VOICE SOURCE PARAMETERS ESTIMATION BY FITTING THE GLOTTAL FORMANT AND THE INVERSE FILTERING OPEN PHASE

Thomas Drugman, Thomas Dubuisson, Nicolas D'Alessandro, Alexis Moinet and Thierry Dutoit

TCTS Laboratory / Electricity Department, Faculté Polytechnique de Mons
31, Boulevard Dolez, 7000, Mons, Belgium
phone: + (32) 65 37 47 49, fax: + (32) 65 37 47 29
email: thomas.drugman@fpms.ac.be, web: www.tcts.fpms.ac.be

ABSTRACT

This paper presents two approaches to the problem of extracting the parameters of the LF source model directly from the speech waveform. The first approach relies on the glottal formant estimated from the anticausal contribution of speech. Indeed the ZZT technique has recently shown its ability to deconvolve speech into its causal and anticausal components. The second method is based on the glottal open phase obtained by inverse filtering. The notion of *unanalyzable* frames and the way to detect and correct them are also presented. Once source parameters are extracted, the coefficients of the ARX speech production model are estimated by spectral division. Decomposition on both synthetic and natural speech, as well as an analysis-synthesis test confirm the accuracy of methods exposed.

1. INTRODUCTION

Using a high quality parametric representation of the speech signal has become a major issue for model-based speech synthesis (notably using HMMs), as well as for pathological voice analysis. In the first case an efficient vocoder can reduce the "buzziness" of the produced speech, which is the main drawback of statistical parameter synthesizers [1]. In the second one, analyzing the evolution of glottal parameters could allow us to detect, or even identify speech pathologies. Furthermore expressive voice synthesis or voice conversion applications can easily take advantage of this modeling by tuning parameters in a suitable way.

Most analysis methods consider speech as produced by a linear time-varying filter excited by a source signal. The goal of this paper is to address the problem of deconvolving these two contributions, i.e to extract the vocal tract and glottal components. Among the techniques described in litterature, some use iteratively the inverse filtering method [2] so as to remove the vocal tract contribution from the speech signal, while others apply Linear Prediction (LP) analysis only during the closing phase of the glottal signal in order to minimize the voice source effect on vocal tract estimation [3]. Other approaches use the Auto-Regressive with eXogenous input (ARX) model proposed in [4] in order to jointly estimate glottal source and vocal tract parameters [5, 6]. Finally a new model of speech (the mixed-phase model) using the Zeros of the Z-Transform (ZZT) representation is proposed in [7]. This technique decomposes speech into its causal and anticausal components, where the latter contribution corresponds to the glottal source open phase.

The paper is structured as follows. In Section 2, models of speech production are briefly presented and the decomposition technique based on ZZT representation is also introduced. In Section 3, we give more details on the methods we have developed. Glottal source parameters are first estimated using two error measures. The first one relies on a Spectral Fitting on the Glottal Formant (SFGF). As for the second one, it is based on a Temporal Fitting on the Open Phase (TFOP). For some voiced frames however, decomposition may be erroneous. By inspecting the detected glottal formant evolution, such frames are easily detected, and corrected. Once the source parameters are extracted, filter coefficients are estimated. Section 4 describes experimental and implementation details. Section 5 presents the results we obtained on synthetic and natural connected speech. Finally Section 6 concludes the paper and proposes possible perspectives.

2. BACKGROUND

In this Section, we first introduce the speech production models we adopted. The concept of speech deconvolution using ZZT is presented afterwards.

2.1 ARX speech modeling

In this paper speech is considered as produced by an Auto-Regressive with eXogenous input (ARX) model and expressed as a time-varying IIR system [4]:

$$\sum_{i=0}^{p} a_i(n)s(n-i) = \sum_{j=0}^{q} b_j(n)u(n-j) + \varepsilon(n), \qquad (1)$$

where $s(n)$ and $u(n)$ respectively denote the speech waveform and the glottal source. In the above equation, $a_i(n)$ and $b_j(n)$ are the time-varying filter coefficients and $\varepsilon(n)$ is the prediction error. From now on we will restrict ourselves to an all-poles and non-zeros filter ($q = 0$). In the Z-domain, the ARX model becomes:

$$S(z) = \frac{b_0 U(z)}{A(z)} + \frac{E(z)}{A(z)}. \qquad (2)$$

Notice that, as all other source-filter models, ARX neglects non-linear interactions between the source signal and the vocal tract.

2.2 LF glottal source modeling

Recurrence equation 1 explicitly involves the source component. For voiced speech, the glottal flow derivative

consists of two phases. During the *open phase*, vocal folds are progressively displaced from their initial state because of the increasing subglottal pressure. When the elastic displacement limit is reached, they suddenly return to this position during the so called *closing phase*.

In [8], Liljencrants and Fant proposed to model the glottal signal by a four-parameter representation illustrated in Figures 1 and 2 for both time and frequency domains. Under its normalized form, a LF wave is entirely characterized by the *open quotient* $O_q = T_c/T_0$, the *asymmetry coefficient* $\alpha_m = T_z/T_c$ and the cut-off frequency F_c. O_q and α_m govern the open phase, determining the glottal formant frequency (F_g) and bandwidth [9]. F_c has an impact on the return phase and imposes the spectral tilt.

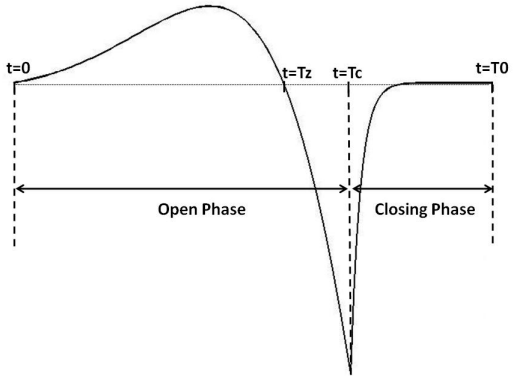

Figure 1: Temporal evolution of the LF glottal flow derivative

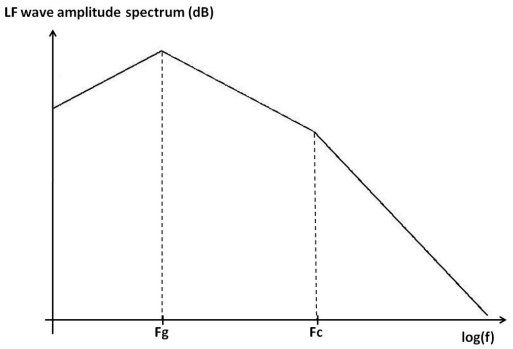

Figure 2: Asymptotic amplitude spectrum of the LF glottal flow derivative

2.3 ZZT-based decomposition of speech

For a series of N samples $(x(0), x(1), ...x(N-1))$ taken from a discrete signal $x(n)$, the *ZZT* representation is defined as the set of roots (zeros) $(Z_1, Z_2, ...Z_{N-1})$ of the corresponding Z-Transform $X(z)$:

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} = x(0) z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \qquad (3)$$

In order to decompose speech into its causal and anticausal contributions [7], *ZZT* are computed on frames centered on each Glottal Closure Instant (*GCI*) and whose length is twice the fundamental period at the considered *GCI*. The spectrum of the glottal source open phase is then computed from zeros out of the unit circle (anticausal component) while zeros with modulus lower than 1 give the vocal tract transmittance modulated by the source spectral tilt (causal component).

3. OUR PROPOSED METHODS

In this Section, techniques we developed are explained. For voiced segments, glottal source parameters are first estimated using the LF model. For this purpose two approaches, SFGF and TFOP, are proposed (3.1). Frames for which decomposition may be irrelevant are then detected and corrected (3.2). Finally a method based on spectral division extracts the ARX filter coefficients (3.3). A flow chart summarizing the overall scheme is presented in Figure 3.

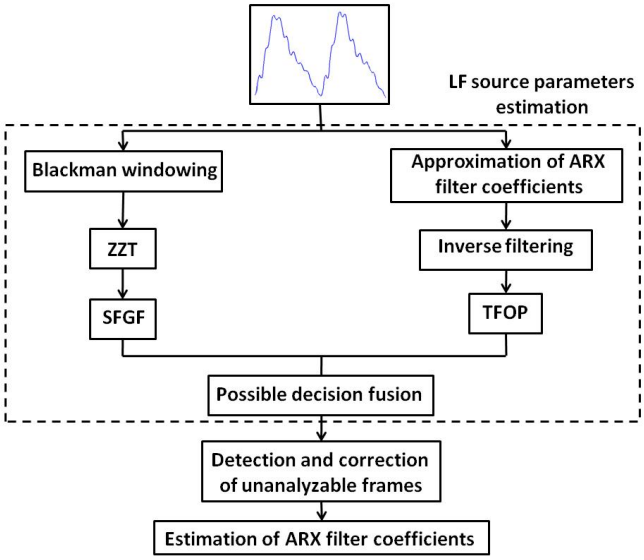

Figure 3: Flow chart for voiced speech frames centered on the current GCI

3.1 Estimation of LF parameters

We hereby present two methods estimating the open quotient O_q and the asymmetry coefficient α_m directly from the speech waveform. For a given voiced frame, both techniques compute a fitting error for each couple (O_q,α_m). As a fitting criterion, the Mean Square Error between a reference and the generated LF wave is considered. Decisions can then possibly be merged from both measures. In the following, an abrupt return phase is considered ($F_c \rightarrow \infty$), assuming that the spectral tilt will be taken into account during the ARX estimation (see Section 3.3).

3.1.1 Spectral Fitting on the Glottal Formant (SFGF)

As previously mentioned (Section 2.2), the glottal formant is entirely characterized by O_q and α_m. The key idea of the first technique is to generate the best LF wave which spectrally fits on the glottal formant. For this purpose, ZZT is first computed on the speech frame, and its anticausal part (corresponding to the open phase of the glottal source) is

isolated. Unfortunately, this signal is not generally temporally exploitable because it contains a high-frequency parasitic noise [7, 10]. Nevertheless the glottal formant is most of the time well-defined and can consequently be used as reference for the fitting. An example of SFGF is illustrated on Figure 4. Glottal formant similarity with the target spectrum is almost perfect. A major advantage of the SFGF method is that it should guarantee a good decomposition even when the glottal formant frequency is greater than the first speech formant ($F_g > F_1$).

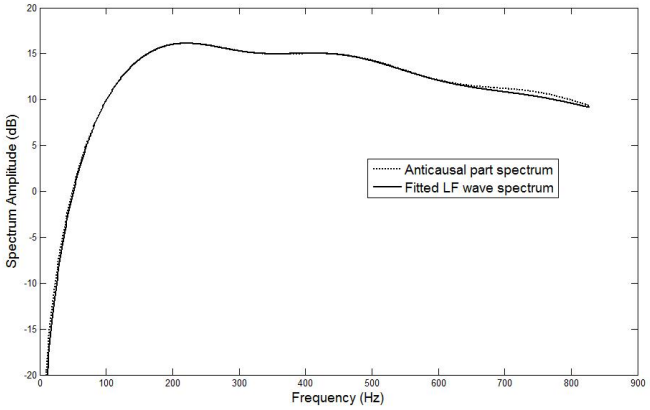

Figure 4: Example of SFGF for a given voiced speech frame

3.1.2 Temporal Fitting on the Open Phase (TFOP)

The extraction of (O_q,α_m) can also be based on the glottal source waveform during the open phase. As discussed in the previous Section, the anticausal component of speech can not serve as a reference due to the presence of a parasitic decomposition noise. However a good estimation of the glottal signal can be obtained by inverse filtering.

For this, a first approximation of the vocal tract response is obtained by dividing an "average" (i.e realistic) LF wave and the speech frame spectra. Vocal tract formants are generally well modeled, so that a reliable estimation of the source open phase is reached by inverse filtering. In Figure 5, this signal is compared with the anticausal component filtered in [0-2kHz]. This low-pass filtering is required to minimize decomposition noise effects and is not restrictive since the open phase characteristics only concern frequencies below 1kHz.

3.1.3 Decision fusion

The two previous techniques act independently. This means that we can take advantage of their possible complementary information by merging their decision, as a function of their reliability. Although complex decision fusion algorithms exist, error measures from both SFGF and TFOP techniques were simply multiplied in this work.

3.2 Unanalyzable frames detection and correction

For some voiced frames, decomposition may be uncorrect. This is generally due to an inaccurate Glottal Closure Instant (GCI) localization or to a large amount of noise in speech. For such frames, glottal formant tracking is erroneous. By inspecting the ratio between detected F_g and the fundamental

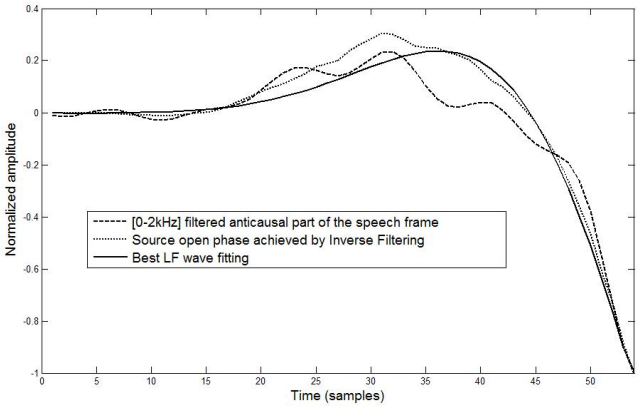

Figure 5: Example of TFOP and comparison between open phases obtained by inverse filtering and by the anticausal part of the speech frame

frequency F_0 , such frames are easily highlighted (see Figure 6) and considered as being unanalyzable. Their source parameters are thereby corrected by linear interpolation between neighbouring analyzable frames. However this linear interpolation does not make much sense when several successive frames are unanalyzable (e.g see frames 51 to 56 in Figure 6). Solving this issue is currently the object of ongoing research (possibly using a Kalman filtering).

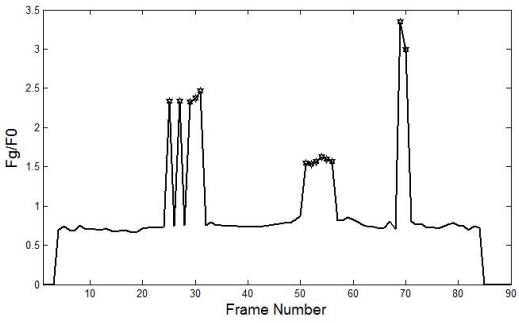

Figure 6: Unanalyzable frame detection for a voiced segment using the $F_{g,detected}/F_0$ ratio

3.3 Estimation of ARX parameters

Once source parameters are estimated, the ARX filter coefficients can be extracted. In [6], they are computed by solving a linear system. Unfortunately some cases lead to a lack of high frequencies in the estimation. To overcome this problem, we propose the following approach. First, source signal and speech frame are windowed by a Hanning function and their Power Spectral Density (PSD) is computed from the FFT. The autocorrelation function of the vocal tract is then obtained by IFFT of the division between the two previous PSDs. Finally, ARX filter coefficients are achieved by running the Durbin-Levinson algorithm (which is a fast implementation of LP analysis) directly from the autocorrelation function.

4. EXPERIMENT DETAILS

Our methods were first validated on synthetic signals and then tested on natural connected speech. Utterances of both male and female speakers were used with a sampling frequency varying from 16 kHz to 44.1 kHz. For voiced segments, analysis was conducted on $2T_0$-long speech frames centered on current GCI. GCIs were localized using the Snack pitch extraction algorithm [11] and the Center Of Gravity (COG) method derived from [12]. During unvoiced regions, LPC was performed on 25 ms long and 10 ms interspaced frames. For the source parameter estimation, speech frames were windowed by a Blackman function as suggested in [7]. To verify the accuracy of parameters extracted on natural speech, an analysis-synthesis application was carried out. Synthesis was performed by a Pitch Synchronous Overlap Add (PSOLA) technique.

5. RESULTS AND DISCUSSION

Two experiments were conducted in order to assess the efficiency of our system. The first experiment acted on synthetic signals so as to confirm the validity of our method, while the second one was run on natural speech. Results for both tests are exposed hereafter.

5.1 Validation on synthetic signals

Synthetic signals were obtained by passing a train of LF waves with known parameters through an auto-regressive filter. In all cases, our method converged towards the expected (O_q,α_m) couples. As shown in Figure 7, source parameters are easily discriminated. The illustrated error is the logarithm of the product between both SFGF and TFOP error measures in order to take both time and frequency domains into account. This combined error measure was used through all our experiments.

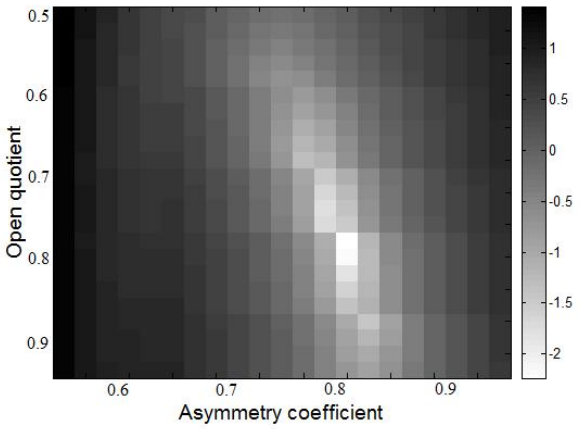

Figure 7: Combined SFGF and TFOP error measure in (O_q,α_m) space for a given synthetic signal frame

5.2 Analysis-Synthesis on natural connected speech

First of all, our system was applied to sustained vowels so as to test its "stability", i.e the consistency of extracted parameters. Indeed, glottal characteristics are almost constant during such pronunciations. We then applied our method to two phonetically balanced datasets. Both contain French natural speech, while speaker gender and sampling frequency differ. Keeping in mind our analysis-synthesis application, and our goal of incorporating them into a HMM-based synthesis system, parameters generally need to be smoothed. For this, a Poisson-windowed average centered on the current time was computed for each frame. In Figure 8, the evolution of the glottal open quotient and asymmetry coefficient is illustrated for the utterance *"Samedi soir"*.

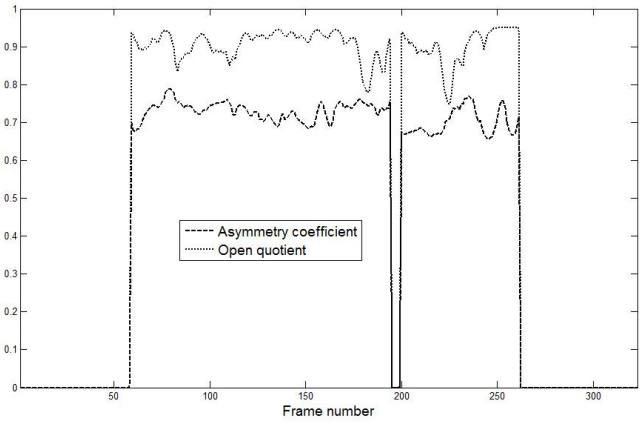

Figure 8: Glottal source parameters evolution for the French utterance *"Samedi soir"*

A way to evaluate the decomposition accuracy is to resynthesize speech directly from the extracted coefficients. This was done by overlapping two period-long Hanning-windowed reconstructed speech frames. An example of waveforms for a pronunciation of phoneme */a/* is presented in Figure 9. A strong similarity is observed. This is also noticed in the spectral domain on Figure 10 for phoneme */l/*, where a good concordance till 6 kHz is noted. These considerations were also reported on a perceptual point of view (examples can be found in [13]). Speaker's prosody and timber are well conserved although high frequencies (beyond 6 kHz) could be better modeled. A Mean Opinion Score (MOS) test assessing the subjective quality of analysis-synthesis was also submitted to 15 persons among whom 6 audio experts. Results are presented in Figure 11.

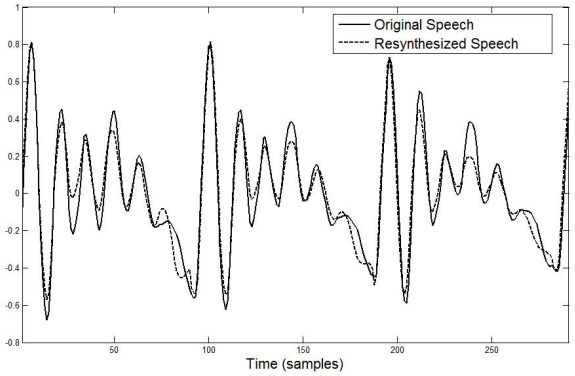

Figure 9: Comparison between original and resynthesized speech waveforms for a pronunciation of phoneme */a/*

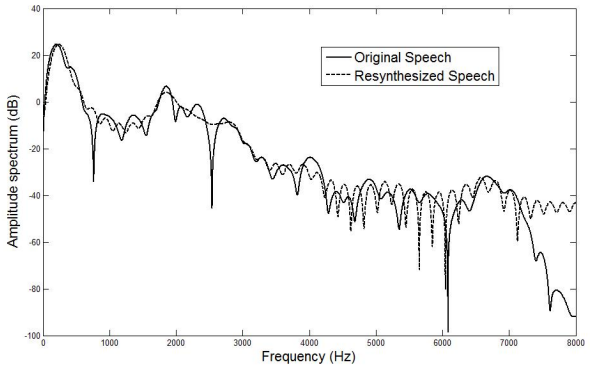

Figure 10: Comparison between original and resynthesized spectra for a speech frame of phoneme /l/

	Audio Experts	All listeners
Average	3	3.132
Standard Deviation	0.756	0.811

Figure 11: Results of the MOS test

6. CONCLUSIONS AND FUTURE WORK

This paper focused on the decomposition of speech into its glottal source and vocal tract components. For estimating the voice source parameters, we proposed two novel approaches which basically act as follows:

- SFGF first computes the ZZT on windowed speech frames in order to isolate the anticausal contribution of speech. LF coefficients are then extracted by fitting the glottal formant.
- In TFOP technique, a first approximation of the ARX filter is obtained by spectral division. LF parameters are then estimated by fitting the inverse filtering open phase.

For some voiced frames, deconvolution can not be efficiently carried out due to ZZT decomposition noise. These frames were detected and corrected by inspecting the evolution of the tracked glottal formant. Once the source is assumed to be known, the vocal tract autocorrelation function is estimated by spectral division. ARX filter coefficients are finally obtained by LP analysis. The validity and consistency of our methods were confirmed on synthetic signals. An application of Analysis-Synthesis on natural connected speech also showed good formal and perceptual results.

Finally, let us suggest some possible improvements we plan to make:

- As a technique for smoothing the source parameters, a Viterbi algorithm could be implemented to find best parameter trajectories over utterances.
- Since our method only considers voiced or unvoiced sounds, only the harmonic component of semi-vowels is modeled, which means that their high-frequency complexity is lost. To overcome this problem, a mixed excitation modeling the noisy residual could be integrated.
- A more complex decision fusion method considering the reliability of each LF parameter extraction technique (based on its error measure) could be carried out.

7. ACKNOWLEDGMENTS

Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique" (FNRS) and Nicolas D'Alessandro by the FRIA fundings. The authors also would like to thank the Walloon Region for its support (ECLIPSE WALEO II grant #516009 and IRMA RESEAUX II grant #415911).

REFERENCES

[1] K. Tokuda, H.Zen and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis 02*, Le Santa Monica, USA, September 11-13. 2002, pp. 227-230, 2002.

[2] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive invese filtering," *Speech Communication*, vol. 11, pp. 109-118, 1992.

[3] E. Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without precise glottal closure information ," *Proc. ICASSP 2004*, vol. 14, pp. 492-501, 2004.

[4] W. Ding, H. Kasuya and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," *IEICE Trans. Inf. Syst.*, vol. E78-D, no. 6, pp. 738-743, June 1995.

[5] Q. Fu and P. Murphy, "Adaptive Inverse Filtering for High Accuracy Estimation of the Glottal Source," in *Proc. NOLISP 03*, Le Croisic, France, May 20-23. 2003, paper 018.

[6] D. Vincent, O.Rosec and T. Chovanel, "Estimation of LF Glottal Source Parameters Based on an ARX Model," in *Proc. INTERSPEECH 05*, Lisbon, Portugal, September 4-8. 2005, pp 333-336.

[7] B. Bozkurt, L. Couvreur and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Comm.*, vol. 49, issue 3, pp. 159-176, 2007.

[8] G. Fant, J. Liljencrants and Q. Lin, "A four parameter model of glottal flow," *STL-QPSR4*, pp. 1-13, 1985.

[9] B. Doval, C. D'Alessandro and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 95, no. 6, pp. 1026-1046, 2006.

[10] T. Dubuisson and T. Dutoit, "Improvement of the source-tract decomposition of speech using analogy with LF model for glottal source and tube model for vocal tract," in *Proc. MAVEBA 07*, Firenze, Italy, December 13-15. 2007, pp 119-122.

[11] K. Sjolander, "The snack sound toolkit version 2.2b1," *http://www.speech.kth.se/snack/*, 2002.

[12] H. Kawahara, Y. Atake and P. Zolfaghari, "Auditory event detection based on a time domain fixed point analysis," *Proc. ISCA ICLSP 2000*, vol. 4, pp. 669-672, 2000.

[13] *http://tcts.fpms.ac.be/∼drugman/.*