Using the Glottal Source in Voice Technology Applications

Thomas Drugman, University of Mons, Belgium

thomas.drugman@umons.ac.be

http://tcts.fpms.ac.be/~drugman/

From artificial voices in GPS to automatic systems of dictation, from voice-based identity verification to voice pathology detection, speech processing applications are nowadays omnipresent in our daily life. By offering solutions to companies seeking for efficiency enhancement with simultaneous cost saving, the market of speech technology is forecast to be particularly promising in the next years.

This presentation will focus on new advances in glottal analysis that have been made in the frame of my PhD thesis, in order to incorporate new techniques within speech processing applications.

While current systems are usually based on information related to the vocal tract configuration, the airflow passing through the vocal folds, and called glottal flow, is expected to exhibit a relevant complementarity. Unfortunately, glottal analysis from speech recordings requires specific complex processing operations, which explains why it has been generally avoided.

The main goal of this work is to provide new advances in glottal analysis so as to popularize it in speech processing. First, new techniques for glottal excitation estimation and modeling are proposed and shown to outperform other state-of-the-art approaches on large corpora of real speech. Moreover, proposed methods are integrated within various speech processing applications: speech synthesis, voice pathology detection, speaker recognition and expressive speech analysis. They are shown to lead to a substantial improvement when compared to other existing techniques.

More specifically, this study covers three separate but interconnected parts. In the first part, new algorithms for robust pitch tracking and for the automatic determination of Glottal Closure Instants (GCIs) are developed. This step is necessary as accurate glottal analysis

requires to process pitch-synchronous speech frames. The proposed method of pitch tracking leads to an appreciable reduction of the F0 frame error in noisy conditions (0dB of SNR with 5 types of noise). As for the algorithm of GCI detection, it outperforms the state-of-the-art in terms of reliability (low rates of false alarms and misses), timing accuracy and robustness (to both additive noise and reverberation).

In the second part, a new non-parametric method based on Complex Cepstrum is proposed for glottal flow estimation. This technique exploits the anticausality of the speech signal, which is a phase property during the production of the glottal source. In addition, a way to achieve this decomposition asynchronously is investigated. A comprehensive comparative study of glottal flow estimation approaches is also given. Our study shows that the proposed Complex Cepstrum-based decomposition and the closed-phase inverse filtering approach give on both synthetic and natural speech the best results.

Relying on this expertise, the usefulness of glottal information for voice pathology detection and expressive speech analysis is explored. The complementarity of the glottal excitation with features from the vocal tract is emphasized for detecting voice disorders. Combining these two sources of information allows a reduction of the patient misclassification rate. As for the analysis of expressive speech, significant modifications of the glottal behavior in Lombard and hypo- or hyperarticulated speech are highlighted.

In the third part, a new excitation modeling called Deterministic plus Stochastic Model (DSM) of the residual signal is proposed. DSM consists of two contributions acting in two distinct spectral bands delimited by a maximum voiced frequency. The deterministic part models the low-frequency contents and arises from an orthonormal decomposition, while the stochastic component is a high-frequency noise modulated both in time and frequency.

This model is applied to HMM-based speech synthesis and voice transformation where it is shown to enhance the naturalness and quality of the delivered voice. Finally, glottal signatures derived from this model are observed to lead to an increase of identification rates for speaker recognition purpose. In this way, a speaker identification rate of 96.35% is reached for 630 speakers using the proposed glottal signatures, when the best approach in the literature using glottal information only achieves 87.05%.