A source-filter separation algorithm for voiced sounds based on an exact anticausal/causal pole decomposition for the class of periodic signals.

Thomas Hézard[1], *Thomas Hélie*[1], *Boris Doval*[2]

[1]Institut de Recherche et de Création Acoustique-Musique
(IRCAM - CNRS UMR 9912 - UPMC), Paris, France
[2]Lutherie-Acoustique-Musique (LAM) - Institut Jean Le Rond dAlembert (IJLRA), Paris, France
thomas.hezard@ircam.fr, thomas.helie@ircam.fr, boris.doval@upmc.fr

Abstract

This paper addresses the source-filter separation problem in the context of causal/anticausal linear filter model of voice production. An algorithm based on standard signal processing tools is proposed for the class of quasi-periodic signals (voiced sounds with quasi-stationary pitch). At first, a one-period frame of an equivalent stationary infinitely periodic signal is built. A particular attention is given to the problems of windowing and temporal aliasing. Secondly, an exact pole decomposition of this signal is computed within the class of T_0-periodic signals. Finally, the glottal closure instant (GCI) and the causal-anticausal factorization of the initial frame are jointly estimated from the latter decomposition. The performance of this algorithm on synthetic signals is demonstrated and the performance on real speech is discussed. In conclusion, application of this new algorithm in a complete voice analysis-synthesis system is discussed.

Index Terms: speech analysis, source-filter separation, causal-anticausal decomposition

1. Introduction

The source-filter model of voice production, based on acoustic theory (see for example [1]), is composed of a source -the glottal flow-, a filter corresponding to the vocal tract and a filter corresponding to the radiation at the lips. Generally, in signal processing applications, the radiation filter is a "derivative" filter and can be combined to the source signal. The source-filter model is then reduced to the pair derivative glottal flow signal model and vocal tract filter model.

The literature presents a wide variety of models for the derivative glottal flow. One can read [2] or [3] for a recent overview of these models. Most source models are temporal parametric models, the most common one being the Liljencrants-Fant (LF) model [4]-[5]. However, the glottal source can also be modelled by a linear filter. In this case the glottal flow is considered to be the response of the glottal filter to an excitation made of Dirac pulses. In our study, we focus on this family of models. As for the vocal tract model, most models can be expressed as an all-pole filter.

As it exists a wide variety of parametric and non-parametric source-filter models of voice production, it also exists a wide variety of estimation methods. [1] and [6] present a quick overview of such methods. The most common approach to estimate the parameters of an all-pole model is the linear prediction analysis [7].

In this paper, we investigate an approach to perform an exact source-filter deconvolution based on an all-pole causal/anticausal model, inside the space Π of T_0-periodic signals. The model and the projection operator $\mathcal{P}$ on Π is presented in section 2. Then, we introduce an operator $\mathcal{H}$ which converts poles into zeros on Π (for the Z-transform). In section 3, operator $\mathcal{H} \circ \mathcal{P}$ is used as a first step of an algorithm to estimate the full set of (non parametric) poles, from which a subset of significant poles is selected jointly to the GCI estimation. The exact reconstruction is verified on synthetic signals in section 4. Finally, in section 5, after an illustration on real signals, we give perspectives to build a robust method based one this approach.

2. Causal/anticausal model

2.1. Source model

The CALM model [8] describes the glottal source as an all-pole filter, composed of one pair of complex conjugate anticausal poles and one real causal pole. The anticausal part of the CALM filter impulse response, which corresponds to the open phase, is an exponentially increasing sinusoid. The causal part, which corresponds to the return phase, is a decreasing exponential.

In our study, we consider the Z-transform of the glottal filter

$$H(z) = \frac{1}{(1-az^{-1})(z-b)(z-\bar{b})}\,, \tag{1}$$

where a ($|a|<1$) is the real causal pole and $\{b,\bar{b}\}$ ($|b|>1$) is the pair of complex conjugate anticausal poles.

2.2. Vocal tract model

The vocal filter model considered here is composed of pairs of complex conjugate causal poles. As usual in source-filter analysis (see for instance [7]), we choose the order of the vocal filter such as the number of pair of complex conjugate poles corresponds to the the Shannon frequency divided by 1000. In other words, the filter response contains one pair of poles (one formant) for every $1000\,Hz$.

In our study, we write the Z-transform of the vocal filter

$$V(z) = \frac{1}{\prod_{k=1}^{K}(1-\alpha_k z^{-1})}\,. \tag{2}$$

In the following, $\tilde{u}(z)$ stands for the Z-transform of $u(n)$.

2.3. Complete model

As we consider only infinitely T_0-periodic signals (space Π), the Z-transform of the complete signal model can be written

$$S(z) = GV(z)H(z)\widetilde{\text{III}}_{T_0,t_i}(z)\,, \tag{3}$$

where $\text{Ш}_{T_0,t_i}(n)$ stands for the Dirac comb of period T_0 centered on time t_i, and G is the gain. t_i defines the location of the commonly called glottal closure instants (GCIs). $V(z)H(z)$ is an all-pole z function with $K+3$ poles and can be decomposed into its causal and anticausal part

$$\frac{1}{(1-az^{-1})\prod_{k=1}^{K}(1-\alpha_k z^{-1})} \text{ and } \frac{1}{(z-b)(z-\bar{b})}.$$

The parameters of the complete model are

$$\theta = [G, a, b, \{\alpha_k\}_{k\in[1,K]}, K, T_0, t_i]^T. \quad (4)$$

3. An algorithm for all-pole causal/anticausal decomposition

Our goal is to retrieve parameters θ that best describe a finite-length extract of a speech signal in the sense of the model described in section 2. We suppose that this extract is quasi-stationary, meaning we work on short frames of speech on which we can consider that glottal source and vocal filter parameters are invariant, typically 20 ms segments. Here is the description of the algorithm we developed for this problem. The performances of this algorithm will be discussed in section 4.

3.1. General description of the algorithm

The main difficulty of poles estimation comes from the infinite-length of the support of pole-type signals. Finite-length signals have all-zeros Z-transforms. Hence, as we work in practice with finite-length signals, poles estimation seems doomed to fail. Another way of seeing the problem is that, as it has been highlighted in [9] and [10], windowing the signal has a drastic influence on the Z-transform that is extremely difficult to study analytically. The algorithm we propose offers a way to solve this problem by transforming poles into zeros in the class of periodic signals, making the support length finite.

The first step is to build an infinitely periodic signal s from the original extract. The second step is to turn the poles into zeros with an appropriate operator $\mathcal{H}$. The third step is to factorize the Z-transform of $\mathcal{H}[s]$ to compute its zeros, which are the poles of the Z-transform of s. Then, a selection of the meaningful zeros is performed. Finally, we can extract from the factorization an estimation of the parameter t_i. A schematic representation of this algorithm is presented in Figure 1.

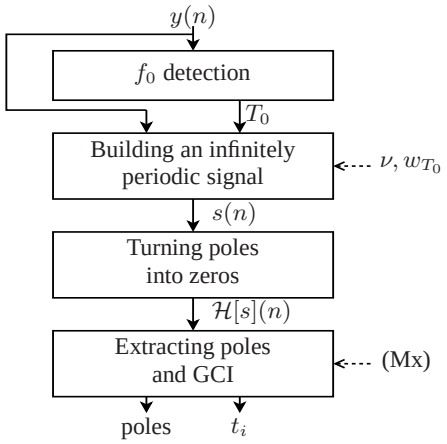

Figure 1: General scheme of the all-pole causal/anticausal decomposition algorithm

3.2. Class $\mathcal{C}$ of infinitely periodic signals

In order to build an infinitely periodic signal $s(n)$, we need to know the periodicity of the input signal $y(n)$. This can be done with the autocorrelation technique or any other f_0 estimation algorithm. In the actual version of the algorithm, we consider only integer periods T_0 (expressed in samples).

Signal $s(n)$ of class $\mathcal{C}$ is defined from y by

$$s(n) = \mathcal{P}[y](n) \underset{def}{=} \big(y(n) \times w_{T_0}(n)\big) * \text{Ш}_{T_0,\nu}(n) \quad (5)$$

where $*$ stands for the convolution operator and ν is a chosen instant representing the beginning of one period. $w_{T_0}(n)$ is a window that can be chosen to "mean" several periods of the signal $y(n)$ but must verify the property

$$\sum_{k=-\infty}^{+\infty} w_{T_0}(n-kT_0) = 1 \quad \forall\, n \in \mathbb{Z}. \quad (6)$$

This property ensures that, if $y(n)$ is an truncated version of an infinitely periodic signal, $s(n)$ retrieves the exact original infinitely periodic signal. For example, one can choose

(W1) $w_{T_0}(n) = \mathbb{1}_{[\nu,\nu+T_0-1]}(n)$ or

(W2) $w_{T_0}(n) = \cos^2\left(\frac{(n-\nu)\pi}{2T_0}\right)\mathbb{1}_{[\nu-T_0,\nu+T_0-1]}(n)$.

Note that (W1) selects one period of the signal beginning at the instant ν, (W2) averages two periods of the signal around the instant ν.

3.3. Turning poles into zeros

Turning poles into zeros can be done using the operator

$$\mathcal{H} \,:\, s \in \Pi \mapsto DFT^{-1}\left(\frac{1}{DFT(s)}\right) \in \Pi \quad (7)$$

where s is the infinitely periodic signal (5), DFT stands for the Discrete Fourier Transform and DTF^{-1} for its inverse. It is obvious that $\widetilde{\mathcal{H}[s]}(z)$ is the inverse of $\tilde{s}(z)$. Hence, poles of $\tilde{s}(z)$ are the zeros of $\widetilde{\mathcal{H}[s]}(z)$ and vice versa.

3.4. Extracting the desired poles

Computing the zeros of the signal $\mathcal{H}[s](n)$ is possible with numerical methods. As we do not want that the periodicity of the signal $s(n)$ interfere with the poles research, we simply have to factorize the Z-transform of one period of $\mathcal{H}[s](n)$, noted $\widetilde{\mathcal{H}[s]}(z)$. In practice, $\mathcal{H}[s]$ is computed over one period of $s(n)$ using the FFT algorithm. The resulting signal $\mathcal{H}[s](n)$ is then factorized with a numerical roots finder to compute its zeros. These zeros are the poles of $\tilde{s}(z)$.

The only question left is how to select the poles. The complete factorization of $\widetilde{\mathcal{H}[s]}(z)$ gives $T_0 - 1$ poles. However, the model we proposed has $K+3$ poles. Three methods are proposed to reduce and/or impose the number of poles.

(M1) The factorization is performed on $K+4$ consecutive coefficients in $\widetilde{\mathcal{H}[s]}(z)$. The selection of these $K+4$ coefficients is done by minimizing the reconstruction error, which is defined as the 2-norm of the difference between the complex spectra of the original signal and the reconstructed signal.

(M2) The factorization is performed on $K+4$ consecutive coefficients in $\widetilde{\mathcal{H}[s]}(z)$ (consecutive in the sense of circular permutations), ensuring to retrieve $K+3$ poles. The selection of these $K+4$ coefficients is done by minimizing the norm (n-norm, $n \in \mathbb{N}^*$) of the unselected coefficients in $\widetilde{\mathcal{H}[s]}(z)$. This method amounts to select the most influential coefficients.

(M3) The factorization is performed on the whole $\widetilde{\mathcal{H}[s]}(z)$. The $K+3$ poles with the maximum residues are selected, which amounts to select the most influential poles.

Note that (M2) and (M3) gives us the liberty to let the algorithm decide the number of poles. It simply needs to replace the minimization of the remaining coefficients (or residues) by thresholding the remaining coefficients (or residues).

Finally, the separation between the causal and the anticausal component is automatically achieved by selecting the pole inside the unit circle for the causal component and outside the unit circle for the anticausal component. Hence, it is possible to separate the anticausal part of the glottal source from the rest of the signal.

3.5. Estimating the parameter t_i

Estimating the parameter t_i amounts to detecting the position of the unique GCI inside the period of $s(n)$. It appears that this is automatically done by the algorithm in the previous step. The factorization of $\widetilde{\mathcal{H}[s]}(z)$ gives M_c causal poles (of absolute value smaller than one) and M_a anticausal poles (of absolute value greater than one). Using the definition of the causality, we can recover t_i,

$$t_i = \nu + d_{opt} + M_a \,, \tag{8}$$

where ν is the window position chosen in section 3.2 and d_{opt} is the position of the first coefficient selected by the (Mx) method in section 3.4.

3.6. General remarks on the algorithm

As we'll see on section 4, this algorithm lets us exactly recover the parameters for signals corresponding to the model, in the "ideal case". Moreover, one can easily show that, in this case, the algorithm results is independent of the choice of

- w_{T_0} as long as it verifies (6),
- ν as long as the support of y contains the support of w_{T_0}.

However, these choices (and the choice between (M1), (M2) and (M3)) can be very important in other cases. In particular, we observed that (M1) gives the best results but it's also the most resources-consuming method. It is interesting to highlight that in the ideal case, $\tilde{s}(z)$ is precisely the same as $GV(z)H(z)$ (for the right choice of ν). Another way of saying this is that one period of $s(n)$ is a periodic summation of the impulse response to the filter $GV(z)H(z)$ and then contains the whole information about the filter.

4. Tests on synthetic signals

We build an "almost ideal case" by filtering a very long Dirac comb of period T_0 and then extract a few periods in the middle of the filtered signal. This ensures that the part of the infinite-response of the filter which is not taken into account is negligible. The glottal source and filter parameters are chosen within classic human speech values. Complex conjugate pairs of poles are generated from frequency and Q factor values. Figure 2 illustrates the behaviour of the algorithm with choices

- ν set at the half of the length of signal y,
- (W2),
- (M2) with 2-norm,

on a synthetic signal with the parameters

- $Fs = 10$ kHz, $f_0 = 200$ Hz, $G = 1$, $K = 2 \times 4$
- $a = 0.8$, $f_b = 300$ Hz, $Q_b = 3$,
- $\{f_k\} = \{0.9, 1.2, 3, 4\}$ kHz, $\{Q_k\} = \{5, 15, 40, 15\}$.

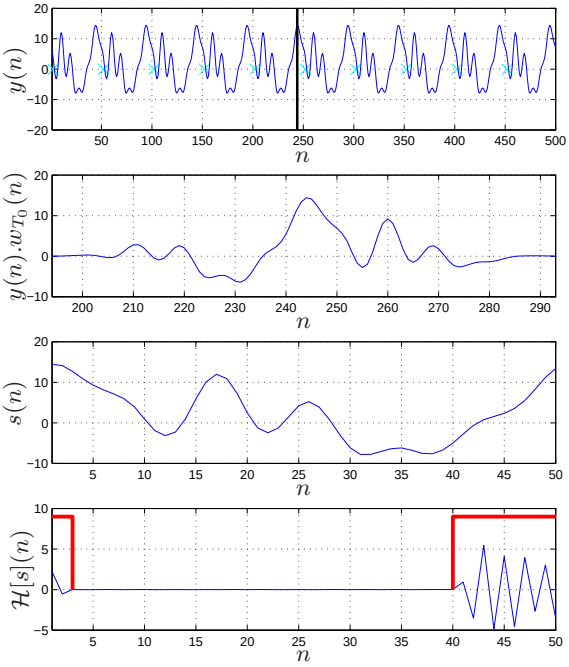

Figure 2: Illustration of the algorithm. From top to bottom: 10 periods of an "almost ideal case" signal with GCIs (x) and the choice of ν (|), $y(n).w_{T_0}(n)$ with the choice (W2), one period of $s(n)$, $\mathcal{H}[s](n)$ with the optimal choice of $K+3$ coefficients.

Results of the algorithm are presented on Figure 3. One can see that the reconstruction is perfect for this almost ideal case. Poles are exactly estimated and the signal is very precisely reconstructed. Note that the GCIs are perfectly estimated.

The algorithm is working perfectly in the ideal case, we tested the algorithm with slight shifts from the ideal case. Still in synthetic speech, we tested the influence of

- a bad estimation of the fundamental period T_0,
- the presence of noise in the signal,
- the number of estimated poles.

Unfortunately, the two first bring drastically down the performances of the algorithm. A slight error in the T_0 estimation makes the poles estimations far from the truth. Introducing some noise in the signal makes the reconstructed signal tend to a flat-spectrum signal. As for the last point, the algorithm can't reconstruct the signal if the the number of estimated poles is smaller than the real number of poles but gives perfect results if the number of estimated poles is greater than the real number

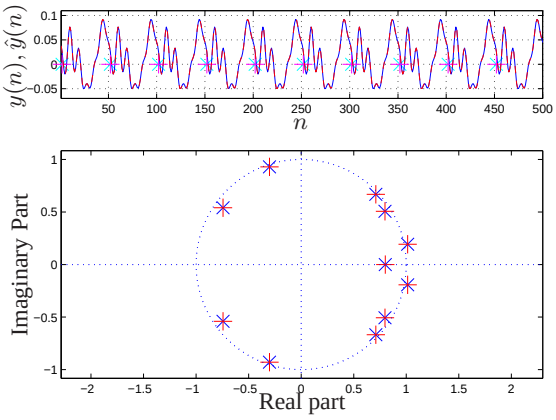

Figure 3: Results of the algorithm. From top to bottom: signal $y(n)$ (-) and reconstructed signal $\hat{y}(n)$ (- -) along with GCIs (x) and estimated GCIs (+), poles of the model (x) and estimated poles (+).

of poles. In this latter case, the algorithm finds some poles that are not in the original signal but with either very small absolute value or very small residue, so that their influence on the reconstruction is negligible.

5. Discussion on robustness

As one can guess given the results with non ideal cases, our algorithm is still very uncertain for real speech signals. We tested it on sustained vowels pronounced by a male speaker in a low register. Figures 4 and 5 present results of the analysis on a segment of a vowel /e/ with a fundamental frequency of 73 Hz. Figure 4 presents the results obtained using a model with 33 poles and Figure 5 presents the results obtained using a model with 13 poles. This latter corresponds to the classic choice described in section 2. If the performances are globally poor, we can observe several informative behaviours of the algorithm. At first, with a low order model, the causal/anticausal decomposition tends to correspond to a low frequency / high frequency decomposition, which is globally coherent with the glottal/source decomposition. Then, we can observe that the glottal formant frequency seems to be well estimated in each case. Finally, we can observe that the reconstruction is much better for the low frequencies than for the high frequencies. This is probably due to the sensibility of the algorithm to the noise.

6. Conclusion and perspectives

We presented a new algorithm for parametric causal/anticausal decomposition of speech signals. We demonstrated that the algorithm is perfectly effective for signals of class $\mathcal{C}$. This algorithm led us to define a new operator on $\mathcal{C}$: operator $\mathcal{H}$ which turns poles into zeros and is easy to implement. Unfortunately, we saw that the algorithm is severely sensitive to noise and errors on the T_0 estimation. However, some regularization techniques are under consideration and may improve the performances of this algorithm. Firstly, the operator $\mathcal{H}$ could be computed using a Wiener deconvolution to decrease the sensitivity to noise. Another perspective to make the method more robust would consist of regularizing $\mathcal{H} \circ \mathcal{P}$ by considering a model with a reduced number of poles from the beginning. At last, a variant of this algorithm taking into account the noise in the signal and using a likelihood minimization on the complex cepstrum is currently being developed.

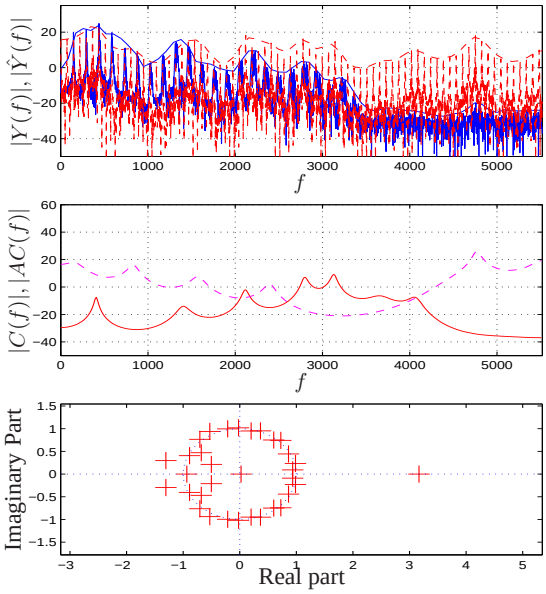

Figure 4: Results of the algorithm for real speech with high order model. Top: Spectrum of original (-) and reconstructed signal (- -). Middle: Spectrum of causal (- -) and anticausal (-) components. Bottom: estimated poles.

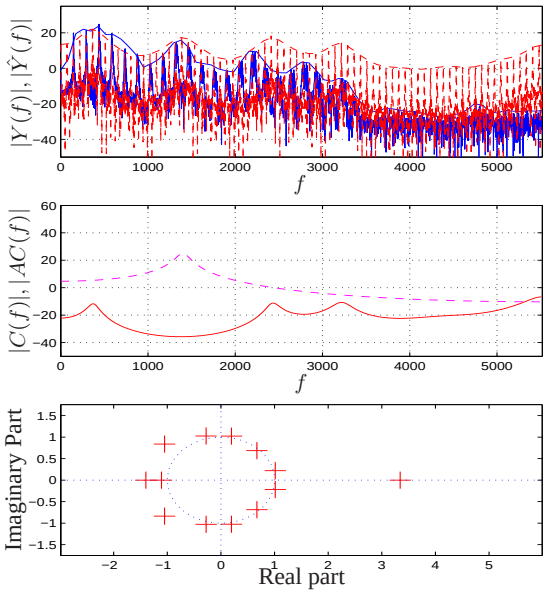

Figure 5: Results of the algorithm for real speech with low order model. See Fig. 4.

7. Acknowledgements

Author thank Erkki Bianco and Gilles Degottex from IRCAM who built the speech audio databases used in this article. Thomas Hézard wishes to thank René Caussé from IRCAM for his accompaniment in this work.

8. References

[1] T. F. Quatieri, *Discrete-time speech signal processing*. Prentice Hall, 2002.

[2] B. Doval, C. D'Alessandro, and N. Henrich, "The Spectrum of Glottal Flow Models," *Acta Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[3] G. Degottex, "Glottal source and vocal-tract separation," Ph.D. dissertation, 2010.

[4] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[5] G. Fant, "The LF-model revisited. Transformations and frequency domaine analysis," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[6] L. Rabiner and R. Schafer, *Theory and applications of digital speech processing*. Pearson Education, 2011.

[7] J. D. Markel and A. H. Gray, *Linear prediction of speech*, B. Springer-Verlag, Ed., 1976.

[8] B. Doval, C. D'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *Voice Quality: Functions, Analysis and Synthesis VOQUAL'03*, 2003, pp. 16–20.

[9] B. Bozkurt, "Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals," PhD thesis, Faculté Polytechnique de Mons, 2005.

[10] T. Drugman, "Advances in Glottal Analysis and its Applications," PhD thesis, University of Mons, 2011.