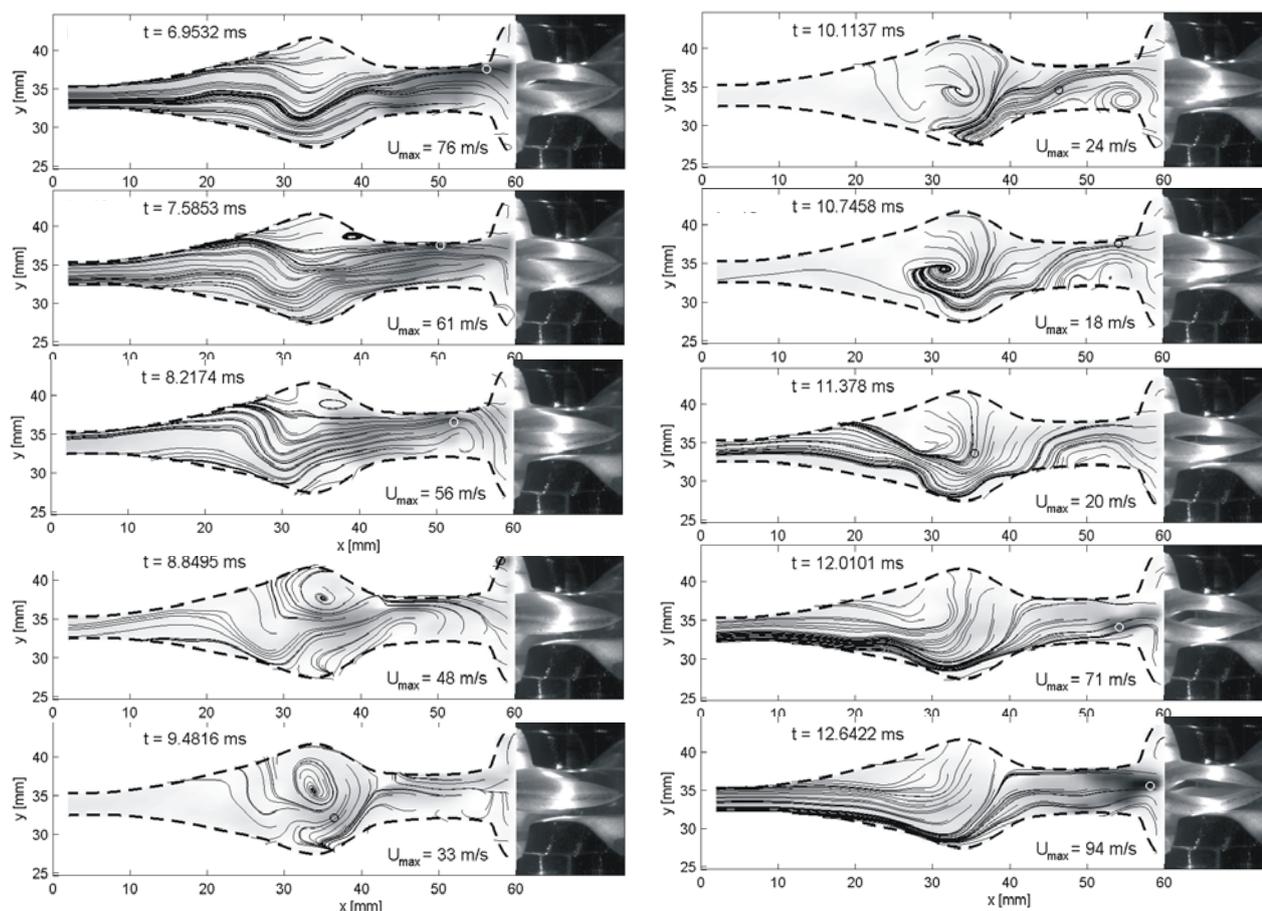
Fig. 5 Streamlines and vocal folds self-oscillations measured at 10 time instants during one oscillation cycle.

channel (in the node denoted by 2 in Fig. 3) is ca 10 times smaller then the maximum axial flow velocity v_x (see Fig. 2).

V. CONCLUSION

An asymmetric jet at the glottis and the formation of large-scale vortices in the larynx are possible to see in the experiment as well as in the numerical simulations. The similar values of maxima and minima of the airflow velocities were obtained both in the experiment and in the numerical simulations. The results are also in a reasonable qualitative agreement with the measurements performed for a driven glottis–shaped orifice [5] and for airflow measured downstream of the artificial vocal folds but without modelling the vocal tract [6]. Further numerical simulations should respect the 3D effects, which were found important in the experiments [2-6].

REFERENCES

[1] S. Zörner, M. Kaltenbacher, G. Link, R. Lerch, M. Döllinger, "Numerical study of the human phonation process by the Finite Element Method" in *Proc. NAG/DAGA 2009* (ed. M. Boone), Rotterdam: NAG and DEGA, 2009, pp. 1718-1721. ISBN 978-3-9808659-6-8.
[2] M. Triep, W. Mattheus, M. Stingl, C. Brücker, "Three-Dimensional Unsteady Flow Nature in the Vocal Tract during Human Phonation" in *Proc. NAG/DAGA 2009* (ed. M. Boone), Rotterdam: NAG and DEGA, 2009, pp. 1741-1743. ISBN 978-3-9808659-6-8.
[3] J. Horáček, P. Šidlof, V. Uruba, J. Veselý, V. Radolf, V. Bula, "PIV Measurement of Flow-Patterns in a Human Vocal Tract Model," in *Proc. Interaction and Feedbacks 2008* (ed. I. Zolotarev), Institute of Thermomechanics ASCR: Prague 2008, pp.27-40.ISBN 978-80-87012-15-4.
[4] J. Horáček, P. Šidlof, V. Uruba, J. Veselý, V. Radolf, V. Bula, "PIV Measurement of Flow-Patterns in a Human Vocal Tract Model," in *Proc. NAG/DAGA 2009* (ed. M. Boone),Rotterdam: NAG and DEGA,2009,pp.1737-1740.
[5] Ch. Brücker, M. Triep, W. Mattheus, R. Schwarze, "Pulsating jet generated by a driven glottis-shaped Orifice," in *Proc. Int. Conf. on Jets, Wakes and Separated Flows*, Technical Univer. of Berlin: 2008, 8p.
[6] Ch. Tao, Y. Zhang, G. Hottinger, JJ. Jiang, "Asymmetric airflow and vibration induced by Coanda effect in a symmetric model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 122, 2007, pp.2270-2278.

FINITE ELEMENT MODEL OF THE HUMAN PHONATION PROCESS

S. Zörner[1,2], M. Kaltenbacher[1], Michael Döllinger[3]

[1] *Alps-Adriatic University of Klagenfurt, Austria, Chair of Applied Mechatronics*

[2] *University of Erlangen-Nürnberg, Germany, Department of Sensor Technology,*

[3] *University of Erlangen-Nürnberg, Germany, Department of Phoniatrics & Pediatric Audiology*

Abstract

The basis of the human phonation process is given by complex interaction of air flow in the larynx together with structural mechanics of the vocal folds. This paper presents a numerical scheme to model the fluid-solid interaction in the human larynx and its resulting acoustic sound.
The scheme is utilised to simulated the phonation process in a 2D-model. Different geometries of the vocal folds have been used to analyse the effect on the fluid field, the vibration of the vocal folds and the sound generation. The results show the self sustained oscillation of the vocal folds and resolve the Coanda effect.
***Keywords:* human phonation, fluid-structure interaction, aeroacoustic, finite element method**

I   Introduction

To simulate the process of human phonation the three physical fields fluid-, solid-mechanics and acoustics are taken into account. Fluid flow describes the airflow through the larynx, which brings the vocal folds to vibrate, and in turn changes the fluid domain. Both, fluid flow and vocal fold vibration, generate sound which propagates through the larynx known as human phonation. The fluid field is modelled with the incompressible Navier-Stokes equations. The solid field is described by the Navier's equation and the acoustic sound propagation is described by the inhomogeneous wave equation based on Lighthill's analogy. The coupling between fluid-solid and solid-acoustic is based on continuum mechanics, while the acoustic source term inside the fluid are computed via Lighthill's analogy. Each of these physical fields is discretised by the finite element method.

Latest finite element laryngeal models have been presented by [9] and [8]. A different approach, based on the immersed boundary method, can be found by [7].

II   Methods

In the following, the relevant physical fields for the phonation process and their coupling will shortly be described. The arising partial differential equations (PDEs) are all solved by applying the Finite-Element method (FEM). For a detailed discussion we refer to [5, 6].

II.1   Fluid mechanics

The governing set of partial differential equations for the fluid mechanics is given by the momentum and mass conservation

$$\rho \frac{\partial \vec{v}}{\partial t} + \rho(\vec{v} \cdot \nabla)\vec{v} + \nabla p - \mu \Delta \vec{v} = 0 \,, \qquad (1)$$

$$\nabla \cdot \vec{v} = 0 \,, \qquad (2)$$

with $\vec{v}$ the flow velocity, ρ the fluid density, p the hydrodynamic pressure and μ the dynamic viscosity. The equations hold for incompressible fluids which may be assumed due to the fact that for the considered application the Mach number is smaller than 0.3. The computational domain of the fluid flow constantly changes since the vocal folds move and hence define the fluid boundary. The difficulty has been tackled by utilising the Arbitrary-Lagrangian-Eulerian (ALE) approach (for details see [1, 2]).

II.2 Solid mechanics

The mechanical displacement $\vec{u}$ of the vocal folds are modelled by Navier's equation

$$\nabla \cdot \sigma_s = \rho_s \frac{\partial^2}{\partial t^2} \vec{u}, \tag{3}$$

where σ_s denotes the Cauchy stress tensor and ρ_s, the density of the solid. Introdcuing the tensor of elasticity $[c]$ and tensor of linear strain $[S]$, allows us to express Hook's law by

$$\sigma_s = [c][S] \tag{4}$$

and the linear strain-displacement by

$$[S] = \nabla^{sym} \vec{u}. \tag{5}$$

Substituting (4) and (5) into (3) results in the final PDE for linear elasticity

$$\mathcal{B}^T [c] \mathcal{B} \vec{u} = \rho_s \frac{\partial^2}{\partial t^2} \vec{u} \tag{6}$$

with the differential operator $\mathcal{B}$ (here given explicitly for the 2D plane case

$$\mathcal{B} = \begin{pmatrix} \frac{\partial}{\partial x} & 0 \\ 0 & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} \end{pmatrix}. \tag{7}$$

II.3 Fluid-solid interaction

The air and vocal folds share a common interface Γ_{fs} so that the nodes for both fields must coincide, given by

$$\vec{x}_f = \vec{x}_s \qquad \text{on } \Gamma_{fs}. \tag{8}$$

Fluid velocity and the first time derivative of the solid displacement are identical since the fluid adheres at the body resulting in the following condition

$$\vec{v} = \frac{\partial}{\partial t} \vec{u} \qquad \text{on } \Gamma_{fs}. \tag{9}$$

This implies for solid mechanics the following inhomogeneous Neumann boundary condition

$$[\sigma_s] \cdot \vec{n} = [\sigma_f] \cdot \vec{n} \quad \text{on } \Gamma_{fs} \tag{10}$$

describing the equivalent of fluid stress $[\sigma_f]$ and solid stress $[\sigma_s]$ in normal direction $\vec{n}$. The fluid stresses can be written explicitly by the hydrodynamic pressure p

and fluid velocity $\vec{v}$ as

$$\vec{\sigma}_s = \quad \rho_f \underbrace{\int_{\Gamma_{fs}} -pI \cdot \vec{n} dx}_{pressure} \tag{11}$$

$$+ \underbrace{\int_{\Gamma_{fs}} \mu \left(\nabla \vec{v} + (\nabla \vec{v})^T \cdot \vec{n} \right) dx}_{shear}. \tag{12}$$

Having Dirichlet boundary condition for the fluid and Neumann boundary conditions for solid mechanics, the fluid-solid interaction is also called Dirichlet-to-Neumann problem.

II.4 Acoustic field

As a basis taking the equation of continuity and momentum, Lighthill's equation in pressure form is derived (for details see [4])

$$\frac{1}{c^2} \frac{\partial^2 p'}{\partial t^2} - \Delta p' = \nabla \cdot (\nabla \cdot \mathbf{T}), \tag{13}$$

with c is the speed of sound and $\mathbf{T}$ the Lighthill tensor

$$T_{ij} = \underbrace{\rho_f v_i v_j}_{Reynolds\ stress} + \underbrace{\tau_{ij}}_{Viscous\ stress} \tag{14}$$

$$+ \underbrace{[(p - p_0) - c^2(\rho_f - \rho_0)]\delta_{ij}}_{Heat\ conduction}. \tag{15}$$

Thereby, p_0 denotes the mean pressure, ρ_f the fluid density and ρ_0 its mean density. Viscous stress may be neglected [4] and the heat conduction is assumed to be zero, which leads to the following approximation of (15)

$$T_{ij} \approx \rho_f v_i v_j. \tag{16}$$

The oscillation of the vocal folds induce sound, which is a surface coupled phenomenon. Along the moving boundary Γ_{fs} the following relation for the mechanical surface and the acoustic pressure needs to be fulfilled

$$\frac{\partial}{\partial t} \vec{u} \cdot \vec{n} = \vec{v}_a \cdot \vec{n} \qquad \text{on } \Gamma_{fs}. \tag{17}$$

Condition (17) forces that the acoustic particle velocity $\vec{v}_a$ are identical to the surface velocity in normal direction. For the considered case it is assumed, that there is no back reaction of the acoustic onto the solid. Using the linearised Euler equation

$$\frac{\partial}{\partial t} \vec{v}_a \cdot \vec{n} = -\frac{1}{\rho_f} \frac{\partial}{\partial n} p' \tag{18}$$
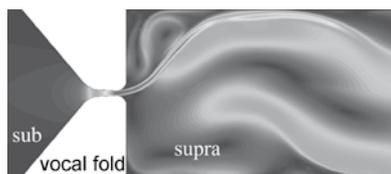
Figure 1: Asymmetric air flow, the jet attaches to the trachea wall - coanda effect.

the source term in acoustic pressure formulation is

$$\frac{\partial}{\partial n}p' = -\rho_{\mathrm{f}}\frac{\partial^2}{\partial t^2}\vec{u}\cdot\vec{n} \quad \text{on } \Gamma_{\mathrm{fs}}. \tag{19}$$

For the FE formulation and its verification we refer to [3].

Results

The geometric setup of the vocal folds have been adopted from the model presented in [8] and inserted into our computational domain. A fluid pressure condition is given at inflow and outflow of the domain.
The resulting simulations shows the development of the Coanda effect - the air jet at the glottis randomly attaches to either side of the trachea wall, as shown inf Fig. 1. Furthermore, the occurring fluid flow forces realistic self-sustained vocal fold oscillation. The vibrations of the vocal folds have been analysed and show for different forms different frequencies in their movement. An eigenfrequency analysis shows that the vibrational frequency correlates with the first eigenfrequency of the vocal folds. The generated sound showed dominant peaks in the frequency domain, which vary with the geometric form of the vocal folds.

Conclusion

A computational scheme has been presented to simulate the human phonation process with all relevant physical fields. To the author's best knowledge this fully coupled scheme is novel. The model is applicable for a parameter study, to analyse the effect of different forms of vocal folds and different pressure conditions for in and outflow on the acoustic sound.

References

[1] T. Belytschko, W.K. Liu, and B. Moran. *Nonlinear Finite Elements for Continua and Structures*. J. Wiley & Sons Ltd, Chichester, 2000.

[2] J. Donea and A. Huerta. *Finite element methods for flow problems*. Wiley, 2003.

[3] M. Kaltenbacher, M. Escobar, I. Ali, and S. Becker. Numerical simulation of flow-induced noise using les/sas and lighthill's acoustics analogy. *International Journal for Numerical Methods in Fluids*, 2009.

[4] M. J. Lighthill. On sound generated aerodynamically I. General theory. *Proceedings of the Royal Society of London*, 211:564–587, 1951.

[5] 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. *Computer Methods in Applied Mechanics and Engineering*, 2009.

[6] Gerhard Link. *A Finite Element Scheme for Fluid-Solid-Acoustics Interactions and its Application to Human Phonation*. PhD thesis, Der Technischen Fakultät der Universität Erlangen-Nürnberg, 2008.

[7] Haoxiang Luo, Rajat Mittal, Xudong Zheng, Steven A. Bielamowicz, Raymond J. Walsh, and James K. Hahn. An immersed-boundary method for flow–structure interaction in biological systems with application to phonation. *Journal of Computational Physics*, 227:9303–9332, 2008.

[8] S.L. Thomson, L. Mongeau, and S.H. Frankel. Physical and numerical flow-excited vocal fold models. In *3rd International Workshop MAVEBA*, pages 147–150. Firenze University Press, 2003. ISBN 88-8453-154-3.

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

RECORDING SPEECH DURING MRI: PART II

J. Malinen, P. Palo

Dept. of Mathematics and Systems Analysis, Helsinki Univ. Tech., Espoo, Finland

Abstract: **We design and construct a recording arrangement for speech during an MRI scan of the speakers vocal tract. We concentrate on the acoustic environment around the test subject inside the MRI machine. The data thus obtained is used for construction and validation of a numerical model of the vocal tract.**
Keywords: **Speech recording, MRI, acoustic wave guides**

I. Introduction

We model vowel production by the wave equation, see [5]. For the vocal tract, we use a boundary controlled wave equation and the corresponding resonance model (i.e., the Helmholtz equation) which we solve with FEM. A similar approach has been taken in, e.g., [9]. As a source for the wave equation, we use a flow mechanical glottis model introduced in [1], [2]. To validate and tune the vocal tract model, we need to extract formants and their bandwidths from speech and singing signals. These signals must be recorded in high quality during an MRI scan.

The present article is the latter part of a description of the sound recording arrangement based on an acoustic sound collector and wave guides; see [6] for the first part. We now concentrate on the acoustics around the test subject inside the MRI machine in the measurement configuration shown in Fig. 1b.

The acoustic signals are captured by the sound collector, see Fig. 1a, whose position with respect to the test subject is as in Fig. 1b. The collector is a two-channel device that operates by the same principle as a differential microphone. It has separate horns for speech and noise — one on each side — that lead to the wave guides. As explained in [6], both sound channels are acousticly transmitted by wave guides to a microphone assembly inside a Faraday cage. Because the noise cancellation is realized by analog electronics, the de-noised signal can be fed back into subject's earphones without delay.

The recording equipment has been designed for recording speech and singing signals for research purposes. Further engineering effort is required to make the device suitable for clinical practice.
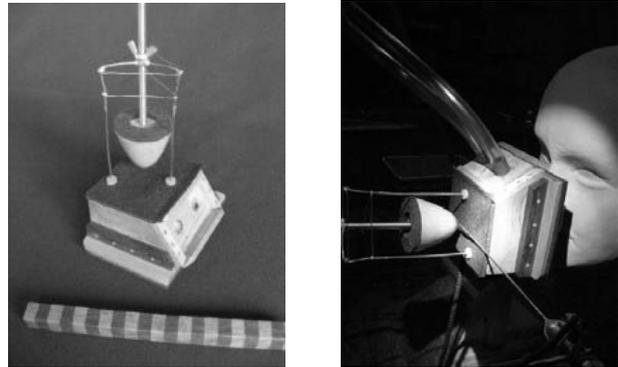
(a)   (b)

Figure 1: (a) The sound collector and a reflector paraboloid suspended from a temporary measurement suspension (shown with a cm scale ruler), (b) Measurement arrangement for near field acoustics of the sound collector

II. Challenges and solutions

A. Engineering challenges

The MRI room is a quite challenging sound recording environment. There is acoustic noise of about 90 dB(SPL) during the imaging sequence. The noise arrives to the sound collector with different delays because of multi-way propagation. A Siemens Magnetom Avanto 1.5 T MRI machine produces a static 1.5 T magnetic field, and an imaging sequence produces an electromagnetic field at 64 MHz with a peak power of several kW. Because of safety and image quality considerations, no metal or electronics can be taken near the test subject. For speech naturality, comfortability of the test subject is important [4].

B. Technical solutions

The sound collector is completely passive, metal free, and without moving parts. The wave guides detach from the sound collector so as not to hinder taking the subject out from the machine. The collector is fully compatible with MR safety requirements and does not cause any artefacts in the images. It fits on

the head coil of a Siemens Magnetom Avanto 1.5 T machine. The test subject lies in supine position inside the MRI machine. The sound collector is about 30 mm away from the lips.

The ambient noise is partly removed by the two-channel recording arrangement (as shown in [Fig. 5, 6]), and partly by attenuation material carefully positioned inside the MRI machine. For acoustic impedance adjustment, both horn surfaces of the collector are covered with attenuation material that also takes care of exhalation noise.

On top of the sound collector in Fig. 1a, there is a reflector paraboloid that widens the incoming noise beam by shadowing it in the middle. Without such a reflector, the noise sample gets collected in an undesirably narrow angle. The test subject affects the acoustic impedance (hence, the frequency response) of the speech channel, and the form and distance of the paraboloid are tuned to approximate the same effect on the noise channel side. The noise cancellation is succesful when the acoustic impedances are close to each other.

See [6] and [8] for details of the components not described here.

C. Measurement approximations.

To simplify analysis, we divide the acoustic space into two parts. We regard a sphere of radius 8 cm around the center point of the sound collector as the *near field*. The characteristic curves (such as Fig. 2) of the whole recording equipment are particularly sensitive to near field phenomena. In the *far field*, noise and its reflections from the MRI machine require most attention.

To further simplify analysis, we divide the frequency range according to the near field length scale: *low frequencies* under 2 kHz ($\lambda/2 > 8$ cm), *middle frequency range* 2 – 4.4 kHz (4 cm $< \lambda/2 <$ 8 cm), and *high frequencies* above 4.4 kHz ($\lambda/2 < 4$ cm). At low frequencies, the two channel noise cancellation is most effective when the inconvenient longitudinal resonances of the wave guides are properly controlled by wave guides' impedance terminations. At high frequencies, propagation of noise can be understood by the ray optical approximation but severe complications in active noise cancellation are caused by phase differences due to multi-way propagation. The middle frequency range has none of the good and all of the bad qualities. In such case, the only reasonable solution is the placement of damping material around the test subject, based on trial and error.
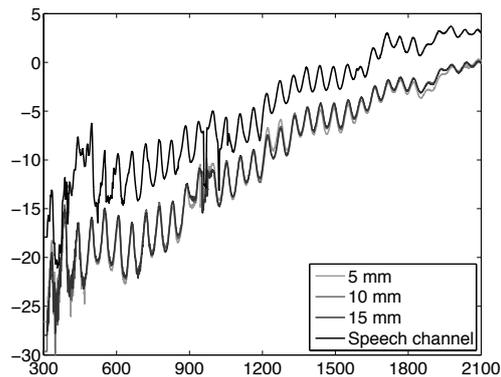
Figure 2: Frequency responses of the speech channel (above) and the noise channel with various reflector positions (below) in mm with attenuation in dB and frequency in Hz. Note that the amplifications of the speech and noise channels are different to improve readability.

III. Characteristic curves and tuning

In the measurements we have a two fold objective: First, to understand the near field behaviour in terms of acoustic impedances and frequency responses. Second, to understand the constraints posed by near field engineering on the technical solutions of the far field problems.

To tune the equipment and to obtain necessary frequency responses for sound post-processing by DSP, we carried out several near field measurements in an anechoic chamber from physical models, see Fig. 1b. We return to acoustic measurements with physical models when weeding out artefacts from acoustic data measured from a test subject.

A. Frequency responses

The frequency responses of the speech and noise channels are given in Fig. 2. The noise channel response is measured with several paraboloid positions. The frequency responses of both channels are very similar as expected from the symmetrical construction of the channels. This is a prerequisite for the noise cancellation to work by analogue signal subtraction.

The data in Fig. 2 has been measured using the experimental setting of Fig. 1b. The tip of the reference microphone probe, see Fig. 4b, is placed at the distance of 5 mm above the surface level of the sound collector at the center of the corresponding horn. The frequency responses of the channels have been determined with respect to the sound pressure at these reference points. Note that the speech channel is measured using the point source in Fig. 4a, and the noise
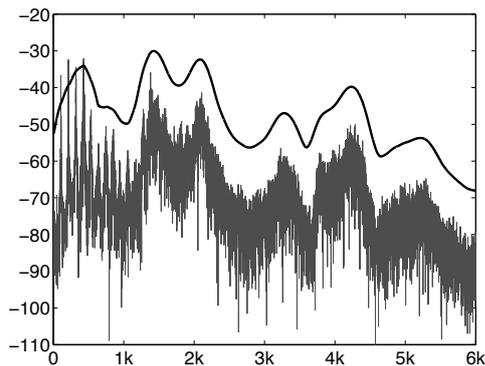
Figure 3: Original and averaged (raised by 20 dB) spectrum of a long production of [øː] recorded within an anechoic chamber

(a) (b)

Figure 4: (a) Acoustic point source, (b) Reference microphone probe (right), microphone units of type Panasonic WM-62 (middle)

channel using a plane wave source that resembles the interior surface of the MRI machine.

B. Tuning the reflector paraboloid

As can be seen from Fig. 2, the paraboloid position does not significantly affect the low frequency response of the noise channel. Thus, the paraboloid position and size can be optimised according to the requirements of the middle frequency range 2 – 4.4 kHz without compromising the noise cancellation properties in the low frequencies. The optimisation objectives in the middle frequency range are both the incoming noise beam shape and noise cancellation (to the extent it is feasible).

IV. Speech recording experiment

We recorded a full set of Finnish vowels produced by both authors. We did not use any model of the MRI device or its noise. The formant peaks at 0.42 kHz, 1.41 kHz, 2.09 kHz, 3.28 kHz, 4.23 kHz, and 5.22 kHz are very clearly visible in spectrograms even without any DSP compensation, see Fig. 3. The glottis pulse is easily recoverable, too.

A typical MRI device produces a sparse and spiky noise spectrum — i.e., the sound energy is restricted to few fairly narrow frequency bands and their superharmonics. Given the linearity of the recording equipment (the microphones are the dominating source of nonlinearity), it is possible to separate speech from the residual noise in frequency domain. This is carried out by recording *a priori* noise spectrum data when the silent test subject lies in the MRI machine during an imaging sequence.

V. Construction of laboratory equipment

In addition to the equipment described below, we use a loudspeaker assembly for simulating the ambient (noise) field. Custom Matlab 7.4 code was written to generate weighted sweeps, and to estimate and compensate frequency responses.

A. Sound source and face model

We constructed an acoustic point source and a natural size face model shown in Fig. 4a and Fig. 1b, respectively. The horn of the point source can be placed at the mouth of the face model.

A practically constant, sufficiently high amplitude sound pressure can be obtained above 300 Hz when the point source is fed by a properly weighted sinusoid sweep signal. Then the virtual source point is at the center of the exponential horn opening on the right in Fig. 4a. Because of the dimensions of the source, acceptable signal cannot be produced under 300 Hz. Fortunately, the recording equipment requires little attention at these low frequencies.

For validation, we measured the polar patterns of the source at the distance of 35 mm from the virtual source point with and without the face model. The patterns are presented in Fig. 5 at frequencies 0.5, 1, 2, and 4 kHz. Even with the face model, the amplitude variation stays within an acceptable 3 dB range.

B. Reference probe

In the near field measurements, even the small ∅ 9 mm reference microphone requires a special probe to avoid considerable distortion in the results. The probe and some microphone units are shown

214

Point source / Point source with face model
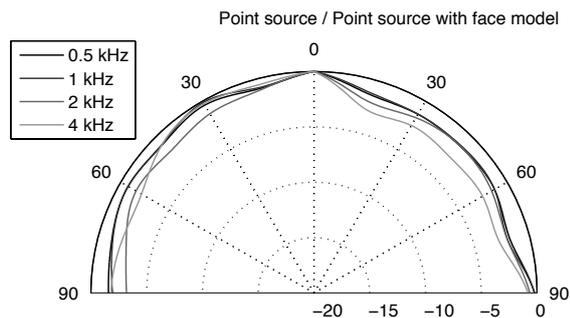
Figure 5: Polar pattern of the point source at 35 mm with (right) and without (left) the face model for angles between 0° and 90° and attenuation in dB

in Fig. 4b. The probe can be seen extending from the lower right corner in Fig. 1b. It is 150 mm long and of $\varnothing$ 2 mm. Its frequency response was measured using the sound source shown in [Fig. 6, 6].

VI. Conclusions

We have described equipment and characteristic curve measurements for a recording arrangement of speech during an MRI scan. Up to 4.4 kHz, the presented equipment performs essentially like a pair of microphones in dipole configuration. For a detailed discussion, see [8].

The space around the subject in the MRI machine contains reflecting surfaces, and the sound collector receives strong echoes of high frequency noise in different phases. For this reason, the active dipole based noise cancellation cannot be expected to work well for frequencies over 2 kHz. Instead, passive attenuation material must be placed inside the MRI machine. The position of the damping material is most easily determined empirically in the imaging situation, rather than by using mathematical or physical models.

There are comparable systems that are based on fiber optics (e.g., [3], [7]). Some of the challenges in acoustic and optical solutions are the same, such as the multi-way propagation of noise. Acoustic equipment is larger than optical, and additional complications from various acoustic impedances require more attention. With acoustic equipment, however, linearity is always guaranteed if the microphones are used within their operational limits; the non-microphonic sound collector is immune to vibrations; and the recording arrangement can be easily modified to meet a great variety of practical situations.

References

[1] Aalto, A. (**2009**). A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load, Master's thesis, TKK, Helsinki. available at http://math.tkk.fi/research/sysnum/.

[2] Aalto, A., Alku, P., and Malinen, J. (**2009**). "A LF-pulse from a simple glottal flow model," in "MAVEBA 2009," Florence, Italy.

[3] Branderud, P. *et al.* (**2009**). Personal communication.

[4] Engwall, O. (**2006**). Speech production: Models, Phonetic Processes and Techniques, chap. Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation., Psychology Press, New York, 301 – 314.

[5] Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (**2007**). "Vowel formants from the wave equation," Journal of the Acoustical Society of America Express Letters **122**, EL1–EL7.

[6] Lukkari, T., Malinen, J., and Palo, P. (**2007**). "Recording speech during magnetic resonance imaging," in "MAVEBA 2007," Florence, Italy, 163 – 166.

[7] Optoacoustics, Ltd., accessed Aug. 25th, (**2009**). http://www.optoacoustics.com/

[8] Palo, P. (**2009**). A wave equation model for vowels: Validation and parameter estimation, Manuscript for Licentiate thesis, xii + 50 pp., TKK, Institute of Mathematics.

[9] Švancara, P., Horáček, J., and Pešek, L. (**2004**). "Numerical modelling of production of Czech wowel /a/ based on FE model of the vocal tract," in "Proceedings of International Conference on Voice Physiology and Biomechanics," .

AUTHOR INDEX