GLOTTAL SOURCE ESTIMATION ROBUSTNESS
A comparison of sensitivity of voice source estimation techniques

Thomas Drugman, Thomas Dubuisson, Alexis Moinet, Nicolas D'Alessandro, Thierry Dutoit
TCTS Lab, Faculté Polytechnique de Mons, 31 Boulevard Dolez, 7000 Mons, Belgium
firstname.lastname@fpms.ac.be

Keywords: Speech Processing, Speech Analysis, Voice Source, Glottal Formant

Abstract: This paper addresses the problem of estimating the voice source directly from speech waveforms. A novel principle based on Anticausality Dominated Regions (ACDR) is used to estimate the glottal open phase. This technique is compared to two other state-of-the-art well-known methods, namely the Zeros of the Z-Transform (ZZT) and the Iterative Adaptive Inverse Filtering (IAIF) algorithms. Decomposition quality is assessed on synthetic signals through two objective measures: the spectral distortion and a glottal formant determination rate. Technique robustness is tested by analyzing the influence of noise and Glottal Closure Instant (GCI) location errors. Besides impacts of the fundamental frequency and the first formant on the performance are evaluated. Our proposed approach shows significant improvement in robustness, which could be of a great interest when decomposing real speech.

1 INTRODUCTION

Source-filter modeling is one of the most widely used in speech processing. Its success is certainly due to the physiological interpretation it relies on. In this approach, speech is considered as the result of a glottal flow filtered by the vocal tract cavities and radiated by the lips. Our paper focuses on the glottal source estimation directly from the speech signal. Typical applications where this issue is of interest are voice quality assessment, statistical parametric speech synthesis, voice pathologies detection, expressive speech production,...

The goal of this paper is twofold. First a simple principle based on anticausality domination is presented. Secondly, different source estimation techniques are compared according to their robustness. Their decomposition quality is assessed in different conditions via two objective criteria : a spectral distortion measure and a glottal formant determination rate. Robust source estimation is of a paramount importance since final applications have to face adverse decomposition conditions on real continuous speech.

The paper is structured as follows. In section 2 a theoretical background on source estimation methods is given. The experimental protocol we used for the comparison is defined in Section 3. In Section 4 results are exposed and the impact of different factors on the estimation quality is discussed. Section 5 concludes the paper and proposes some guidelines for future work.

2 SOURCE ESTIMATION TECHNIQUES

We here present two popular voice source estimation methods, namely the Zeros of the Z-Transform decomposition (ZZT) and the Iterative Adaptive Inverse Filtering technique (IAIF). ZZT basis relies on the observation that speech is a mixed-phase signal (Doval et al., 2003) where the anticausal component corresponds to the vocal folds open phase, and where the causal component comprises both the glottis closure and the vocal tract contributions (see Figure 1). As for the IAIF method, it isolates the source signal by iteratively estimating vocal tract and source parts. After this brief state of the art, our approach based on Anticausality Dominated Regions (ACDR) is explained.

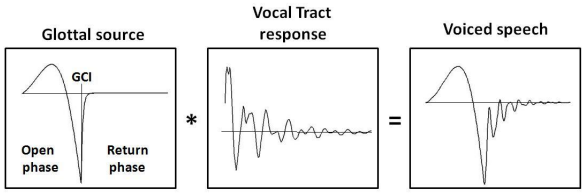

Figure 1: Illustration of the source-filter modeling for one voiced period. The Glottal Closure Instant (GCI) has the particularity to allow the separation of glottal open and closed phases, corresponding respectively to anticausal and causal signals.

2.1 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 $X(z)$:

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

In order to decompose speech into its causal and anticausal contributions (Bozkurt et al., 2007), ZZT are computed on frames centered on each Glottal Closure Instant (GCI) and whose length is twice the fundamental period at the considered GCI. These latter instants can be obtained either by electroglottographic (EGG) recordings or by extraction methods applied on the speech signal (see (Kawahara et al., 2000) for instance). The spectrum of the glottal source open phase is then computed from zeros outside the unit circle (anticausal component) while zeros with modulus lower than 1 give the vocal tract transmittance modulated by the source return phase spectrum (causal component).

2.2 Iterative Adaptive Inverse Filtering

The inverse filtering technique aims at removing the vocal tract contribution from speech by filtering this signal by the inverse of an estimation of the vocal tract transmittance (this estimation being usually obtained by LPC analysis). Many methods implement the inverse filtering in an iterative way in order to obtain a reliable glottal source estimation.

One of the most popular iterative method is the IAIF (Iterative Adaptive Inverse Filtering) algorithm proposed in (Alku et al., 1992). In its first version, this method implements LPC analysis so as to estimate the vocal tract response and use this estimation in the inverse filtering procedure. Authors proposed an improvement in (Alku et al., 2000), in which the LPC analysis is replaced by the Discrete All Pole (DAP) modeling technique (El-Jaroudi and Makhoul, 1991), more accurate than LPC analysis for high-pitched voices.

The block diagram of the IAIF method is shown in Figure 2 where $s(n)$ stands for the speech signal and $g(n)$ for the glottal source estimation. The 1^{st} block

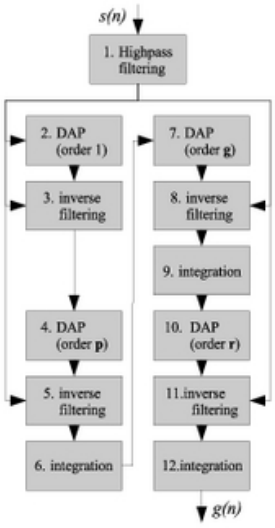

Figure 2: Block diagram of the IAIF method (from the documentation of TKK *Aparat* (Aparat, 2008)).

performs a high-pass filtering in order to reduce the low frequency fluctuations inherent to the recording step. The 2^{nd} and 3^{rd} blocks compute a first estimation of the vocal tract, which is used in the 4^{th} and 5^{th} blocks to compute a first estimation of the glottal source. This estimation is the basis of the second part of the diagram (7^{th} to 12^{th} blocks) where the same treatment is applied in order to obtain the final glottal source estimation.

Based on this method the TKK *Aparat* (Airas, 2008) has been developed as a sofware package providing an estimation of the glottal source and its model-based parameters. We used the toolbox available on the TKK *Aparat* website (Aparat, 2008) for our experiments.

2.3 Causality/anticausality Dominated Regions

As previously mentioned, analysis is generally performed on two-period long GCI-centered speech frames. Since GCI can be interpreted as the starting point for both causal and anticausal phases, it dermar-

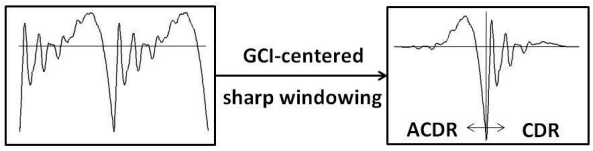

Figure 3: Effect of a sharp GCI-centered windowing on a two-period long speech frame. The Anticausality Dominated Region (ACDR) approximates the glottal source open phase.

cates the boundary between causality/anticausality dominated regions. As the domination zone of influence is limited around the GCI, a sharp window (typically a Hanning-Poisson or Blackman window) is applied to the analysis frame (see Figure 3). Since the causal contribution (comprising the source return phase and the vocal tract components) from the previous period is generally negligible just before the current GCI, the Anticausality Dominated Region (ACDR) makes a good approximation of the source open phase. As long as the window is centered on a GCI and is sufficiently sharp, this simple principle is applicable directly to the speech signal and even more on a first source estimation (obtained by IAIF for example). The dependency on the GCI detection for both ZZT and ACDR techniques will be discussed in Section 4.2.

3 EXPERIMENTAL PROTOCOL

The experimental protocol we opted for is close to the one presented in (Sturmel et al., 2007). Decomposition is achieved on synthetic speech signals for different test conditions. The idea is to cover the diversity of configurations one could find in continuous speech by varying all parameters over their whole range. Synthetic speech is produced according to the source-filter model by passing a known train of Liljencrants-Fant glottal waves (Fant et al., 1985) through an auto-regressive filter extracted by LPC analysis on real sustained vowel uttered by a male speaker. As the mean pitch during these utterances was about 100 Hz, it reasonable to consider that the fundamental frequency should not exceed 60 and 240 Hz in continuous speech. Perturbations are modeled in two ways: by adding a white Gaussian noise on the speech signal and by making an error on the GCI location (see Sections 4.1 and 4.2). Figure 4 summarizes all test conditions (which makes a total of 59280 experiments).

Four source estimation techniques are here compared : ZZT, IAIF, ACDR principle applied to both

Source characteristics	Open Quotient	0,3 : 0,05 : 0,9
	Asymmetry coefficient	0,6 : 0,05 : 0,85
	Pitch	60 : 20 : 240 Hz
Filter	Vowel	/a/,/e/,/i/,/u/
Perturbations	SNR	10 : 10 : 80 dB
	GCI location error	-1,25 : 0,25 : 1,25 ms

Figure 4: Table of parameter variation range.

speech and IAIF source frames. In order to assess the decomposition quality we used two objective measures:

- **Spectral distortion** : Many frequency-domain measures for quantifying the distance between two speech frames x and y arise from the speech coding litterature. Ideally the subjective ear sensitivity should be formalised by incorporating psychoacoustic effects such as masking or isophone curves. A simple relevant measure is the spectral distortion (SD) defined as:

$$SD(x,y) = \sqrt{\int_{-\pi}^{\pi} (20\log_{10}|\frac{X(\omega)}{Y(\omega)}|)^2 \frac{d\omega}{2\pi}} \quad (2)$$

  where $X(\omega)$ and $Y(\omega)$ denote both signals spectra in normalized angular frequency. In (Paliwal and Atal, 1993), authors argue that a difference of about 1dB (with a sampling rate of 8kHz) is rather imperceptible. In order to have this point of reference between estimated and targeted sources we used the following measure:

$$SD(x,y) \approx \sqrt{\frac{2}{8000} \sum_{20}^{4000} (20\log_{10}|\frac{S_{estimated}(f)}{S_{reference}(f)}|)^2} \quad (3)$$

- **Glottal formant determination rate** : The amplitude spectrum for a voiced source (as shown in Figure 1) generally presents a resonance called *glottal formant*. As this latter parameter is an essential feature, an error on its determination after decomposition should be penalized. An example of relative error on the glottal formant determination is displayed in Figure 5 for $SNR = 50dB$. Many attributes characterizing a histogram can be proposed to evaluate a technique performance. The one we used for our results is the underlying surface between $\pm 10\%$ of relative error, which is an image of the determination rate given these bounds.

In the next Section results are averaged for all considered frames.

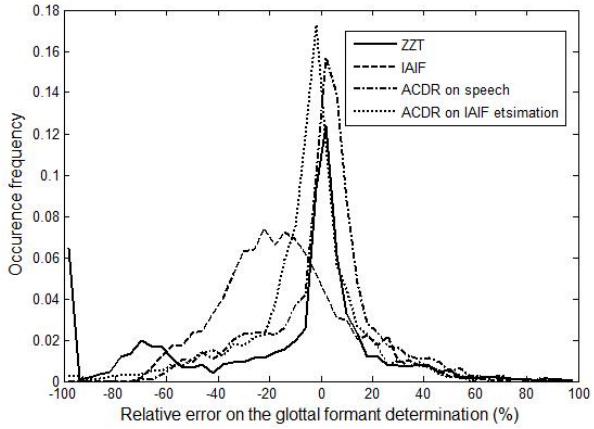

Figure 5: Histogram of relative error on the glottal formant determination (SNR=50dB).

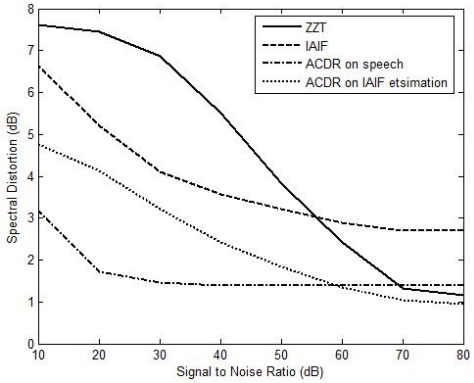

Figure 6: Impact of noise on the spectral distortion.

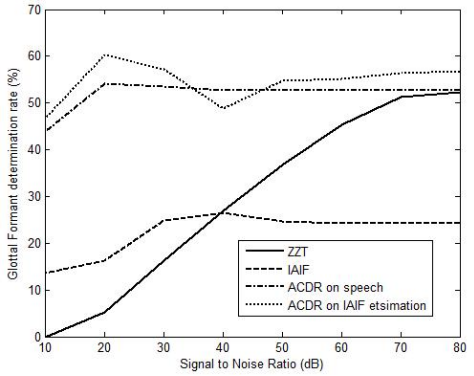

Figure 7: Impact of noise on the glottal formant determination rate.

4 RESULTS

A quantitative comparison between described methods is here presented. More precisely results are oriented so as to answer the two following questions: "*How techniques are sensitive to perturbations such as noise or GCI location error?*" and "*What is the impact of factors such as the fundamental frequency or the first formant on the decompostion quality?*".

4.1 Noise sensitivity

As a reminder a white Gaussian noise has been added to the speech signal at different SNR levels. This noise models not only recording or production noise but also every little deviation to the theoretical framework which distinguishes real and synthetic speech. Results according to both spectral distortion and glottal formant determination rate are displayed in Figures 6 and 7. Among all techniques, ZZT turns out to be the most sensitive. This can be explained by the fact that a weak presence of noise may dramatically perturb the roots position in the Z-plane, and consequently the decomposition quality. Interestingly the utility of applying our proposed ACDR concept is clearly highlighted (see notably the improvement when applied to the IAIF estimation). Even when directly performed on the speech signal, ACDR principle clearly yields robust and efficient results.

4.2 GCI location sensitivity

Another perturbation that could affect a method accuracy is a possible error made on the GCI location. Detecting these particular events directly on speech with a reliable precision is still an open problem although some interesting ideas have been proposed (Kawahara et al., 2000). Consequently it is rare that detected GCIs exactly match their ideal position when analyzing real speech. To take this effect into account we have tested the influence of a deviation to the real GCI location (GCIs are known for synthetic signals). Results for the glottal formant determination rate are shown in Figure 8 for clean conditions (no noise added). As mentioned in (Bozkurt et al., 2007), the ZZT technique is strongly sensitive to GCI detection, since this latter pertubation may affect the whole zeros computation. A similar performance degradation is also observed for ACDR-based methods due to their inherent way of operating. Nevertheless this effect occurs to a lesser extent.

4.3 The influence of pitch

Female voices are known to be especially difficult to analyze and synthesize. The main reason is their high fundamental frequency which implies to treat shorter periods. As a matter of fact the vocal tract response has not the time to freely return to its initial state be-

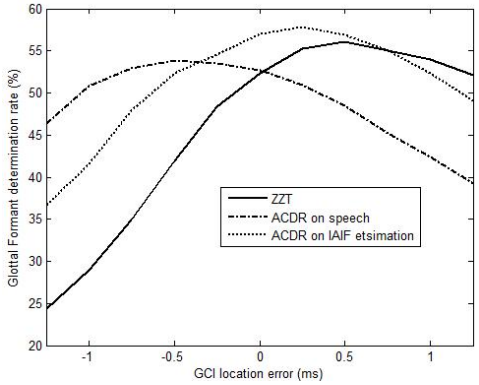

Figure 8: Impact of a GCI location error on the glottal formant determination rate (clean conditions).

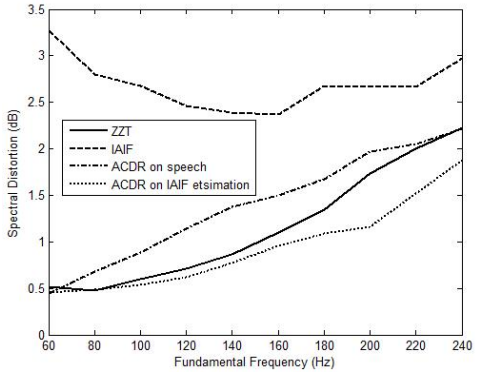

Figure 9: Impact of the fundamental frequency on the spectral distortion (clean conditions).

tween two glottal sollication periods. Consequently the performance of ACDR method applied to high-pitched speech will intrinsically degrade, as it relies on the assumption that the vocal tract response is negligible in the ACDR. This hypothesis turns out to be acceptable in a certain extent and might be reconsidered for high pitch values. Figure 9 presents the evolution of spectral distortion with respect to the fundamental frequency. Unsurprinsingly all methods degrade as the pitch increases, and this in a comparable way.

4.4 The influence of the first formant

In (Bozkurt et al., 2004), authors already reported erroneous glottal formant detection due to incomplete separation of F_1. As argued in previous subsection, particular configurations may lead to reconsider the assumption of ACDR applied on speech. More precisely, decomposition quality mainly depends on the 3 following parameters relative values: the pitch (F_0), the first formant (F_1) and the glottal formant (F_g). The

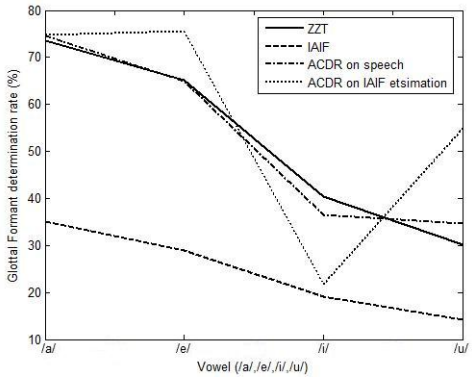

Figure 10: Impact of the first formant on the glottal formant determination rate (clean conditions).

greater is F_0 with regard to F_1 and F_g and the more severe will be the decomposition conditions. Intuitively this latter case can be interpreted as an ever increasing interference between causal and anticausal parts.
In our experiments filter coefficients were extracted by LPC analysis on four sustained vowels. Even though the whole spectrum may affect the decomposition, it is reasonable to consider that the effect of the first formant is preponderant. To give an idea, here are the corresponding first formant values: */a/:728Hz , /e/:520Hz , /i/:304Hz , /u/:218Hz*. The impact of the vowel on the decomposition accuracy is plotted in Figure 10. As expected a clear tendency of performance reduction as F_1 diminishes is observed.

5 CONCLUSION AND FUTURE WORK

This paper addressed the problem of source estimation robustness. A comparison between four different techniques was carried out on a complete set of synthetic signals. These latter methods were the Zeros of the Z-Transform (ZZT), the Iterative Adaptive Inverse Filtering (IAIF), and our proposed concept of Anticausality Dominated Region (ACDR) applied either directly on speech, or on a first source estimation (thanks to IAIF in our case). Two formal criteria were used to assess their quality of decomposition: the spectral distortion and the glottal formant determination rate. Robustness was first evaluated by adding noise to speech. In a general way this noise modeled every little deviation to the ideal production scheme. Interestingly both ACDR-derived methods were the most robust and efficient. Another perturbation we considered was a possible error made on the GCI location. In a second step the influence of the pitch (F_0)

and the first formant (F_1) was analyzed. Decomposition quality was interpreted as a trade-off between three amounts: F_0, F_1 and the glottal formant (F_g). In all our experiments ACDR-based techniques gave the more promising results.
As future work we plan to investigate the incorporation of these methods in the following fields:

- **Statistical parametric speech synthesis:** Hidden Markov models (HMM) have recently shown their ability to produce natural sounding speech (Tokuda et al., 2002). We already adapted this framework for the French language. A major drawback of such an approach is the "buzziness" of the generated voice. This inconvenience is typically due to the parametric representation of speech. Including a more subtle modeling of the voice source could lead to enhanced naturalness and intelligibility.
- **Expressive voice:** User-friendliness is one of the most important demand from the industry. Since expressivity is mainly managed by the source, an emotional voice synthesis engine should take into account realistic glottal source model parameters. Techniques presented in this paper could be used to estimate these parameters on speech samples extracted from an expressivity-oriented speech database.
- **Pathological speech analysis:** Speech pathologies are most of the time due to the irregular behaviour of the vocal folds during phonation. This irregular vibration can be induced by nodules or polyps on the folds and should result in irregular values of model parameters. Methods here presented could hence be used to estimate the glottal source and its features on pathological speech in order to quantify the pathology level.

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

Airas, M. (2008). *TKK Aparat: An environment for voice inverse filtering and parameterization*, volume 33, pages 49–64. Logopedics Phoniatrics Vocology.

Alku, P., Svec, J., Vilkman, E., and Sram, F. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–117.

Alku, P., Svec, J., Vilkman, E., and Sram, F. (2000). Analysis of voice in breathy, normal and pressed phonation by comparing inverse filtering and videokymography. In *ICSLP 2000, Proceedings of the International Conference on Spoken Language Processing*, pages 885–888.

Aparat (2008). Tkk aparat main page. `http://aparat.sourceforge.net/index.php/Main_Page`.

Bozkurt, B., Couvreur, L., and Dutoit, T. (2007). Chirp group delay analysis of speech signals. *Speech Communication*, 49(3):159–176.

Bozkurt, B., Doval, B., and Dutoit, T. (2004). A method for glottal formant frequency estimation. In *Proc. ICSLP, International Conference on Spoken Language Processing, Jeju Island (Korea)*.

Doval, B., d'Alessandro, C., and Henrich, N. (2003). The voice source as a causal/anticausal linear filter. In *Proceedings ISCA ITRW VOQUAL03, Geneva, Switzerland*.

El-Jaroudi, A. and Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Transactions on signal processing*, 39(2):411–423.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. In *STL-QPSR4*, pages 1–13.

Kawahara, H., Atake, Y., and Zolfaghari, P. (2000). Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay. In *ICSLP 2000, Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 664–667.

Paliwal, K. and Atal, B. (1993). Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Trans. Speech Audio Processing*, 1(1):3–14.

Sturmel, N., D'Alessandro, C., and Doval, B. (2007). A comparative evaluation of the zeros of z transform representation for voice source estimation. In *INTERSPEECH 2007, Antwerp, Belgium*, pages 558–561.

Tokuda, K., Zen, H., and Black, A. (2002). An hmm-based speech synthesis system applied to english. In *Proc. IEEE Workshop on Speech Synthesis 02, Santa Monica, USA*, pages 227–230.