IMPROVEMENT OF SOURCE-TRACT DECOMPOSITION OF SPEECH USING ANALOGY WITH LF MODEL FOR GLOTTAL SOURCE AND TUBE MODEL FOR VOCAL TRACT

T. Dubuisson[1], T. Dutoit [1]

1. Circuit Theory and Signal Processing Lab (TCTS Lab), Faculté Polytechnique de Mons, Belgium

Abstract: **In this paper we propose improvements to a recent algorithm of speech decomposition into glottal source and vocal tract contributions. This algorithm is based on the Zeros of the Z-Transform (*ZZT*) representation and requires restrictive conditions about the analysis window. Inaccurate results of decomposition can occur if these conditions are not fulfilled. The improvement method consists in considering an analogy with the LF model for the glottal source and a tube model for the vocal tract. Results are presented for a sustained vowel /a/ in both time and spectral domain. Future developments are also proposed.**

Keywords: **Zeros of the Z-Transform, glottal source, vocal tract, speech decomposition, Glottal Closure Instant**

I. INTRODUCTION

Analysis of the glottal source has been investigated by researchers because it has applications in different fields like speech recognition or voice quality modification. Among glottal source estimation techniques described in literature, some use iteratively the inverse filtering method [1] in order to remove the vocal tract contribution in speech while other apply the LP analysis only during the closed-phase of the glottal source [2] in order to minimize its effect on vocal tract estimation. Another method uses the ARX model [3] in order to jointly estimate glottal source model and vocal tract model parameters. Finally some methods focus on the estimation of glottal source parameters like Open Quotient [4] or Glottal Closure Instants (*GCI*) [5 ,6].

Recently another technique of decomposition of speech into glottal source and vocal tract contributions was proposed. This technique uses the *ZZT* representation of speech [7] and is particularly sensitive to *GCIs* localization. Applied to real speech signals, errors on the estimation of these instants can sometimes lead to noisy decomposition results. That is why improvements of this decomposition are presented here.

This paper is organized as follows. In section II the *ZZT* representation is defined, the *ZZT*-based decomposition and its improvements are described. In section III the results of improvements are presented for a sustained vowel /a/ and compared with results obtained without correction. In section IV the results are discussed and the perspectives are presented.

II. METHODS

A. Database

Test signals were recorded (16 kHz-16 bits) in TCTS Lab and are real sustained vowels /a/, /e/, /o/ and real transitions between these vowels.

B. ZZT representation and decomposition algorithm

For a *N* samples signal *x(n)*, the *ZZT* representation [7] is defined as the set of roots Z_m of the z-transform *X(z)* of the signal x(n):

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

In order to decompose speech into glottal source (glottal flow derivative) and vocal tract impulse response [7], *ZZT* are computed on frames centered on each *GCI* (computed by the algorithm described in [5]) and where length is twice the fundamental period at the considered *GCI*. The glottal source spectrum is then computed from zeros with modulus greater than 1 (maximum-phase components) and the vocal tract spectrum from zeros with modulus lower than 1(minimum-phase components).

C. Improvement of the decomposition

Due to errors on the estimation of *GCIs*, decomposition results can sometimes be noisy, and thus not suitable for accurate analysis of the glottal source. Experiments showed that, if *ZZT*-based decomposition is computed for several frames whose center is shifted by few samples around a *GCI*, better results can be obtained for an instant close but not identical to this *GCI*. The method considers here, for a range of shifts around *GCIs* in voiced island of speech, the *vocal tract candidate* (*VTC*) and the *glottal source candidate* (*GSC*) obtained from *ZZT*-based decomposition in order to determine which shift provides the best results for each *GCI*.

Concerning the glottal source, an analogy is made with the LF model [8]. Indeed, considering *GSCs* obtained for shifts around a given *GCI*, inaccurate decompositions are mainly characterized by a lot of energy located in frequencies higher than 2 kHz, contrary to LF model in which energy is mainly located below 2 kHz. Each *GSC* is therefore characterized by the energy ratio between the 0-2000 Hz band and the whole spectrum:

$$Feature\ GSC = \frac{Energy\ \ [0-2000\ Hz]}{Energy\ \ [0-8000\ Hz]} \qquad (2)$$

The vocal tract being a physical system with its own structure and elasticity, it is assumed that, during the production of a sustained vowel, it has to be as continuous as possible in terms of geometry. To express this continuity, the tube model [9] for the vocal tract is used and the radiuses of this model are computed by LP analysis [10] of the vocal tract impulse response (order set to 18). Each *VTC* is therefore characterized by a vector of 19 radiuses.

Around each *GCI*, the shift corresponding to the best decomposition must be a compromise between two criterions:

- *GSC*: among all the candidates, the elected one must be characterized by the biggest energy ratio between the 0-2000 Hz band and the whole spectrum. The criterion is thus the minimization of the energy in high frequencies.
- *VTC*: during the production of a sustained vowel, the geometry of the vocal tract cannot vary too much between two consecutive *GCIs*. Among all the *VTCs*, the elected candidate must be the one for which the vector of radiuses is the closest to the one corresponding to the candidates for the past and the next *GCI*. The criterion is thus the maximization of the continuity of the vocal tract geometry.

A dynamic programming algorithm is therefore implemented to optimize these criterions on the whole voiced island of speech (see Fig. 1).

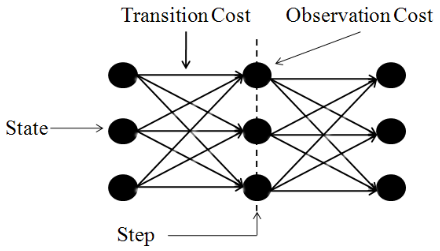

Fig. 1 Dynamic programming algorithm (1 shift before and after each *GCI* – 3 states)

In this algorithm each *step* corresponds to a *GCI* and each *state* corresponds to a particular shift around this *GCI*. The goal of this algorithm is to find the best path among all the shifts by minimizing a cost function on the whole voiced island of speech:

$$Cost(i,j) = Cost(i-1,k) + T\,C_{k/j}^{i-1/i} + O\,C(i,j) \qquad (3)$$

where *i* stands for the step index, *j* for the state index at step *i*, *k* for the state index at step *i-1*, *TC* (*Transition Cost)* stands for the difference of the radiuses between *VTC* at state *k* and *VTC* at state *j*, *OC* (*Observation Cost)* stands for the inverse of the feature defined for the *GSC* at state *j*. At the end of the voiced island of speech, the best path is chosen as the one with the lowest cumulated cost. The position of the *GCIs* can thus be corrected according to this choice.

III. Results

As explained in Section II, the dynamic programming algorithm determines the best shift around each *GCI* according to constraints defined by the cost function. Fig. 2 shows the evolution of its decision for the sustained vowel /a/ and for 4, 6 and 8 samples of shift before and after each *GCI* (9, 13 and 17 states).

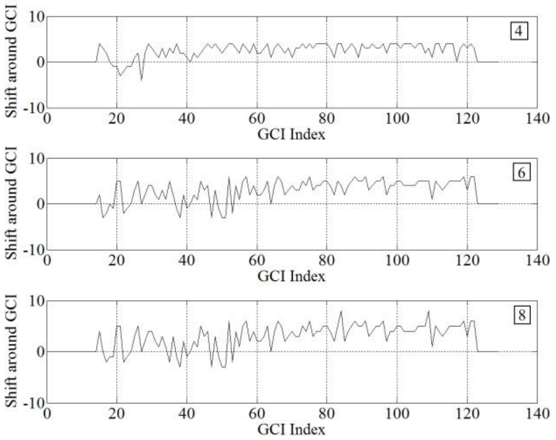

Fig. 2 Decision of the algorithm (from top to bottom: 4, 6 and 8 samples of shifts before and after each *GCI*)

One may see in this figure that considering 4 samples of shift is not enough (saturation is visible in the decision of the algorithm) while computing the decomposition for 8 samples is not necessary (the decision is nearly the same than for 6 samples). However we will show in the next subsection that the results obtained for 4 samples of shift are accurate enough. A shift of 4 samples before and after each *GCI* is therefore considered as a good choice because the improvement obtained for more samples of shift does not justify the increasing cost of computation. From now on the results are presented for 4 samples of shift.

A. Results of improvements in time domain

Fig. 3 shows the glottal sources obtained with and without correction. One may see in this figure that the noisy components are corrected and that the accurate ones before correction remain unaltered. The vocal tract responses are not displayed because the spectral domain is more suitable in order to observe the improvement on this component.

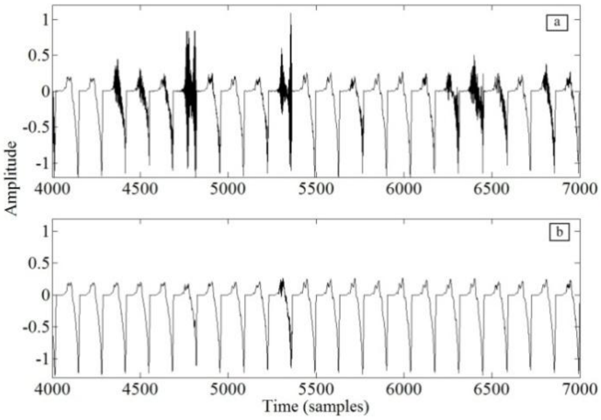

Fig. 3 Improvement of the glottal source in time domain (a: without correction; b: with correction)

B. Results of improvements in spectral domain

In this study a *GCI*-synchronous spectrogram is computed. This representation shows the evolution of the normalized spectrum of each *GCI*-centered period of glottal source and vocal tract response. Fig. 4 shows the *GCI*-synchronous spectrogram for the glottal source without and with correction.

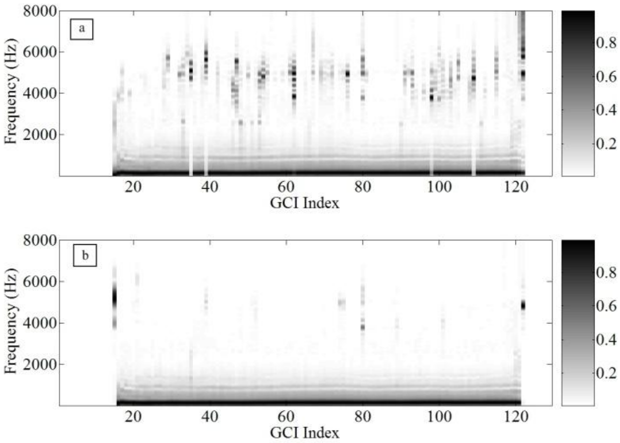

Fig. 4 Improvement of the glottal source in spectral domain (a: without correction; b: with correction)

Accurate glottal sources are characterized by a resonance in low frequencies (the glottal formant) and energy located below 2 kHz while the noisy ones have more energy in higher frequencies. After correction the noisy glottal sources are closer to the other accurate ones.

Concerning the vocal tract impulse response, the formants detected by Wavesurfer [11] on the speech signal are superimposed on the spectrogram in Fig. 5 (dotted lines). The correlation between the trajectory of the formants and the ones detected by Wavesurfer is good before correction although there are discontinuities in the formant trajectories for some *GCIs*. These discontinuities are less present after correction and the energy bursts in high frequencies have disappeared.

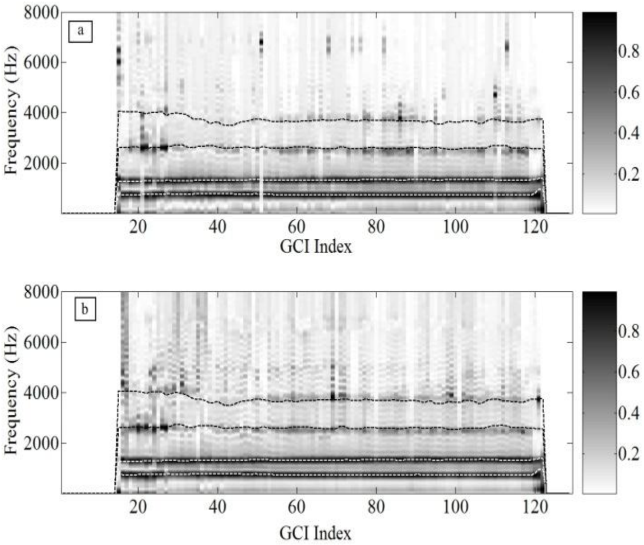

Fig. 5 Improvement of the vocal tract response in spectral domain (a: without correction; b: with correction)

C. Indicators of improvement

In order to quantify the amount of improvement for the two components, indicators are proposed. The glottal source indicator is defined as:

$$100\ x\ \left\{\frac{Mgfaf\ -Mgfbe}{Mgfbe} + \frac{Faf\ -Fbe}{Fbe}\right\} \qquad (4)$$

where M_{gfaf} stands for the magnitude of the glottal formant (spectral resonance detected in the 0-250 Hz band) after correction, M_{gfbe} stands for this magnitude before correction, F_{af} stands for the energy ratio of the glottal source after correction and F_{be} for this ratio before correction. Fig. 6 shows this indicator for the whole sustained vowel /a/.

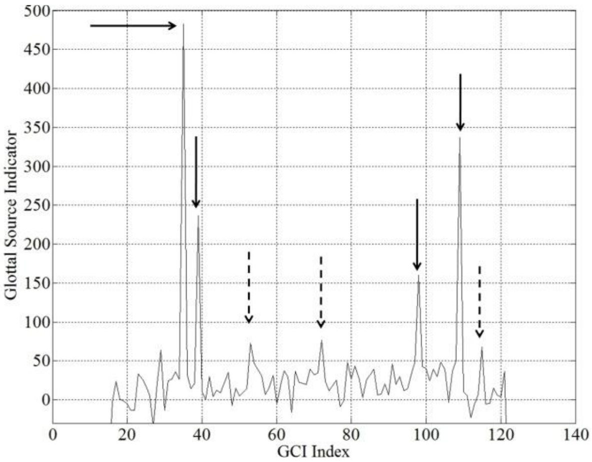

Fig. 6 Glottal source indicator

This indicator shows strong peaks (full arrows) at the *GCIs* for which the resonance in low frequency is not strong enough before correction and smaller peaks (dotted arrows) at those for which the glottal sources have resonance in low frequencies before correction and are less noisy after correction.

The vocal tract indicator uses the information from Wavesurfer in order to quantify the improvement in an objective way. The formant indicator is defined as:

$$100\ x\ \frac{Maf - Mbe}{Mbe} \qquad (5)$$

where M_{af} stands for the magnitude at the formant frequency in the vocal tract spectrum after correction and M_{be} stands for this magnitude before correction. The indicator for the vocal tract is the sum of the formant indicator for the two first formants detected by Wavesurfer. Fig. 7 shows this indicator for the whole sustained vowel /a/.

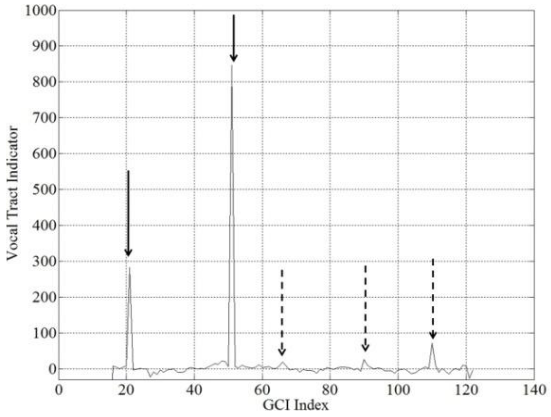

Fig. 7 Vocal tract indicator

This indicator shows strong peaks (full arrows) at the *GCIs* for which the discontinuity in the formant trajectories is important before correction and smaller peaks (dotted arrows) at those for which the energy in high frequencies is more important than for other *GCIs* before correction, but without discontinuities.

IV. Discussion and Conclusion

The method presented here is based on the *ZZT* representation. It thus differs from the inverse filtering based on LP analysis because the estimated LP filter contains both the contributions of glottal source and vocal tract. It also differs from the ARX based methods because the *ZZT*-based decomposition is not based on a glottal source model but only on phase properties of speech signal.

The purpose of this method is the improvement of the decomposition of speech into glottal source and vocal tract response using analogy with the LF model for glottal source and tube model for the vocal tract. These two components are characterized by features used in a dynamic programming algorithm in order to better determine the position of *GCIs* in voiced islands of speech. Accurate results are obtained for sustained vowels. *ZZT*-based decomposition and its improvement can lead to computation of parameters like open quotient or asymmetry coefficient [8] and adequacy between LF model and glottal sources obtained from real speech signals. In case of vocal folds pathology, the observation of *ZZT*-based decomposed sequences could lead to propose a new model for the glottal source.

Acknowledgements

Authors acknowledge support from the Walloon Region, Belgium, grant WALEO II ECLIPSE #516009, the COST Action #2103 ‘Advanced Voice Function Assessment’ and the Interuniversity Attraction Pole program VI-4 DYSCO of the Belgian Science Policy.

References

[1] P. Alku, “An automatic method to estimate the time-based parameters of the glottal pulseform,” *Proc. of ICASSP 1992*, IEEE, vol. 2, pp. 29-32, 1992.
[2] E. Moore and M. Clements, “Algorithm for automatic glottal waveform estimation without precise glottal closure information,” *Proc. ICASSP 04*, IEEE, vol. 14, pp. 492-501, 2004.
[3] D. Vincent, O. Rosec, and T. Chonavel, “Estimation of the LF glottal source parameters based on ARX model,” *Proc. Interspeech 2005*, ISCA, pp. 333-336, 2005.
[4] N. Henrich, B. Doval, and C. d’Alessandro, “Glottal open quotient estimation using linear prediction,” *Proc. MAVEBA 1999*, IEEE, pp. 12-17, 1999.
[5] H. Kawahara, Y. Atake, and P. Zolfaghari, “Auditory event detection based on a time domain fixed point analysis,” *Proc. ICLSP 2000*, ISCA, vol. 4, pp. 669-672, 2000.
[6] A. Kounoudes, P. Naylor, and M. Brookes, “The DYPSA algorithm for estimation of glottal closure instants in voiced speech,” *Proc. ICASSP 02*, IEEE, vol. 1, pp. 820-857, 2002.
[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, “The LF model revisited. Transformation and frequency domain analysis,” *STL-QPSR*, vol. 2-3, pp. 121-156, 1995.
[9] J. Kelly and C. Lochbaum, “Speech synthesis,” *Proc. of 4th International Congress of Acoustics*, pp. 1-4, 1962.
[10] D.G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, 1999, pp. 95-127.
[11] Wavesurfer : http://www.speech.kth.se/wavesurfer.