Zeros of Z-Transform (ZZT) Decomposition of Speech For Source-Tract Separation

Baris Bozkurt, Boris Doval, Christophe D'Alessandro, Thierry Dutoit

Faculté Polytechnique De Mons, TCTS Lab,
Initialis Scientific Park, B-7000 Mons, Belgium, {bozkurt,dutoit}@tcts.fpms.ac.be,
LIMSI,
CNRS, Po Box 133 – F91403 Orsay, France
{boris.doval,cda}@limsi.fr

Abstract

This study proposes a new spectral decomposition method for source-tract separation. It is based on a new spectral representation called the Zeros of Z-Transform (ZZT), which is an all-zero representation of the z-transform of the signal. We show that separate patterns exist in ZZT representations of speech signals for the glottal flow and the vocal tract contributions. The ZZT-decomposition is simply composed of grouping the zeros into two sets, according to their location in the z-plane. This type of decomposition leads to separating glottal flow contribution (without a return phase) from vocal tract contributions in z domain.

1. Introduction

Source-tract separation has been studied by many researchers with various techniques among which the most common is inverse filtering ([1]). Inverse filtering techniques aim removing the vocal tract contribution from speech signals to obtain the glottal flow. The vocal-tract contribution is often modeled as an all-pole filter and estimated with linear prediction (LP)[2] from speech signals. Although LP analysis is a widely used technique, many inefficiencies have been reported, like non-linear source-tract interaction, or dependency on the degree of linear prediction.

This study targets spectral decomposition of speech signals into glottal flow and vocal tract components without a specific model. Our spectral decomposition method is based on a new representation called Zeros of Z-Transform (ZZT), which is an all-zero representation of the z-transform of the signal unlike all-pole modeling. ZZT is rather a representation (not a model). It is simply a representation of the z-transform polynomial through its roots. Therefore, for obtaining the ZZT representation, one only needs to find the roots of a high degree polynomial with some numerical method.

Our study of ZZT representation for speech signals showed that separate patterns for glottal flow and vocal tract contributions can be observed. We will show in this paper that, for a glottal closing instant (GCI) synchronously windowed speech frame, all the zeros outside the unit circle are due to the first phase of the glottal flow and the zeros inside the unit circle are due to the vocal tract filter (plus the spectral tilt component). Our decomposition algorithm simply classifies zeros according to their distance from the origin in the z-plane. The result is: separate ZZT representations for glottal flow and vocal tract filter. From these ZZT representations, discrete Fourier transform (DFT) can also be calculated. It should be noted that the decomposition results in a vocal tract filter response including the spectral tilt component and a glottal flow without a return phase.

In the following sections, we first study ZZT patterns of synthetic speech obtained by the source-filter model and then present our decomposition algorithm. Then we provide the results for a real speech example. Due to space limitations complete testing on synthetic speech (which provides convincing outputs) cannot be demonstrated in this paper. In addition, effective formant estimation and glottal formant parameter estimation methods have been implemented based on ZZT-decomposition. The results of these studies are submitted to the same conference (ICSLP 2004) in two other papers[3,4].

2. ZZT Representation of Speech Signals

2.1. Definition

For a discrete time signal *x[n]*, the Zeros of Z-Transform (ZZT) representation is defined as the set of roots (zeros), *Zm*, of the corresponding z-transform polynomial *X(z)* (where *N* is the length of the time series) :

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

ZZT representation can be presented on the z-plane in cartesian or polar coordinates, as shown in Fig.1.
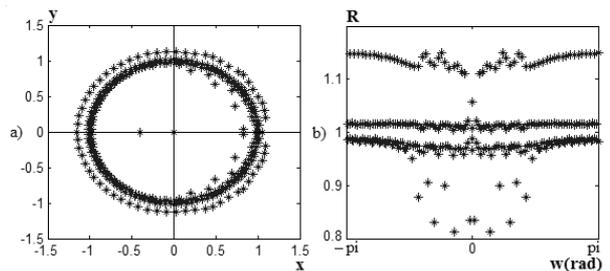

Figure 1: ZZT representation of a signal, a) cartesian coordinates, b) polar coordinates.

Thanks to recent developments in computing technology, the roots of a high degree polynomial can easily be obtained with enough precision to carry spectral analysis in acceptable amount of time. In all zero calculations in this study, the Matlab function *roots*, which finds the eigen values of the associated companion matrix, is used.

2.2. ZZT representation and source-filter model of speech

2.2.1. ZZT of glottal flow

According to the well-known source-filter model for speech, voiced speech signals are produced by exciting the vocal tract system by periodic glottal flow signals. The most widely accepted model for the derivative of the glottal flow signal is the LF model [5], where the signal is supposed to be composed of two non-overlapping parts: an increasing exponential multiplied by a sinusoid (Eq. 2) and a decreasing exponential function (Eq. 3) (both functions are truncated to obtain a one pitch period size data).

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t), 0 \le t \le t_e \qquad (2)$$

$$g(t) = -\frac{E_e}{\varepsilon t_a} \left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)} \right], t_e \le t \le t_c \le T_0 \qquad (3)$$

In Fig. 2, ZZT representations for an LF model (derivative of glottal flow) signal with and without the return phase are presented.
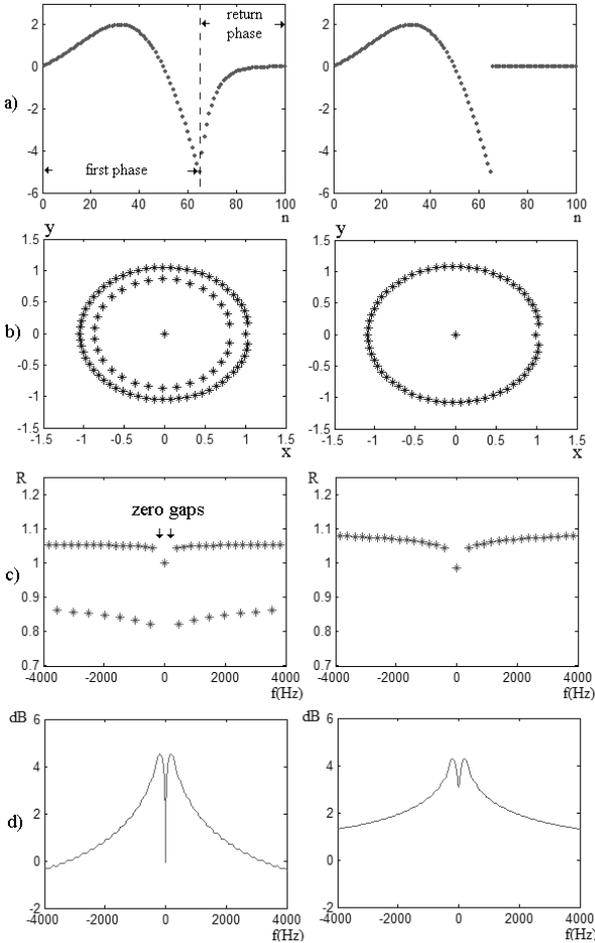

Figure 2: Typical differential LF signal with (left) and without (right) return phase, a) waveform, b) ZZT representation in cartesian coordinates, c) ZZT representation in polar coordinates, d) amplitude spectrum.

A study of the location of zeros for exponential functions is useful for studying ZZT plots of the LF signal. Analytically, for a simple exponential function, all the roots, Z_m (Eq. 6), of the z-transform polynomial $X(z)$ (Eq. 5) calculated for the signal $x(n)$ (Eq. 4) are equally spaced on a single circle at radius $R=a$ (and the zero on the real axis is cancelled by the pole at the same location).

$$x(n) = a^n, n = 0,1...N-1 \qquad (4)$$

$$X(z) = \sum_{n=0}^{N-1} a^n z^{-n} = \frac{1 - (\frac{a}{z})^N}{1 - (\frac{a}{z})} \qquad (5)$$

$$Z_m = a e^{j2\pi m/N}, m = 1,2...N-1 \qquad (6)$$

For an increasing exponential, a>1, the zeros are outside the unit circle and for a decreasing exponential, a<1, the zeros are inside the unit circle.

The ZZT representation of the LF signal, shown in Fig. 2 contains two groups of zeros: a circle inside the unit circle and a circle outside the unit circle in cartesian coordinates (Fig. 2b) or a line below $R=1$ and a line above $R=1$ in polar coordinates (Fig. 2c). The group of zeros inside the unit circle is due to the return phase and the group outside the unit circle is due to the first phase of the LF signal.

On the amplitude spectrum of the glottal flow signal (Fig. 2d), the gap due to the sinusoidal component is observed as a peak in amplitude spectrum as discussed in [6] (around 200Hz for this signal). The zero gaps (Fig. 2c) located outside the unit circle on the wing-like ZZT pattern creates an anti-causal resonance-like spectral peak: a weak spectral peak is observed in the amplitude spectrum and a negative peak is observed in the group delay function (Fig 3c), like the effect of an anti-causal pole at the frequency of the gap. This resonance-like peak on the spectrum carries all information about the first phase of the LF signal (expressed in Eq. 2) and is very important for glottal flow estimation from speech signals.

The return phase exponential component of the differential LF function contributes to the ZZT representation by a group of zeros inside the unit circle, aligned in parallel to the unit circle and the distance of these lined zeros to the unit circle is proportional to the exponential decay coefficient. Again, there exists a gap on the real axis. Its effect on amplitude spectrum is a slope (spectral tilt) change for the high frequency part of the amplitude spectrum (and can be observed on Fig. 2d: the amplitude spectrum of LF signal with a return phase has higher spectral tilt due to the zeros inside the unit circle).

2.2.2. ZZT of windowed synthetic speech signal

Once the glottal flow signal is passed through the vocal tract filter, synthetic speech is obtained. The main change introduced by an all-pole causal filter to the ZZT of a glottal flow excitation signal is observed on the ZZT pattern inside the unit circle. This is due to the fact that the vocal tract filter response is a summation of damped sinusoids (a damped sinusoid for each pole-pair) and is a decaying exponential-like function, therefore its zeros are located inside the unit circle. Time-domain convolution of glottal flow excitation signal with vocal tract filter response results in combining the set of zeros due to the glottal flow with the set of zeros due to the

vocal tract filter response since this operation corresponds to multiplication of z-transforms of the two signals.

Windowing has an important influence on ZZT. Though a complete study on windowing effects to ZZT patterns would be very informative, we leave it to our further studies due to space limitations and present here an example which provides sufficient information for ZZT of windowed speech. In Fig. 3, we present a synthetic speech frame windowed synchronously with GCI. The synthetic speech frame is synthesized by filtering an LF pulse with an all-pole filter response (with resonances at 600Hz, 1200Hz, 2200Hz and 3200Hz).
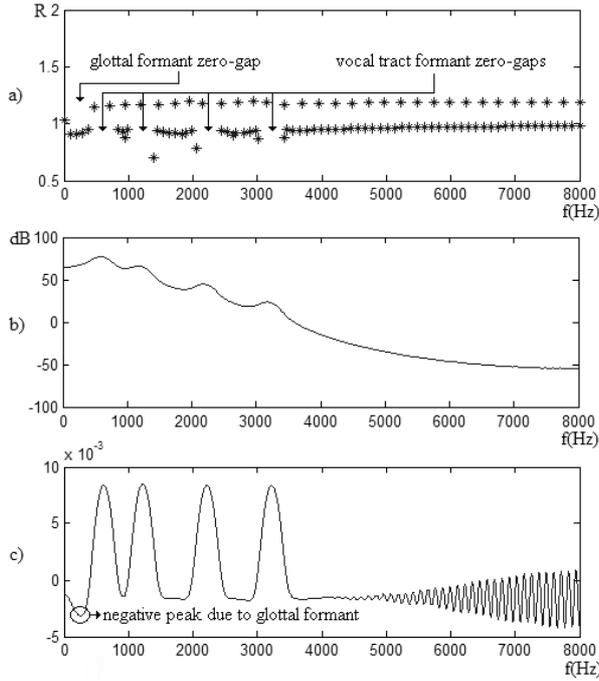

Figure 3: Spectral representations of GCI synchronously windowed synthetic speech frame, a) ZZT representation, b) amplitude spectrum, c) group delay function.

The ZZT representation includes two lines of zeros: one outside the unit circle and one inside the unit circle with gaps creating formant peaks on the spectrum. The reason for this alignment is as follows: once the window is placed such that the increasing exponential part of a single speech frame (due to the first phase (Eq. 2) of glottal flow signal) is multiplied with the first half of the window, which is also increasing, and the decreasing exponential part (due to vocal tract filter response and the return phase of glottal flow) is multiplied with the second half of the window, which is also decreasing, the ZZT of the resulting windowed speech has a pattern close to that of the glottal flow (with additional patterns inside the unit circle due to the vocal tract filter). Zeros of the glottal flow return phase are combined with those of the vocal tract resulting in a single line of zeros. When the window is not centered on the increasing-decreasing function change point, the ZZT-pattern is destroyed, and zeros do not group on the two sides of the unit circle. Therefore, GCI-synchronous windowing is necessary to obtain separate ZZT patterns for glottal flow and vocal tract contributions which provides the opportunity to perform decomposition.

It is also interesting to note at this point that a negative peak due to the glottal formant is observed on the group delay function, at the frequency of the zero gap outside the unit circle. Since the relative distance of the glottal flow zero gap to the unit circle is rather higher than those of vocal tract zero gaps, we cannot observe a peak on the amplitude spectrum. This is one of the reasons that spectral estimation of glottal flow (and even visual observation of it) is not easy. In [3], we present an algorithm for estimating the location of this peak in the context of glottal flow parameter estimation.

3. ZZT-Decomposition For Source-Tract Separation

In Fig. 4, we present our ZZT-decomposition algorithm for source-tract separation based on the characteristics of ZZT of GCI synchronously windowed data.
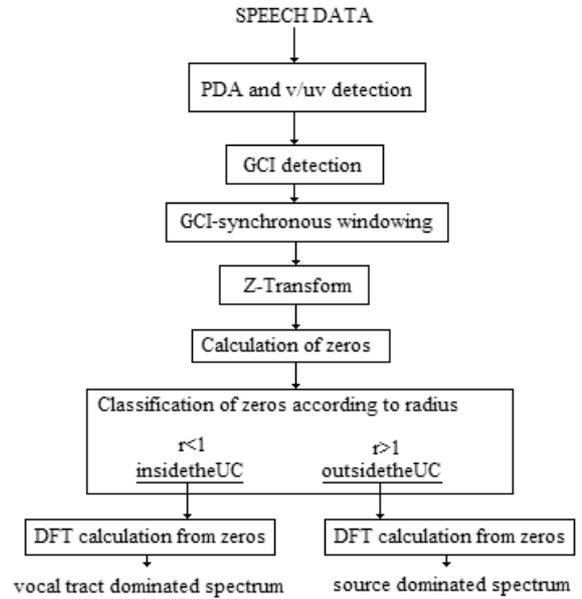

Figure 4: The ZZT decomposition algorithm (PDA stands for pitch detection algorithm, DFT stands for discrete Fourier transform and UC stands for the unit circle).

The decomposition starts with a pitch detection algorithm and voiced/unvoiced decision (decomposition can be performed only for voiced frames). Given first estimate of the pitch mark locations, GCI detection is performed with the technique defined in [7], based on processing of evolution of center of gravity of sliding window analysis. Windowing is very important for effectiveness of the ZZT-decomposition. A Blackman window with a size of two pitch periods and centered at GCI is observed to be a good choice. Zeros are grouped into two by their distance to the origin in the z-plane. Then calculation of DFT for each group is straightforward using the Eq. 7 (N is the number of zeros, G is the gain factor and Zm are the zeros).

$$X(e^{j\varphi}) = Ge^{(j\varphi)(-N+1)}\prod_{m=1}^{N-1}(e^{j\varphi} - Z_m) \qquad (7)$$

The effectiveness of the decomposition method is tested first with synthetic speech and have been confirmed to be of high

quality. Due to space limitations, here we will only present a real speech example. For further tests, the reader is referred to [3] which contains testing of decomposition in a parameter estimation scheme.
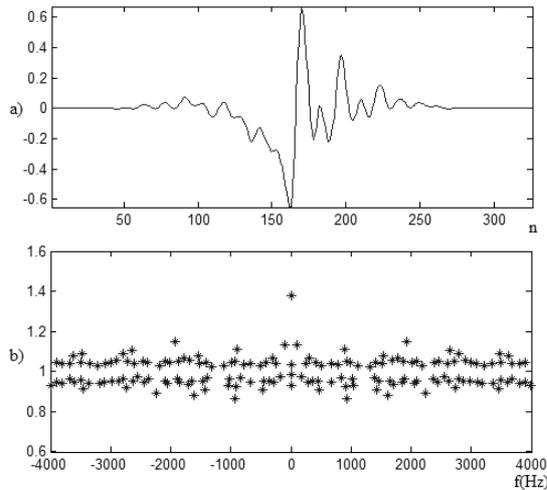

Figure 5: Real speech frame windowed synchronously with GCI, a)waveform, b)ZZT representation.

In Fig. 5, ZZT representation of a real speech frame windowed synchronously with GCI (taken from the Voqual03 database (vowel "*a*", from the word "*party*") and downsampled to 16Khz) is presented. The zero locations are such that a zero gap exists around the unit circle and the zeros are aligned inside and outside the unit circle (as expected).

In Fig. 6, the ZZT-decomposition result for this frame is presented. The decomposition results in separating the first peak as the glottal formant (Fg) peak and the rest of the formant peaks are included in the vocal tract contribution part. This fulfils our expectation for the decomposition of this signal since theoretical values of the formant frequencies for vowel "a" are in agreement with the formant peaks observed in Fig. 6c. An example of vowel "a" with obvious glottal formant peak is presented since it is rather an easy type of signal for visual inspection of formant locations. For sounds with low F1 (first formant) frequency, mid-low open quotient and high pitch, Fg and F1 peak share the same frequency region making visual inspections very difficult. Our decomposition for such examples gives Fg and F1 to be very close but due to lack of reference it is difficult to check reliability of the estimate.

4.   Discussion and Conclusions

In this paper, we have presented a spectral source-tract separation method called the ZZT-decomposition. It is an easy to implement but computationally heavy algorithm due to the need of finding roots of high degree polynomials. Our tests (which could not be completely presented here due to space limitations) showed that ZZT-decomposition successfully separates the glottal formant peak from vocal tract formant peaks in spectral domain. Two companion papers are also submitted to this conference, which test the decomposition algorithm in parameter estimation schemes and show the efficiency of the method.
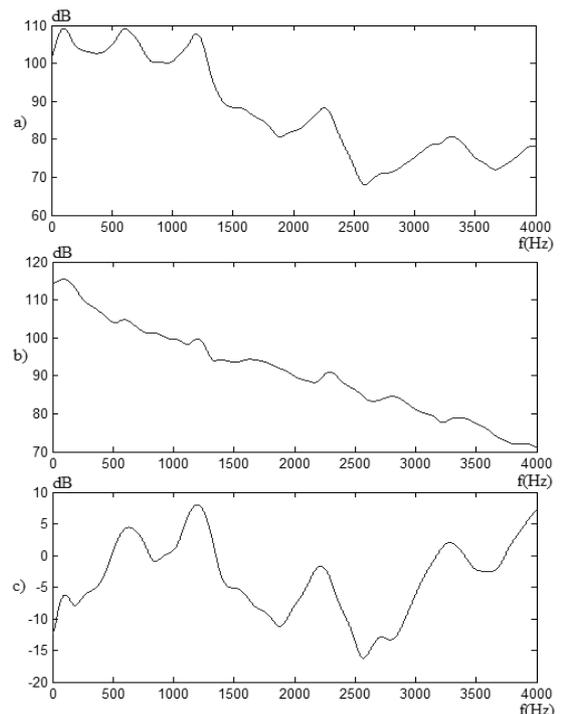

Figure 6: ZZT-decomposition result for a real speech frame ("*a*" from "*party*"), a)amplitude spectrum of the real speech frame, b) Glottal flow dominated amplitude spectrum, c)Vocal tract dominated amplitude spectrum

5.   Acknowledgements

6.   References

[1]   Alku, P., "Glottal Wave Analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109-117, 1992.

[2]   Makhoul, J., "Linear prediction: A tutorial review", *Proc. IEEE*, 63:561-580, 1975.

[3]   Bozkurt, B., Doval, B., d'Alessandro, C. and T. Dutoit, "A method for glottal formant frequency estimation", Submitted to ICSLP 2004, Jeju Island, Korea.

[4]   Bozkurt, B., Doval, B., d'Alessandro, C. and T. Dutoit, "Improved differential phase spectrum processing for formant tracking", Submitted to ICSLP 2004, Jeju Island, Korea.

[5]   Fant, G., "The LF-model revisited. Transformation and frequency domain analysis", *Speech Trans. Lab.Q.Rep., Royal Inst. of Tech. Stockholm*, vol.2-3, pp 121-156,1995.

[6]   Doval, B., d'Alessandro, C. and N. Henrich, "The voice source as a causal/anticausal linear filter,", *Proc. ISCA ITRW VOQUAL 2003, Geneva, Switzerland*, Aug. 2003, pp. 15–19.

[7]   Kawahara, H., Atake, Y., and Zolfaghari, P., "Accurate vocal event detection method based on a fixed-point to weighted average group delay", *Proc. ICSLP*, pp. 664–667, Beijing, 2000.