Thomas Drugman

PhD student in Speech Processing

Affiliation

 TCTS Lab - Université de Mons

 Mons   **Belgium**

Application domain

 Human-computer interaction

Contact

 Boulevard Dolez, 31

 7000 Mons

 +32 65 37 47 49

thomas.drugman@umons.ac.be

## On the Glottal Flow Estimation and its Usefulness in Speech Processing

Speech technologies find an important place within the context of human-computer interactions. Interest for this topic arises from the fact that the speech modality is probably the most natural way of communicating for humans. The following presentation is divided into two parts. The first section addresses a fundamental and crucial problematic in speech processing : given an audio recording of speech, how to extract from it the corresponding glottal flow, i.e the airflow coming from the vocal folds ? Since the glottal flow conveys relevant production information, the second section emphasizes how this signal can be integrated in various concrete applications of speech processing.

Keywords

Speech Processing, Glottal Flow, Speech Synthesis, Voice Pathology, Expressive Speech, Speaker Recognition

1. Estimating the Glottal Flow from the Speech Signal:

The estimation of the glottal flow directly from the speech signal (as captured by microphones) imposes to understand the physical process of speech production. According to the phonation mechanism (see the Figure below), speech results from an airflow evicted from the lungs, arising in the trachea, passing through the glottis, filtered by the vocal tract cavities and finally radiated by the lips. The glottis is defined as the space comprised between the vocal folds. During the production of *voiced sounds*, the airflow arising from the trachea causes a quasi-periodic vibration of the vocal folds. The goal of this first part is precisely to estimate the glottal flow modulated by the vocal folds during voiced sounds. This is a typical problem of *blind source separation* since neither the vocal tract nor the glottal contributions are observable. Based on some properties of the speech signal we proposed an algorithm [1] that showed its potential to reliably and accurately estimate the glottal source. Relying on the speech signal itself is of paramount importance since it allows to be independent of the use of any intruisive (e.g endoscopic cameras) or awkward (e.g laryngograph) device, which is generally avoided in concrete industrial applications.
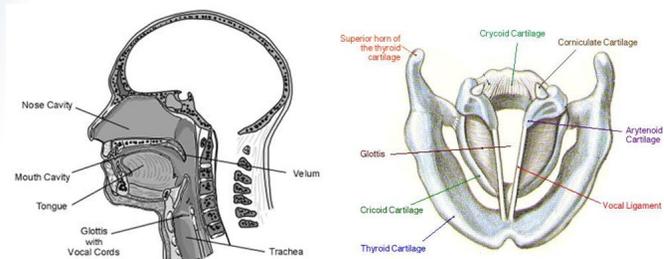
Left plot : Representation of the phonation aparatus. Speech results from an airflow evicted from the lungs, arising in the trachea, passing through the glottis, filtered by the vocal tract cavities and finally radiated by the lips. *Right plot :* Tranversal view of the larynx. Glottis is defined as the space comprised between the vocal folds.

2. *Applications in Speech Processing:*

Since the glottal flow, as estimated in the first Section, is a fundamental signal characterizing the production of speech, its analysis or synthesis may find various applications in speech processing. Regarding the analysis, it is aimed at extracting a set of time-domain or spectral features characterizing the phonation. As for the synthesis aspect, the goal is to find a parametric modeling having the ability to generate realistic glottal signals. Potential applications then include :

– **Speech synthesis :** The goal of speech synthesis is to automatically produce the lecture of an unknown text, which is generally typed by the user. Challenges are typically expressed in terms of naturalness and intelligibility of the generated voice. The main drawback of parametric synthesizers is the buzziness (i.e robotic aspect) in the produced speech, which is due to the difficulty in creating a natural excitation signal (i.e related to the glottal production). We therefore proposed a model of the excitation [2] that showed its effectivenes for increasing significantly the naturalness of the synthetic voice, and consequently its overall quality.

– **Voice pathology detection :** The presence of a voice pathology is known to be related to a dysfunction of the vocal folds behaviour. The glottal flow as estimated in the first part then contains relevant information about the presence or not of a given dysphony. Based on features extracted from the glottal flow, we developed objective tools with the goal of aiding clinicians in their diagnostic [3]. This is possible since the glottal features have distinct distributions for normophonic and dysphonic subjects.

– **Expressive speech analysis :** The production of expressive voice encompasses, among others, modifications taking place in the glottis. It is consequently possible to analyze differences in the glottal production and to integrate them in applications such as emotion recognition or expressive speech synthesis.

– **Speaker identification :** Speaker recognition refers to the automatic task of authentifying the identity of a person using its voice, for the purpose of controling access to information, a place, etc... As the physiology and behaviour of the vocal folds differ from a person to another, it is expected that the glottal flow could be efficiently incorporated into a speaker recognition system.

Bibliography:

– [1] T.Drugman, B.Bozkurt, T.Dutoit, *Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*, Proc. Interspeech, 2009.
– [2] T.Drugman, G.Wilfart, T.Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Proc. Interspeech, 2009.
– [3] T.Drugman, T.Dubuisson, T.Dutoit, *On the Mutual Information between Source and Filter Contributions for Voice Pathology Detection*, Proc. Interspeech, 2009.

Academic career

June-July 2009 : Visitor Researcher
Izmir Institute of Technology, Turkey

From September 2007 : FNRS Research Fellow, PhD Student
TCTS Lab, FPMs, Mons

My PhD thesis aims at incorporating the Hidden Markov Models in voice conversion for speech synthesis

January-June 2007 : Master Thesis
Swiss Federal Institute of Technologies (EPFL), Lausanne, Switzerland

My Master thesis dealt with feature selection and multimodal integration for Audio-Visual Speech Recognition