

The 4th COST 2103

Advanced Voice Function Assessment Workshop

Clinical voice assessment in the future - new horizons

General Chair: **Professor David M Howard (York, UK)**

Co-Chairs: **Mr Andrew Grace (York, UK)**

Professor John Local (York, UK)

Merchant Taylor's Hall, York, UK

19-21 May 2010

ABSTRACTS



THE UNIVERSITY *of York*



The 4th COST 2103

Advanced Voice Function Assessment Workshop

Welcome

It is with great pleasure that I offer you a warm welcome to the City of York for the 4th Advanced Voice Function Assessment Conference sponsored by COST 2103. I hope that your experience here in York will be special for you and that you take away some special memories of your stay here. Our conference venue is one of York's treasures in its own right. The Company of Merchant Taylors in York has its origins in the thirteenth century and records go back to 1387. The Hall is a fine example of a medieval building and its recent restoration has been carefully completed to bring the facilities up to date within the context of proper historical preservation.

The over-arching conference theme is *Clinical voice assessment in the future - new horizons* which fits very closely with the main objective of the EU COST 2103 Action which is *to combine previously unexploited techniques with new theoretical developments to improve the assessment of voice for as many European languages as possible, while acquiring in parallel data with a view to elaborating better voice production models* (COST2103 website). The COST Action, a joint initiative of speech processing teams and the European Laryngological Research Group (ELRG), brings together speech processing engineers and laryngologists as well as voice practitioners and phoniatricians to progress clinical assessment and enhancement of voice quality by sharing findings, interacting with and learning from each other in a spirit of cooperation and friendship.

I hope this Workshop in York deepens this experience for us all both intellectually and socially. You are all very welcome to York.

David M Howard
AVFA '10 General Chair

TUESDAY 18th May 2010		
15:15	Choral Evensong	York Minster
19:15	Welcome to York!	Short walking tour of City of York with our own Guide Meet at the South Door of the Minster , next to the Statue of Constantine (see map inside front cover) for prompt 19:15 start!
WEDNESDAY 19th May 2010		
08:00	REGISTRATION	Merchant Taylor's Hall (see map inside front cover)
08:45	WELCOME: David M Howard (AVFA'10 General Chair)	
INVITED KEYNOTE 1		
08:50	Patrick Naylor	Recent Advances and Future Strategies for Speech Processing [page 7]
WS1: Voice source analysis 1		
09:35	Lionoudaki & Stylianou	On the derivative of the EGG Signal for determining GCIs and GOIs [page 8]
09:55	Kafentzis & Stylianou	A frequency domain approach for the determination of the glottal opening instant in EGG waveforms [page 9]
10:15	Murphy	Normalised Time of Increasing Contact for Different Loudness Levels [page 11]
10:35	COFFEE	
WS2: Voice source analysis 2		
11:00	Libeaux, Ternström & Henrich	Spectrum variations in the electroglottographic signal related to voice sound pressure level [page 13]
11:20	Kafentzis, Stylianou, Alku, & Neumann	Evaluation of inverse filtering approaches using subglottic data [page 14]
11:40	Manfredi, Bocchi, Calisti, G. Cantarella & Peretti	Assessing the performance of a new tool for videokymographic images analysis

12:00	Sáenz-Lechón, Fraile, Godino-Llorente, Fernández-Baíllo, Osma-Ruiz, Gutiérrez-Arriola & Arias-Londoño	Towards objective evaluation of perceived roughness and breathiness based on Mel-frequency cepstral analysis [page 17]
12:20	Astrinaki, Kiagiadaki, Vasilakis & Stylianou	Correlations between subjective and objective assessments of voice quality [page 19]
12:40	LUNCH	
WS3: Voice assessment		
14:00	Schoentgen, Fraj & Grenez	Experimenting with a synthesizer of disordered voices [page 2221]
14:20	Speed, Howard & Murphy	Developing and manipulating models of the vocal tract using VTK [page 24]
14:40	Horacek, Vampola, Laukkanen & Svec	Finite element simulation of vocal exercising effects on phonation [page 26]
15:00	Kiagiadaki, Chlouveraki & Bizakis	The use of quality of life questionnaires for the assessment of patients operated for benign laryngeal lesions: Comparative study of VHI and VoiSS in the Greek language [page 27]
15:20	Alpan, Schoentgen & Grenez	Automatic multi-category classification of disordered voices [page 28]
15:40	TEA	
WP4: Voice analysis methods		
16:00	Fourcin	The future use of hearing in clinical voice measurement [page 30]
16:20	Orlandi, Risaliti, Manfredi, Bocchi & Donzelli	Automatic extraction of cry episodes from newborn infant cry recordings [page 32]
16:40	Pabon	Instantaneous frequency modelling of dynamic F0 and SPL variation [page 34]
17:00	Dejonckere, Moerman, & Martens	Advanced acoustic analysis of substitution voices [page35]
17:20	Pabon	Towards a more flexible assessment of time/frequency grouping [page 36]
17:40	Davis	Improve the voice by using the ear and the brain [page37]

18:00	END OF SESSIONS	
20:00	Workshop Dinner	Merchant Taylors' Hall

THURSDAY 20th May		
INVITED KEYNOTE 2		
08:45	Julian McGlashan	Clinical applications of Speech analysis [page 38]
WP5: Pathological voice assessment		
09:30	Vaičiukynas, Gelžinis, Bačauskienė, Verikas & Vegiene	Fusing GMM and SVM for human voice-based classification of larynx pathology [page 40]
09:50	Arias-Londoño, Godino-Llorente, Markaki & Stylianou	On combining information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for automatic detection of pathological voices [page 41]
10:10	Vicsi, Klara.-Imre & Viktor	Voice disorder detection on the base of continuous speech [page 42]
10:30	COFFEE	
WS6: Professional voice assessment		
11:00	Lindström, Karjalainen, Södersten & Ternström	Voxlog: a new voice accumulator using both accelerometer and microphone sensors [page 42]
11:20	Koutsogiannaki, Pantazis, Stylianou & Laukkanen	Can vocal tremor features indicate vocal loading? [page 44]
11:40	Barlow	Changes in singer performance with different acoustic environment [page 46]
12:00	<i>Svec, Horacek, Vampola, Herbst, Miller, Havlik, Krupa & Lejska</i>	Acoustic and articulatory adjustments in operatic singing: Spectral analysis and magnetic resonance imaging [page 48]
12:20	<i>Vampola & Horáček</i>	Comparison of geometrical and acoustical characteristics of human supraglottal spaces for ordinary and singing voice using finite element models [page 50]
12:40	<i>Kirchhübel, Daffern & Howard</i>	A summary of the acoustical correlates of stress in speech (51)

13:00	LUNCH		
14:00	COST 2103 meetings (COST members only)		
15:30	TEA		
16:00	COST 2103 meetings (COST members only)		
17:00	END		
17:15		Choral Evensong	York Minster
18:30		Public engagement event	Bootham School, (see map inside front cover) followed by group meal in town

	FRIDAY		
08:45	COST 2103 meetings (COST members only)		
10:45	COFFEE		
11:00	COST 2103 Committee Meeting (COST members only)		
13:00	LUNCH		
14:00	End: Have a safe journey home!		

KEYNOTE 1: Recent Advances and Future Strategies for Speech Processing

Patrick Naylor

Imperial College London, UK

The source-filter model of human speech production is the starting point of many speech signal processing methods for analysis, synthesis and recognition. This paper will focus on speech analysis in the context of the classical source-filter model and will present recent advances in signal processing approaches to the study of the voice source. The first part of this paper will present a new technique that operates on the laryngograph (EGG) signal to determine the instants of glottal closure and opening with a very high degree of accuracy. This technique, known as SIGMA, provides a reference identification and segmentation of the larynx cycle in voiced speech. It is compared with two existing techniques in terms of function and performance. The second part of this paper will discuss the case when the laryngograph signal is not available. A new technique will be presented that, operating on the speech signal alone and without the benefit of the laryngograph signal, is able to estimate the glottal closure and glottal opening instants in voiced speech with remarkable accuracy. This non-invasive approach has the potential to be used in the study of closed quotient and pitch, for example. The third part of this paper gives a perspective of signal processing techniques - past, present and future - relevant to the study and analysis of speech. The key achievements to date will be briefly summarized and the future challenges set out in terms of research goals.

On the derivative of the EGG Signal for determining GCIs and GOIs

Christina-Alexandra Lionoudaki, Yannis Stylianou

Computer Science Department, University of Crete, Greece, and Institute of Computer Science - FORTH, Crete, Greece

The determination of GCIs and GOIs, is quite straightforward using Electroglottographic (EGG) signals. In practice, it is the simple time derivative of the Electroglottographic (DEGG) signal which is widely used for this purpose. The positive peak of the derivative corresponds to the closure instant of EGG and the negative peak to the opening instant. However, in several instances the peaks of the derivative are not easily detected, especially the negative ones. In these cases, the derivative (DEGG) forms many negative peaks and then the detection becomes more complicated. In this case, the GCI is firstly determined and then the GOI is then searched inside a predetermined from the GCI time-interval [2]. Moreover, many approaches impose a threshold on the differentiated EGG signal, which provide accurate results during voiced speech but are prone to errors at the onset and end of voicing. The SFS (Speech Filing System) [3] follows a threshold-based technique in DEGG signal. A correlation-based method, called DECOM [1], measures fundamental frequency (F0) and open quotient (Oq) and estimates the distance between two consecutive closing peaks and the distance between an opening peak and the consecutive closing peak. Therefore the derivative of EGG offers a simple way in detecting the important instances during the production of speech; the glottal closing and opening instants. In this work we suggest an alternative method to the simple derivative, that is usually used for generating the derivative of EGG signal, which is based on the spectral methods [5]. Spectral methods provide an elegant way to conduct first and higher order derivatives on discrete time data, with high accuracy. Experiments shown that spectral methods provide slightly better results comparing to the simple derivative approach, in terms of visibility of the major positive and negative peaks used for the detection of GCIs and GOIs.

Furthermore, we suggest a new way to differentiate the EGG signal for estimating the main glottal instants. The gradient of electroglottographic signal is performed with a method referred to as "Slope Filtering", which was proposed in [4]. The EGG signal is filtered in a frame-by-frame basis by an FIR system consisting of a short impulse response, taking into account the vicinity of the samples around the center of each frame. This approach shows to be robust in revealing the major peaks in the slope filtered EGG signal, even in cases where the quality of the EGG recordings is not of good quality. Then, the main glottal instants are easily detected using a simple thresholding approach. Contrary to the simple derivative of the EGG signal, the peaks can be well distinguished and uniquely specified in the slope filtered signal. The proposed method exhibits high accuracy of voiced segments, including the onset and offset regions. The experimental database consists of five sustained vowels (/a/, /e/, /i/, /o/, /u/) recorded by eleven male and five female young speakers. EGG data were recorded with the device from Laryngograph Ltd.

References:

- [1] Henrich, N., D'Alessandro, C., Doval, B., and Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation, *Journal of the Acoustical Society of America*, 1321-1332.
- [2] Howard, D.M., Lindsey, G., A., and Allen, B. Toward the Quantification of Vocal Efficiency. *Journal of Voice*, **4**, (3), 205-212, 1990.
- [3] Huckvale, M. (2008). *Speech Filing System: Tools for speech (SFS)*. Technical report, University College London. <http://www.phon.ucl.ac.uk/resource/sfs/>.
- [4] Turner, C., S. Slope Filtering: An FIR Approach to Linear Regression. *IEEE Signal Processing Magazine*. DSP Tips and Tricks, pages 159-163, November 2008.
- [5] L. N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2000.

this page has been left blank for your own notes

**A frequency domain approach for the determination of the
glottal opening instant in EGG waveforms**

George P. Kafentzis, Yannis Stylianou

*Computer Science Department, Univ. of Crete, Greece and Institute of Computer Science, FORTH,
Crete, Greece*

Glottal Opening and Closure Instants of the glottis are considered as important timings during speech production which have many applications in various speech areas such as voice quality assessment, speech analysis and coding, and speech synthesis and modifications. In the literature, there is a considerable effort from many researchers to automatically estimate those instants [1],[2]. For an accurate detection of the instants, the Electroglottograph (EGG) signal and its derivative (DEGG) provide significant information in that direction [3] and it is considered more robust than the estimators suggested based on the speech signal.

DEGG peaks are generally considered as reliable indicators of glottal opening and closing instants. More specifically, in time domain, the DEGG signal has a large positive peak at the instant of glottal closure (GCI). It is regarded by almost all researchers as a reliable indicator of the pitch onset time. Also, there is a negative peak of the derivative of EGG, and it is regarded as an indicator of glottal opening instance (GOI). The distance in time between the two instants results in the closed phase interval of the glottis. The identification of GCI and GOI are important for voice quality evaluation. Most of the proposed methods in the literature are based in time domain processing, as in [4]. However, there are cases where the peaks are not clearly distinguished in time domain for several reasons, such as poor recording quality of the EGG or misplacement of the electrodes on the skin. A frequency domain approach may be more useful or appropriate in these cases.

Our approach is based on considering the DEGG as a periodic signal and modelling the aforementioned peaks using triangular pulses. A first goal is to find the distance τ between the two peaks of the DEGG. By taking the Fourier Series coefficients of the DEGG model, and the by obtaining the magnitude of the Fourier Transform of the real part of these coefficients, it can be shown that the resulting spectrum has two positive large peaks at frequency points from where the positions in time domain of the two glottal instants can be easily derived. Then the distance τ can be estimated and accurately then estimate GCI and GOI of an EGG signal. The proposed approach has been found to be robust against additive noise and provides ways to determine GOI even in noisy EGG recordings.

Another goal is to examine if the spectral representation of the DEGG waveform of recorded and synthetic EGG signals can be correlated in some way with the speech waveform spectrum. This will have a major impact to speech modification and voice transformation techniques. Experiments will be conducted on a database of normal voices developed in our recording lab.

References:

- [1] Strube, H.W., (1974). Determination of the instant of glottal closures from the speech wave, *Journal of the Acoustical Society of America*, **56**, 1625-1629.
- [2] Smits, R., and Yegnanarayana, B. (1995). Determination of instants of significant excitation in speech using group delay function, *IEEE Trans Speech and Audio Processing*, **3**, 325-333.
- [3] Henrich, N., D'Alessandro, C., Doval, B., and Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation, *Journal of the Acoustical Society of America*, 1321-1332.
- [4] Drugman, T., and Dutoit, T. (2009). Glottal closure and opening instant detection from speech signals, *Proceedings of the International Speech Communication Association, Brighton*, 2891-2894.

this page has been left blank for your own notes

Normalised Time of Increasing Contact for Different Loudness Levels

P. J. Murphy

Department of Physics, University of Limerick, Limerick, Ireland.

The electroglottogram (EGG) signal provides a measure of vocal fold contact area and the first (DEGG) and second (D2EGG) derivatives provide estimates of the speed and acceleration of contact area, respectively. A measure of the closing phase, normalised time of increasing contact (NTIC), is estimated using the cycle peaks of the EGG and D2EGG signals. The point of maximum contact is taken from the EGG signal (this is the positive peak per glottal cycle). The point of the start of increasing contact is estimated using the D2EGG signal. The positive peak in the D2EGG signal occurs during the rapidly rising edge of the peak in the DEGG signal. One sample point back from the D2EGG peak is taken as the beginning of increasing contact – this point was initially determined empirically by viewing the corresponding point on the EGG signal which indicated the beginning of closure. The difference between these times is taken as the time of increasing vocal fold contact, which is used as an approximate estimate of vocal fold collision closing time. This estimate is then time normalised by dividing by the cycle duration to produce the normalised time of increasing vocal fold contact (NTIC). The measure is estimated for EGG recordings taken during the production of the vowel a/ phonated at different loudness levels (soft, neutral, loud and louder). NTIC was greatest for soft phonation and similar for neutral, loud and louder phonations. An interpretation of the results is given in terms of the underlying voice production mechanism.

Spectrum variations in the electroglottographic signal related to voice sound pressure level

Angela Libeaux¹, Sten Ternström¹, Nathalie Henrich²

¹KTH, Dept of Speech, Music and Hearing, Stockholm, Sweden, ²GIPSA-lab, Grenoble, France

Electroglottography (EGG) measures instantaneous vocal-fold contact area and is thus sensitive to the speed of contact and opening. It is easy to acquire, and, lacking the superimposed vocal tract resonances, its spectral envelope is much simpler than that of the airborne signal. Although the signal is not acoustic, one might expect that some of the variation in the spectral tilt of the airborne signal be present also in the EGG. Wideband spectra of EGG signals from sustained phonation were examined under various conditions, including changes of electrode position, vowel, subglottal pressure and SPL.

Recordings were made of subjects producing /pV/ utterances with simultaneous acquisition of EGG and intraoral pressure. Unlike the EGG signal amplitude, the EGG spectrum slope was quite insensitive to electrode placement. The EGG spectrum effects of vowel changes were small. The EGG spectrum envelope was found to be quite linear in dB/octave, from harmonics 2 to about 10, and often much higher. EGG spectrum slope change with SPL was generally greater in soft phonation than in loud phonation. Additionally, recordings from an existing database, of 8 trained male singers with audio and EGG, were analysed for EGG spectrum variation with SPL. The singers performed crescendo tasks on sustained tones, with a typical SPL variation of up to 20 dB from soft to loud. Fitting straight lines to wideband EGG spectra yielded slopes of -14 to -9dB/octave. The slope variation was small but in most cases positively correlated with overall SPL in the range 75-95 dB at 0.3 m. Also, there was in some cases a pronounced ripple in the EGG spectrum, due to double peaks in the time derivative of the closing EGG waveform.

Evaluation of inverse filtering approaches using subglottic data

George P. Kafentzis¹, Yannis Stylianou¹, Paavo Alku², Katrin Neumann³

{kafentz,yannis}@csd.uoc.gr, Paavo.Alku@hut.fi, Katrin.Neumann@em.uni-frankfurt.de

¹ CSD – UOC and ICS-FORTH, Crete Greece, ² LAASP, HUT, Helsinki, Finland, ³ Goethe-University of Frankfurt / Main, Germany

In all proposed models of speech production, Inverse Filtering (IF) is a well known technique for estimating the glottal flow waveform from the speech signal, which acts as a source in the vocal tract system. The estimation of glottal flow is of high interest in a variety of speech areas, such as voice quality assessment, speech coding and synthesis as well as speech modifications. There are several methods in the literature for extracting the glottal flow signal from the speech signal. Most of them are based on Linear Prediction Analysis (LPA). A major obstacle in comparing and/or suggesting improvements in the current state of the art approaches is simply the lack of real subglottic data. In other words, the results obtained from various inverse filtering algorithms, cannot be directly evaluated because the actual glottal flow waveform is simply unknown. The database compiled by K. Neumann [1] containing subglottal, supraglottal, and EGG recordings of sustained vowels /a/, /e/, /i/, /o/, /u/, from a male and a female subject, will be therefore, of a great help in evaluating inverse filtering techniques.

In this work, we present an evaluation of three well known methods of IF. Plumpe and Quatieri [2] suggested a method based on Closed Phase Covariance LPA, using statistics on the first formant frequencies during a pitch period to obtain the closed phase interval. Magi et al [3] suggested an IF method based on Stabilized Weighted LP Analysis, in which a short time energy window controls the performance of the LP model, and Alku et al [4] proposed an IF method based on a Mathematically Constrained Closed Phase Covariance LPA, in which mathematical constraints are imposed on the conventional covariance analysis which results in more realistic root locations on the z-plane.

The three inverse filtering approaches have been applied on the database of subglottic data mentioned above. Then another problem arises. This is the way of comparing the outputs obtained from these approaches with the real subglottic and supraglottic data. A simple Signal to Noise Ratio (SNR) cannot be applied in this case since all signals contain different phase information. Nor an SNR, taking into account magnitude only information is also appropriate in this case because of major differences between the signals. Therefore, we suggest a subjective and a qualitative objective way for the evaluation procedure. The subjective evaluation is based on subjective tests developed for evaluating speech coders such as DMOS (differential Mean Opinion Score) and ABX tests (listeners select if X signal sounds closer to A or B signals). For the objective qualitative evaluation, we consider comparing the probability density functions of parameters such as open quotient (OQ) between the real and estimated data.

References:

- [1] Neumann K, Gall V, Schutte HK, Miller DG (2003) A new method to record subglottal pressure waves: potential applications. *Journal of Voice*, 17, pp 140-159.
- [2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7(5), pp. 569-586, 1999.
- [3] Carlo Magi, Jouni Pohjalainen, Tom Bäckström, Paavo Alku, "Stabilised weighted linear prediction", *Speech Communication*, v.51(5), pp. 401-411, May, 2009
- [4] Paavo Alku, Carlo Magi, Tom Bäckström, "Glottal inverse filtering with the closed-phase covariance analysis utilizing mathematical constraints in modelling of the vocal tract", *Journal of Logopedics, Phoniatrics, Vocology*, Vol. 34, No. 4, pp. 200-209, 2009

this page has been left blank for your own notes

Assessing the performance of a new tool for videokymographic images analysisC. Manfredi*¹, L. Bocchi¹, M. Calisti¹, G. Cantarella², G. Peretti³¹*Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy,*²*Otolaryngology Department, Ospedale Maggiore Policlinico Mangiagalli e Regina Elena, Fondazione IRCCS, Milano, Italy,* ³*Otolaryngology Clinic, Spedali Civili di Brescia, Brescia, Italy*

Videokymography (VKG) is a new technique that delivers high-speed images from a single line selected from the whole videolaryngostroboscopy (VLS) image, allowing for a detailed analysis of vocal fold vibration [1]. Recently, we have developed a new tool, named VKG Analyser, to obtain objective parameters from VKG images, also in case of glottal incompetence. VKG-Analyser is provided with a user-friendly interface to manage patients' data also with the help of plots and pictures and to perform real time analysis of VKG images. It evaluates left-to-right period (Rper), amplitude (Ramp), opening/closing ratios (Roc), and phase symmetry index (PSI) [2] that help the clinician to obtain objective evaluation of results with reliable and reproducible measures. The tool has been successfully tested on synthetic [1] and real images, to assess its reliability and usefulness in the clinical practice. In this work, we focus on results obtained with the recently improved version of VKG-Analyser, where image boundary and snake initialization have been optimised. Also, a zooming feature has been added, to allow for a more precise evaluation of the proper thresholds. The tool has been applied to a set of 26 patients collected in two different clinics (Milano and Brescia, Italy) and cross-checked by local clinicians. Specifically, the data set includes 14 patients affected by vocal fold pathologies (1 vocal fold erythroplasia, 9 vocal fold polyp or cyst, 5 sequelae of cordectomy, 6 males and 8 females, age ranging from 28 to 63 years) and 12 control cases (6 males and 6 females, age 25-56 years) with normal vocal folds. Twenty consecutive frames from the VKG recording of each patient were analyzed.

Results are shown in the Table. Notice higher std values for the pathological group (c.f. controls), corresponding to a higher degree of instability for all parameters. Data were also compared using Student t tests for unpaired data (last row of the Table). A significant difference ($p < 0.05$) between the two groups was found for all four parameters. The objective quantitative analysis of VKG images provided by VKG-Analyser might have a role as a tool to discriminate normal from pathological vibrating vocal folds and to identify subtle abnormalities of vocal fold vibration, especially in case of chronic dysphonia where only mild organic changes are visible.

	Roc		Ramp		Rper		PSI	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
healthy	1,437	0,304	1,052	0,192	1,040	0,468	0,005	0,039
pathological	2,551	1,142	1,191	0,620	1,289	0,943	-0,057	0,122
p value t-test	<<0.0001		0,0169		0,0058		0,0002	

References:

- [1] Švec, J., F. Šram, H. K. Schutte, (2007). Videokymography in Voice Disorders: What to Look For? , *Annals of Otolaryngology, Rhinology & Laryngology*, 116, 3, 172-180.
- [2] Manfredi, C., L. Bocchi, S. Bianchi, N. Migali, and G. Cantarella, (2006). Objective vocal fold vibration assessment from videokymographic images, *Biomedical Signal Processing and Control*, 1, 129–136.

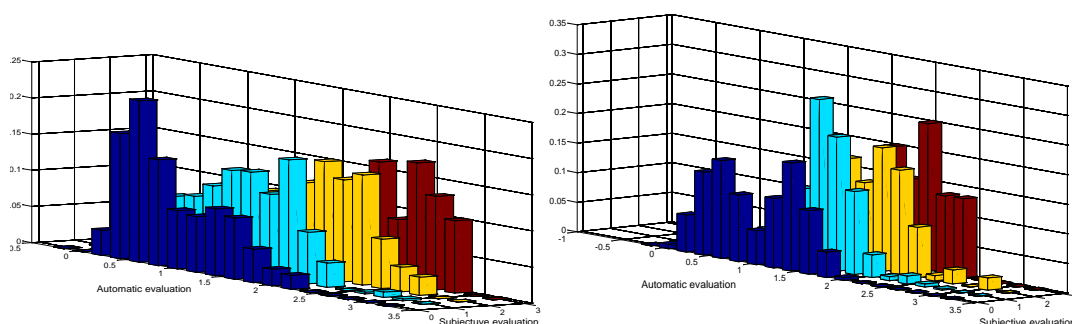
this page has been left blank for your own notes

Towards objective evaluation of perceived roughness and breathiness based on Mel-frequency cepstral analysis

N. Sáenz-Lechón, R. Fraile, J.I. Godino-Llorente, R. Fernández-Baíllo, V. Osma-Ruiz,
J.M. Gutiérrez-Arriola, J.D. Arias-Londoño

Circuits & Systems Engineering Dep. Escuela Universitaria de Ingeniería Técnica de Telecomunicación.
Universidad Politécnica de Madrid – Campus Sur. Carretera de Valencia km 7. 28031 Madrid, SPAIN

The social function of voice for human beings is so relevant that perception is key in the evaluation of voice quality. However, perceptual evaluation has some drawbacks such as time consumption and a non-diminishable degree of variability due to its subjective nature, both among distinct evaluators and even among different evaluations performed by the same listener. Nevertheless, nowadays objective and reliable measures of voice quality are being demanded in order to perform unbiased assessments of therapies, calculation of insurance compensations in case of illnesses or accidents, etc. For this reason, to present a significant effort has been done in analysing correlations among objective acoustic measures of voice quality and labels assigned after perceptual evaluation. To authors' knowledge, most of the published works so far deal with perceptual assessments carried out according to the GRBAS scale. Promising results related to G and B have been reported, while R seems to be more challenging and A and S have seldom been analysed. Within this paper, the authors report on an experiment of automatic labelling of perceived voice roughness (R) and breathiness (B). In contrast to other previous works, the main objective of the experiment has not been to correlate objective measures to perceived R and B, but to automatically evaluate R and B. For this purpose, the MEEI database of voice disorders has been perceptually evaluated by an expert speech therapist listening to records both of sustained vowels and running text. The R and B labels have been later used to train a system that extracts the first Mel-frequency Cepstral Coefficients (MFCC) of the sustained-vowel phonations and, afterwards, it fits a classifier to estimate the corresponding degrees of roughness and breathiness. The first MFCC have been chosen because they fairly model the spectral decay of the glottal source. From the available records, 70% have been used for training and the remaining 30% for testing. The results are summarised in the histograms below. Pearson correlation factors between subjective and automatic labelling of 0.6 have been achieved.



Roughness

this page has been left blank for your own notes

Correlations between subjective and objective assessments of voice quality

Maria Astrinaki¹, Debora Kiagiadaki², Miltiadis Vasilakis¹, and Yannis Stylianou

¹ Computer Science Dept, Univ. of Crete, Greece, ²School of Medicine, Univ. of Crete, Greece

In this work we present correlation results between objective and subjective assessments of voice quality using recordings from 21 patients before and after their treatment. Patients consisted of both men and women, smokers and non-smokers, who suffered from benign laryngeal lesions, such as vocal fold polyps, Reinke's edema, and leukoplakia, a pre-malignant laryngeal lesion. We also studied a case of a male with a verrucous carcinoma of the larynx. All patients underwent microlaryngoscopy and surgical treatment. The Voice Handicap Index (VHI) was used as the first subjective measurement. VHI is a 30-point scale questionnaire, where patients grade the impact of their symptoms in everyday life, before and 3 months after the surgical treatment. The clinician-based GRBAS scale for perceptual evaluation of voice quality was also used. The measurements were conducted by 3 logopedists, before and after treatment, in a randomized blind manner. The sum of the average score of each element in the scale was extracted as the total score per case. Finally, the clinical evaluation, regarding the existence of pathology pre- and post-treatment, was taken into account on this study as well.

Several methods were used to gain objective features of voice quality; specifically, Short-Time Absolute Jitter measure from the Spectral Jitter Estimator (SJE) [1], and Absolute Jitter. Absolute Shimmer and Harmonics-to-Noise Ratio (HNR) estimated from both the Multi-Dimensional Voice Program (MDVP) and PRAAT. HNR quantifies the waveform irregularity of voice signals; jitter and shimmer reflect perturbation in periodicity and amplitude respectively. Jitter is modelled as the movement of one of the two periodic phenomena with respect to the other. The short-term estimates of SJE were further used to calculate the Over metric that depicts the temporal behaviour of jitter, as a percentage of pathological estimates, in relation to a pre-determined threshold [2].

For approximately 95% of the pre-operative cases, all the subjective and objective measurements are correlated. The correlation drops to 38% though, when comparing post-operative assessments. If only Over and WHI, from objective to subjective measurements, correspondingly, are taken into account for the post-operative cases, then the correlation is 78%. In general, for both pre- and post-treatment cases, the perceptual GRBAS scale is poorly correlated with other measurements. On the other hand, the objective Over measurement reflects more accurately both the perceptual evaluation of patients and the clinical evaluation of medical doctors. However, the combination of subjective and objective methods is not always enough to come to a sufficient conclusion. Consider this interesting example from a specific case, where the logopedists and the patient both perceived some kind of voice pathology, while Over on the contrary was measured in the healthy region; according to the clinical evaluation, there was an obstruction that interfered with voice production, but was not associated with the vocal cords, and thus no jitter occurred. In spite of these first encouraging results, further study has to be made, by using a larger sample of patients as well as with a bigger variety of pathologies. Additionally it would be of benefit to gain GRBAS scores from a larger number of evaluators.

References:

- [1] Miltiadis Vasilakis, Yannis Stylianou "Spectral jitter modeling and estimation", Biomedical Signal Processing and Control, Volume 4, Issue 3, July 2009, Pages 183-193, Special Issue on New Trends in Voice Pathology Detection and Classification - M & A of Vocal Emissions
- [2] Emissions2. Miltiadis Vasilakis, Yannis Stylianou "Voice Pathology Detection Based on Short-Term Jitter Estimations in Running Speech", Folia Phoniatria et Logopaedica, Volume 61, Issue 3, June 2009, Pages 153-170

this page has been left blank for your own notes

Experimenting with a synthesizer of disordered voices

J. Schoentgen, S. Fraj, F. Grenez

L.I.S.T., Université Libre de Bruxelles, CP 165/51, 50, Av. F.-D. Roosevelt, B-1050 Brussels, Belgium

The topic of the presentation is the synthesis of disordered voices. Disordered voices here designate voices the timbre, loudness or pitch of which is perceived as anomalous or deviant. Motivations for simulating disordered voices by machine are the discovery of acoustic cues of perceived abnormal timbres, the preparation of boundary markers in the framework of the perceptual assessment of disordered voices as well as the coaching of trainees in the perceptual assessment of voice, and the validation of clinical speech analysis software. The synthesis of disordered voices is a topic that has been studied occasionally only and the simulations that have been presented in the literature have mainly been based on early formant synthesizers.

The synthesizer that is used here, and which has been presented in previous meetings, is constructed to enable controlling sample by sample the vocal frequency, gain and spectral tilt of the glottal excitation signal. The vocal tract is simulated by means of a concatenation of elementary cylinders of unequal section. The simulation of the wave propagation therein enables taking into account the interaction between glottal source and tract via an algebraic model of the glottal airflow proposed by Titze. In addition, the synthesizer comprises stochastic or deterministic models of vocal jitter and shimmer, vocal frequency and amplitude tremor, breathiness, diplophonia, biphonation, random vibrations and asthenia or strain.

The presentation focusses on the perceptual experiments based on simulated stimuli. Experiments that are reported have involved i) mixtures of human and synthetic sustained vowels, and ii) synthetic jitter and additive noise in vowels [a], [i], [u] as well as vowel pairs [ai] and [ia]. The tasks have been: (i) the perceptual discrimination of mixed human/artificial stimuli; (ii) the classification of unknown stimuli into human or synthetic categories; (iii) the discrimination between synthetic stimuli comprising different amounts of vocal jitter, shimmer and additive noise; (iv) the perceptual assessment of the same stimuli by trained speech therapists according to levels of perceived grade, roughness and breathiness.

Results are as follows.

- (i) Naive and expert listeners have been unable to distinguish between sustained human and artificial modal as well as disordered speech sounds
- (ii) Experts in speech synthesis have been more accurate, however, than non-experts when categorizing modal stimuli and clinicians have been more accurate than non-clinicians when categorising disordered stimuli
- (iii) Discrimination experiments involving pairs of synthetic stimuli suggest that modulation noise (i.e. jitter and shimmer) and additive noise (i.e. breathiness) cause timbres that are perceived as the more distinct from each other the larger the difference between the amounts of modulation and additive noise
- (iv) Perceptual evaluation experiments with synthetic stimuli show that speech therapists assign higher levels of "grade" (i.e. hoarseness) to stimuli with larger amounts of additive noise or modulation noise, whereas they assign higher levels of breathiness to stimuli with more additive noise and higher levels of roughness to stimuli with higher levels of jitter and shimmer.

Pitch has been a compounding factor. For a given quantity of modulation noise and additive noise, stimuli have been perceived the less hoarse or rough the higher the vocal frequency. Pitch has appeared to influence the perception of breathiness less than the perception of roughness or hoarseness. Vowel quality has been observed to have a minor influence only and evolving vowel qualities [ai] and [ia] have been rated as reliably as static qualities [a], [i] and [u].

this page has been left blank for your own notes

Developing and Manipulating Models of the Vocal Tract using VTK

Matt Speed, David M Howard, Damian Murphy
Audio Lab, University of York, Heslington, UK
mdas100@ohm.york.ac.uk

Over the past two decades, MRI (magnetic resonance imaging) scanning of the vocal tract has taken on considerable significance and without doubt sprung to the forefront of multiple fields in voice research. It has applications in the extrapolation of parameters for articulatory modelling, the generation of grids for finite-element and finite-difference based acoustics modelling and building visual models of the vocal tract for speech therapy and language tuition.

An inevitable restriction on the accuracy of MR imaging is imposed by the capture times required and the subject's ability to maintain a particular tract arrangement with/without phonation. Capture times have however shown very considerable improvement since 1991 when times were as high as 3 minutes for a single phonation, through around 30 seconds in the 2002 work of Engwall, to sub-second imaging achieved for instance in the recent RT-MRI (Real-time MRI) work of Bresch et al.

While this capability is of course a fantastic resource for the researcher, handling the data produced by MRI scanning can represent a significant undertaking. After collection of the imaging follows a demanding sequence of post-processing. Although tools do exist to process the output of MR imaging, voice specific analysis falls outside the scope of these predominantly neuroimaging focussed applications. Developing bespoke tools for this specialised postprocessing can therefore demand a considerable investment in time and skills.

VTK (The Visualization Toolkit) is an open source, OpenGL-based graphics toolkit, geared towards medical imaging. It is predominantly C++ based, yet features wrappers for Python, Java and tcl. Its design is object oriented, and processing is channelled through a strictly defined graphical pipeline. Whilst powerful and freely available, VTK's greatest advantage is surely the ease with which standalone applications can be developed. With a reasonable grasp of one of VTK's supported programming languages, MRI data can be processed and developed into powerful models in a fairly short space of time.

This presentation provides an overview of the key concepts of VTK development. The toolkit's application to vocal tract modelling is then demonstrated across various stages of MRI postprocessing, from data acquisition through to generating accurate 3D models.

It continues to show how these models can be manipulated to provide other representations of the vocal tract, such as lower dimensionality intersections (midsagittal cross-sections), developing grids for finite-element and finite-difference computation and useful visualisations of simulation.

this page has been left blank for your own notes

Finite element simulation of vocal exercising effects on phonation

Jaromir Horacek¹, Tomas Vampola², Anne-Maria Laukkanen³, Jan G. Svec⁴

¹*Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Prague,* ²*Department of Mechanics, Biomechanics and Mechatronics, Faculty of Mechanical Engineering, Czech Technical University, Prague,* ³*Department of Speech Communication and Voice Research, University of Tampere, Finland,* ⁴*Laboratory of Biophysics, Dept. Experimental Physics, Palacky University Olomouc, Czech Republic*

Earlier observations have shown that SPL tends to increase after vocal exercising on semiocclusions like voiced fricatives and phonation into a tube. Recent CT results of the vocal tract show certain changes during and after phonation into a tube. (Laukkanen et al. and Vampola et al., submitted to JSLHR and JASA). These changes include a widening of the front part of the oral cavity and the lower pharynx (just above the epiglottis), and a narrowing of the region between the lower part of the tongue body and the back wall of the pharynx. Acoustic recordings showed somewhat lower formants F1, F2, F4 and F5 and somewhat higher F3. SPL increased in average 2.3 dB. The present study investigated the effects of the vocal tract changes on sound energy transfer from the vocal tract. Two 3D finite element (FE) computer models were constructed, based on CT measurements of a female subject phonating on [a:] before and after phonation into a tube (30 cm in length, 8 mm in inner diameter, made of glass). First, a short pulse of acoustic flow velocity with a broadband frequency range up to 5 kHz was used as the acoustic excitation of the FE models at the vocal folds level and the acoustic pressure at a distance of 4 cm in front of the mouth was numerically simulated by the transient analysis and the power spectral densities of the response computed. Then the harmonic analysis was performed in the vicinity of the formant frequencies showing the acoustic resonance mode shapes of vibrations in the vocal tract. Average values of SPL generated inside the vocal tract and computed from the average values of the pressure oscillation amplitudes were compared with the SPL simulated in front of the mouth in order to display the efficiency of sound energy transfer at different frequencies. The lower formants F1-F3 represent classical vibration modes also solvable with a 1D vocal tract model. For higher formants, instead, more complicated transversal 3D modes of vibration are prominent and therefore they require a 3D modeling approach modeling the effects of piriform sinuses and valleculae. Comparison of the SPL computed inside and outside the vocal tract showed that formants differ in their cost-efficacy, F4 (at about 3.5 kHz, i.e. at the speaker's or singer's formant region) being the most effective. The human hearing threshold is also relatively low between 2 and 4 kHz. Consequently, a sound energy concentration around 3.5 kHz (F4-F6) region is an effective tool for communication both from the point of view of production and perception. Based on the FE modelling results, the changes observed in the vocal tract are able alone to explain the measured increase in SPL. The results suggest that exercising on semi-occlusions help in optimizing the vocal tract setting for improved energy transfer from the vocal tract and thus improved communication and vocal economy.

The use of quality of life questionnaires for the assessment of patients operated for benign laryngeal lesions: Comparative study of VHI and VoiSS in the Greek language

Debora Kiagiadaki, Gregory Chlouverakis and John Bizakis

School of Medicine, Univ. of Crete, Greece

Subjective evaluation of voice quality is a major part of multidimensional assessment of voice pathology. Numerous questionnaires have been developed and used. Voice Handicap Index is the most popular and has been translated in many languages. It is a 30-item questionnaire, with 3 subscales (10-item): functional-VHI1, physical-VHI2 and emotional-VHI3. Each question is graded by the patient with 0 to 4 according to how severely it affects his everyday life. Maximum score is 40 for each subscale and 120 is the total maximum score.

Voice Symptom Scale (VoiSS) is an enriched form of VHI. It is also a 30-item point scale with 3 subscales like VHI, but with a different diarthrosis/structure: impairment-VoiSS1, physical-VoiSS2 and emotional-VoiSS3, with 15, 7 and 8 items accordingly. The 2 questionnaires have 8 questions in common.

Recently VHI has been translated and validated to the Greek language [1]. In our study, VoiSS was translated in the Greek language and back translated by a native English speaker, teacher of English language. The study included 23 patients with benign laryngeal lesions who underwent microlaryngoscopy. Patients completed both questionnaires before and 3 months after surgery and the results were analyzed.

The statistics reveal a significant correlation between all subscales. The most powerful correlation is this between "functional" subscale of VHI (VHI2) and "physical" subscale of VoiSS (VoiSS1) pre and post treatment, and between "emotional" subscales of both VHI and VoiSS (VHI3 and VoiSS3) post treatment. The correlation is significant comparing the post treatment change in both questionnaires. In emotional subscale changes significantly post operatively in both questionnaires. The overall change of VoiSS score is statistically significant whereas the overall change of VHI score shows a trend and doesn't change significantly.

References:

- [1] M.E. Helidoni, T. Murray, J. Moschandreas, C. Lionis, A. Printza, and G.A. Velegrakis. "Cross-Cultural Adaptation and Validation of the Voice Handicap Index Into Greek", *Journal of Voice*, DOI: 10.1016/j.jvoice.2008.06.005, Aug. 27th, 2008

Automatic multi-category classification of disordered voices

Ali Alpan¹, Jean Schoentgen^{1,2}, Francis Grenez¹

¹ *Laboratoire d'Images, Signaux et Dispositifs de Télécommunications,*

Université Libre de Bruxelles, Brussels, Belgium, ²*National Fund for Scientific Research, Belgium*

The objective is to present results on automatic classification of disordered voices. Experiments have been carried out with two corpora. The first is the Kay Elemetrics Voice Disorders Database. This database is split into normophonic and pathological utterances. Analyses and categorisation have been carried out with vowels [a]. The second corpus comprises sustained vowel [a] produced by 251 normophonic and dysphonic speakers. The perceived degree of hoarseness (grade scale of GRBAS) for each item has been determined by five judges. This corpus is referred to as corpus 2 hereafter. Variogram analysis has enabled calculating inter-cycle dysperiodicities (sample-by-sample). The dysperiodicities have been summarized via a segmental signal-to-dysperiodicity ratio.

The generalized variogram [1-2] enables tracking vocal dysperiodicities in running speech without an a priori knowledge of the average cycle length. It is based on the comparison of two analysis frames that are positioned in neighbouring speech cycles. The analysis is bilateral to avoid comparing the present frame to a frame belonging to a different phonetic segment. The deterministic evolution of the signal amplitude owing to sound onsets and offsets or sound-specific intensities are taken into account by equalizing the signal energy across analysis frames. Segmental signal-to-dysperiodicity ratios (SDR) are calculated to summarize vocal dysperiodicities. The speech and dysperiodicity signal are band-filtered and the SDRs computed for each band. Finite differences have been included to take into account the temporal evolution of the by-band SDRs. The stimuli have been categorized automatically by means of Support Vector Machines (SVMs), which have become a popular tool for categorical classification. Advantages of SVMs are the absence of local minima problems and the feeble number of parameters that must be fixed by the experimenter [3]. Test and training corpora have been distinct (six-fold cross-validation).

The binary classification has been carried out on a subset of the Kay corpus, including 53 normal and 173 pathological vowels [a]. Results show a correct classification rate of 97% when using segmental signal to dysperiodicity ratio in 7 bands and their first and second finite differences. Three categories have been generated for the second corpus. The first includes stimuli with a grade score less than 0.2. The second includes stimuli with a grade score of 1.2, 1.4 and 1.6 and the last one includes those with a grade score higher than 2. Preliminary results show correct classification rates close to 70%.

References

- [1] A. Alpan, A. Kacha, F. Grenez, and J. Schoentgen, "Assessment of vocal dysperiodicities in connected disordered speech", in Proc. Interspeech, Antwerp, Belgium, pp. 1178-1181, August 2007
- [2] Kacha, A., Grenez, F., and Schoentgen, J., "Estimation of dysperiodicities in disordered speech", Speech Com., Vol. 48, pp.1365-1378, 2006.
- [3] Bennett, K. P., Campbell, C., Support vector machines: hype or hallelujah?, ACM SIGKDD Explorations Newsletter, v.2 n.2, p.1-13, Dec. 2000.

this page has been left blank for your own notes

The future use of hearing in clinical voice measurement

Adrian Fourcin

Emeritus Professor, Department of Phonetics and Linguistics, University College London, UK

The most widespread current methods of voice measurement make no use of established relevant knowledge of hearing in spite of the obvious importance of the heard voice to patient and family. Nor indeed is any real note taken of the practical need to measure voice in connected speech. Present signal processing algorithms are largely concerned only with the measurement of, mathematically defined, aspects of sustained vowels. Three aspects are discussed of the ways that auditory processing may prove to be of future clinical importance in voice measurement.

- The temporal span of the auditory monitoring of voice production appears to extend over a considerable range. At the level of tens of seconds, connected fluent speech with auditory control of production, tends to give equal duration proportions to voice and silence-plus-voiceless fricatives. This provides a simple clinical criterion for one aspect of a return to normality.
- The cycle to cycle vocal fold vibrational irregularities of duration, that are termed jitter in sustained vowel measurement, are under quite tight auditory control in normally produced connected speech. Impairment in hearing entails impairment in production. Voice pathology by itself can be usefully measured and quantitatively evaluated using algorithms that are based on hearing parameters. These techniques can then in turn be used to measure clinically relevant aspects of pitch perception in normally produced connected speech that are not readily open to psychophysical determination.
- Phonosurgical interventions sometimes lead to voice improvements that are not associated with tangible changes in standard measures of pitch and loudness. Vocal fold closed (contact) phase is an example of a physiological component of connected voice that is not ordinarily measured but that makes an important contribution to perceived voice quality. Clinically useful auditory criteria for the evaluation of voice contact phase in connected speech can be defined by extending the hearing based techniques developed for pitch measurement.

The measurement examples that are discussed here are all based on a combination of physiological with acoustic data. Our present methods of inverse filtering do not take account of subglottal coupling and the full development of auditorily inspired voice measurement will come when we can apply acoustic algorithms that are able to extract what we hear from what we say.

this page has been left blank for your own notes

Automatic extraction of cry episodes from newborn infant cry recordings

S. Orlandi¹, C. Risaliti¹, C. Manfredi^{1*}, L. Bocchi¹, G. P. Donzelli²

¹Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy, ²Department of Paediatrics, AOU A. Meyer – Università degli Studi di Firenze, Firenze, Italy

Premature babies are tiny and fragile, and are often connected to monitors and sensors that detect changes in a baby's condition and environment. Great efforts are devoted to the development of non-invasive monitoring devices and reliable analysis techniques. Among them, research is carried on concerning newborn infant cry analysis that could reflect the development and possibly the integrity of the central nervous system. Moreover, cerebral blood oxygenation is affected by stress caused by the effort required during crying due to impaired auto-regulation of premature babies [1] and may adversely influence their development [2]. In previous work we have studied the distress occurring during cry in full-term and in pre-term newborn infants to point out possible differences between the two groups of subjects [3]. Recordings, lasting from half an hour up to several hours, were manually analyzed to find out significant cry episodes. This step is time-consuming and implies subjective judgment, thus making results less reliable. Hence, an automatic detection of cry and "noise" (e.g. non-cry) intervals from the whole recording is advisable.

In this work the Otsu method [4] is applied to the Short-Term Energy measure (STE) histogram of the signal. The method provides the optimal threshold t_u for the separation of a bimodal histogram minimizing the intra-class variance of the two resulting classes (e.g. "cry" and "noise"). However, in case of energy oscillation around the threshold level, several disjoint cry episodes are detected instead of a single one. This problem has been solved by iteratively applying the Otsu algorithm to the histogram of the "noise" class to detect a lower threshold t_l , and applying a hysteresis thresholding where t_u is the upper threshold and t_l the lower threshold: 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. This procedure identifies two sets representing the starting and the ending points of the cry episodes.

The method has been applied to a group of 30 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 (Critical Care Unit of the Children Hospital A. Meyer and Nuovo Ospedale S.Giovanni di Dio, Scandicci, Firenze, Italy). Audio signals (0.5-1hour length) were recorded with a unidirectional microphone (Shure SM58) with sampling rate $F_s=44$ kHz and 16 bit resolution. The assessment of the performance of the algorithm has been tested using a set of recordings where cry episodes were manually labelled. We defined a correctly detected cry episode as an episode overlapping for at least 50% with a manually labelled one. An episode that does not overlap a manually identified one is defined as a false positive, while a manually labelled episode that has not been detected is defined as a false negative. The analysis has provided a fairly high percentage of correctly detected cry episodes (85.77%) and a low percentage of false positive events (6.13%).

References:

- [1] Goberman, A.M., Robb, M.P., (1999). Acoustic examination of preterm and full-term infant cries—the long-time average spectrum, *J Speech Lang Hear Res.*, **42**, 850–86
- [2] Lou H.C., Lassen N.A. & Frii-Hansen B., (1979). Impaired autoregulation of cerebral blood flow in the distressed new born infant, *Journal of Paediatrics*, **94**, 118-121.
- [3] Orlandi, S., Bocchi, L., Calisti, M., Manfredi, C., Donzelli, G.P., (2009). Recovery of oxygen saturation level in newborns, *Proc. Int. Workshop MAVEBA 2009*, 14-16 Dec., Firenze, Italy
- [4] Otsu, N., (1979). A threshold selection method from gray-level histograms, *IEEE Trans. Sys. Man & Cyber.*, **9**, 62-66.

this page has been left blank for your own notes

Instantaneous frequency modelling of dynamic F_0 and SPL variation

Peter Pabon,

Royal Conservatory, 2595 CA Den Haag, Voice Quality Systems, Utrecht University.

The Fundamental frequency F_0 and Sound Pressure Level (SPL) delineate the phonetogram or Voice Range Profile. They appear in the graph as static framing parameters that set out the range of the voice, but in real life the F_0 and SPL parameter are far from static and exist only as dynamic modulation "signals". Both signals can be thought to consist of three components that each represent a different aspect of voice quality: 1) an offset or slow trend that marks the average F_0 or SPL, 2) a faster movement in the form of a quasi stationary sinusoidal shaped modulation (the vibrato) and 3) fast irregularly fluctuations (noise); the jitter in the F_0 and the shimmer in the SPL curve. For the characterization of the periodic modulation component an instantaneous frequency (IF) description is easily recognized as the best model, but it is good to realize that this IF model may also apply equally well with the other (slow or fast) aperiodic components. That the IF model is more than a symbolic description, becomes apparent when the modulation signals are transformed to a phasor or analytical signal representation. This basically makes it simple to specify a modulation frequency and modulation amplitude value with only a fraction of a modulation cycle. When we want to assess the dynamic dependency of vibrato and other modulation aspects using a phonetogram mapping, then this excellent short term behaviour makes the complex IF model the best candidate. Central problems with this mapping are (1) the sampling of the modulation depends on the F_0 itself. Period-to-period perturbations thus appear at a variable frequency distance compared to the external clock of trend and vibrato modulations. (2) modulations may proceed while the carrier (voicing) can be intermittently unavailable or disturbed. How to deal with discontinuity?

In this paper the dynamic modelling, the separation and the settling of the three components is explored. Nonlinear filter adaptation schemes are discussed that can jump and clinch or speed up settling times.

Advanced acoustic analysis of substitution voices

P.H.Dejonckere,^{1,2} M.B.J.Moerman,^{1,3} J.P.Martens,⁴

¹*Institute of Phoniatics, University Medical Centre Utrecht, Utrecht, The Netherlands,* ²*Catholic University of Leuven, Belgium,* ³*Department of ENT/head and Neck Surgery, AZ Jan Palfijn-Maria Middelaers Gent, Ghent, Belgium,* ⁴*Electronics and Information Systems Department, Ghent University, Ghent, Belgium.*

Substitution voicing (SV) – i.e. with a voice that is not generated by 2 vocal folds - cannot be accurately evaluated with the commonly used programs for acoustic voice analysis, which are intended for ‘common’ dysphonias and quasi-periodic voices. An analysis program “Ampex” (Auditory Model Based Pitch Extractor) created and further developed by Van Immerseel & Martens (JASA 1992;91:3511) has proven to be able to extract in a valid way the period in irregular signals with background noise. It also detects low frequency components (< 0.1 KHz) and is suited for running speech. It has been efficiently used for spasmodic dysphonia voices and for (tracheo-)esophageal voices and appears to be basically valid.

This program is now for the first time used on a large database of various kinds of substitution voices. For this multi-center trial, data of 116 patients with surgery for advanced laryngeal cancer were collected from 7 European academic hospitals : 11 cases of fronto-lateral laryngectomy / Tucker’s procedure; 31 cases of total laryngectomy with cricopharyngeal myotomy, 15 cases of total laryngectomy without cricopharyngeal myotomy, 22 cases of cricothyroid(epiglottid)ectomy; 37 cases of cordectomy (from type III on). The acoustic analysis was achieved on running speech (reading a short text, phonetically balanced) in 5 different languages. 8 parameters were measured :

PVF/PVS: the proportion of voiced frames and voiced speech frames. The better the voice, the highest the percentages.

AVE: the average voicing evidence. The more regular (periodic) the voice frames, the higher the AVE.

VL 90 parameter: the 90th percentile of the voicing length distribution (the number of consecutive voiced frames). Phonatory breaks reduce this parameter.

Jitter: Period to period variability.

Corrected jitter: The correction means that only frames with a reliable fundamental frequency (Fo) are taken in account.

PFU: The percentage of frames with “unreliable” Fo. Frequency shifts make Fo unreliable.

T max : The maximum length of speech without pause.

Variance analysis demonstrates that – at group level - each of the 8 acoustic parameters (except the uncorrected Jitter) significantly differs across the 5 different types of laryngeal surgery (each of them being characterized by a different anatomical vibration structure).

As acoustic parameters have different units, a z-transformation was first achieved (and, when relevant, a sign inversion) before creating a global acoustic score combining all 8 parameters without weighting. This global score significantly differs across the 5 different types of laryngeal surgery.

Performing or not a cricopharyngeal myotomy during a total laryngectomy significantly influences the acoustic parameters of the tracheo-esophageal voice.

These conclusions may have practical consequences for making surgical choices when several approaches are equivalent from an oncological point of view.

Towards a more flexible assessment of time/frequency grouping

Peter Pabon,

Royal Conservatory, 2595 CA Den Haag, Voice Quality Systems, Utrecht University.

Spectrum analysis or frequency measurement is synonymous to speed assessment. Accordingly, speed (frequency) change or acceleration reveals best what binds spectrum components together, groups them. The typical grouping phenomena; modulation, filtering or irregular/periodic excitation, all proceed on different time scales, and may encode as either convolved or additive. On the whole, the selective assessment of grouped movement in time and frequency can be seen as a central theme in the modeling of the voice.

When a series of spectral components is calculated on base of the same fixed linear time frame, then especially for the high frequencies, the differential, the rate of change suffers from ambiguities that all come back to the problem of evaluating on a too extended time scope. A constant-Q spectrum analysis offers a much better scale to weight spectral rate of change. The general approach, the wavelet transform, is rigid in its implementation as it fixes any differential change to the clock of its own decomposition scheme. An alternative implementation of a constant-Q spectrum analysis is presented, based on an exponential warping of the time axis. With this method it is possible to zoom in at sections of the frequency domain and dynamically adjust the Q-factor, or to evaluate rate of change along the time axis, along the frequency axis or along a mixed axis. The method uses a pre stage, multi-rate (octave band) down sampling, followed by a conventional convolution and can thus benefit from the fast complex convolution properties of the FFT. Essential is that a continuous, but variable bandwidth time signal always remains the start-off point for the warp, so there will be no confusing scale interpretation, while also the door to straightforward filtering, or direct subtraction remains open. The way the time signal is re-sampled, depends on how time and frequency are warped together according to the chosen analysis settings. A typical feature of the logarithmic representation that comes with the exponential warping, is the opposing but equal proportionality of both signal and spectrum axis. It shows in the variability of the components, but also as a reversion of the harmonic template. The main advantage of the double access, is that cross-de-correlation (grouping) can be done on both scales in parallel. The information centering can be judged in two complementary directions. For periodicity detection, time domain cross correlation and frequency domain log scale harmonic summation merge to an identical, two angle approach. The analysis and its typical controls and constraints will be discussed and demonstrated.

Improve the voice by using the ear and the brain

Dorinne S. Davis

The Davis Center, 19 State Rt 10 E, Ste 25, Succasunna, NJ, 07876 USA

A case study involving a 49 year old female whose larynx had been traumatized during surgery and who was told nothing further could be done was undertaken using The Davis Model of Sound Intervention. VS1 began a program using sound and vibration to support improved voice control, voice production, and speech production. An assessment to determine if a sound-based therapy could support VS1 towards change was administered. Three therapies were implemented and over a period of 6 months, noticeable change was evidenced in VS1's breath control, vocal quality and stamina, articulation, use of vocal mechanism, and overall voice production.

The process involved the Diagnostic Evaluation for Therapy Protocol (DETP®) and subsequent therapies as Berard Auditory Integration Training, the Tomatis® Method, and BioAcoustics™. The underlying premise is that there is a connection between the voice, the ear, and the brain as established by both The Tomatis Effect and The Davis Addendum® to the Tomatis Effect. The voice reveals the irregular frequencies of the body and by reintroducing these frequencies to the body, the voice regains coherence. With The Davis Addendum to the Tomatis Effect, the specific body frequencies, such as the tongue muscle in VS1's case, were reintroduced to the person's ear. After a series of sound-based therapies, stimulation to the nerves innervating the larynx was created, early emotional trauma was released, and the surgically impaired nerves to the tongue muscles were re-stimulated by sound frequencies, and the person regained most of the laryngeal, vocal and oral musculature needs for regaining her voice.

This case study suggests a possible new approach for addressing voice problems by first making sure the connection between the voice, the ear, and the brain are functioning and well-balanced.

KEYNOTE 2: Clinical applications of Speech analysis

Julian McGlashan

*Department of Otorhinolaryngology, Queen's Medical Centre,, Nottingham University Hospitals,
Nottingham, UK*

Voice measurement in clinical practice has three main potential uses: a) as an aid for diagnosis b) as an outcome measure following an intervention such as voice therapy or surgery and 3) as an aid to directing therapeutic interventions. However it remains a significant challenge to ensure that the measurements are sensitive, have adequate specificity and are a true reflection of the patient's perception of benefit from treatment.

For diagnosis the gold standard remains a combination of taking a detailed clinical history and performing a laryngeal examination including observation of the pattern of vibration of the vocal fold mucosal wave. Empirical trials of voice therapy or medication and diagnostic microlaryngoscopy are also sometimes required. The inherent difficulty with measurement is that voice problem is often multi-factorial and may be due to one or more of the following conditions: a structural abnormality, inflammation, a neuromuscular disorder and a muscle tension imbalance. In addition factors such as variability of the voice problem with voice use, environmental conditions and the voice sample used for analysis may mean the assessment in the clinic at a particular point in time is not always representative of the voice impairment. Outcome measures can show trends in improvement towards normative values but are not always a reflection in the patient's perception of improvement of their vocal problem. The difficulty remains that each patient will have specific vocal needs ranging from a basic ability to communicate to highly specific complaints tone, range, effort and stamina. There is no one single measure or combination of measures that can be used which shows universal strong correlation with either perceptual or self-reported measures of vocal health change.

Where measures have been particularly helpful is in the objective assessment of specific voice parameters which can then be used to direct the therapeutic intervention. The simplest and easiest example is mean (or modal) speaking fundamental frequency. From a clinical perspective vocal fold contact is also valuable. Perhaps not enough emphasis has been made of micro analysis of voice recordings of during speech, particularly when done in a standardised method such as reading a text and when combining acoustic and laryngographic measures. They can not only be useful for demonstrating consistency and change in the voice over time but also in understanding the links between perceptual and vocal fold vibratory abnormalities.

this page has been left blank for your own notes

Fusing GMM and SVM for human voice-based classification of larynx pathology

Evaldas VAIČIUKYNAS¹, Adas GELŽINIS¹, Marija BAČAUSKIENĖ¹,

Antanas VERIKAS^{1,2}, Aurelija Uloza³

evaldas.vaiciukynas@stud.ktu.lt, adas.gelzinis@ktu.lt, marija.bacauskiene@ktu.lt,

antanas.verikas@hh.se, v_aurelija@yahoo.com

¹*Department of Electrical & Control Equipment, Kaunas University of Technology, Lithuania,*

²*Intelligent Systems Laboratory, Halmstad University, Sweden, ³Department of Otolaryngology,*

Kaunas University of Medicine, Lithuania

In this paper identification of laryngeal disorders using cepstral parameters of human voice is investigated. Mel-frequency cepstral coefficients (MFCC), extracted from audio recordings, are further approximated, using various strategies (sampling, averaging, and GMM-based estimation). The effectiveness of SVM, LS-SVM and GMM techniques in categorizing such pre-processed data into normal, nodular, and diffuse classes is explored. Constructed custom kernels were tested and compared to the non-linear RBF kernel in the SVM-based classification of voice recordings. Also, since it is a multi-class problem, various schemes to combine binary decisions are compared.

A sequence kernel, defined over a pair of matrices, rather than over a pair of vectors and calculating the kernelized principal angle (KPA) between subspaces, was designed. When using this kernel, larynx pathology recognition is done on the premises, that different disorders generate different subspaces and the degree of alignment (principal angles) between them can be measured. Other kernels we tested are distance kernels specifically tailored to GMM by exploiting covariance matrices, while the simple GMM super-vector kernel variant uses only means. The Monte-Carlo sampling (MCS) and the Kullback-Leibler (KL) divergence combined with the earth mover's distance (KL-EMD) are the similarity metrics used here. The KL-divergence approximation from the Monte-Carlo sampling is the distance measure based on the cross likelihood ratio test between samples, generated from patients' GMM models, while the Earth mover's distance is conceptually equivalent to the Mallows or Wasserstein distance between probability distributions. When tested on voice recordings collected from 410 subjects (130 normal, 140 diffuse, and 140 nodular), the MCS kernel, using full covariance matrices, and the KL-EMD kernel, using diagonal covariance matrices, provided the best overall performance. The sequence kernel and the distance kernels outperformed the popular RBF kernel, but the difference is statistically significant only in the distance kernels case. In all the tests, SVM outperformed the LS-SVM. The results indicate that features, estimated with GMM, and SVM kernels, designed to exploit this information, is an interesting fusion of probabilistic and discriminative models for human voice-based classification of larynx pathology.

On combining information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for automatic detection of pathological voices

Julián D. Arias-Londoño*, Juan I. Godino-Llorente, Maria Markaki, Yannis Stylianou

The automatic assessment of voice quality based on acoustic analysis is an efficient tool for the objective support of the diagnosis and the screening of vocal and voice diseases. Most of the works done in this area have been oriented to the study of acoustic parameters and pitch perturbation measures. Nevertheless, approaches based on the characterization of the spectral components for identifying the abnormal glottal activity have shown to be reliable in the detection pathological voices. Two of the techniques used in this context are: Mel-Frequency Cepstral coefficients (MFCC) and Modulation spectral features.

MFCC has been used in the detection of pathological voices with good results. The main advantage of the MFCC parameters is that they do not exhibit dependency on previous pitch estimations. Besides, the alterations related with the mucosal waveform due to an increase of mass are reflected in the low bands used to calculate the MFCC, whereas the higher bands are able to model the noisy components due to a lack of closure. On the other hand, modulation spectral may be seen as a nonparametric way to represent the modulation in speech. Moreover, it offers an implicit way to fuse the various phenomena observed during speech production, in a compact way and it provides important dynamic information related to key parameters in the detection of pathological voices. Some works in the state of the art, have concluded that complementary information between MFCC and Modulation Spectral features exists. However, there are important differences related to the parameterization of both kinds of features, which obstruct their fusion in order to be used in a single system. Due to this fact, this work investigates the use of a combining classifier strategy to fusion information from Modulation Spectra and MFCC.

The advantage of combining outputs of different classifiers instead of merging features is that the structure of the feature space used to feed each classifier is simpler. Furthermore, although one of the classifiers would yield a better performance, the sets of speech registers misclassified by the different classifiers would not necessarily overlap; therefore the combination of their outputs could improve the overall performance of the system. The experiments were carried out by using two different databases. The first database is The MEEI database of voice disorders. The second database was recorded by Universidad Politécnica de Madrid, and it is referred to as Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database. The pattern classifiers used are based on Gaussian Mixtures Models and Support Vector Machines, because they have been used successfully for the same task. The results are summarized in the ROC and DET plots below

Voice disorder detection on the base of continuous speech

Vicsi, Klara.-Imre, Viktor

Laboratory of Speech Acoustics, Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Stoczek u. 2, H- 111 Budapest, Hungary

Generally in voice production, there is a close connection between the mutations of voice generation organs (differences in size, in tissue flexibility, etc.), and the measurable acoustical parameters (pitch, sound pressure, spectrum, etc.) of the generated speech product. This connection was examined during years. The European Laryngological Research Group, formed in 1989, accepted a common protocol for the assessment of voice pathology. Researches, inside and outside of this group, were carried on the base of steady state sounds only, thus the results of the voice analyses were limited. The aim of the presented research work was to develop statistical signal processing methods on the base of a well constructed continuous pathological speech database and to construct an automatic diagnostic system, which helps to detect voice disorders.

For this purpose a big continuous database have been collected, which contains phonetically balanced continuous speech examples of patients, grouped according to pathological cases, with the cooperation of the Institute of Oncology, Department of Head-Neck Surgery. This database has been processed:

Sound perception evaluation scale RBH^{*}, what is a popular scale in the practice of phoniatriy, was used, to differentiate the degree of the voice generation disorders in the database. Different pathological cases were classified on the base of this numeric scale. The usefulness of different acoustical analysing methods was examined to establish which method reflected the degree of the voice generation disorders best. Statistical distributions of the acoustical parameters were examined by measuring these parameters at the middle of the vowels in continuous speech.

It was demonstrated, that the statistical acoustical parameters in the quasy stacioner part of the vowels in continuous speech can represent the perceptual classification of experts much better than in the traditionally used steady state sounds.

**RBH ((Rauhigkeit) (Behauchtheit) (Heiserkeit)): 0 = normal voice quality, 3 = heavy huskiness.*

Voxlog: a new voice accumulator using both accelerometer and microphone sensors

Fredric Lindström, Suvi Karjalainen, Maria Södersten, Sten Ternström

Occupational and environmental medicine, Umeå University, Sweden

Voice accumulators, or voice dosimeters, are wearable devices which estimate and log voice parameters such as phonation time, fundamental frequency, and voice sound pressure level. Voice accumulators have been used to research the vocal behaviour of the wearer in his/her natural environment, particularly in several occupational voice assessment studies. Voice accumulators have also been used by clinicians to assess voice and to increase voice therapy carry-over, although such clinical usage of voice accumulators is still in its very early stages.

Several different voice accumulators have been proposed, both those that register airborne sounds using microphones and those that register tissue-borne vibrations using accelerometers. The new voice accumulator "Voxlog" has been developed under a series of research grants in Sweden. Speaking in high levels of noise has been pointed out as a risk factor for voice problems. The background noise is thus an interesting factor from a voice ergonomics perspective. This voice accumulator can concurrently register the level of the background noise. The dual-sensor system with a microphone and an accelerometer is presented, as well as real scenario evaluations of different approaches to the mounting of the sensors.

this page has been left blank for your own notes

Can vocal tremor features indicate vocal loading?

M. Koutsogiannaki¹, Y. Pantazis¹, Y. Stylianou¹ and A.-M. Laukkanen²

¹*Institute of Computer Science, FORTH, and Multimedia Informatics Lab, University of Crete, Greece*

²*Department of Speech Communication and Voice Research, University of Tampere, Tampere, Finland*

Among the professions in which voice loading is a primary concern are teachers, singers, actors, radio/tv reporters, salespeople and phone operators. This study investigates the effect of specific acoustic features on vocal loading. Earlier studies have shown correlations between vocal loading and acoustic features such as a rise in fundamental frequency and sound pressure level and a decrease of the spectral tilt [1]. These findings either indicate that vocal folds have been warmed up and, thus function more effectively after vocal loading or that the muscle activity of the vocal organ has increased and the production is becoming hyperfunctional. The latter tends to increase vocal loading further. Changes in the laryngeal muscle activity, measured by electromyography, may indicate vocal fatigue [2]. These results also suggest that when the vocal muscles get fatigued there is a peak in the frequency of amplitude tremor in the acoustic signal recorded. Different vocal tremor attributes may result from the fact that as vocal loading strains the laryngeal muscles it may affect their ability to sustain the tension of the vocal folds constant. The present study investigates whether vocal tremor may be found after a short vocally loading test.

There are two attributes that characterize vocal tremor: modulation frequency, which measures how fast the modulations occur, and modulation level, which measures how strong the modulations are. We developed an algorithm for the accurate extraction of modulation frequency and modulation level [3] based on an AM-FM decomposition algorithm [4]. The proposed algorithm consists of: (1) estimation of instantaneous components of the recorded speech, (2) preprocessing where very slow modulations are removed, and (3) estimation of the modulation frequency and modulation level from the processed instantaneous component. For the evaluation of vocal tremor attributes as acoustic features of vocal loading, the following experiment was carried out. Six university students (3 males, 3 females, mean age 23 years) uttered 3 sustained vowels [a:, i:, u:] 3 times before and after a vocally loading test. The samples were recorded in a well-damped studio at the University of Tampere, Finland. The vocally loading task was to shout numbers for five minutes, at a sound pressure level of 90 dB, measured at 1 meter, in the well-damped studio. After loading, the participants were asked to report on a scale from 0 to 3 how tired their throat felt. Additionally, two speech trainers perceptually evaluated the vowel samples, using a scale from -3 (very poor) to +3 (excellent). The vowel samples recorded before and after a vocally loading test will be analyzed for the tremor attributes. Correlation analysis will be carried out to study the relation between vocal tremor attributes and the results obtained in the subjective and perceptual evaluations. Preliminary results showed that statistical measurements of vocal tremor attributes differ slightly before and after vocal loading. We will further report the possibility of connecting vocal tremor and vocal loading using data recorded in field conditions from teachers before and after a working day [1, 5].

References:

- [1] A.-M. Laukkanen, I. Ilomaki, K. Leppanen, and E. Vilkmann. Acoustic Measures and Self-reports of Vocal Fatigue by Female Teachers. *Journal of Voice*, 22:283-289, 2008.
- [2] V.J. Boucher and T. Ayad. Physiological Attributes of Vocal Fatigue and their Acoustic Effects: A Synthesis of Findings for a Criterion-based Prevention of Acquired Voice Disorders. *Journal of Voice*, Article in Press, Available online 24 March 2009.
- [3] Y. Pantazis, M. Koutsogiannaki, and Y. Stylianou. A Novel Method for the Extraction of Vocal Tremor. In *MAVEBA*, Florence, 2009.
- [4] Y. Pantazis, O. Rosec, and Y. Stylianou. AM-FM Estimation for Speech based on a Time-varying Sinusoidal Model. In *Interspeech*, Brighton, 2009.
- [5] A.-M. Laukkanen, E. Kankare. Vocal loading related changes in male teachers' voices investigated before and after a working day. *Folia Phoniatica et Logopaedica*, 58(4):229-2

this page has been left blank for your own notes

Changes in singer performance with different acoustic environment

Christopher Barlow

School of Computing and Communication, Southampton Solent University

There is significant interest in development of new techniques and technology to increase standards of vocal performance, as this could have an impact on the music economy. In recent years there has been increasing interest in the use of technology to analyse performance of singers, in order to improve understanding of how the singer's voice works, and provide biofeedback in order to improve technique and rate of learning.

However, there are a number of key omissions from the majority of the research on vocal performance. A particular problem is the lack of data from real performances and venues, from both an acoustic and 'performance' perspective. Instead, the majority of research has been undertaken in a laboratory environment, where the analysis tools are frequently based on phonetically balanced texts and vowel vocalisations rather than 'real' repertoire. Other areas of performance science – particularly Sport and Dance – make use of real time analysis of the performer under performance conditions, and it is suggested that this is an area that needs considerably more examination for musicians in general, and for singers in particular. Most singers of music in the Western Classical tradition rely on the acoustics of the performance venue to provide aural feedback to control aspects of their vocal performance, and are therefore particularly vulnerable to changes in acoustic environment.

This pilot project analysed a sung phrase of ~50 seconds duration taken from recordings by a solo unaccompanied mezzo-soprano singer recorded in a hemi-anechoic chamber (RT60 = 0), and the same singer recorded in a large Victorian church with RT60 ~ 3.5 s. The phrase was recorded several times in each environment, and was part of a solo well known by the singer, which she had performed in a variety of different concert environments, and using a text which naturally utilised several sustained vowels, which were ideal for spectral analysis. The solo was recorded in its entirety and was intended to be sung unaccompanied, eliminating a number of key issues with conventional research methods in this area.

A number of key parameters were analysed, including Mean frequency over a number of sustained notes, mean tempo, vibrato rate/depth, and pitch of formant frequencies of sustained vowels in the musical phrase evaluated using Long Term Average Spectra. Significant variation was found between the recordings in each venue for the parameters of tempo, vibrato rate and depth, and the frequencies of formants F1-F3. Pitch variation was exhibited but not at a significant level.

Howard and Brereton (2008) analysed singers in a similar manner, examining different parameters, and found raised vocal fold closed quotients, a shallower spectral slope and increased intensity for the anechoic recordings, suggesting heightened effort and increased vocal load and a resultant change in singer tone quality. This study supports these findings, indicating that significant changes in tone quality, and other key aspects of performance are highly likely where singers are assessed in anechoic spaces, and that researchers need to take into account this consideration when assessing vocal properties.

this page has been left blank for your own notes

Acoustic and articulatory adjustments in operatic singing: Spectral analysis and magnetic resonance imaging

J. G. Svec¹, J. Horacek², T. Vampola³, C. Herbst¹, D. G. Miller⁴, R. Havlik⁵, P. Krupa⁶, M. Lejska⁵

¹ Laboratory of Biophysics, Dept. Experimental Physics, Palacký University Olomouc, tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic, ² Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Dolejskova 5, 182 00 Prague 8, Czech Republic, ³ Department of Mechanics, Biomechanics and Mechatronics, Faculty of Mechanical Engineering, Czech Technical University Prague, Karlovo nám. 13, 121 35 Prague 2, Czech Republic, ⁴ EGGs for Singers, Achterste Kamp 9, 9301 RB Roden, the Netherlands, ⁵ Audio-Fon Centrum, Obilní trh 3-4, 602 00 Brno, Czech Republic, ⁶ Department of Medical Imaging, St. Anne's Faculty Hospital, Masaryk University, Pekarska 53, 656 91 Brno, Czech Republic

Singer's formant, i.e., energy boost of the spectral components of voice around 3 kHz, has been recognized as an important factor for operatic singing. Latest modelling studies suggest that such energy boost can be achieved through various articulatory adjustments, but detailed in vivo data are needed. This study offers magnetic resonance (MR) images and acoustic spectral data from two male operatic singers who produced voice both in naïve and operatic singing quality at the same fundamental frequency on vowels [a] and [ε]. In both the singers the MR images revealed a) lowered larynx, b) raised soft palate, c) flattened tongue, d) adjusted head position, and e) straightened spine in operatic singing when compared to the naïve singing. Anatomical differences (e.g., epiglottis shape and position, overall shape and volume of the vocal tract) were found between the two singers. This was related to differences in the voice spectrum, where clustering and lowering the formant frequencies was identified but the strategy was different for the two vowels and the two singers. The data suggest that different anatomical dispositions in singers lead to different possibilities for optimizing voice quality. Recognizing these inter-individual differences can play an important role for the science-based singing voice pedagogy in future.

this page has been left blank for your own notes

Comparison of geometrical and acoustical characteristics of human supraglottal spaces for ordinary and singing voice using finite element models

T. Vampola¹, J. Horáček²

tomas.vampola@fs.cvut.cz, jaromirh@it.cas.cz

¹ Department of Mechanics, Biomechanics and Mechatronics, Faculty of Mechanical Engineering, Czech Technical University, Karlovo nám. 13, 121 35 Prague 2, ² Institute of Thermomechanics ASCR, v.v.i., Dolejškova 5, 182 00 Prague 8

Professional singers use the so-called “singer’s formant” technique. The “singer’s formant” belongs to the basic qualities of the operatic singer’s voice and is formed by clustering the formants F3-F5. When singing with an orchestra, the existence of a singer’s formant is necessary in order to be heard over the sounds of the orchestral instruments, even if a first impression is that the orchestra must overwhelm the human voice. A well-trained (male) singer’s voice with the singer’s formant has a significant resonance maximum in the frequency region near 3000 Hz, where the voice level is significantly higher than the sound level of the instruments [1]. Different singers and various singing styles utilize different resonance strategies. These strategies have not been properly and fully mapped yet. The origin of the singer’s formant is usually attributed to larynx lowering and pharynx-widening that results in a new independent resonance cavity [2].

Better knowledge of the shape, space and structural changes of the human vocal tract due to the existence of the singer’s formant during singing is enabled by using magnetic resonance (MR) investigations. Especially, the MR enables scanning of the vocal tract spaces of the subjects in phonation position of resonant cavities for vowels with and without the singer’s formant.

The contribution deals with the design of the volume and subsequently finite element (FE) models of the human vocal tract for phonation of the vowel /a:/ without and with the singer’s formant. The models were developed from the MR measurements. The vocal cavities of two voluntaries were scanned by MR technique to determine the main shape and dimensional changes of the vocal tract that contribute to the singer’s formant. The MR investigation of acoustic vocal spaces was performed for two male (bass-baritone and baritone) singers singing vowel /a:/ with and without the singer’s formant. The created primary four volume models developed from the MR images were then transformed into the 3D FE computer models of the vocal tract, similarly as in [3]. The FE programming code ANSYS was used for frequency-modal analyses of the acoustic spaces of the vocal tracts and transient analysis for the numerical simulations of phonation.

The results show substantial increase of the volume and length of the vocal tract and changes in the acoustic frequency–modal characteristics of the FE models of the acoustic spaces due to the existence of the singer’s formant. Comparison of the results for the two singers investigated shows their different strategies resulting in the appearance of the singer’s formant.

References:

- [1] J. Sundberg, “Research on the singing voice in retrospect”, Speech, Music and Hearing, KTH, Stockholm, TMH-QOSR Vol. 45, 1974, pp.11-22.
- [2] J. Sundberg, “Articulatory interpretation of the singing formant”, JASA. 55(4), 1974, pp.838-844.
- [3] T. Vampola, J. Horáček, J.G. Švec. “FE Modeling of Human Vocal Tract Acoustics. Part I: Production of Czech Vowels.” *Acustica United with Acta Acustica*. Vol. 94, 2008, pp.433

A summary of the acoustical correlates of stress in speech

Christin Kirchhübel, Helena Daffern and David M. Howard

Department of Electronics, University of York, York, UK

{ck531@york.ac.uk|helenadaffern@hotmail.com|dmh8@ohm.york.ac.uk}

Modelling the influence of stress on speech and voice has been of interest to researchers from many different disciplines including psychology, psychiatry, medicine, engineering and forensic speech science. The aim of this paper is to summarise the acoustical correlates of stress in speech by reviewing relevant empirical studies and to suggest how this might be relevant to data capture and analysis in the clinic.

Research into 'speech under stress' faces several challenges including the difficulty in defining the concept of stress, the limitations in collecting stressed speech and the problem of testing the type and level of stress induced (Murray et al. 1996). As a result, past studies often lack comparability. When summarizing the evidence these conceptual and methodological differences are addressed.

Changes in respiration and muscle tension have repeatedly been evidenced to be physiological correlates of stress. Based on this, it may be expected that acoustic parameters such as fundamental frequency (f_0), intensity and speaking/articulation rate (SR/AR) are affected by stress (Scherer 1981a). In general, empirical findings support these predictions but it is also apparent that inter-speaker variability must not be overlooked (Streeter et al. 1982, Jessen 2006). The paper argues that there is a need to move beyond the three parameters mentioned above. Looking at vowel formants and voice quality, for example, might be of potential interest. It is also suggested that future research should aim to control the data in a way to allow for comparability of results between studies.

References

- Jessen, M. (2006). Einfluss von Stress auf Sprache und Stimme. Unter besonderer Berücksichtigung polizeidienstlicher Anforderungen. Idstein: Schulz-Kirchner Verlag GmbH.
- Murray, I. R., Baber, C. & South, A. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication*, 20, 3-12.
- Scherer, K.R. (1981a). Vocal indicators of stress. In J. Darby (Ed.). *Speech Evaluation in Psychiatry*, pp. 171-187. New York: Grune & Stratton
- Streeter, L.A., Macdonald, N.H., Apple, W., Krauss, R.M. & Galotti, K.M. (1983) "Acoustic and perceptual indicators of emotional stress." *The Journal of the Acoustical Society of America* 73, (4): 1354-1360.