A comparative evaluation of the Zeros of Z Transform representation for voice source estimation

Nicolas Sturmel, Christophe d'Alessandro and Boris Doval

LIMSI-CNRS
BP 133, F-91403
ORSAY, FRANCE
{nicolas.sturmel,cda,boris.doval}@limsi.fr

Abstract

A new method for voice source estimation is evaluated and compared to Linear Prediction (LP) inverse filtering methods (autocorrelation LPC, covariance LPC and IAIF [1]). The method is based on a causal/anticausal model of the voice source and the ZZT (Zeros of Z-Transform) representation [2] for causal/anticausal signal separation. A database containing synthetic speech with various voice source settings and natural speech with acoustic and electro-glottographic signals was recorded. Formal evaluation of source estimation methods are based on spectral distances. The results show that the ZZT causal/anticausal decomposition method outperforms LP in voice source estimation both for synthetic and natural signals. However, its computational load is much heavier (despite a very simple principle) and the method seems sensitive to noise and computation precision errors.

Index Terms: speech, lpc, zzt, inverse filtering, benchmark

1. Introduction

Voice source estimation is still a challenging problem for speech processing applications. Two broad classes of methods have been proposed so far: digital inverse filtering and source-tract deconvolution using the Fourier transform. On the one-hand, the Linear Predictive (LP) based methods take advantage of the autoregressive structure of the vocal tract acoustic filter. This filter structure is given by the linear acoustic model of speech production. LP can be extended to autoregressive and moving average (ARMA) inverse filter structures, but without clear improvement in source estimation. On the other hand Fourier Transform (FT) based methods take advantage of another feature of the speech production process, namely the multiplicative combination of source and filter in the frequency domain. For instance, homomorphic deconvolution methods (cepstrum) use the multiplication to addition transformation property of the logarithm for spectral envelope estimation. The Group Delay decomposition method is another Fourier transform based deconvolution method. Because of difficulties in phase unwrapping, these methods are scarcely used for voice source signal estimation.

Recently, another feature of speech production process has been pointed out and a new deconvolution algorithm has been proposed for exploiting this feature [3, 2].

The first keypoint of this new source estimation is the so-called Causal Anticausal Linear Model (CALM) of the voice source. It is shown in [3] that the glottal signal can be considered as an impulse train filtered by a causal/anticausal linear filter. Then the source-tract separation problem can be considered as a causal and anticausal filters identification problem. As for digital filters, causal (resp. anticausal) poles are placed inside (resp. outside) the unit circle, a simple criterion can be applied for sorting causal and anticausal contributions to the spectrum.

The second key point of the separation algorithm is a method for causal and anticausal component estimation and separation: the so-called Zero of the Z transform (ZZT) signal representation [2]. In this method, a Z-transform polynomial is formed using the (windowed) signal samples of a few speech periods. Precise centering of the analysis frame on the Glottal Closing Instant (GCI) is of paramount importance. The roots of this polynomial are the zeros of the Z-transform. It can be shown that roots outside (resp. inside) the unit circle correspond to the anticausal part of the voice source (resp. the causal part of source and vocal tract). A simple algorithm for source/filter separation has been proposed in [2]. This new method gave very promising results. However, to go a step further, a formal comparative evaluation of the method is needed. The aim of the present work is a formal evaluation of the performances of the new method for voice source estimation, and a comparison with the results of the best inverse filtering method currently used in speech processing applications.

The paper is organized as follows. In the next section, the ZZT source estimation algorithm is presented. Three LP-based inverse filtering methods commonly used for source estimation are described in section 3. Section 4 reports on our experiments, based on a database of synthetic and natural speech signals with variable source parameter settings, and distance measures for source estimation. Section 5 discusses the results obtained and section 6 concludes.

2. ZZT source/tract separation

The implementation of the ZZT inverse filtering is based on the algorithm presented in [2], illustrated on figure 1. The different steps of the algorithm are as follows :

0. First, GCI points need to be detected, either by synchronous electro-glottographic (EGG) recording analysis or by direct extraction from the speech signal itself.
1. The signal is truncated in two periods frames, using GCI informations.
2. Each frame is windowed using a two period long blackman window.

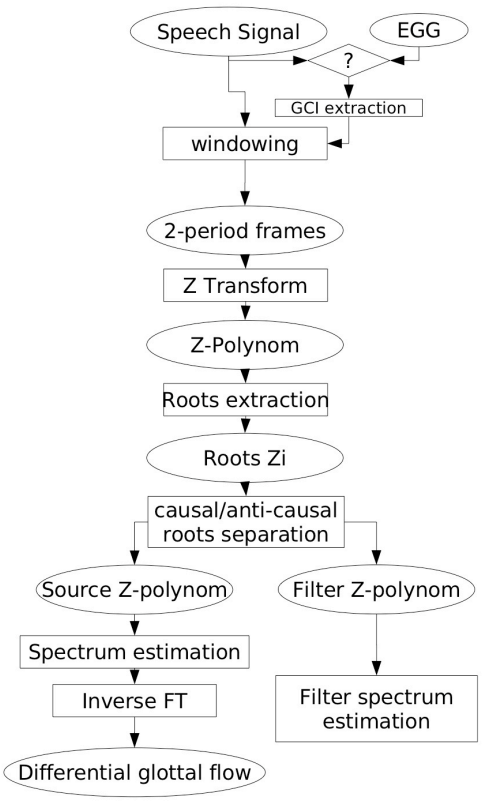

Figure 1: ZZT algorithm, for differential glottal flow waveform estimation

3. Roots Z_i of the associated polynomial P (the frame Z-transform) are computed :

$$P(z) = \sum_{i=0}^{N-1} s(i)z^{-i} = z^{1-N} \prod_{i=1}^{N-1} (z - Z_i) = z^{1-N}\tilde{P}(z)$$

4. The zeros are sorted for separation of the anticausal (G) and causal (F) components of the signal, note that $P(z)|_{|Z_i|=1}$ corresponds to a periodic component, unlikely to happen in non harmonic spectrum signals :

$$\tilde{P}(z) = \prod_{i=1}^{K} (z - Z_i)|_{|Z_i|>1} \prod_{j=1}^{N-1-K} (z - Z_j)|_{|Z_j|<1} = G(z)F(z)$$

5. Causal and Anticausal spectra are computed
6. Glottal flow waveform can therefore be obtained by inverse FT.

3. LP based inverse filtering

The other tested methods are based on the Linear Prediction of speech [4, 5] :

1. The autocorrelation LP algorithm performed on the whole speech sample as described in [5]. The autoregressive filter order used here is $2 + 2\frac{Fs}{1000}$. (18 at $Fs = 16kHz$)

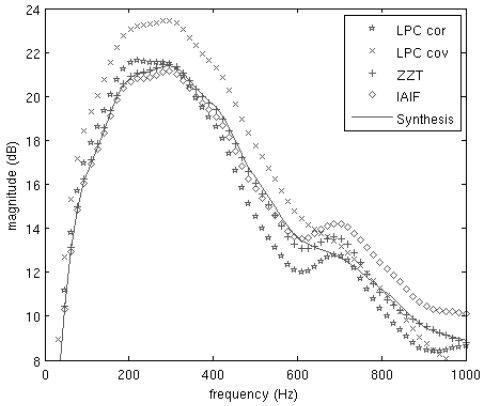

Figure 2: Close up on the differential glottal spectra from a synthesized speech sample analysis.

The differential glottal flow is obtained by filtering the speech signal using the inverse autoregressive (AR) filter

$$g(n) = \sum_{i=0}^{N} A(i)s(n-i) \quad \text{from [5]}$$

where $A(i)$ are the coefficients of the estimated filter.

2. The covariance LP algorithm. Since this algorithm can be performed on short speech segments, we implemented a pitch synchronous closed phase covariance estimation [6] while assuming O_q at an average closed phase duration of about 50% of the period. Like the first one, this algorithm gives the autoregressive vocal tract filter coefficients then used to obtain the differential glottal flow as above. The order used here was limited by the number of samples during the glottal closed phase : never higher, but sometimes lower than 18.

For those two methods a pre-emphasis filter is used to improve estimation, with transfer function $1 - c_1 z$ ($c_1 = 0.98$)

3. Alku's Iterative Adaptive Inverse Filtering algorithm [1], which directly estimates the derivative glottal flow. Knowing that the choice of the pre-emphasis coefficient is determinant to perform the best vocal tract's filter estimation, an iterative adjustment of the c_1 pre-emphasis coefficient is the key of this method, leading theoretically to a better estimation of the glottal flow.

4. Experiments

Synthetic test signals are generated using the LF model [7] rewritten as in [8] and a formant synthesizer with three implemented filters : a synthetic /a/ computed according to common formant frequencies and bandwidths and two natural filters obtained by lpc analysis : */i/* and */u/*.

Open quotient (O_q), Fundamental period (T_0) and Asymmetry (α_m [8], corresponding to $\frac{Tp}{Te}$ parameter in the LF model) are varying among test condition. The remaining parameter, the return phase quotient, is set to zero as it belongs to the causal part of the differential glottal flow and can therefore not be estimated by ZZT. Preliminary tests showed that the effect of the return phase quotient on lp-based source estimation is negligible. This parameter is then not considered in this study.

The test is performed with the following parameter variations:

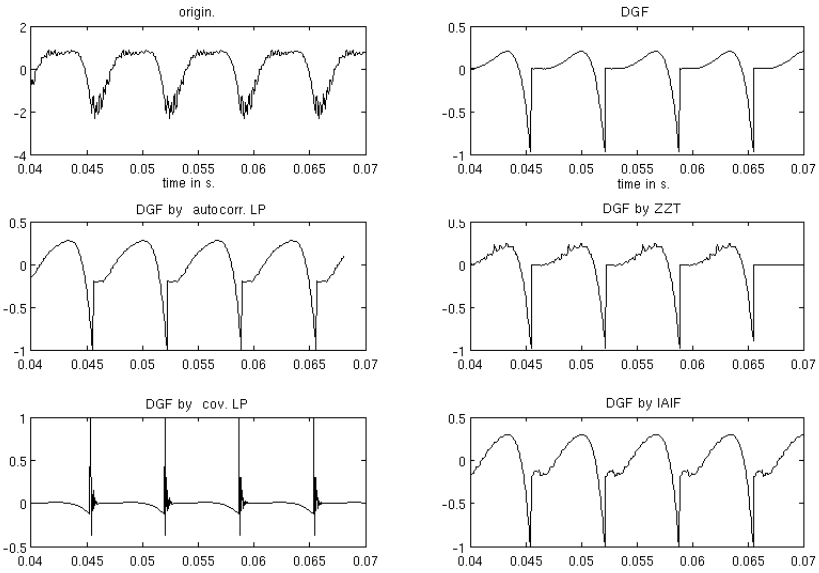

Figure 3: Comparison of the 4 methods on a synthetic signal. A synthetic signal (top left) was synthesized by filtering the glottal flow waveform (top right) by an /i/. The parameters are : $F_0 = 150Hz$, $O_q = 0.8$ and $\alpha_m = 0.8$. The middle and low rows show the differential glottal flow estimates for each method.

- F_0 (90, 110, 150, 190, 230, 270, 330 Hz)
- O_q from 0.3 to 0.9 by 0.05 steps (13 values)
- α_m from 0.6 to 0.85 by 0.05 steps (6 values)
- Vowels among the three chosen (/a/, /i/, /u/)
- Noise/Signal ratio : $-300dB$ (noiseless signal), -60dB (recording white noise simulation)

A total of 3276 test conditions are computed according to these parameter variations.

The comparison criterion is based on the quadratic spectral energy difference (E) between the synthesized differential glottal flow log-magnitude spectrum (G_{synth} in dB) and the log magnitude spectrum obtained by inverse filtering (G_{if} in dB) on a given frequency range and for a given method as illustrated on figure 2.

Lp-based methods lack accuracy for high frequency estimation of the differential glottal flow ; so a reduced frequency range ($0Hz$ to $4000Hz$) is used for computing this error, otherwise, the ZZT decomposition would have been advantaged, because of its intrinsic absence of return phase :

$$e = \sqrt{\sum_{f=0}^{4000} (\|G_{if}(f)\| - \|G_{synth}(f)\|)^2}$$

Spectrums are computed on two periods (GCI centered) multiplied by a Blackman window. On figure 2, we can see an example of error between estimated and original differential glottal waves spectra. The corresponding errors e are : 10 for autocorrelation LP, 180 for covariance LP, 6 for ZZT, and finally 23 for IAIF method.

In order to summarize the results obtained by the 3276 tests, cumulative errors C are computed for a given set of parameters. The cumulative error is the mean of e depending on O_q, α_m, f_0, vowel, noise (N) and decomposition algorithm.

$$C = \frac{\sum e(O_q, \alpha_m, f_0, vowel, noise, algorithm)}{card(e)}$$

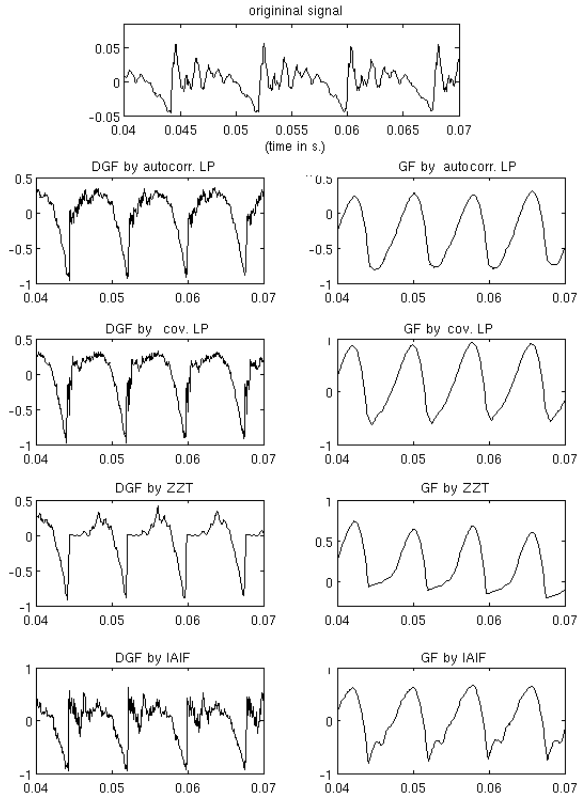

Figure 4: Comparison of the 4 methods on a real signal. From the real speech sample (top), the differential glottal flow estimate (left) and glottal flow estimate (right) is reported for each method. Male speaker, vowel /a/, $F_0 = 120Hz$, EGG measurements gave $O_q \approx 0.5$.

vowel /a/					
O_q	α_m	autocorr.	cov. LP	ZZT	IAIF
0.3	0.65	56.2	130.	**12.6**	53.1
0.3	0.8	53.3	114.	**14.0**	40.3
0.5	0.65	44.6	224.	**12.8**	52.0
0.5	0.8	36.1	206.	**11.3**	37.7
0.7	0.65	46.4	145.	**29.3**	50.2
0.7	0.8	28.0	137.	**11.0**	44.1

vowel /i/					
O_q	α_m	autocorr.	cov. LP	ZZT	IAIF
0.3	0.65	72.0	74.8	**24.2**	53.3
0.3	0.8	76.6	85.2	**15.8**	35.7
0.5	0.65	51.6	69.0	**16.9**	51.5
0.5	0.8	49.1	64.6	**18.7**	28.1
0.7	0.65	67.1	122.	**22.0**	69.0
0.7	0.8	34.0	108.	**16.4**	28.2

vowel /u/					
O_q	α_m	autocorr.	cov. LP	ZZT	IAIF
0.3	0.65	66.3	81.3	**24.2**	68.4
0.3	0.8	73.9	90.7	**24.1**	64.3
0.5	0.65	58.0	73.9	**31.0**	71.3
0.5	0.8	48.6	74.6	**23.6**	39.9
0.7	0.65	63.9	61.2	**42.6**	82.4
0.7	0.8	35.9	63.9	**25.2**	31.2

Table 1: Cumulative errors of each method for different values of open quotient O_q and asymmetry coefficient α_m for the three vowels. Errors are averaged on every fundamental frequency and noise level. *The best results are in bold*

5. Results

Direct results of the benchmark are the cumulative spectral distances presented on table 1. Lowest values represent the closest spectrum to the original used for synthesis ; note that in every condition ZZT gives the best result by far, and that IAIF and autocorrelation LP show both good performances too.

Source estimation examples are presented in figures 3 and 4. Figure 3 presents the estimated source waveforms for a synthetic speech signal /a/. Note that the original synthetic source is known, and can be compared directly to the estimated source waveforms. Figure 4 presents source estimations for a real speech signal. Both glottal flow and its derivative are shown for each method. An Electroglottographic reference is available for this example, showing that the open quotient is about 0.5 (i.e. the closed phase of the source is about half of the period). In the example, the ZZT is the only method giving a closed phase of about 0.5.

Spectral distance results and visual inspection of the waveforms are leading to the following observations:

– The pitch synchronous covariance linear prediction seems the worst differential glottal wave estimator. Since it is performed on a very short signal segment, the autoregressive filter order may probably be too small for accurate estimation of the vocal tract filter. Nevertheless, the overall low frequency restitution of the glottal formant is realist.

– The IAIF methods seems the most robust one tested in this paper in the sens that it gives good results in almost every case : the adaptive part of the algorithm appears to be useful for fitting even the worst signals. However noise and ripples on the estimated differential glottal waveform make it hardly usable for parameter extraction or analysis.

– The auto-correlation linear prediction is surprisingly the better LP-based source estimation in this benchmark. However, tests on signals are using long analysis window, exploiting the time invariance assumed by the method. This is not always realistic for actual time-varying real speech signals. Furthermore, it can be seen on table 1 that the worst cases are those were the pre-emphasis does not completely suppress the glottal formant : low O_q values leading to a glottal formant at two or three times F_0, and low values for α_m leading to a more resonant formant.

– The ZZT inverse filtering outperforms lp-based methods both in spectral measurements and time-domain observations. The absence of ripples in the glottal closed phase together with the very good benchmark results are the strongest arguments in favor of this method. On real signals, it is the only one to present a clearly visible closed phase on glottal flow waveforms (figure 4). The low error values achieved during benchmark make ZZT the best choice for glottal parameter estimation by model fitting. However, the method relies heavily on precise glottal closing instants determination, and it seems also relatively weak for low signal to noise ratio. Computational load is heavier than for LP based methods, because it is based on roots extraction from a high degree polynomial.

6. Conclusion

In this paper a new deconvolution method has been evaluated, based on an all zero estimation of the speech causal/anticausal linear model. It was compared with linear prediction based inverse filtering. The results showed that ZZT is a promising deconvolution method, able to outperform LP inverse filtering in every speech condition. Moreover the ZZT seems to be powerful for estimation of the glottal flow and its parameters, such as O_q. The main drawbacks being iare computational cost, lack of robustness for noise corrupted signals and the paramount importance of accurate glottal closing instant detection.

7. References

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

[2] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE*, vol. 12, pp. 344–347, 2004.

[3] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," *proc. VOQUAL'03, ISCA Workshop, Geneva*, Aug. 2003.

[4] J. D. Markel and A. H. Gray Jr, *Linear Prediction of Speech*, Springer Verlag, Berlin, 1976.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63(5), pp. 561–580, 04 1975.

[6] Wong D., Markel J., and Gray A. Jr, "Least squares glottal inverse filtering from an acoustic speech wave," *IEEE trans.*, vol. 35, pp. 350–355, 1979.

[7] G. Fant, J. Liljencrants., and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[8] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *acta acustica united with acustica*, vol. 92, pp. 1026–1046, 2006.