Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 49 (2007) 159–176

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Chirp group delay analysis of speech signals ☆

Baris Bozkurt *, Laurent Couvreur, Thierry Dutoit

TCTS Lab., Faculté Polytechnique De Mons, Initialis Scientific Parc, B-7000 Mons, Belgium

Received 13 December 2005; received in revised form 19 December 2006; accepted 20 December 2006

Abstract

This study proposes new group delay estimation techniques that can be used for analyzing resonance patterns of short-term discrete-time signals and more specifically speech signals. Phase processing or equivalently group delay processing of speech signals are known to be difficult due to large spikes in the phase/group delay functions that mask the formant structure. In this study, we first analyze in detail the z-transform zero patterns of short-term speech signals in the z-plane and discuss the sources of spikes on group delay functions, namely the zeros closely located to the unit circle. We show that windowing largely influences these patterns, therefore short-term phase processing. Through a systematic study, we then show that reliable phase/group delay estimation for speech signals can be achieved by appropriate windowing and group delay functions can reveal formant information as well as some of the characteristics of the glottal flow component in speech signals. However, such phase estimation is highly sensitive to noise and robust extraction of group delay based parameters remains difficult in real acoustic conditions even with appropriate windowing. As an alternative, we propose processing of chirp group delay functions, i.e. group delay functions computed on a circle other than the unit circle in z-plane, which can be guaranteed to be spike-free. We finally present one application in feature extraction for automatic speech recognition (ASR). We show that chirp group delay representations are potentially useful for improving ASR performance.

Keywords: Group delay processing; Phase processing; Windowing; Spectral analysis; Automatic speech recognition

1. Introduction

1.1. Motivations

Most of speech processing methods use Fourier spectrum, i.e. some processing of the Fourier transform of speech signals. The magnitude spectrum is classically the preferred part of Fourier spectrum although it carries only part of the available information. Due to the difficulties involved in phase processing, the phase spectrum is commonly ignored.

Early investigations on the perceptual relevance of phase information came up with the conclusion that the human ear is phase-deaf (von Helmholtz, 1912). After a long period of time, the issue was restudied and falsifying results were obtained (Schroeder, 1959; Schroeder and Strube, 1986; Patterson, 1987). Recently, although in a very limited number, more and more researches have studied the perceptual relevance of phase information and show evidences about the importance of phase information in speech perception. Through human perception experiments, Liu et al. (1997) and Paliwal and Alsteris (2003) showed that the short-time phase spectrum contributes to speech intelligibility as much as the corresponding power spectrum. Pobloth and Kleijn (1999) showed in a speech coding and psycho-acoustic research that human beings are able to distinguish between different phase spectra much better than often assumed. The studies of Kawahara et al. (2001)

☆ Section 5 of this manuscript was partially published in (Bozkurt and Couvreur, 2005).

* Corresponding author. Present address: Izmir Institute of Technology, Department of Electrical Engineering, Gulbahce Koyu, Urla, Izmir, Turkey. Tel.: +90 232 750 6505; fax: +90 232 750 6274.

E-mail addresses: barisbozkurt@iyte.edu.tr (B. Bozkurt), Laurent.Couvreur@fpms.ac.be (L. Couvreur), Thierry.Dutiot@fpms.ac.be (T. Dutoit).

URL: http://www.tcts.fpms.ac.be

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.12.004

showed that phase information plays an important role in high quality speech synthesis. Banno et al. (2001) proved that the human auditory system is sensitive to the difference between zero and non-zero phase signals. These studies showed that phase information plays an important part in perception. However, due to the difficulties in analyzing the phase content of signals, perception studies can be conducted with only synthetic signals and it is very difficult to utilize these results for designing algorithms for processing real speech data.

Another topic where phase processing is essential is sinusoidal/harmonic coding-modification-synthesis of speech signals. Sinusoidal/harmonic modeling is largely and effectively used in speech coding applications (Marques, 1989; Marques et al., 1990; McAulay and Quatieri, 1991). Reliable estimation and proper coding of harmonic phases is essential for speech coding applications and is reported to be a difficult problem. Often the phase related problems are tried to be avoided by various methods: using the zero-phase or minimum-phase phase spectrum obtained from magnitude spectrum information (Oppenheim, 1969) deriving the mixed-phase signal phase spectrum through complex cepstrum (Quatieri, 1979) compensating phase with some all-pass filtering at the speech reconstruction stage (Hedelin, 1988; Sun, 1997) etc. Most of such methodologies target avoiding the mismatch between what is expected and what is measured, using compensation methods.

Phase processing is also essential when sinusoidal/harmonic modeling is to be used in the context of concatenative synthesis; i.e. synthesis of speech signals by concatenation of pre-recorded speech segments. For such a task, the database of recorded speech segments needs to be transformed into a parametric database by sinusoidal analysis, i.e. each short-time speech frame has to be represented by harmonic amplitude and phase parameters. During speech synthesis, speech segments are re-constructed with modified prosody (only pitch and duration are modified in most of the concatenative synthesis systems) and concatenated in such a way that no audible discontinuity exists at concatenation points. Such operations require effective algorithms to estimate and modify the phase of harmonics (for example, (Stylianou, 1996)).

Phase processing is essential for speech processing; yet its use is not limited to this field. It is also used for many other signal processing fields. For instance, let us mention radar signal processing (Costantini et al., 1999; Chen and Zebker, 2002) medical imaging (Chavez et al., 2002; Frolova and Taxt, 1996) source localization (Andersen and Jensen, 2001; Li and Levinson, 2002) as well as many other research fields like optics, solid state physics, geophysics, holography, etc. (Vyacheslav and Zhu, 2003). Nevertheless, this study is focused in analyzing phase spectrum of speech signals.

By its nature, the phase spectrum is in a wrapped form and the negative first derivative of its unwrapped version, the so-called group delay function is generally preferred since it is easier to study and process. Unfortunately, the group delay function often presents spurious large spikes that make its processing very problematic.

Yegnanarayana and Murthy proposed various methods (Yegnanarayana et al., 1988; Murthy et al., 1989; Murthy & Yegnanarayana, 1991a) to remove these spikes and perform formant tracking from the spike-free group delay function. The common steps involved in their methods are: obtaining the magnitude spectrum for a short-time windowed speech frame, smoothing the magnitude spectrum via cepstral smoothing and computing smooth minimum-phase group delay from this representation through cepstrum. Note that the resulting group delay function is a kind of smoothed magnitude spectrum and the conversion to minimum-phase destroys part of the phase information available in the signal. The advantage of this representation compared to the magnitude spectrum is that the formant peaks appear with better resolution.

Other recent studies address the spike problem and propose group delay based features: the modified group delay function (MODGDF) (Hegde et al., 2004a) and the product spectrum (PS) (Zhu and Paliwal, 2004). These new representations are obtained by modifying one term in the group delay computation, which is supposed to be the main source of spikes in the group delay function. These representations are further discussed in the following sections.

In this paper, we aim at getting more insight about the sources of the spikes in group delay functions of speech signals and study this issue from the perspective of the patterns of the zeros of their z-transform in the z-plane. We show that the existence of zeros close to the unit circle at frequencies where the group delay function is computed (which are known to be responsible for the spikes) is highly dependent on how windowing is performed. Based on these observations, we propose techniques to avoid the zeros close to the frequency points where the group delay function is computed. These techniques consists of using appropriate windowing in terms of window location, width and shape, and computing chirp group delay function, i.e. on circles other than the unit circle. The effectiveness of these representations is demonstrated for feature extraction in automatic speech recognition experiments.

1.2. Plan

In Section 2, we formerly define the group delay function of short-term discrete-time signals. We then discuss the difficulties involved in group delay analysis and present recent approaches to tackle these difficulties, namely the modified group delay function (Hegde et al., 2004a) and the product spectrum (Zhu and Paliwal, 2004). In Section 3, we study the zero patterns of z-transform of short-term speech signals and discuss the sources of spikes on group delay functions with respect to windowing operation. Based on our observations, we propose an appropriate windowing technique to enhance formant information in group delay functions. In Section 4, we present another group delay representation based on chirp z-transform, namely the chirp group delay representation. In Section 5, we apply all the group delay representations as well as the power spectrum for reference in an automatic speech recognition experiment and compare the performances of these representations.

2. Difficulties in group delay analysis and proposed solutions

2.1. Difficulties in group delay analysis

For a short-term discrete-time signal $\{x(n)\}$, $n = 0, 1, \ldots, N-1$, the group delay function is defined as the negative first-order derivative of the Fourier transform phase function $\theta(\omega)$,

$$\mathrm{GDF}(\omega) = \frac{-\mathrm{d}\theta(\omega)}{\mathrm{d}\omega}. \tag{1}$$

Eq. (1) can also be expressed as

$$\mathrm{GDF}(\omega) = \frac{X_{\mathrm{R}}(\omega)Y_{\mathrm{R}}(\omega) + X_{\mathrm{I}}(\omega)Y_{\mathrm{I}}(\omega)}{|X(\omega)|^2}, \tag{2}$$

where $Y(\omega)$ is defined as the Fourier transform of $nx(n)$ (Oppenheim et al., 1999). The notations R and I refer to real and imaginary parts, respectively. The advantage of this representation is that explicit phase unwrapping operation, which is generally a problematic task, is not necessary in order to compute the group delay function.

Group delay analysis is known to be difficult because large spikes may be present (Yegnanarayana et al., 1984). This is simply explained by noting that the term $|X(\omega)|^2$ in Eq. (2) can get very small at frequencies where there exist one or more zeros of the z-transform of the signal $x(n)$ very close to the unit circle (Hegde et al., 2004a).

We can also add a geometric explanation at this point. For a given short-term discrete-time signal $x(n)$, the z-transform polynomial $X(z)$ can be expressed using an all-zero representation as

$$X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} = x(0)z^{-N+1}\prod_{m=1}^{N-1}(z - Z_m), \tag{3}$$

where Z_m are the roots of the z-transform polynomial $X(z)$. The Fourier transform, which is simply the z-transform computed on the unit circle, can be expressed as

$$X(\omega) = x(0)(\mathrm{e}^{\mathrm{j}\omega})^{(-N+1)}\prod_{m=1}^{N-1}(\mathrm{e}^{\mathrm{j}\omega} - Z_m) \tag{4}$$

provided that $x(0)$ is non-zero. Each factor in Eq. (4) corresponds to a vector starting at Z_m and ending at $\mathrm{e}^{\mathrm{j}\omega}$ in the z-plane. In practice, the Fourier transform is evaluated at frequencies regularly spaced along the unit circle, the so-called frequency bins. As illustrated in Fig. 1, the vector $\mathrm{e}^{\mathrm{j}\omega} - Z_m$ changes drastically its orientation when going from a frequency bin to the next one around the root Z_m. The change is amplified as the root gets closer to the unit circle. Hence the group delay function, i.e. the rate of change in the phase spectrum, is very high at frequency bins close to a root/zero of the z-transform polynomial and becomes ill-defined when the zero coincides with a frequency bin. It results in spikes in the group delay function that get larger as the roots come closer to the unit circle.

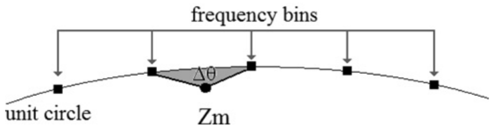

Fig. 1. Geometric interpretation for spikes in the group delay function of a signal at frequency locations on the unit circle close to a zero of its z-transform polynomial.

For speech signals, many zeros of the z-transform appear to be very close to the unit circle. The effect of zeros close to the unit circle can be easily observed both on magnitude spectra and group delay functions. In Fig. 2, we present a windowed real speech frame with its zeros and spectrum plots. The zeros close to the unit circle are shown in between dashed lines in Fig. 2b and they are also superimposed on spectrum plots to draw attention to their relation with the spectral dips in the magnitude spectrum (see Fig. 2c) and the spikes on the group delay function (see Fig. 2d). We observe that the group delay function is merely dictated by the zeros close to the unit circle and appears like a DC function with spikes due to these zeros (see Fig. 2d). This domination of spikes conceals other spectral information and the formant structure is hardly observed in such group delay function. The spikes in the group delay function appear as an important obstacle in speech processing: the often cited and unsolved "phase unwrapping problem" is mainly due to these spikes on the group delay function. For the magnitude spectrum, the effect of zeros close to the unit circle (spikes) is much smaller and we can still observe formants.

2.2. Recently proposed methods for group delay analysis

2.2.1. Modified group delay function (MODGDF)

As exposed previously, spikes in group delay can be explained by the very low values of the $|X(\omega)|^2$ term as defined in Eq. (2). The basic idea behind the modified group delay function (MODGDF) (Hegde et al., 2004a) consists in smoothing the magnitude spectrum $|X(\omega)|$ in order to avoid extremely low values. Such smoothing can be easily performed through cepstral processing. However, this modification alone cannot remove all spikes but some formant peaks can be enhanced. To further reduce spikes, two new parameters have been introduced: α and γ which need to be fine-tuned according to the environment. In all tests/plots of this study, we have set the parameters as in (Hegde et al., 2004b) namely $\alpha = 0.4$ and $\gamma = 0.9$. The modified group delay function is eventually defined as

$$\mathrm{MODGDF}(\omega) = \left(\frac{\tau_p(\omega)}{|\tau_p(\omega)|}\right)(|\tau_p(\omega)|)^{\alpha} \tag{5}$$

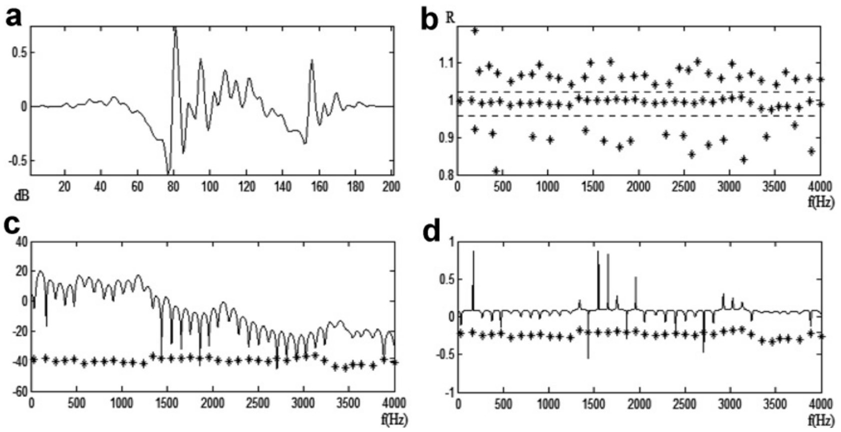

Fig. 2. Effects of zeros to spectra of a signal: (a) Hanning windowed real speech frame (phoneme /a/ in the word "party"), (b) zeros in polar coordinates, (c) magnitude spectrum and (d) group delay function. The zeros close to the unit circle are superimposed on spectrum plots to show the link between the zeros and the dips/spikes on the spectra.

with

$$\tau_p(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}\right), \tag{6}$$

where $S(\omega)$ is the cepstrally smoothed version of $|X(\omega)|$.

2.2.2. Product spectrum (PS)

In (Zhu and Paliwal, 2004) another group delay function is proposed, the so-called product spectrum (PS). It is a version of Eq. (2) where the denominator $|X(\omega)|^2$, which is considered to be a source of spikes, is simply removed. The product spectrum $\mathrm{PS}(\omega)$ is then defined as the product of the power spectrum and the group delay function:

$$\mathrm{PS}(\omega) = |X(\omega)|^2\mathrm{GDF}(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega). \tag{7}$$

2.2.3. Our approach

In this study, we propose alternative paths to tackle the spike problem. We first aim at understanding the sources of zeros close to the unit circle in speech signals by studying the zero patterns of their z-transform and show how appropriate windowing can effectively reduce the spike effect. Besides, we investigate the use of chirp z-transform to guarantee certain distance of zeros from the frequency points where Fourier transform is computed. These ideas are presented in detail in the next two sections.

3. ZZT representation

In this section, we discuss the usefulness of representing short-term speech signals by the zeros of their z-transform (ZZT) for reliable estimation of group delay function. There are mainly two useful points of the ZZT representation for speech signals: (i) it sheds light into many difficulties involved in group delay processing and for this reason provides us with the opportunity to design better algorithms, (ii) patterns exist in the ZZT of speech signals which make it possible to design a new spectral decomposition method for source-tract separation. The latter issue has been already presented in (Bozkurt et al., 2005) and will not be discussed in this paper. We primarily focus on the former issue in the following.

3.1. Definition

In Section 2, we have already stated that the z-transform polynomial $X(z)$ of a short-term discrete-time signal $\{x(n)\}$, $n = 0, 1, \ldots, N-1$, can be expressed using an all-zero representation (see Eq. (3)). The zeros of the z-transform (ZZT) representation is defined as the set of roots (zeros) $\{Z_1, Z_2, \ldots, Z_m\}$ of the z-transform polynomial $X(z)$.

The analytical study of the ZZT representation of speech signals is a complex problem since finding analytically the roots of high order polynomials (higher than 4) is shown to be impossible (Abel-Ruffini theorem, (Abel, 1826)). For this reason, our methodology for studying the ZZT representation of speech signals is more based on observation of the locations of the roots (on the z-plane) found by numerical methods. In this study, roots were estimated as the eigenvalues of the associated companion matrix (Edelman and Murakami, 1995) like it is implemented in the Matlab ROOTS function. Other efficient algorithms for finding roots of a polynomial can be found in (Sitton et al., 2003).

The ZZT representation can be presented on the z-plane in cartesian or polar coordinates. We generally prefer polar coordinates since visual comparison with amplitude or phase spectrum is much easier as shown in Fig. 3. When the ZZT plot is given on polar coordinates (Fig. 3d), we can already observe that the peaks in the spectra correspond to zero-gaps on the ZZT plot.

We should also note here that reconstruction of the original time-domain signal from its ZZT representation is not trivial. One could think, "we just multiply the roots

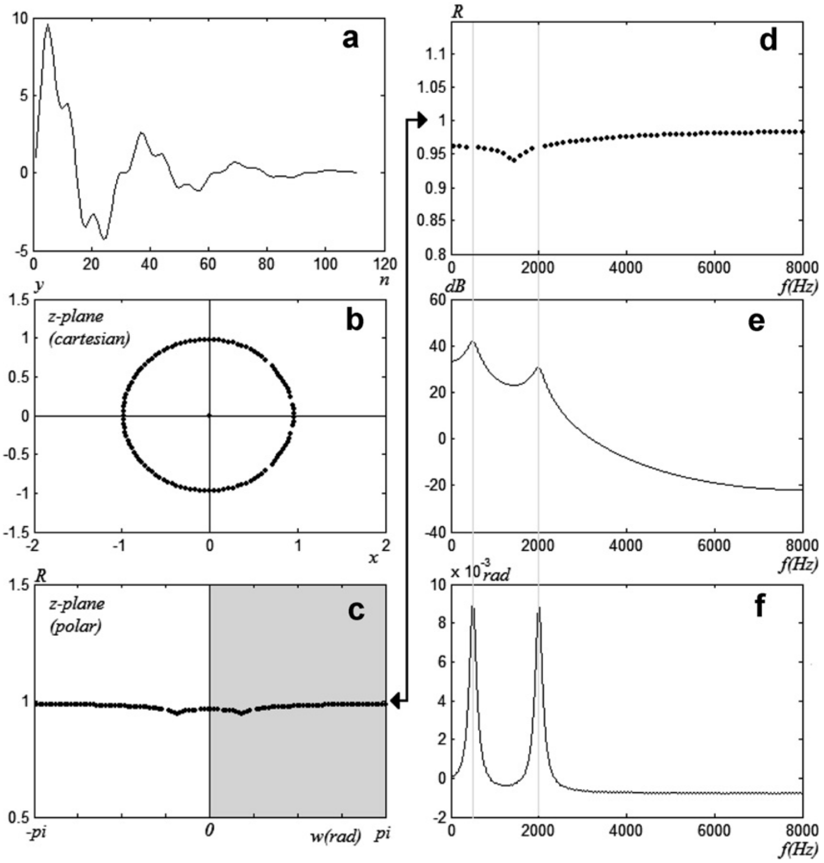

Fig. 3. ZZT plots on the z-plane: (a) the time-domain signal, (b) ZZT plot in cartesian coordinates, (c) ZZT plot in polar coordinates, (d) ZZT plot in polar coordinates (right half-plane), (e) magnitude spectrum and (f) group delay function.

and we have the signal'' but actually such an operation usually results in reconstructed signals being completely different from the original signals due to the fact that we have limited precision zeros/roots and the degree of the polynomial is high (159 for a 10 ms frame at 16,000 Hz).

3.2. ZZT and the source–filter model of speech

In the source–filter model of speech (Fant, 1960) continuous time voiced speech signals can be expressed as $s(t) = d(I(t)^* U_g(t)^* v(t))/\mathrm{d}t$ where $I(t)$ stands for an impulse train, $U_g(t)$ stands for the glottal flow signal, $v(t)$ stands for the vocal tract filter impulse response and the lip radiation component is approximated by a derivation operation. Equivalently, we can re-write this as: $s(t) = I(t)^*(d(U_g(t))/\mathrm{d}t)^* v(t)$ including the derivative operation in the glottal flow part.

In Fig. 4, we present the source–filter model of voiced speech in various domains: the first row is in the time domain, the second row is in the ZZT domain and the third row is in the log-magnitude spectrum domain. In each domain, the different contributions of the source–filter model of speech interact with each other through some operator: convolution (∗), union (U) and addition (+), respectively. This figure has been discussed in detail in (Bozkurt et al., 2005). Due to space considerations we only provide a short review here.

Consider the second row of Fig. 4. The ZZT pattern for the 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, which create the harmonic peaks on the magnitude spectrum (third row of Fig. 4).

The ZZT representation of the differential glottal flow signal (LF model (Fant, 1985)) which is shown in second column of Fig. 4, contains two groups of zeros: a group of zeros below $R = 1$ (inside the unit circle) and a group of zeros above $R = 1$ (outside the unit circle) in polar coordinates. The group of zeros inside the unit circle is due to the return phase of the glottal flow excitation waveform and the group outside the unit circle is due to the first phase of the LF signal.

The zeros of the vocal tract filter response are mainly located inside the unit circle due to the decreasing exponential shape of this signal and there are gaps for the formant locations, which create formant spectral peaks. We observe a wing-like shape for the ZZT pattern of the vocal tract response depending on the location of the truncation point for the time-domain response (demonstrated in movie 2 on our demo website (www-zzt)).

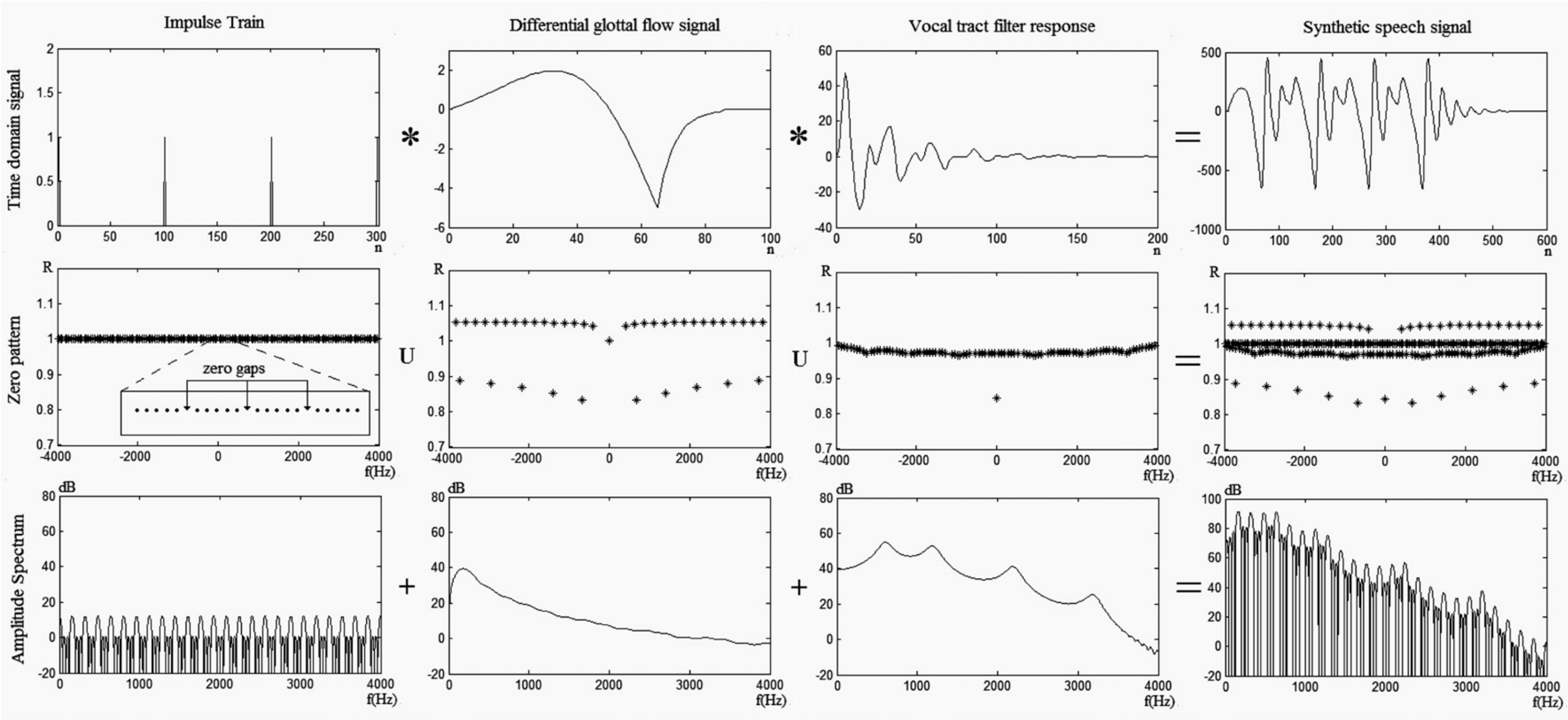

Fig. 4. ZZT and source–filter model of speech. Note that magnitude spectra added are in dBs.

It is interesting to note here that the ZZT set of speech is just the union of ZZT sets of its three components. This is due to the fact that the convolution operation in time-domain corresponds to multiplication of the z-transform polynomials in z-domain. What is interesting is that the ZZT of each component appear at a different area on the z-plane and have effect on the magnitude spectrum relative to their distance to the unit circle. The closest zeros to the unit circle are the impulse train zeros and they cause the spectral dips on the magnitude spectrum, which give rise to harmonic peaks. Vocal tract zeros are the second closest set and the zero-gaps due to formants contribute to the magnitude spectrum with formant peaks on the spectral envelope. Differential glottal flow ZZT are further away from the unit circle and their contribution to the magnitude spectrum is rather vague and distributed along the frequency axis. For more details, the reader is referred to (Bozkurt et al., 2005).

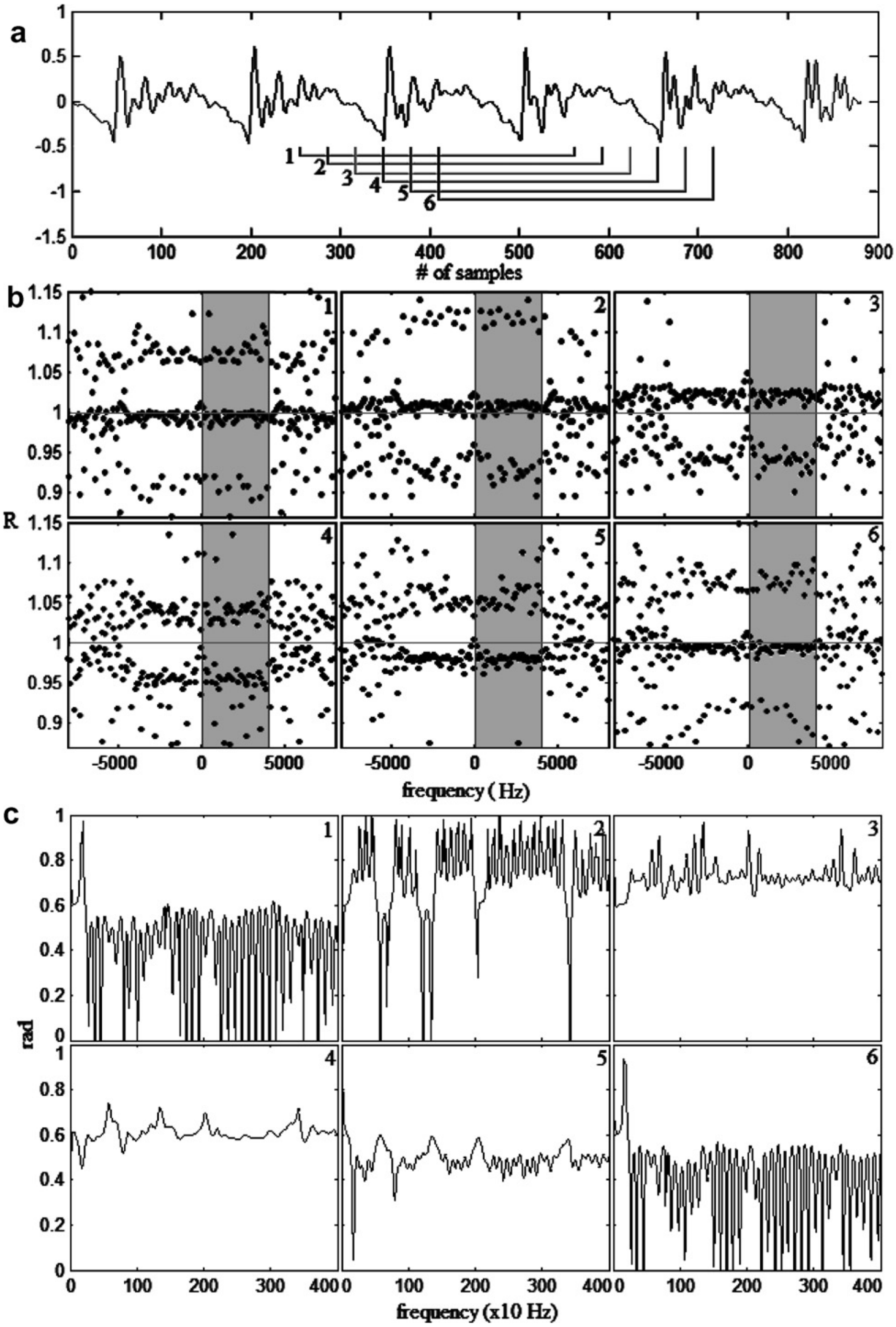

Fig. 5. Effect of window location on ZZT and group delay function of a real speech signal: (a) time-domain signal and window locations, (b) corresponding ZZT representations and (c) corresponding group delay functions. Note that a Blackman window is used.

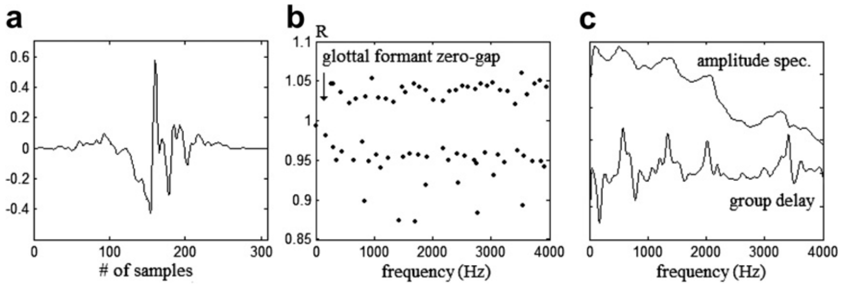

Fig. 6. GCI synchronous windowing: (a) time-domain waveform of the windowed signal, (b) corresponding ZZT representation and (c) amplitude spectrum and group delay scaled plotted together.

3.3. ZZT of windowed speech signals

In real-life applications, the windowing operation is essential especially for the short-term spectral analysis of signals. The effects of windowing on phase spectrum estimation have been addressed in a few recent studies (Zolfaghari et al., 2003; Alsteris and Paliwal, 2004). Especially the study of Alsteris and Paliwal (2004) is a good example of the importance of windowing in phase estimation: the authors show that Liu's extensive study (Liu et al., 1997) on phase contribution to speech intelligibility can provide quite different results when the window shape and window shift is modified in the procedure. Still the studies mentioning the importance of the windowing operation lack background explanation and the preferences are often based on trial-error methodologies. In this section, we study the effects of windowing on the ZZT representation, which happens to be a very appropriate representation to understand the windowing effects on group delay functions.

The size, location and shape of the window play an important role on the resulting ZZT patterns for windowed speech signals. It is very difficult to analytically study the effects of windowing on the ZZT patterns since we face the difficult problem: the windowing operation is a term-wise multiplication operation in the time-domain, and there is no known methodology for estimating roots of the resulting z-transform polynomial directly from the roots of the two polynomials that are subject to term-wise coefficient multiplication. For this reason, our discussions in this section are based only on experimental observations of ZZT patterns on the z-plane. One of the best ways of studying such variations is to create movies and observe changes due to variations in windowing size, location and function.

More especially, we have created movies by shifting windows on signals and observing ZZT and group delay function using: synthetic and real speech signals, window of size T0 (pitch period), 2T0 and 3T0, and window of type Rectangular, Hamming, Hanning, Blackman, Gaussian, Hanning–Poisson. Candidate window shapes are chosen according to their popularity and their spectral characteristics (the reader is referred to (Harris, 1978) for window shape definitions). Some of these movies are available on our demo website (www-zzt).

3.3.1. *Effect of window location on ZZT patterns and group delay*

In Fig. 5, we demonstrate the effect of windowing location on ZZT patterns and the group delay function for a real speech frame (vowel /a/). A Blackman window of two pitch period size (2T0) is slided in six steps within a pitch period and the resulting ZZT are presented. Each window position is indicated on the signal on the top figure with reference numbers. The ZZT representation and the group delay function of the resulting windowed data for each window are presented with the window index indicated on the right-top corner of the figure. The frequency axis range of the ZZT representation is -8 kHz to 8 kHz to present an overall view of zero locations. The group delay functions are presented in the 0–4 kHz range and this range is shaded on the ZZT representations.

The six ZZT representations in Fig. 5 show that the influence of window location on ZZT patterns and the group delay functions is indeed very important. Among the six possibilities presented, the fourth window, centered on the glottal closure instant (GCI[1]) is special: there exists a zero-free region around unit circle in the (0–4000 Hz) frequency region.

In Fig. 6, we present a zoomed plot of this complicated picture and show the windowed signal waveform, ZZT representation and the corresponding magnitude spectrum and group delay function. The group delay contains peaks, which corresponds to the formant peaks observed on the magnitude spectrum. In addition, for the lowest frequency peak in the magnitude spectrum, the group delay peak has a negative direction. This is due to the fact that this peak is not a vocal tract formant peak but is due to the first phase of the glottal flow component, the so-called *glottal formant* (Doval et al., 2003). The glottal formant refers to a spectral peak with anti-causal characteristics, therefore in the ZZT

[1] In all experiments presented in this manuscript, the GCI detection is achieved by processing the center-of-gravity evolution signal obtained by shifting an analysis window on the speech signal as described in (Kawahara et al., 2000). Any other GCI estimation method can be used. It has been observed (as further detailed in (Bozkurt, 2005)) that sensitivity to GCI location is not too high: i.e. small shifts in GCI estimations (around %5 of the actual pitch period) do not introduce considerable errors.

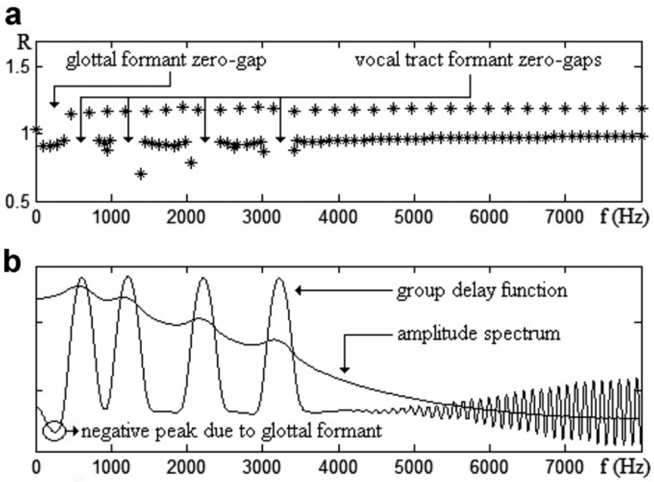

Fig. 7. Single excitation synthetic speech signal: (a) zeros on the *z*-plane in polar coordinates, (b) magnitude spectrum and group delay function.

representation there exists a zero-gap outside the unit circle. This results in a negative peak (or a dip) in the group delay function. This issue is discussed in (Bozkurt and Dutoit, 2003) where the mixed-phase model for speech is presented.

The reason for a zero-free region on the unit circle is as follows. As we have presented in the previous section, the impulse train contributes to the ZZT representation of speech with zeros on the unit circle. When the window size is not larger than 2T0, we can assume that there is no impulse train component and expect to have only zeros inside the unit circle due to the vocal tract and the glottal flow return phase, and zeros outside the unit circle due to the glottal flow first phase. For a single excitation speech signal, our expectation of the ZZT pattern and the group delay function is presented in Fig. 7. The ZZT pattern in Fig. 7 is well-kept once the window is placed such that the increasing exponential part of the time-domain speech signal (mainly due to the first phase of the glottal flow) is multiplied with the first half of the window, which is also increasing, and the decreasing exponential part is multiplied with the second half of the window, which is also decreasing. This results in a zero-gap on the unit circle and smooth group delay functions with mixed-phase characteristics. When the window center is not synchronized

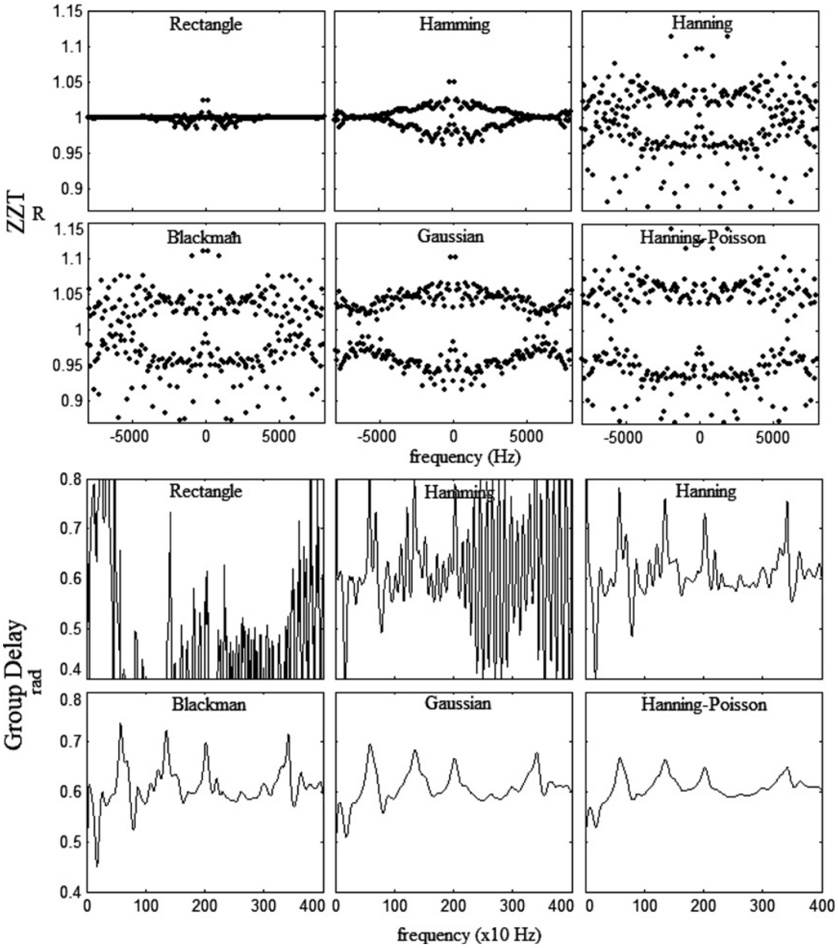

Fig. 8. Effect of window shape to ZZT patterns and group delay functions. GCI synchronous windowing using several window shapes: rectangle, Hamming, Hanning, Blackman, Gaussian and Hanning–Poisson.

with the glottal-closure instant, the ZZT patterns are destroyed and zeros appear on the unit circle resulting in large spikes on the group delay function.

We conclude that GCI synchronous windowing is necessary to obtain windowed signals with ZZT patterns that match the theory presented. In addition, the group delay functions obtained in such a way reveals formant peaks clearly together with a spectral dip for the glottal formant, which is the maximum phase component of the mixed phase signal.

3.3.2. *Effect of window shape on ZZT patterns and group delay*

The window shape has also an important influence to the ZZT patterns and to the group delay functions. In Fig. 8, we present the ZZT patterns and group delay functions obtained by windowing the real speech data in Fig. 6 by various 2T0-size window shapes centered at GCI.

The resulting ZZT patterns are quite different for the different window shapes used. For obtaining spike-free group delay functions, it is important to have a zero-free region around the unit circle and given this criterion Blackman, Gaussian and Hanning–Poisson windowing functions are advantageous.

When GCI synchronous windowing is concerned, it is possible to consider the window being composed of two parts, an increasing first half boosting the glottal flow first phase component of the signal and a decreasing second half boosting the vocal tract response and return phase of glottal flow. This is mainly due to the specific localization of the events in the speech signal, and it only holds for speech signals for which phase spectrum is not modified by some filtering operation. We can consider two parts independently and adjust the contribution of each half which leads to ‘‘asymmetric windowing’’ of the speech frame. In Fig. 9 we present the effect of two different asymmetric windows on the ZZT representation and spectrum of speech frame used in Fig. 6.

The second half of the window in Fig. 9a has a higher decaying coefficient than its first half (and vice versa for the window in Fig. 9b). Multiplication with an exponential shifts the ZZT in the radial direction (Bozkurt, 2005). When we compare the two ZZT plots, we observe that for the first window, ZZT inside the unit circle due to the

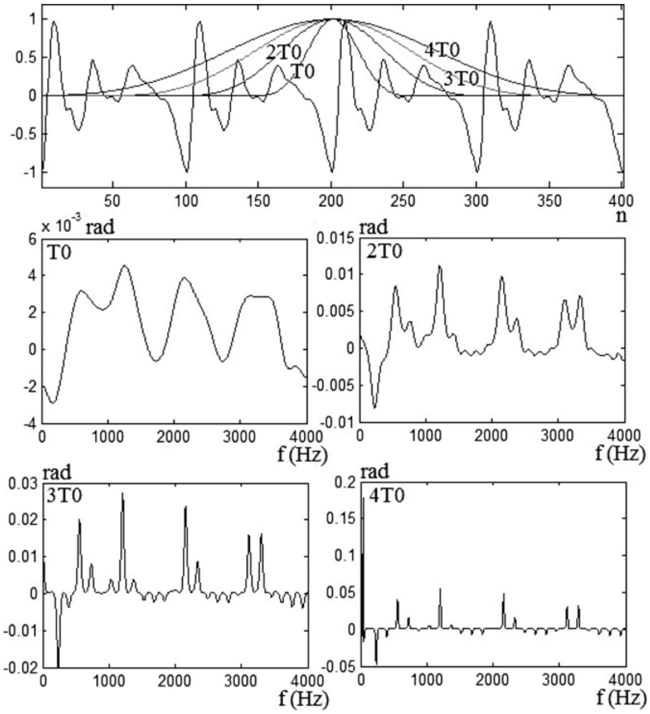

Fig. 10. Effect of windowing size to group delay of a synthetic speech signal. Each (Blackman) window size is indicated on the window waveform on the top figure. The group delay function of the resulting windowed data for each window is presented with the window size indicated on the left-top corner of the figure.

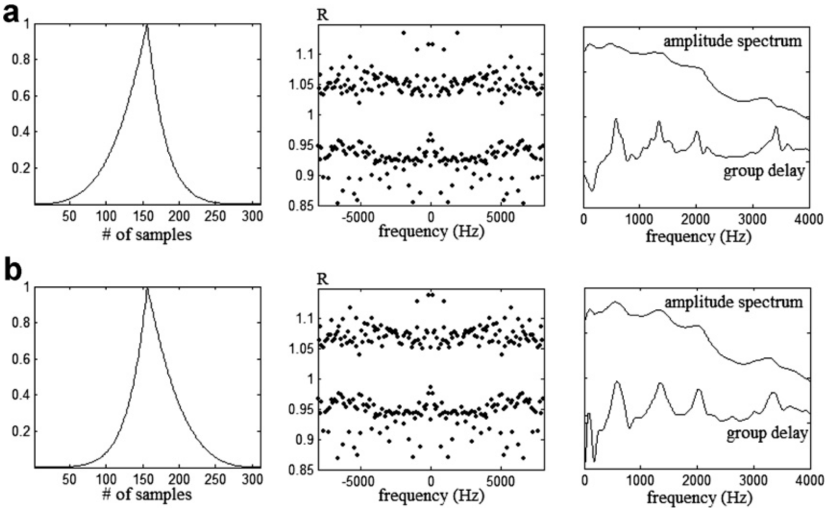

Fig. 9. GCI synchronous Hanning–Poisson asymmetric windowing. Each row includes the window shape used, the ZZT representation, the group delay function and the amplitude spectrum of the windowed signal.

vocal tract response is shifted further away from the unit circle. Therefore, the formant peaks are less prominent. Equally, for the second window, the glottal flow ZZT are shifted away from the unit circle resulting in a less prominent glottal formant peak. This example shows that asymmetric windowing can be applied to adjust the level of contribution of the glottal flow or the vocal tract in the resultant magnitude spectrum once the window is centered at GCI and the window size is less than 2T0.

3.3.3. Effect of window size on ZZT patterns and group delay

The window size is also important. It determines the number of zeros in the ZZT representation. There is especially a significant difference in group delay functions obtained with a window size shorter or longer than 2T0. For windows larger than two pitch periods, the signal contains several periods, which means an impulse train component has to be considered. This results in the ZZT of the impulse train to appear close to the unit circle introducing spikes in the group delay function. This is demonstrated in Fig. 10 where the window center is at GCI and we only vary the window size. A window size in the T0–2T0 range appears to be a good choice for group delay processing.

3.3.4. Group delay spectrogram

We have shown that group delay functions can provide the formant structure of a real speech signal when proper windowing is applied (2T0-size Hanning–Poisson window centered at GCI instants). By gathering group delay functions on successive frames, we can obtain spectrogram-like plots. Fig. 11 shows such a plot as well as the classical spectral magnitude spectrogram for the uttered sentence “*she has left for a great party today*” with modal phonation (www-Voqual03). For the group delay spectrogram, only the positive part of the group delay functions computed for voiced frames, which are characteristic of vocal tract resonances, are rendered. The correlation between the two spectrograms is obvious for the formant tracks. This figure shows that the group delay functions indeed carry resonance information of the signal once windowing is properly performed.

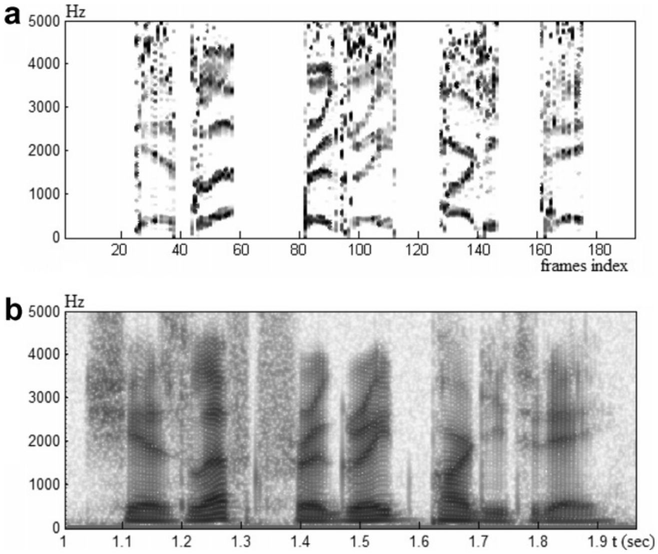

Fig. 11. Group delay spectrogram: (a) Group delay and (b) magnitude spectrogram for the sentence “she has left for a great party today”.

3.4. Appropriate windowing for group delay function computation

In this section we have shown that windowing plays a very important role in reliable group delay estimation. The interesting outcome of this observation is that smooth group delay signals with clear formant peaks can be obtained when the window size is T0 or 2T0, the window center is synchronized with the glottal closure instant and the window shape is Blackman, Gaussian or Hanning–Poisson. For almost all other cases the group delay function includes large spikes hiding resonance information. In the following, we refer to this representation as the Group Delay of GCI-synchronously windowed data (GDGCI).

Until this point we have discussed how to get rid of masking spikes due to inappropriate windowing that hide the speech characteristics (like formants) in the actual group delay functions. Although the group delay functions computed after appropriate windowing are quite smoother and spike-free, it is still difficult to design robust parameter estimation algorithms. The robustness of such an algorithm to additive noise is likely to be low since the noise contribution is likely to add zeros that will fall close to the unit circle thereby introducing spikes. In addition, GCI and pitch detection are rarely very robust which may result in inappropriate size and location of the window. These factors reduce the robustness of group delay based parameter extraction algorithms. In the next section, we further target new group delay representations that are more robust and easier to process for obtaining resonance information. We introduce chirp group delay processing methods as alternative ways of obtaining phase related spectral information. Such representations are potentially useful in various speech applications like feature extraction for automatic speech recognition (ASR), formant tracking, etc.

4. Chirp group delay processing

4.1. Definition

We define the chirp group delay[2] (CGD) as the negative derivative of the phase spectrum (the group delay function) computed from chirp *z*-transform (Rabiner et al., 1969)

[2] Initially the term “differential phase spectrum” has been used in our early papers. After a suggestion by Kuldip Paliwal, we have decided to use “chirp group delay” instead.

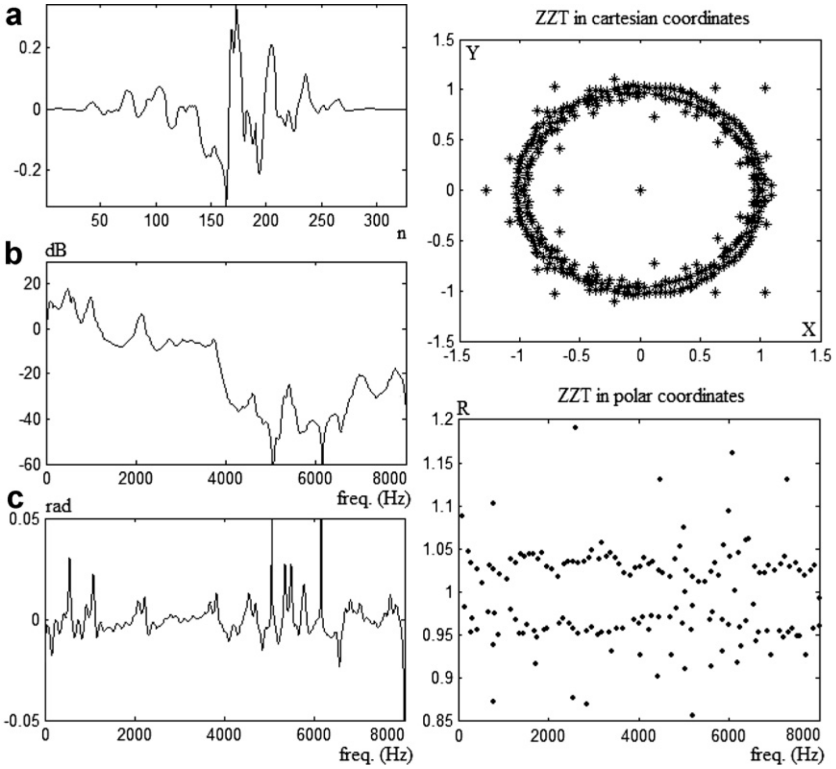

Fig. 12. Remaining problems for group delay of GCI synchronously windowed data: (a) GCI synchronously windowed speech data, (b) magnitude spectrum, (c) chirp group delay function. ZZT representation is presented on the right column both in cartesian and polar coordinates.

that is z-transform computed on a circle/spiral other than the unit circle. Given a short-term discrete-time signal $\{x(n)\}$, $n = 0, 1, \ldots, N-1$, the chirp Fourier transform $\widetilde{X}(\omega)$ is defined as

$$\widetilde{X}(\omega) = X(z)|_{z=\rho e^{j\omega}} = \sum_{n=0}^{N-1} x(n)(\rho e^{j\omega})^{-n} = |\widetilde{X}(\omega)| e^{j\tilde{\theta}(\omega)} \qquad (8)$$

where ρ is the radius of the analysis circle. Similarly to Eq. (1), the chirp group delay function CGD(ω) is defined as

$$\mathrm{CGD}(\omega) = -\frac{d(\tilde{\theta}(\omega))}{d\omega}. \qquad (9)$$

Interestingly enough, a fast Fourier transform (FFT) can be used to compute $\widetilde{X}(\omega)$ by re-writing Eq. (8) as:

$$\widetilde{X}(\omega) = \sum_{n=0}^{N-1} (x(n)\rho^{-n})(e^{j\omega})^{-n} = \sum_{n=0}^{N-1} \tilde{x}(n)(e^{j\omega})^{-n}. \qquad (10)$$

Therefore, for the computation of the CGD(ω) of a given short-term signal, it is sufficient to term-wise multiply the signal samples with an exponential time-series and compute the group delay with direct formula in Eq. (2).

Equivalently, given a ZZT representation for a signal and provided that $x(0)$ is non-zero, we can compute its chirp Fourier Transform by the following equation:

$$\widetilde{X}(\omega) = x(0)(\rho e^{j\omega})^{(-N+1)} \prod_{m=1}^{N-1} (\rho e^{j\omega} - Z_m). \qquad (11)$$

4.2. CGD of speech signals

In Section 3, we have mentioned that appropriate windowing results in avoiding spikes in the group delay function. This is merely true for most of the clean voiced speech frames. However, for some examples (especially those containing additive noise), all spikes cannot be avoided. In Fig. 12, we present one such example of real speech. The formant structure is observed on group delay and most of spikes are avoided. But still we cannot guarantee that no zero will be close to the unit circle and the group delay contains some noise and two sharp spikes. For most of the speech applications, this is undesirable.

For applications like formant tracking and speech recognition, we propose to use CGD with some rough control on zero locations so that group delay is computed on a zero-free region. Two new representations are proposed here for this purpose: a GCI synchronous method and an asynchronous method.

4.2.1. Chirp group delay of GCI-synchronously windowed speech signals

Two steps are necessary in the computation of the Chirp Group Delay of GCI-synchronously windowed speech signals (CGDGCI): suppression of the zeros outside the unit circle on GCI synchronously windowed data and then computing the CGD outside the unit circle from zeros inside the unit circle. This representation contains only the phase information of the minimum-phase component of the data. For GCI synchronously windowed speech signals, the minimum-phase component is due to vocal tract and return phase of the glottal flow, therefore such a representation can be used for formant tracking and speech recognition applications successfully.

In Fig. 13, we present the CGDGCI representation computed on the example in Fig. 12. The ZZT after suppression of zeros outside the unit circle and the analysis circle are presented on the left part of Fig. 13. Clearly, the CGDGCI representation is much smoother than the GDGCI representation in Fig. 12.

In (Bozkurt et al., 2004) we have presented a high-quality formant tracking algorithm, which simply picks the peaks on the CGDGCI representation. There are actually two drawbacks of such an algorithm: (i) GCI detection is necessary and it is sensitive to noise and (ii) roots computation is required for removal and it is computationally heavy. We developed another representation that is presented in the next section to get rid of these difficulties.

4.2.2. Chirp group delay of the zero-phase version of speech signals

Again the procedure contains two steps for the computation of Chirp Group Delay of the Zero-Phase version of a given signal (CGDZP): computation of the zero-phase version of the signal and computation of the CGD on a circle outside the unit circle using the chirp z-transform. The zero-phase version of a short-term signal is obtained by taking the inverse Fourier transform of its magnitude spectrum $|X(\omega)|$. The conversion to zero-phase guarantees that all of the zeros occur very close to the unit circle (Bozkurt, 2005) therefore the resulting chirp group delay representation is very smooth with well-resolved formant peaks. However, the phase information is destroyed for this case, therefore the representation contains only the information available in the magnitude spectrum as in the group delay representations by Yegnanarayana et al. (1988); Murthy et al. (1989); Murthy and Yegnanarayana (1991a), Murthy and Yegnanarayana (1991b) but formant peak resolutions appear with higher resolution than the magnitude spectrum. In Fig. 14, we provide the CGDZP representation computed on the example in Fig. 12. Our Matlab code for CGDZP computation is shared on (www-cgd).

4.3. Discussion

The main motivation for processing CGD computed on circles other than the unit circle is to get rid of spikes created by zeros of the z-transform close to the unit circle, which mask formant peaks on group delay functions. By both manipulating the ZZT and adjusting the analysis circle radius for CGD computation, we can guarantee certain distance of zeros to the analysis circle. The choice of the circle radius is definitively a sensitive issue and its value should be tuned for every application. When the analysis circle is set too far away from the areas on z-plane concentrated with zeros, the resolution of formant peaks become

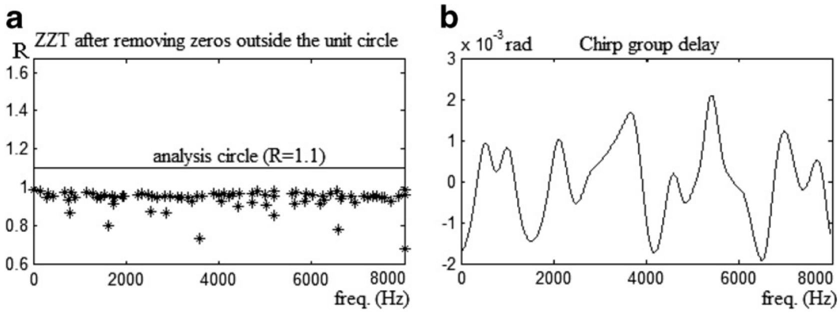

Fig. 13. GCI-synchronous chirp group delay function estimation: (a) ZZT after suppression of ZZT outside the unit circle and (b) chirp group delay.

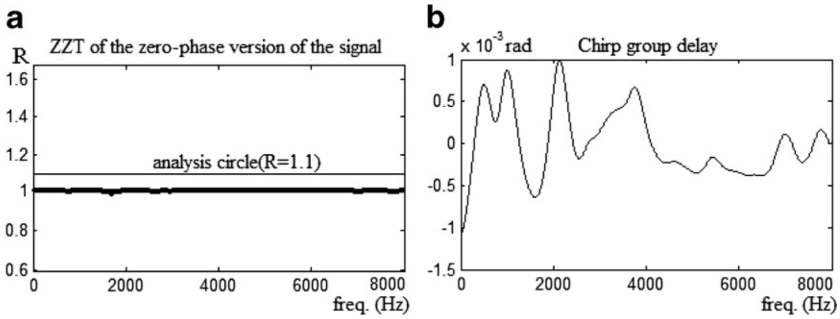

Fig. 14. GCI-asynchronous chirp group delay function estimation: (a) ZZT after suppression of ZZT outside the unit circle and (b) chirp group delay (CGDZP).

poorer (i.e., CGD gets too smooth). As the analysis circle gets closer to those areas, the resolution of formant peaks gets higher but also the risk of having spikes. For this reason, in our experiments with formant tracking and ASR feature estimation, we first searched for some optimum value for the radius (by incrementing the radius by 0.1 in the range $[0.9, 2]$ and checking the system performance) and found that $\rho = 1.12$ appears to be a good choice for 30 ms. window size and 8 kHz sampling rate. In our study we have observed that ρ needs to be tuned according to the number of samples in data window (since this also defines number of zeros).

The asynchronous version is more advantageous than the synchronous method in terms of: (i) computational efficiency, (ii) independency from GCI synchronization and (iii) robustness to noise (see ASR tests in the next section). However, the actual phase information is destroyed in the asynchronous version, since it contains only the information available in the magnitude spectrum.

5. Application to speech recognition

There are various applications where group delay representations, and especially the proposed ZZT-based and chirp group delay representations, can be successfully used. In (Bozkurt et al., 2005) we have presented a source-tract separation algorithm and in (Bozkurt et al., 2004) we have presented a glottal flow parameter estimation method using the ZZT representation. In (Bozkurt, 2005) we have proposed a high quality formant tracker based on processing the CGDZP representation. Due to space limitation in this paper, we have chosen to present only the results of our tests in feature extraction for automatic speech recognition (ASR) to demonstrate the usefulness of group delay representations.

5.1. ASR feature extraction

The first step in automatic speech recognition consists in feature extraction. It aims at representing the time course of speech signals in a more compact, less redundant and less variable form. This is classically performed by chopping speech signals into possibly overlapping finite-length frames and extracting a few acoustic coefficients for every frame, which are well conditioned for pattern recognition. A typical setup consists in using 30 ms frames shifted by 10 ms.

Most of the techniques for computing the acoustic coefficients use Fourier transform to estimate the amplitude/power spectrum and apply some processing to capture its essential shape, and more especially the resonances when they are present. For example, the Mel-warped Frequency Cepstral Coefficient (MFCC) algorithm (Huang et al., 2001) applies a filterbank to the power spectrum of any given frame and then derives cepstral coefficients from the filter outputs. In order to better capture the speech dynamics, the MFCC's are generally augmented with the delta coefficients, which approximate their first-order time derivative (Huang et al., 2001). The set of acoustic coefficients is finally completed with the delta and delta–delta coefficient of the frame log-energy.

In this section, we investigate the possibilities of using group delay representation for ASR feature extraction. The idea that comes naturally to mind is to replace the

Table 1
Methods for ASR feature extraction based on power spectrum or group delay function

Feature Extraction	Description
MFCC	Mel-warped Frequency Cepstral Coefficients: 1st–13th order cepstral coefficients obtained by inverse discrete cosine transform of the log-compressed outputs of a Mel-spaced 24-band filterbank applied to the power spectrum. The feature vector is augmented by delta coefficients and the delta(−delta) frame log-compressed energy
MODGDF-CC	Modified Group Delay Function – Cepstral Coefficients: 1st–13th order cepstral coefficients obtained by inverse discrete cosine transform of the outputs of a Mel-spaced 24-band filterbank applied to the modified group delay function. The feature vector is augmented by delta coefficients and the delta(−delta) frame log-compressed energy
PS-CC	Product Spectrum–Cepstral Coefficients: 1st–13th order cepstral coefficients obtained by inverse discrete cosine transform of the outputs of a Mel-spaced 24-band filterbank applied to the product spectrum. The feature vector is augmented by delta coefficients and the delta(−delta) frame log-compressed energy
GDGCI-CC	GCI-synchronous Group Delay – Cepstral Coefficients: 1st–13th order cepstral coefficients obtained by inverse discrete cosine transform of the outputs of a Mel-spaced 24-band filterbank applied to the group delay of windowed data synchronously with glottal closure instants. The feature vector is augmented by delta coefficients and the delta(−delta) frame log-compressed energy
CGDGCI-CC	GCI-synchronous Chirp Group Delay – Cepstral Coefficients: 1st–13th order cepstral coefficients obtained by inverse discrete cosine transform of the outputs of a Mel-spaced 24-band filterbank applied to the chirp group delay ($\rho = 1.12$) of windowed data synchronously with glottal closure instants. The feature vector is augmented by delta coefficients and the delta(−delta) frame log-compressed energy
CGDZP-CC	Chirp Group Delay of the zero phased version – Cepstral Coefficients: 1st–13th order cepstral coefficients obtained by inverse discrete cosine transform of the outputs of a Mel-spaced 24-band filterbank applied to the chirp group delay ($\rho = 1.12$) of zero-phase version of frame signal. The feature vector is augmented by delta coefficients and the delta(−delta) frame log-compressed energy

All methods are applied on frames of length 30 ms and shifted by 10 ms.

power spectrum by any group delay function in the MFCC feature extraction algorithm. Two recent studies have already considered this approach using the modified group delay function (Hegde et al., 2004b) or the product spectrum (Zhu and Paliwal, 2004) which are reviewed in Section 2.2. We suggest here to compare these two representations as well as the standard power-based MFCC with our proposed group-delay-based representations, which are described in Sections 3 and 4. In Table 1, we list all the feature extractions that are considered in the ASR experiments in next section.

Fig. 15 presents a typical time-domain speech signal and its classical group delay function. As expected, the group delay function computed directly on the speech frame contains mainly spikes and resonance information cannot be observed. In Fig. 16, we show the five group delay based representations together as well as the power spectrum for this speech frame. The formant peaks appear with high resolution in GDGCI, CGDGCI and CGDZP where in MODGDF the spectral envelope appears to be blurred and PS is actually very similar to the power spectrum (as in (Zhu and Paliwal, 2004)). GDGCI includes a spike at high frequencies due to a zero, which cannot be avoided by only GCI-synchronous windowing. As more noise is added to signals, such spikes occur more frequently, therefore the robustness of GDGCI to noise is expected to be low. Thanks to zero removal techniques and zero-phasing, CGDGCI and CGDZP are more robust to noise.

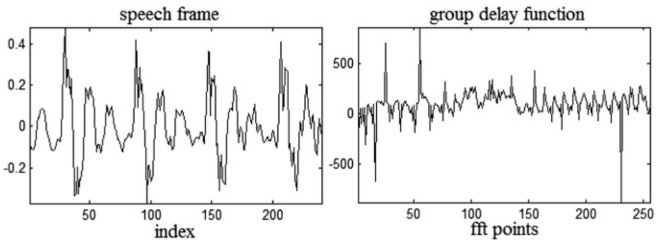

Fig. 15. Time-domain signal of a 30 ms speech frame and its group delay function. The frame example is extracted from the noise-free utterance “mah_4625” of the test set A of the AURORA-2 (Hirsch and Pearce, 2000) and corresponds to vowel/i/ in word “six”.

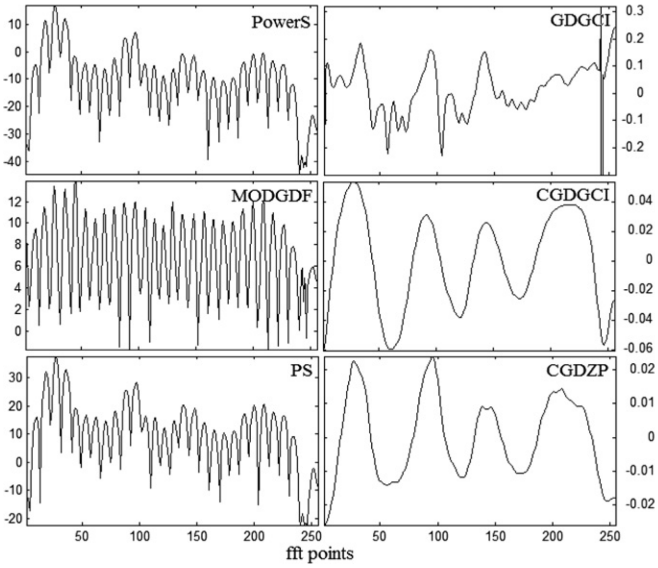

Fig. 16. Power spectrum (PowerS) and group delay representations for the speech signal frame in Fig. 15.

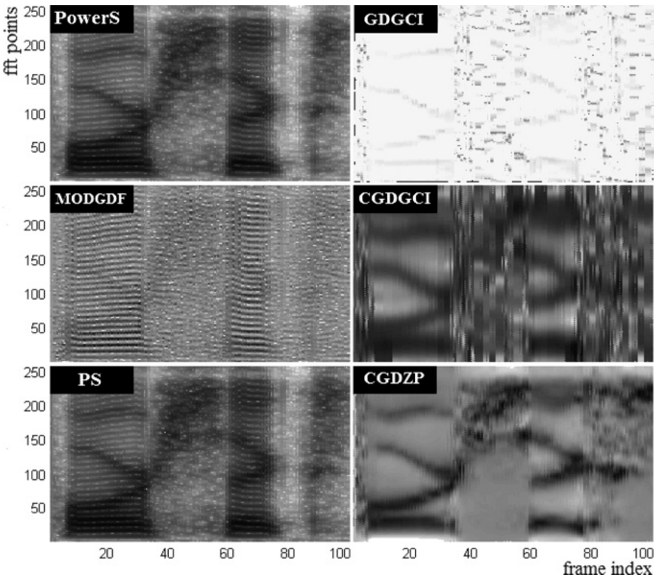

Fig. 17. Spectrogram plots of a noise-free utterance. Only the first half of the file, “mah_4625a”, that contains the digit utterance “4 6” is presented.

In Fig. 17, we also present spectrogram-like plots obtained using the described group delay representations as well as the classical power spectrum. The formant tracks can be well observed on all of the spectrograms except for MODGDF, and PS is again very close to PowerS (as in Fig. 1 of (Zhu and Paliwal, 2004)). GDGCI representation is vague to some level. This is mainly due to the fact that unvoiced frames include spikes with large amplitudes that force a low contrast on the plots. Actually, the group delay functions computed on unvoiced frames mostly do not contain resonance information but random spikes. GDGCI and CGDGCI are actually the two representations that really suffer from this problem.

These observations suggest that the representations have some potential for ASR feature extraction. The main concern is if they can provide complementary information to the power spectrum and improve performance.

5.2. ASR experiments

The ASR system that is considered in this work is based on the STRUT toolkit (www-Strut). It relies on the hybrid Multi Layer Perceptron/Hidden Markov Model (MLP/HMM) technology (Bourlard and Morgan, 1994) where the phonemes of the language under consideration are modeled by HMM’s whose observation state probabilities are estimated as the outputs of a MLP. Such an acoustic model is trained beforehand in a supervised fashion on a large database of phonetically segmented speech material and is naturally dependent on the nature of the extracted acoustic features. Therefore, an acoustic model was built for every feature extraction of Table 1.

The AURORA-2 database (Hirsch and Pearce, 2000) was used in this work. It consists of connected English digit utterances sampled at 8 kHz. More exactly, we used the clean training set, which contains 8440 noise-free utterances spoken by 110 male and female speakers, for building our acoustic models. These models were evaluated on the test set A. It has 4004 different noise-free utterances spoken by 104 other speakers. It also contains the same utterances corrupted by four types of real-world noises (subway, babble, car, exhibition hall) at various signal-to-noise ratios (SNR) ranging from 20 dB to -5 dB. During the recognition experiments, the decoder was constrained by a lexicon reduced to the English digits and no grammar was applied.

Table 2 gives the word error rates (WER) for the ASR system tested with the feature extractions described in Table 1. Errors are counted in terms of word substitutions, deletions and insertions, and error rates are averaged over all noise types. In Table 3, the results are also provided when combining MFCC feature extraction with the others. The combination is simply performed by taking a weighted geometric average of the probability outputs of the combined acoustic models:

$$p_{12} = p_1^{\lambda} \cdot p_2^{1-\lambda}. \tag{12}$$

where p_{12}, p_1 and p_2 denote the combined probability and the probability provided by the two combined acoustic model, respectively. The combination parameter λ takes its value in the range $(0, 1)$ and is optimized for every combination.

Table 2
ASR performances for various feature extractions on the AURORA-2 task

Feature extraction	SNR (dB)							
	∞	20	15	10	5	0	−5	∞
MFCC	1.9	6.7	18.6	45.2	75.1	88.8	91.5	1.9
MODGDF-CC	3.2	19.0	41.7	68.7	86.1	91.0	92.3	3.2
PS-CC	2	6.7	19.4	45.3	75.5	89	92.2	2
GDGCI-CC	8.8	32.8	49.4	69	88.3	98.6	100	8.8
CGDGCI-CC	3.2	12.3	25.6	50.8	80.8	97	99.8	3.2
CGDZP-CC	1.8	5.8	12.2	29.4	62.6	88.7	97.6	1.8

Results are given in terms of word error rate (WER) in percent.

Table 3
ASR performances for features combined with MFCC on the AURORA-2 task

Feature extraction	SNR (dB)							
	∞	20	15	10	5	0	−5	∞
MODGDF-CC	2.1	8.5	23.9	52.7	79.5	89.5	91.5	2.1
PS-CC	1.9	6.7	18.6	44.4	74.6	88.5	91.6	1.9
GDGCI-CC	2.1	7.8	16.8	36	64.4	88	96.1	2.1
CGDGCI-CC	1.8	5.8	12.2	29.1	58	83.8	93.8	1.8
CGDZP-CC	1.7	5	10.4	24.8	52.7	82.3	91.1	1.7

Results are given in terms of word error rate (WER) in percent.

Our main target in these experiments was to test whether a phase/group delay representation carries complementary information to that of the power spectrum in the framework of feature extraction for ASR systems. The results presented in Table 2 shows that the group delay representations have generally this potential. In our in-detailed analysis, we have observed that the GDGCI-CC, which is the pure group delay function computed on GCI-synchronous data without further processing, mainly suffers from window size problems (including several pitch periods result in zeros on the unit circle). In addition, GDGCI-CC and CGDGCI-CC do not carry reliable information for unvoiced frames. Besides, the values in the last two rows of Table 3 compared to the MFCC-only results on the first row in Table 1 are in all cases lower except for the extreme noise setting SNR $= -5$ dB. This demonstrates that CGDGCI-CC and CGDZP-CC features extraction provide complementary information to MFCC.

5.3. Discussion

In this section, we list the five group delay based representations used for feature extraction and we compare them as well as the power spectrum for reference in an ASR experiment. The results show that two of the representations that we propose provide good results, outperform the other group delay representations and contain equivalent and even complementary information to the power spectrum that is potentially useful for improving ASR performance. It should be noted here that for computation of MODGDF-CC we used the same parameter values as in (Hegde et al., 2004b). The authors mention that the parameters need to be tuned for the specific system and data. For this reason our comparison is limited in the sense that only one set of parameters is used for comparison with MODGDF-CC. Our Matlab code for CGDZP computation is shared on (www-cgd) so that further comparisons can be made or the tests can be repeated.

Noise robustness is definitively a sensitive issue and neither MFCC nor the group delay representations are effectively robust to additive noise. The degradation of the recognition performances in the presence of noise is primarily due to the mismatch between the training conditions (clean speech) and the test conditions (noisy speech). Several approaches can be adopted to reduce these acoustics discrepancies (Gong, 1995; Junqua, 2000). First, the speech signal can be captured with as less noise as possible by using spatially selective microphone or arrays of microphones. Further techniques can be applied to enhance speech signals. These techniques concentrate on enhancing amplitude spectrum and can be hardly generalized to phase spectrum. Other techniques aim at adapting the acoustic models to noisy conditions and could be used for group delay representations.

6. Conclusions

This study proposed new group delay based spectral representations and demonstrated their use in speech recognition.

First, the difficulties in group delay processing were discussed. Through a systematic study of the zeros of z-transform (ZZT) of windowed speech signals, we showed that windowing lies at the very heart of the problem of spikes in the derivative of phase spectrum, i.e. the group delay function, due to zeros close to the unit circle. We showed that avoiding these spikes is possible by performing the windowing appropriately: glottal closure instant synchronous windowing with a size of two pitch periods and with one of the following three windowing functions: Blackman, Gaussian or Hanning–Poisson.

Although the group delay one can obtain after appropriate windowing reveals formant peaks, it is not possible to guarantee absence of zeros close to the unit circle for noisy speech even when windowing is appropriately performed. Chirp group delay representation is proposed as an alternative representation to group delay. The chirp group delay function simply corresponds to the group delay function computed on a circle in z-plane other than the unit circle. It is the negative derivative of the phase component of the chirp z-transform. The advantage is that the analysis circle can be chosen in such a way that certain distance from zeros (therefore spike-freeness) is guaranteed.

Due to the strong link between zeros and spikes in the group delay functions, zero locations and phase characteristics need to be studied together. The combination of the ZZT representation with the chirp group delay processing algorithms provides an effective framework for the study of the resonance characteristics of speech signals.

We have demonstrated an application in automatic speech recognition (ASR), more specifically in acoustic feature extraction. The speech recognition tests we have handled were rather limited though sufficient to demonstrate the potential: chirp group delay representations contain equivalent information to that of the power spectrum and are potentially useful for improving speech recognition by combining them with classical acoustic features.

Actually, phase processing is not only necessary for speech processing. We hope that the discussions presented in this work will be also accessible and useful to researchers from different fields like: radar signal processing, medical imaging, sound source localization, optics, solid state physics, geophysics, holography, etc.

Acknowledgements

Baris Bozkurt was funded by Region Wallonne/Belgium, grant FIRST EUROPE #215095. Laurent Couvreur was funded by Region Wallonne/Belgium in the project WIST IC&C. We would like to thank to Christophe d'Alessandro and Boris Doval for their valuable contribution to the previously published ZZT study which served as a basis in this study.

References

Abel, N.H., 1826. Beweis der Unmöglichkeit, algebraische Gleichungen von höheren Graden als dem vierten allgemein aufzulösen. J. Reine Angew. Math 1, 65.

Alsteris, L., Paliwal, K.K., 2004. Importance of window shape for phase only reconstruction of speech. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), May, pp. 573–576.

Andersen, T.H., Jensen, K., 2001. On the importance of phase information in additive analysis/synthesis of binaural sounds. In: Proc. of International Computer Music Conference (ICMC), August.

Banno, H., Takeda K., Itakura, F., 2001. A study on perceptual distance measure for phase spectrum of stimuli. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), May, pp. 3297–3300.

Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publisher.

Bozkurt, B., Dutoit, T., 2003. Mixed-phase speech modeling and formant estimation, using differential phase spectrums In: Proc. of ISCA Turorial and Research Workshop on Voice Quality (VOQUAL), August, pp. 21–24.

Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T., 2004. Improved differential phase spectrum processing for formant tracking. In: Proc. of International Conference on Spoken Language Processing (ICSLP), October.

Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T., 2005. Zeros of z-transform representation with application to source–filter separation in speech. IEEE Signal Process. Lett. 12 (4), 344–347.

Bozkurt, B., Couvreur, L., On the use of phase information for speech recognition. In: Proc. of European Signal Processing Conference (EUSIPCO)'05, September.

Bozkurt, B., 2005. Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals, Ph.D. thesis, Faculté Polytechnique De Mons, Belgium.

Chavez, S., Xiang, Q.S., An, L., 2002. Understanding phase maps in MRI: a new cutline phase unwrapping method. IEEE Trans. Medical Imaging 21 (8), 966–977.

Chen, C.W., Zebker, H.A., 2002. Phase unwrapping for large SAR interferograms: statistical segmentation and generalized network models. IEEE Trans. Geosci. Remote Sensing 40 (8), 1709–1719.

Costantini, M., Farina, A., Zirilli, F., 1999. A fast phase unwrapping algorithm for SAR interferometry. IEEE Trans. Geosci. Remote Sensing 37 (1), 452–460.

Doval, B., d'Alessandro, C., Henrich, N., 2003. The voice source as a causal/anti-causal linear filter. In: Proc. of ISCA Turorial and Research Workshop on Voice Quality (VOQUAL), August, pp. 15–19.

Edelman, A., Murakami, H., 1995. Polynomial roots from companion matrix eigenvalues. Math. Comput. 64 (210), 763–776.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton and Co., Netherlands.

Fant, G., 1985. The LF-model revisited transformation and frequency domain analysis. Speech Trans. Lab. Q. Rep., Royal Inst. Tech 2–3, 121–156.

Frolova, G.V., Taxt, T., 1996. Homomorphic deconvolution of medical ultrasound images using a Bayesian model for phase unwrapping. In: Proc. of Ultrason. Symp. 2, 1371–1376.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Commun. 16 (3), 261–291.

Harris, F.J., 1978. On the use of windows for harmonic analysis with the Discrete Fourier Transform. Proc. IEEE 66 (1), 51–83.

Hedelin, P., 1988. Phase compensation in all-pole speech analysis. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 339–342.

Hegde, R.M., Murthy H.A., Gadde, V.R., 2004a. The modified group delay feature: a new spectral representation of speech. In: Proc. of International Conference on Spoken Language Processing (ICSLP), October.

Hegde, R.M., Murthy H.A., Gadde, V.R., 2004b. Continuous speech recognition using joint features derived from the modified group delay function and MFCC. In: Proc. of International Conference on Spoken Language Processing (ICSLP), October.

Hirsch, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluation of speech recognition Systems under noisy conditions. In: Proc. of ISCA Turorial and Research Workshop on Automatic Speech Recognition (ASR), September.

Huang, X., Acero, A., Hon, H.W., 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall.

Junqua, J.C., 2000. Robust Speech Processing in Embedded Systems and PC Applications. Kluwer Academic Publishers.

Kawahara, H., Estill J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: Proc. of International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), September.

Kawahara, H., Atake, Y., Zolfaghari, P., 2000. Accurate vocal event detection method based on a fixed-point to weighted average group delay. In: Proc. of International Conference on Spoken Language Processing (ICSLP), Beijing, China, October.

Li, D., Levinson, S.E., 2002. A linear phase unwrapping method for binaural sound source localization on a robot. Proc. Int. Conf. Robotics Automation (ICRA) 1, 19–23.

Liu, L., He, J., Palm, G., 1997. Effects of phase on the perception of intervolic stop consonants. Speech Commun. 22 (4), 403–417.

Marques, J.S., 1989. Sinusoidal modeling of speech: application to medium to low bit rate coding, Ph.D. thesis, Technical University of Lisbon, Portugal.

Marques, J. S., Almeida, L. B., Tribolet, J. M., 1990. Harmonic coding at 4.8 kb/s. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 17–20.

McAulay, R.J., Quatieri, T.F., 1991. Sinusoidal coding. In: Kleijn, W., Paliwal, K. (Eds.), Speech Coding and Synthesis. Marcel Dekker, pp. 165–172.

Murthy, H.A., Murthy, K.V., Yegnanarayana, B., 1989. Formant extraction from phase using weighted group delay function. Electron. Lett. 25 (23), 1609–1611.

Murthy, H.A., Yegnanarayana, B., 1991a. Formant extraction from group delay function. Speech Commun. 10 (3), 209–221.

Murthy, H.A., Yegnanarayana, B., 1991b. Speech processing using group delay functions. Signal Process. 22, 259–267.

Oppenheim, A.V., 1969. A speech analysis-synthesis system based on homomorphic filtering. J. Acoust. Soc. Amer. (JASA) 45 (2), 458–465.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing, second ed. Prentice-Hall.

Paliwal, K. K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proc. of European Conference on Speech Communication and Technology (EUROSPEECH), September, pp. 2117–2120.

Patterson, R.D., 1987. A pulse ribbon model of monoaural phase perception. J. Acoust. Soc. Amer. 82 (5), 1560–1586.

Pobloth H., Kleijn, W. B., 1999. On phase perception in speech. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 29–32.

Quatieri, T.F., 1979. Minimum and mixed-phase speech analysis-synthesis by adaptive homomorphic deconvolution. IEEE Trans. Acoustics, Speech Signal Process. 27 (4), 328–335.

Rabiner, L.R., Schafer, R.W., Rader, C.M., 1969. The chirp z-transform algorithm and its application. Bell Syst. Tech. J 48 (5), 1249–1292.

Schroeder, M.R., 1959. New results concerning monoaural phase sensitivity. J. Acoust. Soc. Amer. 31 (11), 1597.

Schroeder, M.R., Strube, H.W., 1986. Flat-spectrum speech. J. Acoust. Soc. Amer. 79 (5), 1580–1583.

Sitton, G.A., Burrus, C.S., Fox, J.W., Treitel, S., 2003. Factoring very-high-degree polynomials. IEEE Signal Process. Mag. 20 (6), 27–42.

Stylianou, Y., 1996. Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modification, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, France.

Sun, X., 1997. Phase modeling of speech excitation for low bit-rate sinusoidal transform coding. In: Proc. of International Conference on Acoust., Speech Signal Process. (ICASSP), vol. 3, pp. 1691–1694.

von Helmholtz, H.L.F., 1912. On the Sensations of Tone, London.

Vyacheslav, V., Zhu, Y., 2003. Deterministic phase unwrapping in the presence of noise. Opt. Lett. 28 (22), 2156–2158.

Yegnanarayana, B., Saikia, D.K., Krishnan, T.R., 1984. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Trans. Acoust., Speech Signal Process. 32 (3), 610–623.

Yegnanarayana, B., Duncan, G., Murthy, H. A., 1988. Improving formant extraction from speech using minimum-phase group delay spectra. In: Proc. of European Signal Processing Conference (EUSIPCO), vol. 1, pp. 447–450.

Zhu, D., Paliwal, K. K., 2004. Product of power spectrum and group delay function for speech recognition. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 125–128.

Zolfaghari, P., Nakatani, T., Irino, T., Kawahara, H., Itakura, F., 2003. Glottal closure instant synchronous sinusoidal model for high quality speech analysis/synthesis. Proc. of European Conference on Speech Communication and Technology (EUROSPEECH), 2441–2444.

Boite, J.-M., Couvreur, L., Dupont, S., Ris, C., Speech Training and Recognition Unified Tool (STRUT), <http://tcts.fpms.ac.be/asr/project/strut>.

Speech Material for the 2003 workshop on Voice Quality – Function, Analysis and Synthesis, <http://www.limsi.fr/VOQUAL>.

Introduction page for Chirp Group Delay processing: <http://tcts.fpms.ac.be/demos/zzt/cgd.html>.

Demo Page for Zeros of the Z-Transform (ZZT) Representation: <http://tcts.fpms.ac.be/demos/zzt>.