Biomedical Signal Processing and Control 37 (2017) 1–4

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Editorial

Models and devices for a multimodal study of the human phonatory apparatus: Technological results and clinical applications

Speech is the primary means of communication between humans and results from complex interaction between the vibration of the vocal folds at the larynx and the movements of the voluntary articulators (mouth tongue, jaw, etc.).

A primary purpose of voice analysis is to extract features or parameters that represent relevant characteristics of the acoustic waveform.

First studies go back to the 18th century but strong interest was shown in speech characteristics in the early 20th century along with the growth of the telecommunications industry. Efforts were made to develop tools for speaker identification and for background noise removal. Voice quality assessment was firstly developed for speech coding or synthesis and for speech enhancement. The need for such testing methods became apparent in the 1950s along with the development of new analogue communication systems.

More recently research has focussed on biomedical applications. The aim of voice analysis in the biomedical field being quite different from that of telecommunications, different approaches were developed. Specifically, the acoustical analysis and modelling of the voice source and the vocal tract in healthy and pathological voices are among the main fields of research. The aim is that of extracting the voice characteristics (fundamental frequency and vocal tract resonance frequencies) together with their deviation from 'healthy conditions'. Degradation of voice, generally called hoarseness, is in fact one of the major symptoms of benign laryngeal diseases such as polyps or nodules but is often the first symptom of neoplastic diseases such as laryngeal cancer.

The acoustical properties of pathological voice are commonly investigated in terms of perturbations in the fundamental frequency (jitter) and amplitude (shimmer) as well as looking for spectral irregularities through measures of harmonics-to-noise-ratio (HNR). Though all such parameters occur to some extent in normal voices, beyond specified thresholds they can be identifiers of vocal pathology. As for the acoustical analysis, the need for quantifiable measures of voice as related to treatments and medical or surgical intervention has led to a search for acoustic correlates of dysphonic severity. An objective measure of the 'noise level', intended as a measure of voice quality, is of great relevance in biomedical applications. In their daily practice, laryngologists require an objective measure of hoarseness and a scale to compare results of various treatments. Perceptual scales are commonly used, but their correlation to objective indexes is still an open problem. Reliable models of the human voice can contribute to solve this problem.

Models are inherently a simplified version of the real physical mechanism, derived from physical principles (mechanical, electrical) and acoustical properties (e.g. multitube resonances) of voice and speech, which are in turn based on anatomy and physiology of the vocal apparatus. Recently, sophisticated versions of biomechanical models have been developed to describe the shape and dynamics of the vocal apparatus more accurately. Thanks to the increasing computing capabilities, the resulting mathematics became tractable and, if the boundary conditions are properly chosen, reliable parameters of the model can be obtained.

Voice analysis is mainly performed in the time, frequency, wavelet and cepstral domains. Classical methodologies have been and are applied for voice analysis but new ones were developed and compared. The so-called non-parametric frequency domain analysis is the most widely used, as it allows fast computations by means of FFT-based algorithms. However, its applicability and resolution capability are linked to the length of the data window under analysis: the shorter the time window the lower the frequency resolution. This is particularly critical for irregular voice signals that are stationary on very short data frames of some tens of milliseconds or high-pitched ones such as e.g. newborn infant cry and the singing voice. Time domain analysis aims at describing the signal dynamics by means of appropriate linear or non-linear models. Due to its high-resolution characteristics, the linear approach that makes use of AR models (linear prediction (LP)) overcomes problems encountered with classical frequency domain analysis giving reliable models of the speech waveform dynamics, though generally at the expense of more cumbersome algorithms. Also, both continuous and discrete wavelet transforms (CWT and DWT respectively) can be successfully applied to the estimation of F0 and formants. The advantage of the wavelet-based methods is their capability of performing a multiresolution analysis with a very low computational burden, thus they are particularly suited for quasi-stationary voice signals. The drawback is the semi-empirical choice of the mother wavelet. Cepstral analysis allows the excitation sequence at the glottis to be separated from the vocal tract impulse response, which are convolved in time. Thus, fundamental frequency and formants can be recovered from the transformed signal in the quefrency domain. The

http://dx.doi.org/10.1016/j.bspc.2017.06.013
1746-8094/

method is fast and simple but suffers from the same drawbacks of the frequency domain approach as far as resolution is concerned. Moreover, the choice of the lifter is critical. Finally, the application of other non-linear analysis techniques to vocal emissions place more realistic assumptions on the voice production mechanism being based on non-linear dynamical systems theory at the expense of more complex algorithms. Advanced analysis techniques include chaotic models, hidden Markov models, neural networks, and other sophisticated tools.

In addition to studying the audio signal recorded with a microphone, it is worth mentioning the somewhat indirect measure given by the electroglottograph (EGG), a device that provides the noninvasive measurement of the degree of contact between the vibrating vocal folds during voice production. Finally, new numerical methods and devices for image analysis can give more information than usual stroboscopy, that provides the standard view of the larynx. In fact, due to its limited frame rate stroboscopy cannot provide enough details to evaluate highly irregular vocal fold vibrations. Videokymography (VKG) registers the movements of the vocal folds with a high time resolution on a line perpendicular to the glottis, thus it was shown very useful in severely dysphonic patients with strong aperiodic vocal folds vibration. An increased use of other high-speed device in clinics and research laboratories, as well as improved technical capabilities in computer software and hardware and imaging techniques, has allowed detailed views of the vocal folds in motion. Quantitative evaluation of asymmetries and dysperiodicities is made available through increasingly refined software tools.

Along with traditional statistical analysis, new and powerful classification techniques are increasingly applied to differentiate voice signals coming from subjects of different gender, age, native language and different pathologies, to automatically assess the voice quality of patients after surgery or treatment.

The extensive research in modelling and analysis of the human voice gave rise to a huge number of indexes, both perceptual and objective, as well as different measures of the same quantity (such as e.g. jitter and noise). This still prevents a unified comparison of results and there is a need for a standardisation of acoustic measures. Moreover, standard data bases for voice analysis and pathology classification are still missing. Recently reliable synthetic signals are being developed to test software tools.

In addition to adult's voice analysis, it is worth noting that all the above mentioned approaches are increasingly applied in relatively new research areas such as non-speech vocal emissions like newborn infant cry, coughing and snoring, related to physiological and linguistic development, obstructive apnoea and asthma, respectively.

From this short introduction, it emerges that the field of speech and voice analysis is an inherently multidisciplinary one. Signals and images come both from newborns and adults and for a large number of pathologies or diseases useful information can be gained from accurate models and analysis. Hence, co-operation between biomedical engineers, clinicians, physicians, mathematicians and psychologists, is not only desirable but necessary as research not only focus on speech production, hearing and linguistic structure but explores and probes the complex interrelations between these areas.

The MAVEBA biannual series of Workshops held in Firenze, Italy, and never discontinued over the years, focuses on all these themes. It came into being in 1999 from the need, particularly felt by the organizers to share know-how, aims and results between areas that until then seemed quite distinct such as bioengineering and medicine. Therefore, its first aim was to stimulate contacts between specialists from different fields active in research and industrial developments in the area of voice signal analysis for biomedical applications. The scope of the Workshop includes all aspects of voice modelling and analysis, ranging from basic research to all kinds of biomedical applications, devices and related established and advanced technologies.

Over the years MAVEBA has reached full maturity and the initial issues of the workshop have grown and spread also in other aspects of research such as occupational voice disorders, singing voice, neurology, rehabilitation, image and video analysis with applications ranging from the newborn to adult, elderly and singers. In fact, there has been a continuous parallel expansion both in clinical research and in technology devoted to this field leading to an increasing interaction between researchers in technological and clinical disciplines, aiming at providing a common basis of knowledge for future research in this exciting area of biomedical investigation.

Not only multidisciplinary but also multimodality is a major keyword of MAVEBA: in fact different methodologies that relate both to the analysis of signals and of images for the study of the human vocal system are increasingly developed and successfully compared. Combining methods and skills is thus the key to the best results.

Under these perspectives, over the years well-known specialists from all over the World and dealing with any discipline related to voice come to Firenze and attend the MAVEBA Workshop giving their contribution with free papers, invited lectures and special session, offering participants a deeper insight into relevant aspects and results. Young researchers have the opportunity to discuss with specialists and their best paper are awarded and some granted by the BSPC Journal.

As for the past editions, this SI collects peer reviewed contributions that are extended versions of papers presented at the MAVEBA 2015 Workshop whose proceedings are available since 2003 both in printed and open access format (with ISBN code) at http://www.fupress.com/ricerca?q=maveba.

Exceptionally, in 2015 the MAVEBA workshop was held together with the Pan European Voice Conference (PEVOC) and the Collegium Medicorum Theatri meeting (CoMeT), whose abstract book is printed by FUP: http://www.fupress.com/ricerca?q=pevoc.

Specifically, this Special Issue collects ten papers that deal with the main aspects of biomedical applications of voice analysis, both methodological and applicative. An overview is presented here.

The first two papers mainly concern methodological and modelling aspects, tested both with synthesized voices and using human voice signals.

The first paper from Alzamendi et al. "Modeling and joint estimation of glottal source and vocal tract filter by state-space methods" investigates state-space methods to enhance the joint estimation of the glottal source and the vocal tract information. First, a state-space voice model is formulated considering the stochastic glottal source ruled by a stochastic difference equation that allows to accurately capture perturbations occurring at glottal level. Then, combining this voice model and the state-space framework, an inverse filtering method is developed that allows to jointly estimate both glottal source and vocal tract filter. The performance of this method is tested both with synthesized voices and using human voice signals. The results demonstrate that accurate estimates of the glottal source and the vocal tract filter can be obtained over several scenarios. Moreover, the method is shown to be robust with respect to different phonation types.

The paper from Silvia Orlandi et al. "Testing software tools for newborn cry analysis using synthetic signals" concerns the challenging issue of fundamental frequency (F0) and formant frequencies estimation in the high-pitched newborn cry. The acoustical analysis of the infant cry has the advantage of being a cheap