Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 283504, 4 pages
doi:10.1155/2009/283504

*Editorial*

# Analysis and Signal Processing of Oesophageal and Pathological Voices

**Juan Ignacio Godino-Llorente,[1] Pedro Gómez-Vilda (EURASIP Member),[2] and Tan Lee[3]**

[1] *Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid, Carretera Valencia Km 7, 28031, Madrid, Spain*
[2] *Department of Computer Science & Engineering, Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla del Monte, 28660, Madrid, Spain*
[3] *Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong*

Correspondence should be addressed to Juan Ignacio Godino-Llorente, igodino@ics.upm.es

Received 29 October 2009; Accepted 29 October 2009

## 1. Introduction

Speech not only is limited to the process of communication but also is very important for transferring emotions, it is a small part of our personality, reflects situations of stress, and has a cosmetic added value in many different professional activities. Since speech communication is fundamental to human interaction, we are moving toward a new scenario where speech is gaining greater importance in our daily lives. On the other hand, modern styles of life have increased the risk of experiencing some kind of voice alterations. In this sense, the National Institute on Deafness and Other Communication Disorders (NIDCD) pointed out that approximately 7.5 million people in the United States have trouble using their voices [1]. Even though providing statistics on people affected by voice disorders is a very difficult task, as reported in [2], it is underlined that between 5 and 10% of the US working population have to be considered as using their voice in an intensive way. In Finland, these statistics are estimated close to 25%. Still in [2], the conclusions point out that the voice is the primary tool for about 25 to 33% of the working population. While the case of teachers has been largely studied in literature [2, 3], singers, doctors, lawyers, nurses, (tele-)marketer people, professional trainers, and public speakers also make great demands on their voices and, consequently, they are prone to experiencing voice problems [1, 4–6]. Therefore, in addition to medical consequences in daily life (treatment, rehabilitation, etc.), some voice disorders have also severe consequences regarding professional (job performance, attendance, occupation changes) and economical aspects but also far from being negligible, regarding social activities, and interaction with others [2–4].

However, despite many years of effort devoted to developing algorithms for speech signal processing, and despite the elaboration of automatic speech recognition and synthesis systems, our knowledge of the nature of the speech signal and the effects of pathologies is still limited. In spite of this, voice scientists and clinicians take profit of the simple models and methods developed by speech signal processing engineers to build up their own analysis methods for the assessment of disorders of voice (DoV).

Yet, the limitations of existing models and methods are felt in both areas of expertise, that is, speech signal processing applications and assessment of DoV. For example, the intervals within which signal model parameters must remain constant to represent signals with timbre that is perceived as natural are unknown. Moreover, such efficient control of voice quality has important applications in modern text-to-speech synthesis systems (creating new synthetic voices, simulating emotions, etc.). Voice clinicians, on the other hand, have expressed their disappointment with regard to the performance of existing methods for assessing voice quality, with a special focus on the forensic implications. Major issues with current methods include robustness against noise, consistency of measurements, interpretation of estimated features from a speech production point of view, and correlation with perception.

So there exist a need for new and objective ways to evaluate the speech, its quality, and its connection with other phenomena, since the deviation out of the patterns considered of normality can be correlated with many different symptoms and psychophysical situations. As previously commented, research to date in speech technology has focussed the effort in areas such as speech synthesis, recognition, and speaker verification/recognition. Speech technologies have evolved to the stage where they are reliable enough to be applied in other areas. In this sense, acoustic analysis is a noninvasive technique which is an efficient tool for the objective support and the diagnosis of DoV, the screening of vocal and voice diseases (and particularly their early detection), the objective determination of vocal function alterations, and the evaluation of surgical as well as pharmacological treatments and rehabilitation. Its application should not be restricted to the medical area alone, as it may also be of special interest in forensic applications, the control of voice quality for voice professionals such as singers, speakers, the evaluation of the stress, and so forth.

In addition, digital speech processing techniques pay a special role dealing with oesophageal voices. The quality of voice and the functional limitations of the laryngectomized patients remain an important challenge for improving their quality of life.

On the other hand, the acoustic analysis reveals as a complementary tool to other methods of evaluation used in the clinic based on the direct observation of the vocal folds using videoendoscopy. Therefore, a deeper insight into the voice production mechanism and its relevant parameters could help clinicians to improve prevention and treatment of DoV. In this sense, and in order to contribute filling in this gap, during the last ten years, links and co-operation among different research fields have become effective to define and set up simple and reliable tools for voice analysis. As a result, there exists a joint initiative to the European level devoted to the research in this field: the COST 2103 Action [7], funded by the European Science Foundation, is a joint initiative of speech processing teams and the European Laryngological Research Group (ELRG). The main objective of this action is to improve voice production models and analysis algorithms with a view to assessing voice disorders, by incorporating new or previously unexploited techniques, with recent theoretical developments in order to improve modelling of normal and abnormal voice production, including substitution voices. This is an interdisciplinary action that aims to foster synergies between various complementary disciplines as a promising way to efficiently address the complexity of many current research and development problems in the field of DoV. In particular, the progress in the clinical assessment and enhancement of voice quality requires the cooperation of speech processing engineers and voice clinicians.

The aim of this special issue is to contribute with a step-forward filling in the aforementioned gaps.

## 2. Summary of the Issue

For this special issue, 31 submissions were received. After a difficult review process, 12 papers have been accepted for publication. The accepted articles address important issues in speech processing and applications on oesophageal and pathological voices.

The articles in this special issue cover the following topics: methods of voice quality analysis based on frequency and amplitude perturbation and noise measurements; development of acoustic features to detect, classify, or discriminate pathological voices; classification techniques for the automatic detection of pathological voices; automatic assessment of voice quality; automatic word and phoneme intelligibility in pathological voices; analyzing and assessing the speech of cognitive impaired people; automatic detection of obstructive sleep apnoea from the speech; robust recognition of dysarthric speakers; and, automatic speech recognition and synthesis to enhance the quality of communication.

In this issue, two papers describe the methods of voice quality analysis based on frequency and amplitude perturbation (i.e., jitter and shimmer) and noise measurements. Although these measurements have been widely applied in the state of the art for a long time, still present some drawbacks, and further research is needed in this field.

The jitter value is a measure of the irregularity of a quasiperiodic signal and is a good indicator of the presence of pathologies in the larynx such as vocal fold nodules or a vocal fold polyp. The paper by Silva et al. focuses on the evaluation of different methods found in the state of the art to estimate the amount of jitter present in speech signals. Also, the authors proposed a new jitter measurement. Given the irregular nature of the speech signal, each jitter estimation algorithm relies on its own model making a direct comparison of the results very difficult. For this reason, in this paper, the evaluation of the different jitter estimation methods is targeted on their ability to detect pathological voices. The paper shows that there are significant differences in the performance of the jitter algorithms under evaluation.

In addition, with respect to the classic acoustic measurements, since the calculations of Harmonics-to-Noise Ratio (HNR) in voiced signals are affected by general aperiodicity (like jitter, shimmer, and waveform variability), the paper by Ferrer et al. develops a method to reduce the shimmer effects in the calculation of the HNR. The authors proposed an ensemble averaging technique that has been gradually refined in terms of its sensitivity to jitter, waveform variability, and required number of pulses. In this paper, shimmer is introduced in the model of the ensemble average and a formula is derived which allows the reduction of shimmer effects in HNR calculation.

On the other hand, several articles presented in this issue reported works about detecting, classifying, or discriminating pathological voices. Three of them focus on the development of acoustic features.

The paper by Dubuisson et al. presents a system developed to discriminate normal and pathological voices. The proposed system is based on features inspired from voice pathology assessment and music information retrieval. The paper uses two features (spectral decrease and first spectral tristimulus in the Bark scale) and their correlation, leading to correct classification rates of 94.7% for pathological voices and 89.5% for normal ones. Moreover, the system provides a normal/pathological factor giving an objective indication to the clinician.

Ghoraani and Krishnan propose another methodology for the automatic detection of pathological voices. The authors proposed the extraction of meaningful and unique features using adaptive time-frequency distribution (TFD) and nonnegative matrix factorization (NMF). The adaptive TFD dynamically tracks the nonstationarity in the speech, and NMF quantifies the constructed TFD. The proposed method extracts meaningful and unique features from the joint TFD of the speech, and automatically identifies and measures the abnormality of the signal.

In addition, Carello and Magnano evaluated in their paper the acoustic properties of oesophageal voices (EVs) and tracheo-oesophageal voices (TEPs). For each patient, some acoustic features were calculated: fundamental frequency, intensity, jitter, shimmer, and noise-to-harmonic ratio. Moreover, for TEP patients, the tracheostoma pressure at the time of phonation was measured in order to obtain information about the "in vivo" pressure necessary to open the phonatory valve to enable speech. The authors reported noise components between 600 Hz and 800 Hz in all patients, with a harmonic component between 1200 Hz and 1600 Hz. Besides, the TEP have better acoustic characteristics and a lower standard deviation. To investigate the correlation between the pressure and the TEP voice signals, the cross spectrum based on the Fourier transform was evaluated. The most important and interesting result pointed out by this analysis is that the two signals reported equal fundamental frequency and the same harmonic components for each TEP subject considered.

Two more papers in this issue discussed different classification techniques for the automatic detection of pathological voices. The paper by Kotropoulos et al. compares two distinct pattern recognition approaches: the detection of male subjects who are diagnosed with vocal fold paralysis against male subjects who are diagnosed as normal; the detection of female subjects who are suffering from vocal fold edema against female subjects who do not suffer from any voice pathology. Linear prediction coefficients extracted from sustained vowels were used as features. The evaluation was carried out using a Bayes classifier with Gaussian class conditional probability density functions with equal covariance matrices.

Fredouille et al. address the important task of voice quality assessment. They proposed an original back-and-forth methodology involving an automatic classification system as well as knowledge of the human experts (machine learning experts, phoneticians, and pathologists). The automatic system was validated with a dysphonic corpus, rated according to the GRBAS perceptual scale by an expert jury. The analysis showed the interest of the (0–3000) Hz frequency band for this classification problem. Additionally, an automatic phonemic analysis underlined the significance of consonants and more surprisingly of unvoiced consonants for the same classification task. Submitted to the human experts, these observations led to a manual analysis of unvoiced plosives, which highlighted a lengthening of voice onset time (VOT) according to the dysphonia severity validated by a preliminary statistical analysis.

Four more papers deal with the analyzing and assessing of different types of impaired or disordered speech.

The paper by Saz et al. presents the results in the analysis of the acoustic features (formants and the three suprasegmental features: tone, intensity, and duration) of the vowel production in a group of young speakers suffering different kinds of speech impairments due to physical and cognitive disorders. A corpus with unimpaired children's speech is used to determine the reference values for these features in speakers without any kind of speech impairment within the same domain of the impaired speakers; that is, 57 isolated words. The signal processing to extract the formant and pitch values is based on a linear prediction coefficient (LPC) analysis of the segments considered as vowels in a hidden Markov model- (HMM-) based Viterbi forced alignment. Intensity and duration are also based in the outcome of the automated segmentation. As main conclusion of the work, it is shown that intelligibility of the vowel production is lowered in impaired speakers even when the vowel is perceived as correct by human labelers. The decrease in intelligibility is due to a 30% of increase in confusability in the formants map, a reduction of 50% in the discriminative power in energy between stressed and unstressed vowels, and a 50% increase of the standard deviation in the length of the vowels. On the other hand, impaired speakers kept good control of tone in the production of stressed and unstressed vowels.

Likewise, it is commonly acknowledged that word or phoneme intelligibility is an important criterion in the assessment of the communication efficiency of a pathological speaker. Middag et al. developed a system based on automatic speech recognition (ASR) technology to automate and objectify the intelligibility assessment. This paper presents a methodology that uses phonological features, automatic speech alignment (based on acoustic models trained with normal speech), context-dependent speaker feature extraction, and intelligibility prediction based on a small model that can be trained on pathological speech samples. The experimental evaluation of the new system revealed that the root mean squared error of the discrepancies between perceived and computed intelligibilities can be as low as 8 on a scale of 0 to 100.

Morales and Cox modelled the errors done by a dysarthric speaker and attempt to correct them using two techniques: a) a set of "metamodels" that incorporate a model of the speaker's phonetic confusion-matrix into the ASR process; b) a cascade of weighted finite-state transducers at the confusion-matrix, word, and language levels. Both techniques attempt to correct the errors made at the phonetic level and make use of a language model to find the best estimate of the correct word sequence. The experiments showed that both techniques outperform standard adaptation techniques.

Pozo et al. proposed the use of ASR techniques for the automatic diagnosis of patients with severe obstructive sleep apnoea (OSA). Early detection of severe apnoea cases is important so that patients can receive early treatment, and an effective ASR-based detection system could dramatically reduce medical testing time. Working with a carefully designed speech database of healthy and apnoea subjects, they describe an acoustic search for distinctive apnoea voice characteristics. The paper also studies abnormal nasalization in OSA patients by modelling vowels in nasal and nonnasal phonetic contexts using Gaussian mixture model (GMM) pattern recognition on speech spectra.

Finally, the paper by Selouani et al. proposes the use of assistive speech-enabled systems to help both French and English speaking persons with various speech disorders. The proposed assistive systems use ASR and speech synthesis in order to enhance the quality of communication. These systems aim at improving the intelligibility of pathologic speech making it as natural as possible and close to the original voice of the speaker. The resynthesized utterances use new basic units, a new concatenating algorithm, and a grafting technique to correct the poorly pronounced phonemes. The ASR responses are uttered by the new speech synthesis system in order to convey an intelligible message to listeners. An improvement of the perceptual evaluation of the speech quality (PESQ) value of 5% and more than 20% was achieved by the speech synthesis systems dealing with substitution disorders (SSD) and dysarthria, respectively.

To conclude, this special issue aims at offering an interdisciplinary platform for presenting new knowledge in the field of analysis and signal processing of oesophageal and pathological voices. From these papers, we hope that the interested reader will find useful suggestions and further stimulation to carry on research in this field.

## Acknowledgments

The authors are extremely grateful to all the reviewers who took time and consideration to assess the submitted manuscripts. Their diligence and their constructive criticism and remarks contributed greatly to ensure that the final papers have conformed to the high standards expected in this publication. Moreover, we would like to thank all the authors who submitted papers to this special issue for their patience during the always hard and long reviewing process, especially to those that unfortunately had no opportunity to see their work published. Last, but not least, we would like to thank the Editor in-Chief and the Editorial Office of EURASIP Journal on Advances in Signal Processing for their continuous efforts and valuable support.

*Juan Ignacio Godino-Llorente*
*Pedro Gómez Vilda*
*Tan Lee*

## References

[1] National Institute on Deafness and Other Communication Disorders (NIDCD), ANR2008—Document B/anglais VoxAcCom Page 6/39, October 2009, http://www.nidcd.nih.gov/health/statistics/vsl.asp.

[2] *La voix. Ses Troubles Chez Les Enseignants*, INSERM, 2006.

[3] American Speech-Language-Hearing Association, October 2009, http://www.asha.org/default.htm.

[4] E. Smith, M. Taylor, M. Mendoza, J. Barkmeier, J. Lemke, and H. Hoffman, "Spasmodic dysphonia and vocal fold paralysis: outcomes of voice problems on work-related functioning," *Journal of Voice*, vol. 12, no. 2, pp. 223–232, 1998.

[5] Medline Plus, October 2009, http://www.nlm.nih.gov/medlineplus/voicedisorders.html.

[6] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[7] M. Kob and P. H. Dejonckere, ""Advanced voice function assessment"—goals and activities of COST action 2103," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 173–175, 2009.