Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech

Baris Bozkurt, Boris Doval, Christophe D'Alessandro, *Member, IEEE*, and Thierry Dutoit, *Member, IEEE*

***Abstract*—We propose 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. A decomposition method for source-tract separation is presented based on ZZT. The ZZT-decomposition consists in 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 contribution in the z domain.**

***Index Terms*—Glottal flow estimation, source-filter separation, spectral representation, zeros of z-transform (ZZT).**

I. Introduction

THIS LETTER introduces a new spectral representation for a signal: the zeros of z-transform (ZZT) representation. It is defined as the set of zeros of the z-transform polynomial of a discrete time signal. A systematic study of the ZZT of voiced speech indicates that some patterns exist for the zeros. Here, we show that separate patterns for the glottal flow and vocal tract contributions appear on a ZZT representation of speech signals. Then, one can design a spectral source-tract separation algorithm based on ZZT-decomposition without any spectral modeling.

The sections are organized as follows. In Section II, we define the ZZT representation and discuss the patterns for ZZT representation of the source-filter model of speech and windowed short-time speech frames. Section III introduces an application of the ZZT representation, a source-tract separation algorithm named the ZZT-decomposition, and Section IV is dedicated to conclusions and future work.

II. ZZT Representation of Speech Signals

A. Definition of ZZT and the ZZT of Glottal Flow

For a discrete time signal $x(n)$ with length N, the ZZT representation is defined as the set of roots (zeros) Z_m of the corresponding z-transform polynomial $X(z)$

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

Manuscript received June 22, 2004; revised November 20, 2004. This work was supported by Region Wallonne, Belgium under Grant First Europe 215095. Parts of this letter were presented in [5]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

B. Bozkurt and T. Dutoit are with the Faculté Polytechnique de Mons, TCTS Laboratory, B-7000 Mons, Belgium (e-mail: bozkurt@tcts.fpms.ac.be; thierry.dutoit@fpms.ac.be).

B. Doval and C. D'Alessandro are with LIMSI–CNRS, 91403 Orsay, France (e-mail: boris.doval@limsi.fr; Christophe.d'Alessandro@limsi.fr; cda@limsi.fr).

Digital Object Identifier 10.1109/LSP.2005.843770

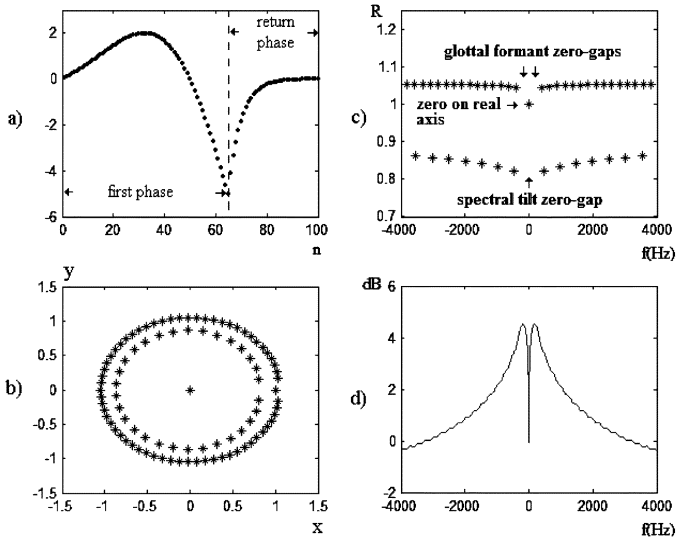

Fig. 1. Typical differential LF signal (a) Waveform. (b) ZZT representation in Cartesian coordinates. (c) ZZT in polar coordinates. (d) Amplitude spectrum.

provided that $x(0)$ is nonzero.

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 Liljencrants–Fant (LF) model [1], where the signal is supposed to be composed of two nonoverlapping parts: an increasing exponential multiplied by a sinusoid (2) and a decreasing exponential function (3) (both functions are truncated to obtain a one-pitch-period-size data).

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

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

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 (6) of the z-transform polynomial $X(z)$ (5) calculated for the signal $x(n)$ (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). 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

$$x(n) = a^n, \quad n = 0, 1, \ldots, N-1 \quad (4)$$

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

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

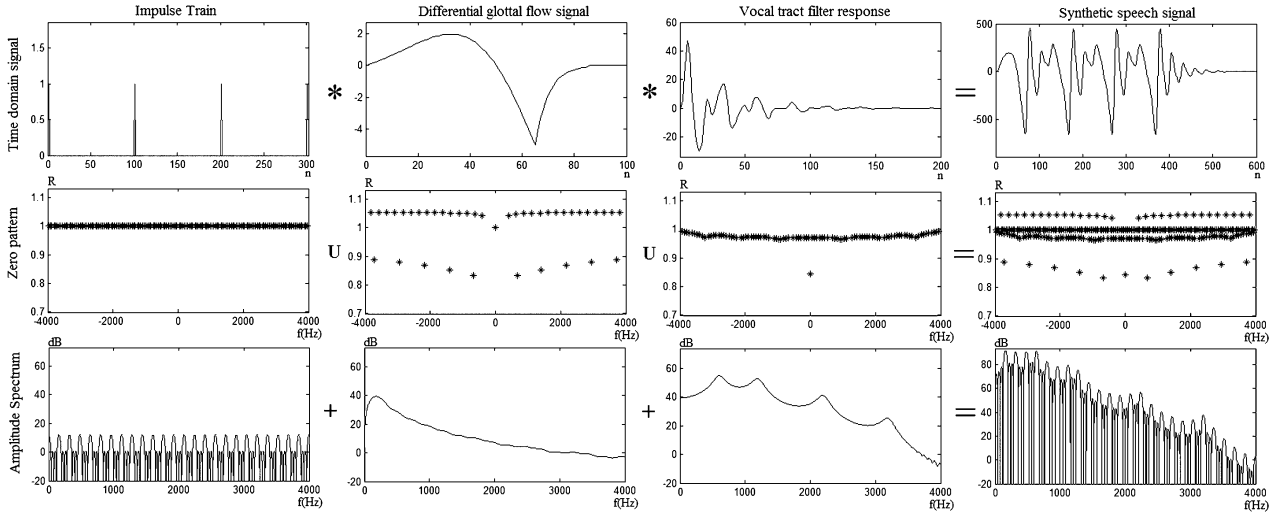

Fig. 2. ZZT patterns for source-filter model of voiced speech.

The ZZT representation of the LF signal, shown in Fig. 1, contains two groups of zeros: a circle inside the unit circle and a circle outside the unit circle in Cartesian coordinates [Fig. 1(b)] or a line below $R = 1$ and a line above $R = 1$ in polar coordinates [Fig. 1(c)]. 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.

The sinusoidal component of the first phase is responsible for the zero gaps located outside the unit circle on the wing-like ZZT pattern [glottal formant zero-gaps on Fig. 1(c)]. These gaps create, in turn, an anticausal resonance-like spectral peak that can be observed on the amplitude spectrum [see Fig. 1(d), at around 200 Hz for this signal], as discussed in [2], and on the group delay [see Fig. 3(b)] as a negative peak. This is like the effect of an anticausal 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 (2)] and is called the glottal formant (Fg) in [2].

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 [spectral tilt zero-gap on Fig. 1(c)]. Its effect on amplitude spectrum is a slope (spectral tilt) change for the high-frequency part of the amplitude spectrum.

B. ZZT Representation and Source-Filter Model of Speech

In Fig. 2, we present the ZZT patterns for the source filter model of speech for voiced speech. Each row presents the model in one domain: time domain in the first row, ZZT representation (z-plane) in the second row, and log-amplitude spectrum in the third row. The operators are convolution $(*)$, union (U), and addition $(+)$. For simplicity, the lip radiation component is included in the source signal as a derivation, resulting in a differential glottal flow signal on the second column. We now discuss in detail the second row: the ZZT-representations.

The ZZT pattern for an impulse train is such that zeros are equally spaced on the unit circle, with the exception that there exist gaps at all harmonics of the fundamental frequency that create the harmonic peaks on the amplitude spectrum. The location of zeros for a truncated impulse train (7) with period P can be analytically found by finding the roots of its z-transform (8). The roots of the denominator in (8) are expressed in (9), and the roots of the numerator are expressed in (10). P roots of the denominator cancel P roots of the numerator, resulting in $P(M-1)$ zeros for the impulse train z-transform, located on the same circle, and P zero gaps exist on the zero-circle at multiples of the fundamental frequency 2π m/P

$$x(n) = \sum_{k=0}^{M-1} \delta(n - kP) \quad n = 0, 1, \ldots, N-1, \quad N \geq P(M-1) \tag{7}$$

$$X(z) = \sum_{n=0}^{N-1} \sum_{k=0}^{M-1} \delta(n - kP) z^{-n} = \sum_{k=0}^{M-1} z^{-kP} = \frac{1 - \left(\frac{1}{z}\right)^{\mathrm{PM}}}{1 - \left(\frac{1}{z}\right)^{P}} \tag{8}$$

$$Z_d = e^{j2\pi m/P}, \quad m = 1, 2, \ldots, P-1 \tag{9}$$

$$Z_n = e^{j2\pi m/PM}, \quad m = 1, 2, \ldots, PM-1. \tag{10}$$

The zeros of the vocal tract filter response are mainly inside the unit circle, due to the decreasing exponential character of this signal, and there exist gaps for the formant locations that create formant spectral peaks. Again, we observe the wing-like character for ZZT patterns of the vocal tract response, depending on the location of the truncation point for the time-domain response.

C. ZZT of Windowed Synthetic Speech Signal

As for many other spectral analysis methods, the use of a weighting window allows for a better visualization of spectral features. A complete study on windowing effects to ZZT representation would be very informative but is out of the scope of this letter. We present in Fig. 3 an example of a synthetic speech frame windowed synchronously with the glottal closure instant (GCI) by a two-pitch-period Blackman window. The synthetic

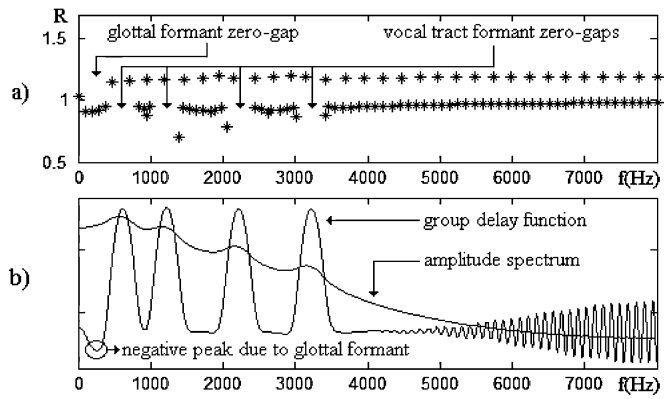

Fig. 3. Spectral representations of GCI synchronously windowed synthetic speech fram. (a) ZZT representation. (b) Amplitude spectrum and group delay function (scaled to be plotted together).

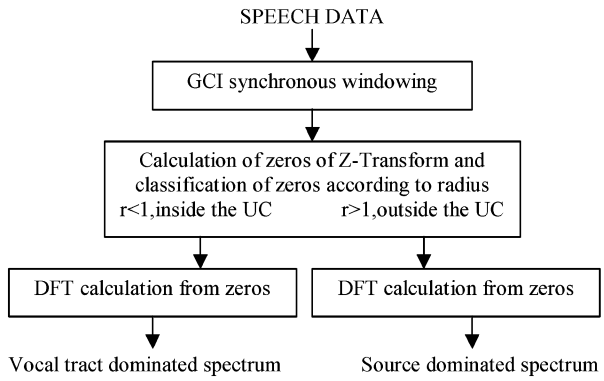

Fig. 4. ZZT decomposition algorithm (DFT stands for discrete Fourier transform, and UC stands for the unit circle).

speech frame is synthesized by filtering an LF pulse with an all-pole filter response.

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 (2) of the glottal flow signal] is multiplied with the first half of the window, which is also increasing, and the decreasing exponential part (due to the vocal tract filter response and the return phase of the 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 turning 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 first phase and vocal tract (plus the return phase) contributions, which provides the opportunity to perform decomposition. In addition, 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 [see Fig. 3(b)]. Since the relative distance of the glottal flow zero gap to the unit circle is much higher than those of vocal tract zero gaps, we cannot observe a peak on the amplitude spectrum.

III. 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. GCI detection is performed with the technique defined in [3]. A Blackman window with a size of two pitch periods and centered at GCI is observed to be a good choice for the GCI synchronous windowing operation. Zeros are separated into two subsets based on their radius. Computing DFTs for each group is straightforward using (11) as follows:

$$X(e^{j\varphi}) = Ge^{(j\varphi)(-N+1)} \prod_{m=1}^{N-1} (e^{j\varphi} - Z_m). \qquad (11)$$

The most important detail for the ZZT-decomposition algorithm is a single zero on the real axis due to the anticausal portion of the signal, which, in some cases, falls inside the unit circle. This zero can be observed in Fig. 1(c) and the second row second column of Fig. 2. No governing rule has been found for the classification of this zero (inside or outside?), and a heuristic approach is used for this problem; if no zero has been found on the real axis in the range $\mathrm{R} = [1\ 1.1]$, then the closest zero on the real axis to point $(\mathrm{R} = 1, \varphi = 0)$ is included in the set of zeros outside the unit circle.

The effectiveness of the decomposition method was first tested with synthetic speech. We have observed that the original and estimated spectra for source and vocal tract contributions are very close, and ZZT-decomposition is indeed capable of separating source and vocal tract contributions to a high extent (not completely, though, mainly because the vocal tract response of the previous pitch cycle is also partially present in the actual period analyzed; small variations due to vocal tract formants are observable both on the time domain glottal flow signal and its amplitude spectrum), and parameter estimation can be performed effectively on the resulting signals. Here, we present a real speech decomposition example in Fig. 5. The speech frame is taken from vowel /a/, from the word "party." The actual amplitude spectrum of the windowed speech frame is presented in Fig. 5(a), and the amplitude spectra of glottal flow and vocal tract contributions obtained by ZZT-decomposition are presented in Fig. 5(b).

The ZZT-decomposition procedure results in separating the first peak as the glottal formant 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. 5(b). An example of vowel /a/ is presented since it is a rather 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.

The ZZT-decomposition algorithm is tested in a parameter estimation scheme for real speech signals due to the difficulty of obtaining reliable glottal flow reference signals for comparison through actual signals. A sustained vowel /a/ with flat pitch

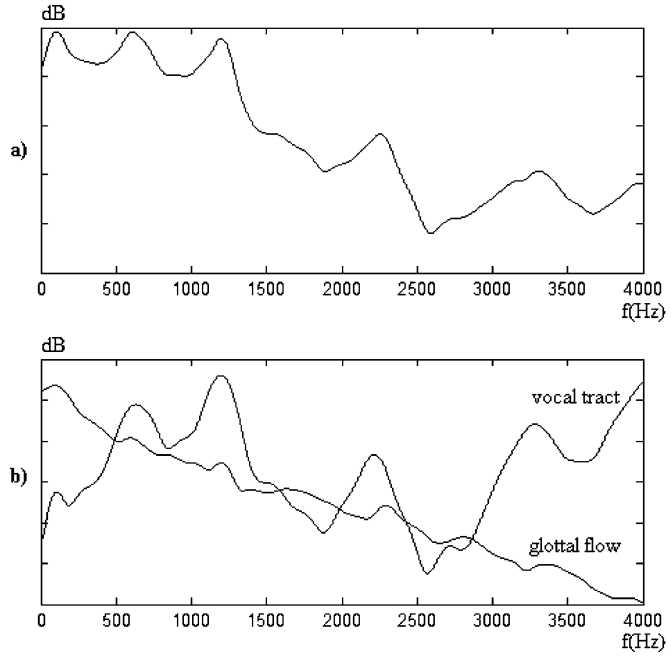

Fig. 5. ZZT-decomposition result for a real speech frame /a/ from "party." (a) Amplitude spectrum of the real speech frame. (b) Glottal flow dominated and vocal tract dominated amplitude spectra.

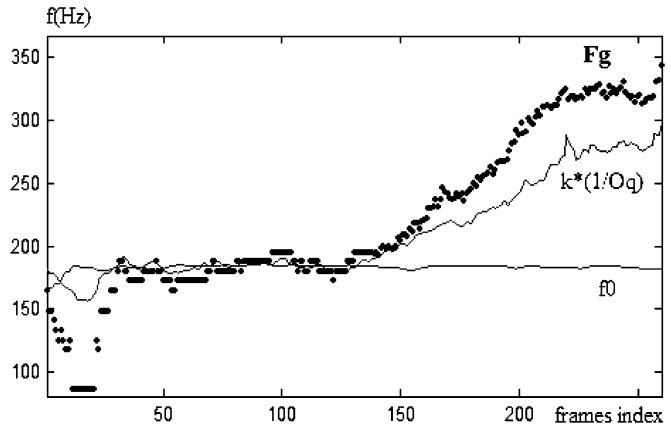

Fig. 6. Comparing glottal formant frequency estimate with inverse of open quotient estimate and f0 estimate ($\mathrm{k} = 115$).

and decreasing open quotient has been uttered by one of the authors, and EGG signals were recorded in parallel to obtain (Oq) estimate (by using the algorithm in [6]) as reference. Glottal formant frequency tracking is performed by picking the low-frequency peak on the glottal flow dominated spectra obtained by ZZT-decomposition. As discussed in [2], Fg location is mainly defined by the pitch period and open quotient, and the effect of asymmetry coefficient variation to Fg variation is rather minor. For this reason, for our speech example with almost constant pitch, we expect the Fg estimate to be highly correlated with the inverse of the open quotient estimate. In Fig. 6, we present our Fg estimate together with the f0 curve and inverse of the open quotient estimate scaled with a constant. This example shows that the ZZT-decomposition has the potential to be used in studying some phonation variations in real speech signals.

IV. Conclusions and Future Work

In this letter, we have presented a new domain of study: an all-zero representation of speech named ZZT representation and its application to source-filter separation of speech signals. Our ZZT-decomposition method provides two spectra: a glottal-flow-dominated spectrum and a vocal-tract-dominated spectrum. The decomposition is of high quality though not complete: Return phase of glottal flow is included in the vocal-tract-dominated spectrum. The contribution of the vocal tract in the glottal-flow-dominated spectrum is observed as ripples of low amplitude, while the contribution of glottal flow in the vocal-tract-dominated spectrum is hardly observed. The ZZT-decomposition was successfully applied in a parameter estimation scheme and shown to be effective [4].

Some of the open problems related to the ZZT representation are as follows: How sensitive is the method to GCI detection errors? How does noise and the aperiodic components of speech contribute to the ZZT representation? What are the criteria for finding an optimal windowing function? Our further work will also include studies on these points.

Acknowledgment

The authors would like to thank N. Henrich for providing the recorded speech file and the corresponding open quotient estimate from EGG for the glottal formant frequency estimation test.

References

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

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

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

[4] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "A method for glottal formant frequency estimation," in *Proc. ICSLP*, Jeju Island, Korea, 2004.

[5] ——, "Zeros of Z-transform (ZZT) decomposition of speech for source-tract separation," in *Proc. ICSLP*, Jeju Island, Korea, 2004.

[6] N. Henrich, B. Doval, C. d'Alessandro, and M. Castellengo, "Open quotient measurements on EGG, speech and singing signals," in *Proc. 4th Int. Workshop Adv. Quantit. Laryngoscopy, Voice, Speech Res.*, Jena, Germany, Apr. 2000.