Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 782–795

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Spectro-temporal analysis of speech signals using zero-time windowing and group delay function

Yegnanarayana Bayya [a], Dhananjaya N. Gowda [b,*]

[a] International Institute of Information Technology, Hyderabad, India
[b] Department of Information and Computer Science, Aalto University, Finland

Received 14 December 2012; received in revised form 14 February 2013; accepted 20 February 2013
Available online 9 April 2013

Abstract

Traditional methods for estimating the vocal tract system characteristics typically compute the spectrum using a window size of 20–30 ms. The resulting spectrum is the average characteristics of the vocal tract system within the window segment. Also, the effect of pitch harmonics need to be countered in the process of spectrum estimation. In this paper, we propose a new approach for estimating the spectrum using a highly decaying window function. The impulse-like window function used is an approximation to integration operation in the frequency domain, and the operation is referred to as zero-time windowing analogous to the zero-frequency filtering operation in frequency domain. The apparent loss in spectral resolution due to the use of a highly decaying window function is restored by successive differencing in the frequency domain. The spectral resolution is further improved by the use of group delay function which has an additive property on the individual resonances as against the multiplicative nature of the magnitude spectrum. The effectiveness of the proposed approach in estimating the spectrum is evaluated in terms of its robustness to additive noise, and in formant estimation.

Keywords: Zero-time windowing; Zero-frequency filtering; Group delay function; NGD spectrum; HNGD spectrum

1. Introduction

During production of speech the excitation source and the shape of the vocal tract change continuously with time, with significant interaction between the source and the system. It is indeed a signal processing challenge to extract features of the changing vocal tract system from the speech signal. Short-time spectrum analysis smears the information of the vocal tract system either in the time domain (as in the narrowband (NB) spectrogram) or in the frequency domain (as in the wideband (WB) spectrogram). Moreover, the averaging effect in the time domain in the NB spectrogram may destroy the critical/useful information of the vocal tract system captured in the signal immediately after the impulse-like excitation, which takes place around the glottal closure instant (GCI) in each glottal cycle. In methods based on modeling the vocal tract system, such as all-pole model in the linear prediction (LP) analysis (Makhoul, 1975), the size of the window and the order of the prediction need to be chosen a priori. If the size of the window is large (> 2 pitch periods), the pitch period affects the estimated LP coefficients (LPCs), and if the size of the window is small, the LPCs are affected due to poor estimation of the autocorrelation coefficients from short segments of data. Moreover, the position of the window relative to the signal waveform also influences the analysis significantly. Pitch synchronous analysis, anchored around the epochs (GCIs), can help overcome the effects of position of the window. Averaging the autocorrelation coefficients over three or four successive glottal cycles will reduce the errors in their estimation. These methods are used to derive the formants in the closed and open phase regions of the glottis in each cycle (Yegnanarayana and

* Corresponding author.
E-mail address: dhananjaya.gowda@aalto.fi (D.N. Gowda).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.02.007

Veldhuis, 1998). But these methods require careful analysis of speech signal to select those regions. For larger size windows, one can obtain an estimate of the overall (averaged or smoothed) spectrum over that window, even though it is known that the vocal tract system may change significantly even within a pitch period due to opening and closing of the glottis within a glottal cycle. Efforts to overcome the influence of the pitch period on the envelope of the short-time spectrum, such as TANDEM-STRAIGHT (Kawahara et al., 2008), are very effective in producing the spectral envelope corresponding to the response of the vocal tract system. But this response also represents the averaged spectral characteristics over the duration of the analysis segment.

In general, most of the attempts to capture the time-varying characteristics of the excitation source and the vocal tract system during speech production involve one of the following:

(a) *Quasistationarity assumption of the production process (Rabiner and Schafer, 2010; Deller et al., 2000)*
In this representation the spectral envelope and the excitation characteristics are interpreted from the short-time spectrum derived through discrete Fourier transform (DFT) or LP analysis using block processing. As mentioned earlier, these analysis methods bring out the average characteristics over the block of data.

(b) *Mono and multicomponent AM–FM sinusoidal model representations of the signals (Gianfelici et al., 2007; Santhanam and Maragos, 2000;* Abe and Honda, 2006*)*
These representations involve estimation of the time-varying model parameters. Here the model parameters are sensitive to block processing. Also this representation does not use any knowledge of speech production.

(c) *Source-system models of speech production (Vargas and McLaughlin, 2008; Welling and Ney, 1998; Deng et al., 2006)*
These models assume that the vocal tract system can be represented as time-varying resonances, and attempt to derive the parameters. For this the resonance information is extracted either by LP analysis or by damped sinusoidal models using block processing.

(d) *Instantaneous fundamental frequency and pitch adaptive analysis (Kawahara et al., 1999; Yegnanarayana and Veldhuis, 1998)*. In this case the instantaneous fundamental frequency is derived, followed by pitch synchronous analysis of the speech segments. But the analysis depends critically on extraction of the pitch epochs.

In this paper we address the issue of extraction of the spectral features of the vocal tract system from very short (effectively much smaller than 5 ms) segment of speech. A new approach called *zero-time windowing* (ZTW) of speech is proposed here to extract the spectral characteristics of the vocal tract system. The ZTW operation involves multiplication of a short duration speech signal with a window, which results in an impulse-like signal with most of the energy at the beginning of the window, i.e., near zero-time. Hence the signal components at the beginning of the window get more emphasis. This is important if the objective is to derive the vocal tract response characteristics at some important events in speech production, such as at the instants of significant excitation of the vocal tract system.

The effect of zero-time windowing operation in time domain is equivalent to smoothing the spectrum by successive integration in the frequency domain. This is analogous to the *zero-frequency filtering* (ZFF) proposed in Murty and Yegnanarayana (2008), Yegnanarayana and Murty (2009) for extraction of the excitation source features, such as epochs and instantaneous fundamental frequency. The features in the spectrum derived using ZTW are highlighted by exploiting the high resolution and additive properties of the group-delay function (Yegnanarayana, 1978; Yegnanarayana and Murthy, 1992; Joseph et al., 2006). These features are extracted effectively when the analysis segment starts around the epoch, which most of the time is the desired information. The resulting spectral information is not affected by the duration of the pitch period. The spectral information is also not dependent on the choice of model parameters, as in the LP analysis. It should be noted that the proposed ZTW method is not intended as an alternative to the existing methods of short-time spectrum analysis. The objective of the proposed method is mainly to explore the possibility of deriving the spectral characteristics at every sampling instant, so that the time varying characteristics of the vocal tract can be captured. In some sense the method may complement (by providing temporal resolution) the existing methods which provide mostly the spectral characteristics of the vocal tract system averaged over the analysis window.

The organization of the paper is as follows: Section 2 discusses the basis for the proposed method of deriving the spectral information at every instant. Section 3 presents the proposed method for extracting the spectral features, and shows the instantaneous nature of the spectral features. The section also discusses the effect of analysis parameters (size and shape of the analysis window used for each segment) on the resolution of the spectral features and on the ripple due to truncation. Section 4 gives analysis of the instantaneous spectral features for different categories of sounds. The robustness of the proposed ZTW method in estimating the spectrum under different degradations, and its ability to extract formants from short segments of speech is discussed in Section 5. A comparison with the popular STRAIGHT and LP based methods is also provided. Section 6 gives a summary and possible applications of the proposed method for speech analysis.

2. Basis for the proposed method of extracting instantaneous spectral features

2.1. Zero-frequency filtering

Filtering speech signals through a *zero-frequency resonator* (ZFR) was proposed recently to extract the epochs in voiced speech (Murty and Yegnanarayana, 2008). The method involves passing the speech signal through an ideal digital resonator located at 0 Hz, whose transfer function is given by

$$H(z) = \frac{1}{(1-z^{-1})^2} = \frac{1}{1-2z^{-1}+z^{-2}}. \quad (1)$$

This is equivalent to the following operation in the time domain:

$$y_1[n] = 2y_1[n-1] - y_1[n-2] + s[n], \quad (2)$$

where $s[n]$ is input (differenced to remove the DC bias) speech signal, and $y_1[n]$ is the output of the resonator. Fig. 1(b) shows the output of the 0 Hz resonator for the synthetic signal shown in Fig. 1(a). The synthetic signal is generated at 8 kHz sampling rate using an impulse train with 10 ms periodicity, and the 10th order LP coefficients derived from a segment of continuous speech corresponding to vowel [i]. For epoch extraction, $s[n]$ is passed through a cascade of two resonators, which is equivalent to integrating the signal four times, to reduce the effects of formants. The resulting output

$$y_2[n] = 2y_2[n-1] - y_2[n-2] + y_1[n] \quad (3)$$

has polynomial growth as shown in Fig. 1(c). The trend in the output is removed by subtracting the running mean computed over a window of size $2P+1$ samples around each instant. The resulting signal is called *zero-frequency filtered* (ZFF) signal, or simply the *filtered signal*, and is given by

$$y[n] = y_2[n] - \frac{1}{2P+1}\sum_{k=-P}^{P} y_2(n+k), \quad (4)$$

where $2P+1$ is the window size in samples corresponding to about 1.5 times the average pitch period. The filtered signal (see Fig. 1(d)) is used to derive the important characteristics of speech production, namely, that of the excitation source (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009). The instants of positive zero-crossing correspond to the locations of the impulse-like excitations, and are called epochs. The locations of epochs are shown (denoted by arrows) in Fig. 1(d), and also in Fig. 1(a).

The frequency domain interpretation of the operations in zero-frequency filtering is illustrated in Figs. 1(e) through 1(h). Fig. 1(e) shows the spectrum of the segment of synthetic speech in Fig. 1(a). Figs. 1(f) and 1(g) show the spectra of the outputs of a single and a double resonator, respectively, obtained by multiplying the spectrum in Fig. 1(e) with the frequency responses (shown by dotted lines in the respective figures) of the resonators. Note the large dynamic ranges of the spectra in Figs. 1(f) and 1(g), relative to the dynamic range of the spectrum (Fig. 1(e)) of the signal. Note also that these spectra cannot be computed directly from Figs. 1(b) and 1(c) due to discontinuity of the signal at the ends of the segments. In Fig. 1(g) most of the spectral information (i.e., formants and spectral roll-off) is deemphasized, and only the region near 0 Hz has significant (in terms of amplitudes) values. The spectrum of the ZFF signal in Fig. 1(d) is shown in Fig. 1(h), which shows a peak around the pitch frequency. In other words,

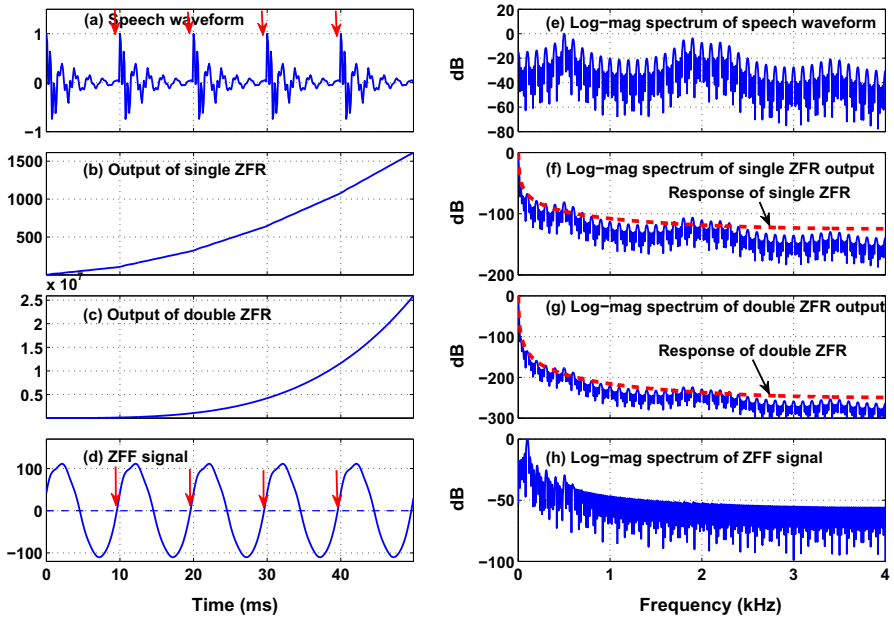

Fig. 1. Illustration of zero-frequency filtering of speech signals on a segment of synthetic speech waveform. The epoch locations are marked by downward arrows in (d).

it appears that the original signal in Fig. 1(a) is filtered by a bandpass (resonance-like) filter. But it is important to note that the peak of this filter, as governed by the choice of window length for trend removal, need not be located at the pitch frequency. The locations of the positive zero-crossings in Fig. 1(d) do not change, as long as the peak of this bandpass filter is around the pitch frequency. This can be verified by taking a segment consisting of sequence of impulses spaced by pitch period, and computing the ZFF using different window lengths (in the range of 1–2 pitch periods) for the trend removal operation in Eq. (4). More discussions and illustrations on the effect of window length in ZFF are given in Murty and Yegnanarayana (2008), Yegnanarayana and Murty (2009).

2.2. Zero-time windowing

Note that the zero-frequency resonator has the following frequency response (Murty and Yegnanarayana, 2008):

$$|H(\omega)| = \left|\frac{1}{(1-z^{-1})^2}\right|_{z=e^{j\omega}} = \frac{1}{2(1-\cos\omega)} = \frac{1}{4\sin^2\omega/2.} \quad (5)$$

In the discrete frequency domain, $\omega = 2\pi k/N$, where N is the number of samples in the DFT computation. The operation of passing a signal through a 0 Hz resonator is equivalent to multiplying the spectrum of the signal with a window function given by the frequency response of the resonator. This is equivalent to integrating the signal twice, which results in a polynomial-type growth/decay function (Murty and Yegnanarayana, 2008). The advantage of using such a window function in the frequency domain is that, unlike any other window it does not smear the discontinuities or abruptness in the signal in time domain. All the discontinuities such as those due the instants of glottal closure are preserved, and can be obtained by successive differencing. In this paper, we propose to use a similar windowing operation, but in the time domain. By choosing a similar function in time domain which gives more weightage to samples around zero-time, we are performing an operation that is closer to integration in the frequency domain, and thereby not smearing the spectral information as much as any arbitrary window would.

Let us now consider the equivalent operations in the time domain, analogous to zero-frequency filtering in frequency domain, i.e., multiplying the time domain signal with a window function similar in shape to the frequency response of the 0 Hz resonator. The window function (analogous to Eq. (5)) is given by

$$w_1[n] = \begin{cases} 0, & n = 0 \\ 1/(4\sin^2(\pi n/(2N))), & n = 1, 2, \ldots, N-1, \end{cases} \quad (6)$$

where N is the window length. The value $w_1[0] = 0$ is chosen to avoid division by zero. This also makes the mean value of the FT of the windowed signal to be zero, but does not alter the spectral peaks or valleys. Note that a window function of the type $1/n$ is equivalent to integration in the frequency domain. The window function $1/(4\sin^2(\pi n/(2N)))$ can be approximated to $1/n^2$ for smaller values of n, if $N \gg M$, where M is the length of the speech segment and N is chosen to be the DFT length or the length of the segment after appending with zeros. Hence the chosen function provides an approximation to integration in frequency domain.

A segment of speech signal starting at an arbitrary epoch location is shown in Fig. 2(a). The windowed signals obtained by using the window function $w_1[n]$ once and twice (equivalent to single and double resonator of the type in Eq. (5) used in obtaining ZFF signal), are shown in Figs. 2(b) and 2(c), respectively. The DFT spectra of the signals in Figs. 2(a), 2(b) and 2(c) are shown in Figs. 2(d), 2(e) and 2(f), respectively. The spectra of the windowed signals in Figs. 2(e) and 2(f) are smoothed versions of the spectrum in Fig. 2(d). This is analogous to Figs. 1(b) and 1(c) with respect to Fig. 1(a). Note the relative scales of the vertical axes in these cases like in Figs. 1(b) and 1(c).

The spectral features such as the peaks due to formants appear to be lost in Fig. 2(f) due to smoothing. But these features are embedded in the fine fluctuations in the smoothed spectrum. One obvious way of extracting these spectral features is by removing the trend, which can be achieved by successive differencing in this case. Fig. 2(g) shows the spectrum obtained by twice successively differencing the plot in Fig. 2(f), and then computing the Hilbert envelope (HE). The reason for computing the HE after successive differencing is explained in Section 3.4. The formant features of Fig. 2(d) can be seen in Fig. 2(g). The spectral rolloff feature is lost in these operations. Also note that the vertical scale in Fig. 2(g) is much smaller than in Fig. 2(f), as the values in Fig. 2(g) are the small fluctuations of the values in Fig. 2(f).

One can exploit the additive and high resolution properties of the group-delay function to highlight the formant features of the spectrum (Yegnanarayana, 1978). Let $x[n]$ be the windowed signal. That is

$$x[n] = s[n]w_1[n], \quad n = 0, 1, 2, \ldots, N-1. \quad (7)$$

The group-delay function is computed as follows (Oppenheim and Schafer, 1975):

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}, \quad (8)$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the discrete-time Fourier transform (DTFT) of $x[n]$, and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ is the DTFT of $y[n] = nx[n]$. The division by the squared magnitude spectrum in Eq. (8) can lead to problems when $|X(\omega)|^2$ is very small due to the presence of zeros in the vocal tract system. In fact the division by $|X(\omega)|^2$ in Eq. (8) can be completely avoided. The features of the spectrum

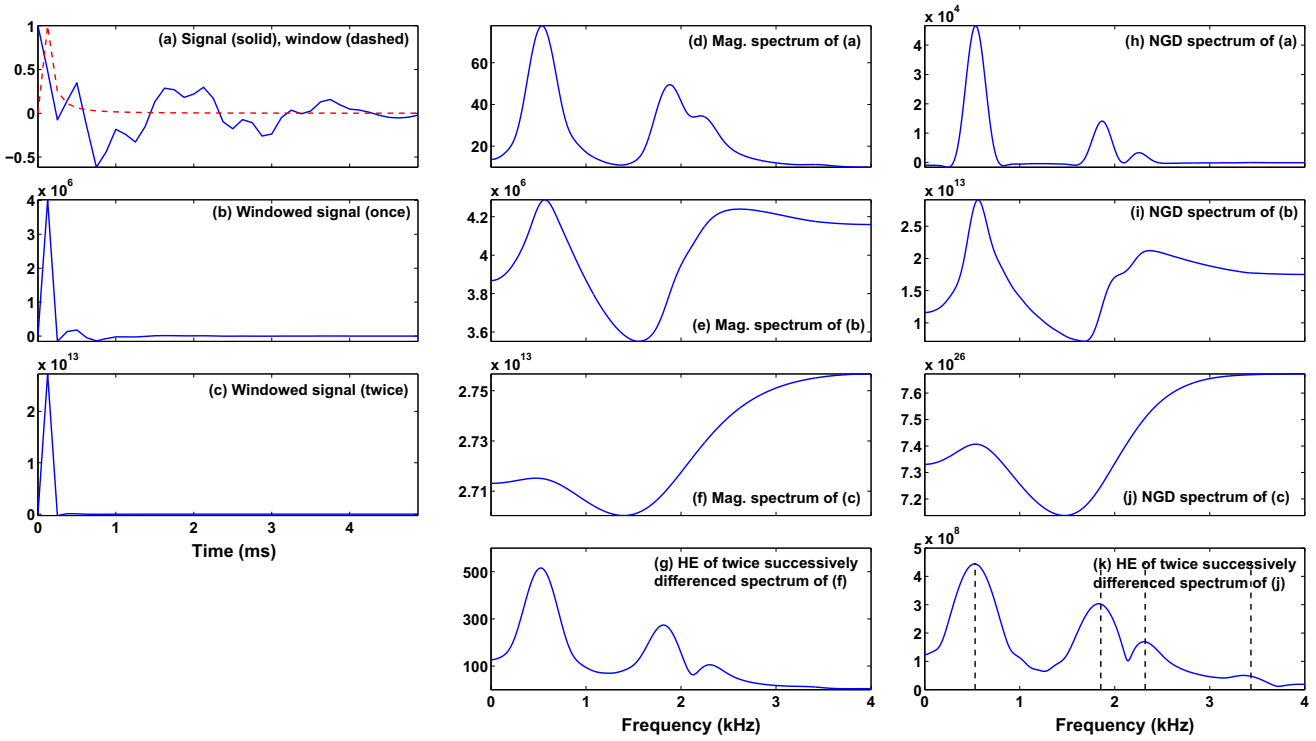

Fig. 2. Illustration of zero-time windowing for a short segment (5 ms) of synthetic speech. (a) Signal (solid line) and the window function (dashed line). The window function is scaled by $\max(w_1[n])$ for display purpose. (b) Windowed signal (once). (c) Windowed signal (twice). (d) Magnitude spectrum of (a). (e) Magnitude spectrum of (b). (f) Magnitude spectrum of (c). (g) HE of twice successively differenced spectrum of (f). (h) NGD spectrum of (a). (i) NGD spectrum of (b). (j) NGD spectrum of (c). (k) HE of twice successively differenced spectrum of (j). The true formant frequencies are shown by vertical lines.

can also be seen in the numerator of the group-delay (NGD) function, i.e.,

$$g(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega). \tag{9}$$

The NGD $g(\omega)$ usually has higher resolution (i.e., $g(\omega) \propto |X(\omega)|^4$) around the formants than even $\tau(\omega)$, since $\tau(\omega) \propto |X(\omega)|^2$ around the formants (Joseph et al., 2006; Yegnanarayana, 1978). This high resolution property of NGD helps to highlight the resonance features of the spectrum.

Figs. 2(h), 2(i) and 2(j) are the NGD plots (i.e., $g(\omega)$) corresponding to signals in Figs. 2(a), 2(b) and 2(c), respectively. The NGD plot in Fig. 2(k) is obtained by twice successively differencing the NGD plot in Fig. 2(j), and then computing the HE. Note that the spectral peak features can be seen better in Fig. 2(k) as compared to Fig. 2(g), especially the higher formants. One can observe even the weak fourth formant in this case. This is primarily due to the multiplicative and additive nature of the individual resonances in the magnitude spectrum and group delay spectrum, respectively (Yegnanarayana, 1978). The true formant frequencies are shown by vertical lines in Fig. 2(k).

The main objective of this paper is to explore the possibility of deriving the instantaneous spectral characteristics at every sampling instant, if possible. In order to achieve this objective the segment of speech signal is multiplied with a window function of the type given in Eq. (6), so that the samples near the time origin get more emphasis than other samples in the segment. The windowed signal $x[n]$ looks almost like an impulse, as most large amplitude values are near the origin $n = 0$. The spectral characteristics derived from the windowed signal indeed correspond to the characteristics of the segment around zero-time. Hence the result can be interpreted as *instantaneous spectral features*. In the next section, we describe the zero-time windowing method for extraction of instantaneous spectral features from speech signals.

3. Zero-time windowing approach for instantaneous spectral features

3.1. Basic steps in ZTW

The basic steps in the proposed approach are as follows:

(a) Consider the differenced signal $s[n]$ at the sampling frequency of f_s Hz. For discussion throughout this section we consider f_s =10 kHz.
(b) Consider M samples of the signal, starting from an arbitrary reference set at $n = 0$. That is $s[n]$ is defined for $n = 0, 1, \ldots, M-1$.
(c) Choose the DFT length $N \gg M$ so that we have sufficient sampling in the frequency domain. Append the signal $s[n]$ with appropriate number of zeros to make its length equal to N.

(d) Compute the windowed signal $x[n] = s[n]w_1[n]$, for $n = 0, 1, \ldots, N-1$, where $w_1[n]$ is given by Eq. (6).
(e) Compute the NGD function of $x[n]$. The NGD function of $x[n]$ is given by

$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], \quad k = 0, 1, \ldots, N-1 \qquad (10)$$

where $X[k] = X_R[k] + jX_I[k]$ is the N-point DFT of the sequence $x[n]$, and $Y[k] = Y_R[k] + jY_I[k]$ is the N-point DFT of the sequence $y[n] = nx[n]$. Note that in these notations $X[k] = X(\omega)|_{\omega=2\pi k/N}$, $Y[k] = Y(\omega)|_{\omega=2\pi k/N}$, and $g[k]$ is obtained through $X[k]$ and $Y[k]$. Note also that $g[k]$ is not the sampled version of $g(\omega)$ in Eq. (9), as it is computed using $X[k]$ and $Y[k]$.

Fig. 3(a) shows a segment of speech waveform. The NGD plots are obtained at every sampling instant of time, and are shown in Fig. 3(b). In this case $M = 50$ samples, corresponding to 5 ms data. Each analysis segment of the signal is appended with $N - 50$ zeros, where $N = 2048$ in this illustration. The NGD plots are given for $(N/2 + 1)$ points, corresponding to the frequency range of $0 - f_s/2$. Fig. 3(c) shows the DFT spectrum plots of the windowed signal $x[n]$ computed at each sampling instant of time. It is evident from Fig. 3 that the NGD plots show the resonance peaks of the vocal tract system better compared to the DFT spectrum plots. The temporal resolution of the spectral features can be observed clearly from Figs. 3(b) and 3(c) due to zero-time windowing. Note that successive differencing of both the DFT spectrum and the NGD spectrum bring out the spectral features better as shown in Figs. 2(g) and 2(k), respectively. But the NGD plots are used to take advantage of the additive and high resolution property of the group delay function (Yegnanarayana, 1978).

To examine the details, the NGD plots within one glottal cycle are shown in Fig. 4(a). The figure shows that the NGD plot at one sampling instant seems to be large within each pitch period. This corresponds to the instant of glottal closure where a large excitation is imparted to the vocal tract system, and hence significant due to its high SNR properties. All other NGD plots at other sampling instants have small to very small values. Note that the windowed signal at epochs has the appearance of an impulse, in the sense that all sample values for n greater than a few samples are very small, due to the behaviour of the window function $w_1[n]$. On the other hand, the NGD plots at many instants seem to be significant in the unvoiced segment, as shown in Fig. 4(b).

3.2. Effect of window size

Fig. 5(a) (first column) shows the NGD plots for different values of M corresponding to 2, 3, 5 and 10 ms, all starting near the epoch. The plots show that the resonance

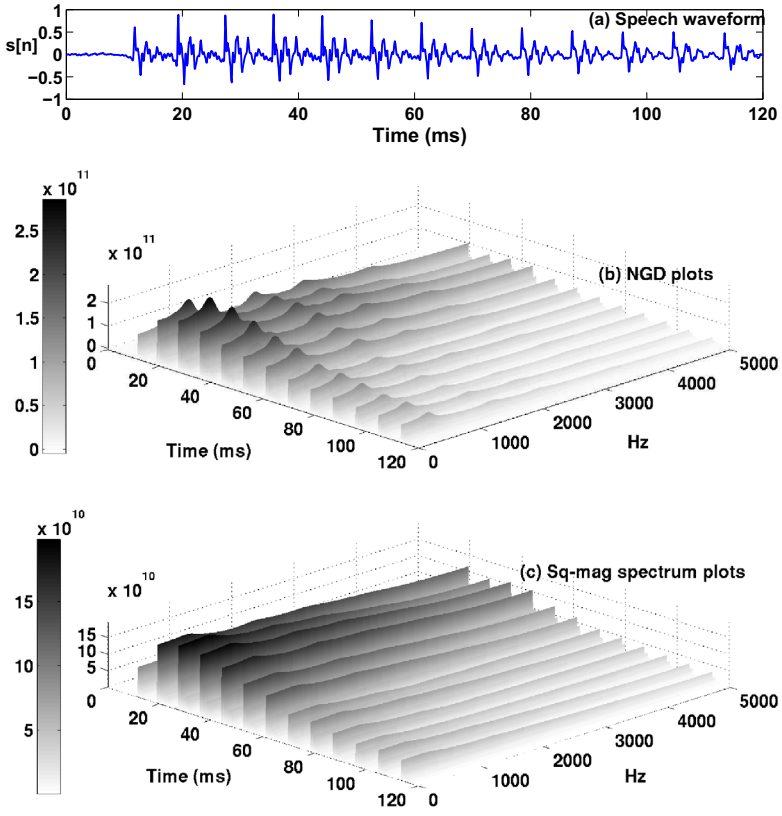

Fig. 3. Illustration of the effect of windowing on NGD and DFT spectra. (a) A segment (120 ms) of voiced speech waveform. 3-dimensional (instantaneous) spectral plots computed over 5 ms windowed segments for every sample shift using (b) NGD function, and (c) squared-magnitude DFT.

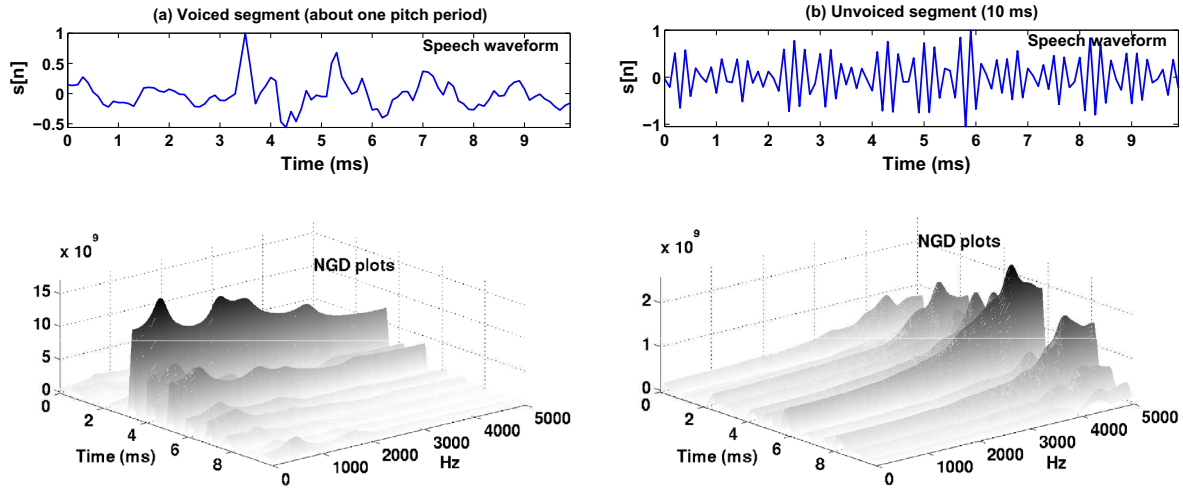

Fig. 4. Instantaneous NGD spectrum for (a) about one pitch period of voiced and (b) 10 ms of unvoiced segments of speech.

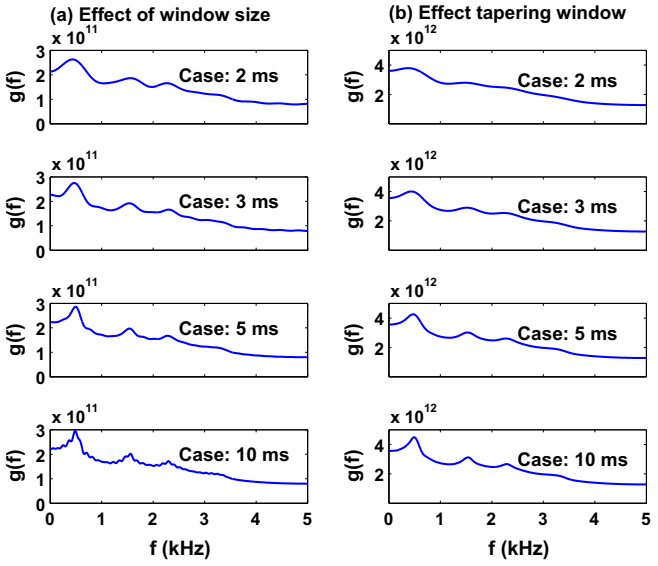

Fig. 5. (a) NGD plots computed for a segment of speech windowed using window sizes of 2, 3, 5 and 10 ms, respectively. (b) Effect of using a tapering window function $w_2[n]$ in Eq. (11).

peaks are obtained even for small window sizes. As expected, for smaller window sizes the spectral peaks are broader, i.e., the frequency resolution is poorer. But as the window size is increased, the ripple due to truncation can be seen in the NGD due to smaller period of the cycles in the frequency domain. Note that the ripple in the NGD for a segment greater than pitch period (Case: 10 ms) is large due to the effect of impulse-like excitation at the next epoch.

The effect of truncation at the end of the window can be reduced by using a tapering window function of the type

$$w_2[n] = 2\left(1+\cos\left(\frac{\pi n}{M}\right)\right) = 4\cos^2\left(\frac{\pi n}{2M}\right), \quad n = 0,1,\ldots,M-1, \tag{11}$$

instead of a rectangular window. This will reduce the ripple in the NGD plots as shown in Fig. 5(b) (second column). This window function performs an equivalent operation of summing the adjacent samples *twice* in the frequency domain. This can be shown as follows by considering the time domain operation of computing the sum of two samples (i.e., $(x[n]+x[n-1])$). The filter transfer function is $(1+z^{-1})$, whose frequency response is $|1+e^{-jw}| = 2\cos(\omega/2)$. In a manner similar to the arguments for Eq. (6), the equivalent time domain operation is multiplying with the window function given in Eq. (11). Other tapering window functions may also be used, but the effect of these window functions may not be as simple an interpretation as adding the adjacent two samples in the frequency domain.

3.3. Effect of applying multiple stages of ZTW

The above discussion shows that the NGD spectrum brings out the spectral features better than the magnitude spectrum, although they may be smeared due to windowing by $w_1[n]w_2[n]$. The spectral features can be highlighted by successively differencing the NGD. But successive differencing may highlight the ripple due to truncation. In order to reduce the ripple effect, several stages of ZTW can be used, followed by successive differencing. The effect of applying two stages of zero-time windowing, i.e., using a window function $w_1^2[n]w_2[n]$ on the NGD plots is shown in Fig. 6(a) for different sizes of window function. Note that the differences in the NGD plots for different window sizes are in the small fluctuations in the curves, which cannot be seen on the scale of the plots. But spectral features can be highlighted by successive differencing of the NGD spectrum. Fig. 6(b) shows the result of differencing the NGD spectrum twice. The plots are inverted (sign-reversed) so as to get the correct orientation of the formants after the double differencing. Note the difference in the vertical

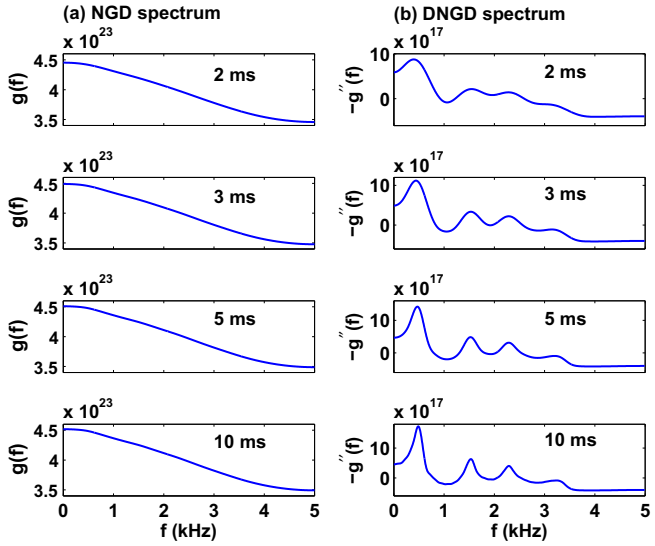

Fig. 6. Effect of using the ZTW function $w_1[n]$ *twice* and the tapering window $w_2[n]$ *once* for different window sizes. (a) NGD spectra. (b) Sign-reversed double-differenced NGD spectra (i.e., DNGD spectra).

scales for the plots in Figs. 6(a) and 6(b). The sign-reversed double-differenced NGD (referred to as DNGD) shown in Fig. 6(b) brings out the spectral details. As expected the spectral peaks are sharper for larger window lengths. The ripple effect is also reduced compared to Fig. 5.

3.4. Hilbert envelope of the DNGD

As seen from Fig. 4 in Section 3, the NGD plot around the epoch is strongest (high amplitude and large dynamic range) compared to the NGD plots in other regions. Interestingly, this is also true in the unvoiced regions, where the epochs are generally located at random instants, and are closer to each other, compared to the epochs in the voiced region. All instants corresponding to the positive zero-crossings in the zero-frequency filtered signal are considered as epochs. The strongest DNGD plots and their instants are identified by the peak values of the DNGD plots around each epoch. The strongest DNGD plots around each epoch are shown in Fig. 7(b).

In the computation of the DNGD spectrum orientation of some of the peaks, especially the higher order resonances, may have been affected due to several integration and differencing operations involved. One such example can be seen in the case of fourth formant in Fig. 7(b). Note that while the integration due to windowing is in the analog domain, the differencing is in the discrete frequency domain. Also, some of the spectral peaks (due to poles) in the NGD may be affected by the nearby spectral valleys (due to zeros). The peaks in the NGD due to large bandwidth formants are more likely to be affected by the nearby spectral valleys. This will make it difficult to identify the location of the peak in the NGD or the DNGD. In order to highlight the peaks of the spectral features, the DNGD is further processed by computing the Hilbert envelope

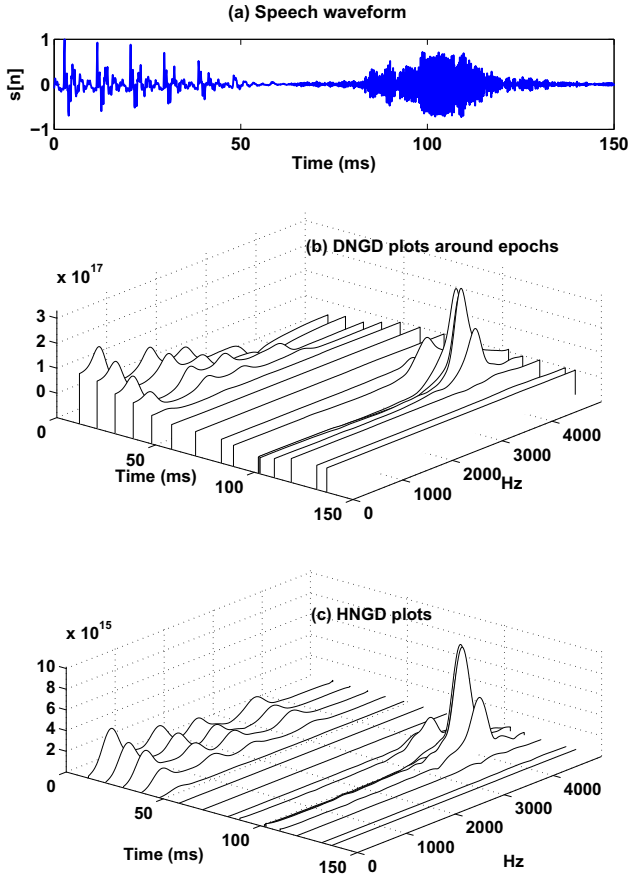

Fig. 7. Selection of DNGD plots around epoch. (a) Speech waveform. (b) DNGD plots selected around epoch locations. (c) HNGD plots selected around epoch locations.

(HE) of the DNGD. These are called HNGD plots. The Hilbert envelope $a[n]$ of a sequence $e[n]$ is obtained as

$$a[n] = \sqrt{e^2[n] + e_h^2[n]}, \tag{12}$$

where $e_h[n]$ is the Hilbert transform of the sequence $e[n]$, and is computed as follows:

$$e_h[n] = IDFT\{E_h(\omega)\}, \tag{13}$$

where

$$E_h(\omega) = \begin{cases} -jE(\omega), & 0 < \omega < \pi \\ jE(\omega), & -\pi < \omega < 0 \end{cases} \tag{14}$$

and $E(\omega)$ is the DTFT of the sequence $e[n]$. Note that for HNGD, we compute the Hilbert envelope of the DNGD sequence. Note also that the HE brings out the peak values of the spectrum from any (differenced or otherwise) spectrum, thus eliminating the uncertainties in the locations of the peaks. The HNGD plots for a segment of speech consisting of a voiced region and an unvoiced region are shown in Fig. 7(c). It can be seen that the HNGD plots eliminate the uncertainties in the location or orientation of the peaks in the DNGD plots, especially the higher resonances (see the fourth formant in Figs. 7(b) and 7(c) in the

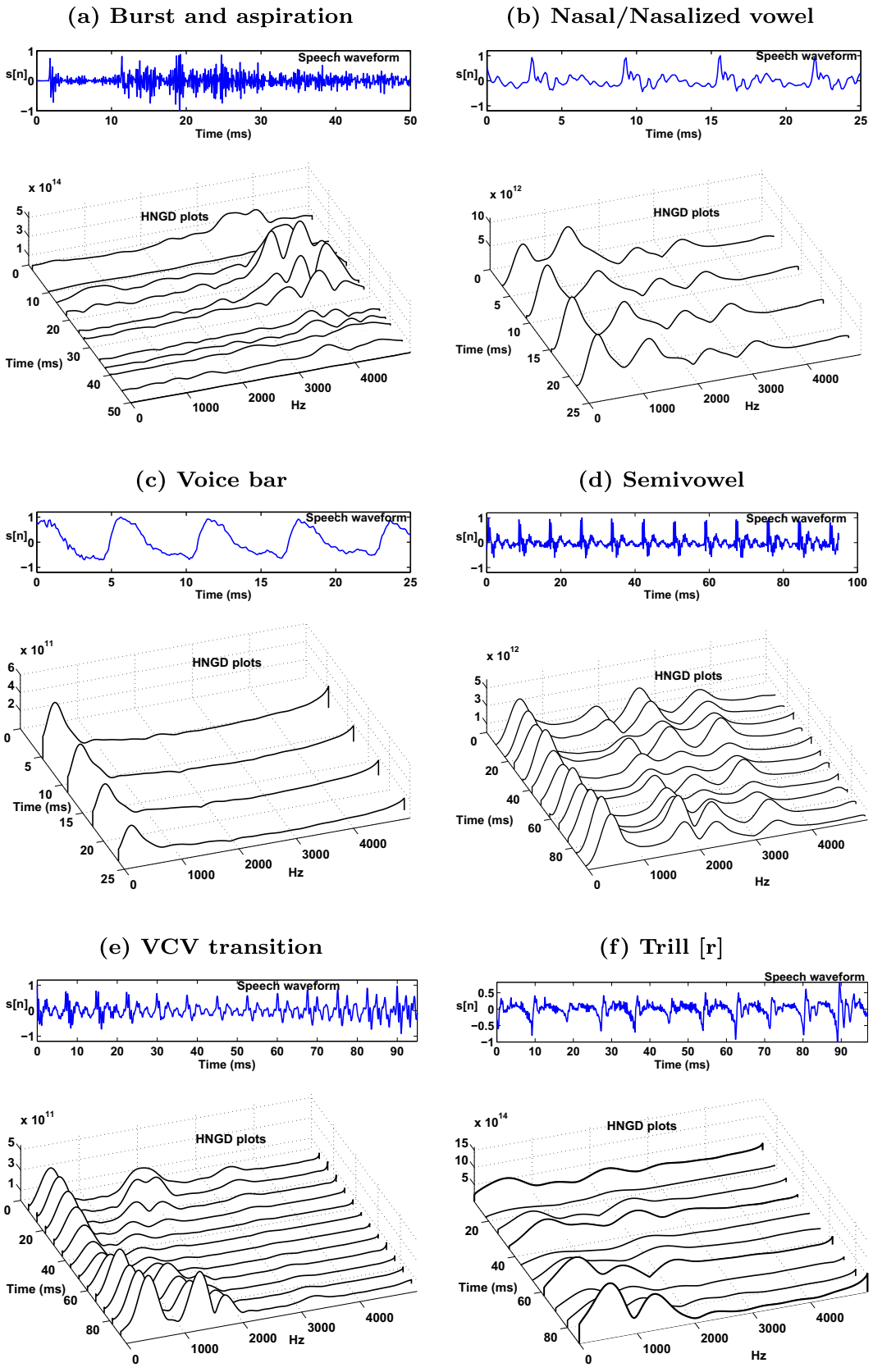

Fig. 8. HNGD plots for different categories of sounds. (a) Burst and aspiration ([t^h] as in 'toast'). (b) Nasal ([n] as in 'dont ask'). (c) Voice bar ([dcl] as in 'had'). (d) Semivowel ([y] as in 'all year'). (e) VCV transition ([ira] as in 'oily rag'). (f) Trill ([r] uttered in isolation).

voiced region). In the unvoiced region, the HNGD plots show peaks in the high frequency region, and these peaks exist at many more instants in the region (compared to voiced region) with somewhat random amplitudes. But the 3-D plot clearly shows the frication region, even better than the waveform, because of the temporal resolution of the spectral features.

4. Analysis of different categories of sounds

In this section we will examine the spectral features for different categories of sounds using the zero-time windowing of speech signals. Throughout this study, we use a 5 ms segment of data at each instant, and the data is multiplied with $w_1^2[n]w_2[n]$ as discussed in the previous section. The spectral features are displayed using the HNGD function of the windowed data. Throughout, 2048-point DFTs are used, by appending the windowed signal with zeroes. This will provide adequate number of samples for displaying the spectral features in the frequency domain.

4.1. Spectral characteristics of different categories of sounds

We discuss briefly the spectral characteristics through the HNGD plots for a few categories of sounds chosen for illustration.

4.1.1. Burst and aspiration

The spectral features for a short burst followed by an aspiration as in [th] is shown in Fig. 8(a). The burst region can easily be seen in the plot. The nonstationary nature of the aspiration region can also be seen clearly in this plot due to the temporal resolution of the spectral features provided by this method of analysis. Note that it is difficult to mark the different regions in the time waveform, as clearly as in the frequency domain.

4.1.2. Nasals and nasalized vowels

For nasal sounds there is a dominant first resonance in the low ($<$ 300 Hz) frequency region, which may mask the presence of other resonances. Fig. 8(b) shows the

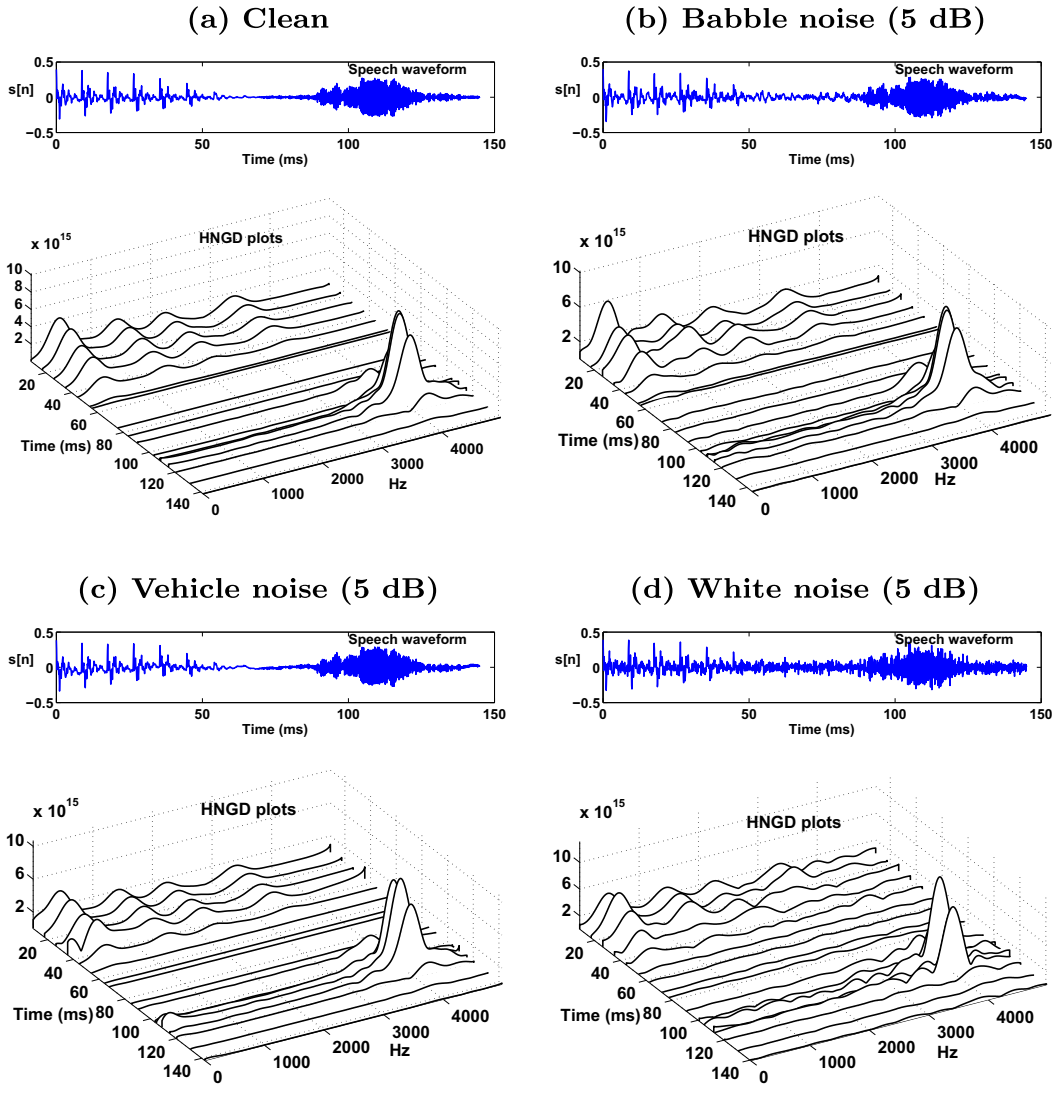

Fig. 9. HNGD plots for a segment of speech ([os] as in 'toast') for different types of additive noise at 5 dB SNR from NOISEX database.

HNGD plots for a few cycles of the nasal sound [n] uttered in the context of 'dont ask'. The low frequency nasal resonance is clearly seen, along with the other resonances at higher frequencies.

4.1.3. Voice bars

In the case of voice bar the spectral peak in the low frequency corresponds to glottal formant (Fig. 8(c)), which is much lower than the low frequency resonance of the nasal sound. Note that the peak is not due to the fundamental frequency, as the analysis window length is less than a pitch period.

4.1.4. Semivowels

The transition behaviour of the resonances of the vocal tract in semivowels can be seen clearly in the instantaneous spectral features of the HNGD as shown in Fig. 8(d).

4.1.5. Transitions

The instantaneous spectral features in the transition regions from vowel to consonant or vice versa can be captured well in the HNGD plots due to good temporal resolution of the spectral features obtained by the zero-time windowing method, as shown in Fig. 8(e). In fact the spectral characteristics in the transition regions for CV and VC sounds are useful to determine the nature of the consonant, even if there is no release of the burst as may be the case for some stop consonants in continuous speech.

4.1.6. Trill

The spectral characteristics of a trill [r] are distinct in the sense that the vocal tract shape is changing continuously due to the vibration of the tip of the tongue. The spectral features at successive epochs are varying as can be seen in Fig. 8(f) for a trill sound uttered in isolation. In the case of trills the vocal tract system and hence the spectral characteristics are changing continuously even within a glottal cycle. Also, the tapping of the tongue tip against the roof of the oral cavity may cause some secondary excitations. These secondary excitations can produce ripples in the estimated spectrum, if they are closer to the instants of glottal excitations. In order to display the spectral features better, only 3 ms segments are considered in this illustration. Note that the amplitude changes at successive epochs are also reflected in these plots.

4.2. Robustness of zero-time windowing approach

We examine the robustness of the zero-time windowing approach for extracting the instantaneous spectral features, when the speech signal is degraded by different types of degradations. We consider babble, vehicle and white noise from the NOISEX database (Varga and Steeneken, 1993) to generate speech signals at different signal-to-noise ratios (SNR). Fig. 9 shows the HNGD plots for clean and noisy data at 5 dB SNR for the three cases of noises. The noise data is added to the speech signal of the complete utterance to obtain the desired SNR. Hence the SNR in the displayed segment may be higher as the energy of the segment is higher relative to the average energy of the utterance. It is interesting to note that even at 5 dB SNR all the features of the voiced and unvoiced segments are preserved in the HNGD plots. Important spectral information is preserved in the voiced and fricative regions in the case of white noise. Some low amplitude formants are affected in the higher frequency region due to lower SNR in those regions.

5. Evaluation of zero-time windowing method

In this section the effectiveness of the zero-time windowing method for extracting the spectral features is discussed. We show that the spectral peaks in the HNGD plots indeed correspond to the formant locations, by applying the ZTW method on synthetic speech. Robustness of the ZTW method for different types of degradation is studied. A comparison with the short-time Fourier spectrum, LP spectrum, and the STRAIGHT spectrum is provided. Performance of the ZTW method in an application such as formant extraction from real speech is also studied.

5.1. Analysis of HNGD plots for synthetic speech

In order to verify that the spectral features obtained from the ZTW method indeed correspond to the formant contours, a synthetic speech signal is generated using the formant contours shown in Fig. 10(a) to represent the vocal tract system and a periodic sequence of impulses representing the excitation. Synthetic signals are generated for a voiced excitation (using impulses at 10 ms interval and 8 kHz sampling rate), as well as for a random noise excitation. The HNGD plots are computed at each impulse location using a 5 ms window. The top four peaks in the

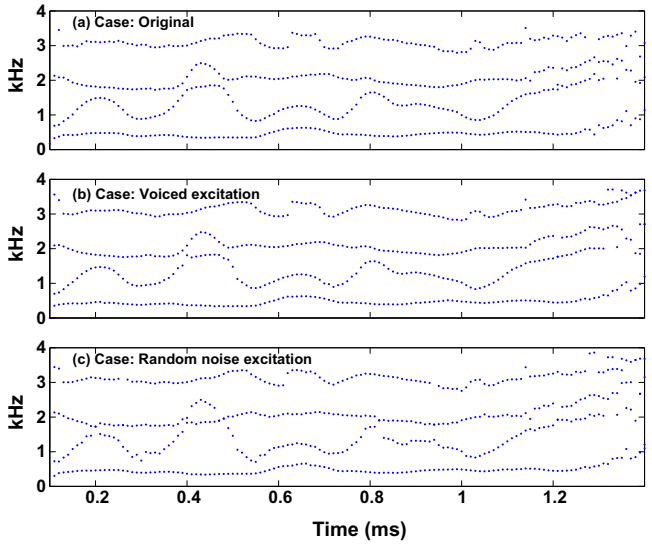

Fig. 10. Formant peaks detected from the HNGD plots for synthetic signals generated for (b) voiced excitation and (c) random noise excitation, using the reference formant contours shown in (a).

HNGD plots are used to plot the formant contours shown in Figs. 10(b) and 10(c) for voiced and random noise excitation cases, respectively. It can be seen that the formant contours indeed match the original formant contours in both the cases.

5.2. Evaluation of robustness of the HNGD plots

Robustness of the zero-time windowing method is evaluated using a 10th order (for a sampling rate of 8 kHz) LP spectrum as reference, as it highlights the spectral peaks better than the STFT spectrum. Fig. 11 shows the spectrograms computed using LP analysis and ZTW analysis for the clean case and for three different SNRs (10, 5 and 0 dB) of the additive noise degradation of the signal. The signal used is an utterance from a male speaker in the TIMIT corpus (Garofolo et al., 1993), downsampled from 16 kHz to 8 kHz before further processing. Both LP analysis and ZTW analysis are performed on the same segment of 5 ms considered around the GCIs. But for LP analysis a Hamming windowed data is used, whereas for ZTW analysis the window function $w_1^2[n]w_2[n]$ is used. It is clear that the formant information degrades faster in LP analysis compared to the ZTW method, as the SNR is reduced.

The spectral distance between clean and degraded signals is used for comparison. The spectral distance between a noisy spectrum $X_{noi}[k]$ and a clean (reference) spectrum $X_{ref}[k]$ is computed as

$$d_{i,j} = \frac{1}{(N/2)+1} \sum_{k=0}^{k=(N/2)+1} \frac{(X_{ref}[k] - X_{noi}[k])^2}{X_{ref}^2[k]}, \tag{15}$$

where $\{i, j\}$ denotes ith utterance and jth epoch location, N is the FFT length. The average spectral distortion over voiced epochs of all utterances is computed as

$$d = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{1}{N_i} \sum_{j=1}^{N_i} (d_{i,j})^{1/2}, \tag{16}$$

where N_u is the number of utterances and N_i is the number of voiced epochs in the ith utterance. The average spectral distortion is computed for 10 male and 10 female utterances chosen arbitrarily from the TIMIT corpus (Garofolo et al., 1993). Fig. 12 shows comparison of ZTW, LP, STFT and STRAIGHT methods using spectral distortion as a measure of robustness. The spectral distortion with respect to clean case is clearly lower for the HNGD spectra, especially at low SNRs, compared to the other spectrum estimation techniques. The spectral distortion is the highest

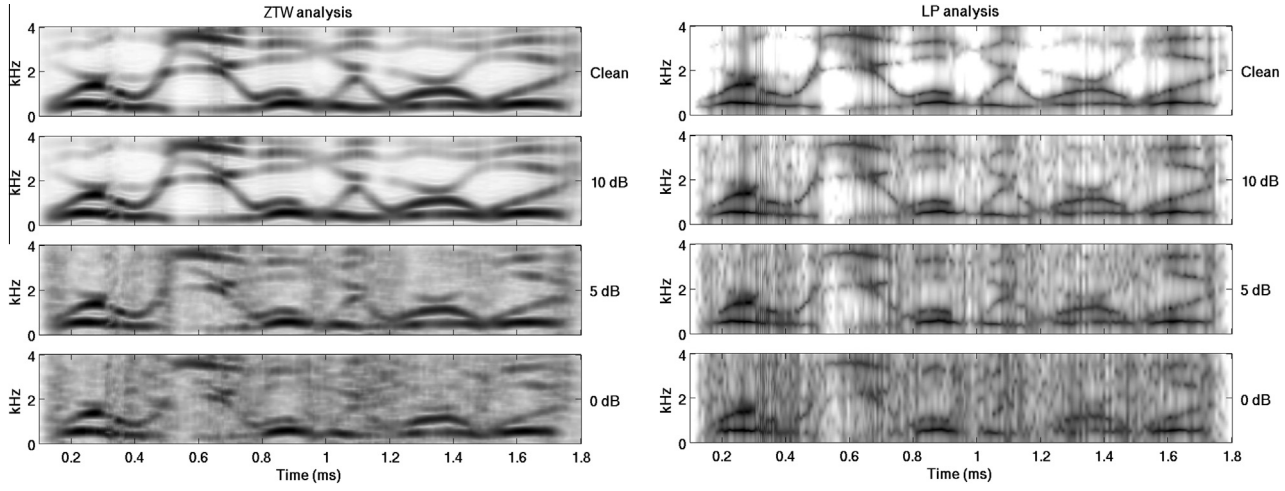

Fig. 11. Spectrogram display for an utterance ('Where were you while we were away?') by a male speaker for additive white noise at different SNRs.

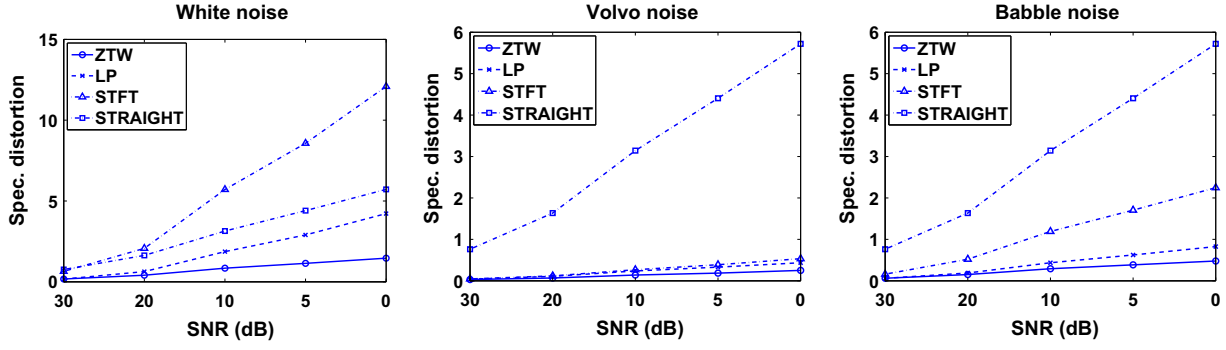

Fig. 12. Comparison of average spectral distortion of noisy spectra with respect to clean spectra obtained from ZTW, LP and STFT methods, for three different types of noise.

Table 1
Performance of formant extraction using different methods. GDR denotes gross detection rate, and MAD denotes mean absolute deviation.

Method	Performance metric	F_1	F_2	F_3
ZTW	GDR (in %)	70.7	89.2	85.8
	MAD (in Hz)	44	92	98
ESPS	GDR (in %)	74.7	80.2	79.9
	MAD (in Hz)	46	105	103
STRAIGHT	GDR (in %)	83.8	95.1	73.1
	MAD (in Hz)	45	71	81

for STRAIGHT spectrum, especially for volvo and babble noise, as the method relies on accurate estimation of the pitch frequency.

5.3. Formant extraction using HNGD plots

The usefulness of the ZTW method in extraction of formants is studied in this section. Peaks in the HNGD spectrum are picked using a differenced Gaussian window of width 100 Hz. Frequencies corresponding to the four largest peaks are taken as the formant frequencies. The accuracies of the formants extracted are tested on the VTR database (Deng et al., 2006). The VTR database has manually corrected first three formant tracks. The evaluation is done on the test set containing a total of 192 sentences uttered by 16 male and 8 female speakers (8 utterances per speaker). This is the core test subset of the TIMIT corpus (Garofolo et al., 1993). The utterances are downsampled from a sampling rate of 16 kHz to 8 kHz before further processing. The accuracies of the first three formants extracted are measured in terms of the mean absolute deviation (MAD) of the detected formant frequency from the reference frequency. The average deviation is computed only for formant values which are within 20% deviation from the reference value or 300 Hz absolute deviation, whichever is smaller. The number of formant values detected within the specified constraint is measured as the gross detection rate (GDR). The average deviation (in Hz) and the gross detection rate (in %) are given in Table 1. The performance of formant extraction is evaluated only for vowel, semivowel and diphthong regions. As a comparison, performance of the formant contours extracted using the popular ESPS package (Sjolander and Beskow, 2000) and the STRAIGHT method (Kawahara et al., 2008) is also given in Table 1. It can be seen that the ZTW method gives overall performance comparable to ESPS method, with better detection accuracy for F_2 and F_3, and somewhat lower detection accuracy for F_1. However, the deviations for all the three detected formants are lower for the ZTW method compared to the ESPS method. The STRAIGHT method performs better than either of the ZTW or LP based methods in detecting the first two formants. It should be noted that the STRAIGHT spectrum is estimated using the version of the code STRAIGHTv40 available at (Kawahara et al., 2008), which uses a default window size of 40 ms. It should also be noted that the ground truth provided with the VTR database are first estimated using an algorithm which computes LP cepstra over the typical short-time analysis window size of around 25 ms, which are later smoothed and corrected manually (Deng et al., 2006). Thus the STRAIGHT method and the ground truth both use averaged formant values estimated over several glottal cycles. On the other hand, the ZTW method gives the formant values corresponding to the vocal tract shape near the GCIs, mostly the closed phase of the glottal cycle. This may be the main reason for lower GDR in the case of ZTW method for first two formants, as these formants, especially the first formant, are due to the effective length of the vocal tract, which is different during the closed and open phase of the glottal cycle. It is interesting to note that the ZTW method performs better than the STRAIGHT method in detecting the third formant. This can be attributed primarily to the additive and high resolution properties of the group delay function.

6. Summary and conclusions

This paper presented a new approach for speech analysis. The approach is based on the concept of zero-time windowing, a term used analogous to zero-frequency filtering. The zero-time windowing of speech signals in time domain is equivalent to smoothing the spectral information by successive integration of the samples in the frequency domain. The features in the spectrum such as formants can be highlighted by removing the trend using differencing. Note that the zero-time windowing operation in the time domain is equivalent to integration with the analog frequency variable, whereas the differencing in the frequency domain to remove the trend is a digital operation.

The high resolution properties of the numerator of the group delay can be used to display the spectral features such as formant locations (in voiced regions) and high energy spectral bands (in unvoiced regions). It is interesting to note that the spectral features can be seen even when the speech segment is less than 5 ms. The resulting spectral information is model-free. These spectral features can be derived at every sampling instant, and hence the zero-time windowing method provides instantaneous spectrum analysis.

Most of the time the spectral information is strongest (in terms of amplitude and dynamic range) for segments near

the instants of significant excitation, which correspond to epochs or GCIs in the voiced segments, and to random instants in the unvoiced regions. Thus one can select the spectral features only around the instants of significant excitation of the vocal tract system. The spectral features of weaker sounds like nasals or voice bars can also be displayed well by the proposed analysis method.

The high temporal resolution provided by this analysis makes it possible to extract the spectral features even from degraded speech. The usefulness of the ZTW method needs to be tested for more practical degradations caused by reverberation, multispeaker data and coded data. The high resolution properties of the zero-time windowing method may help in providing some direction to deal with speech signals in such practical environments.

In the proposed zero-time windowing approach, we can select the region around the epoch, and then average features in those selected regions for feature extraction or enhancement. Thus we can say that with the proposed approach we select the relevant information first, and then average the selected information, if necessary. On the other hand, in the conventional methods of spectrum analysis, we average the information over a segment (like in DFT or LP analysis), and then select the features from the result. It appears that this is a fundamental difference in these approaches, and hence may help in augmenting the information obtained by the current methods of speech analysis.

Acknowledgements

The authors would like to thank the Department of Information Technology, Government of India for supporting this activity through sponsored research projects. The second author would also like to thank The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170) for supporting his stay in Finland as a Postdoctoral Researcher.

References

Abe, T., Honda, M., 2006. Sinusoidal model based on instantaneous frequency attractors. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1292–1300.

Deller Jr., J.R., Hansen, J.H.L., Proakis, J.G., 2000, second ed. In: Discrete Time Processing of Speech Signals Wiley IEEE Press, New York, USA.

Deng, L., Acero, A., Bazzi, I., 2006. Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. IEEE Trans. Audio Speech Lang. Process. 14 (2), 425–434. http://dx.doi.org/10.1109/TSA.2005.855841.

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., 2006. A database of vocal tract resonance trajectories for research in speech processing. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing, Toulouse, France, pp. I-369–I-372.

Garofolo, J.S. et al., 1993. In: TIMIT Acoustic–Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia, USA.

Gianfelici, F., Biagetti, G., Crippa, P., Turchetti, C., 2007. Multicomponent AM & FM representations: an asymptotically exact approach. IEEE Trans. Audio Speech Lang. Process. 15 (3), 823–837.

Joseph, M.A., Guruprasad, S., Yegnanarayana, B., 2006. Extracting formants from short segments of speech using group delay functions. In: Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH), Pittsburgh PA, USA, pp. 1009–1012.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27 (3–4), 187–207.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. Tandem-straight: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing, Las Vegas, USA, pp. 3933–3936. <http://www.wakayama-u.ac.jp/kawahara/STRAIGHTtrial/>.

Makhoul, J., 1975. Linear prediction: a tutorial review. In: Proceedings of the IEEE, vol. 63, pp. 561–580.

Murty, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1602–1613.

Oppenheim, A.V., Schafer, R.W., 1975. In: Digital Signal Processing. Prentice Hall, Englewood Cliffs, New Jersey, USA, pp. 1–585.

Rabiner, L., Schafer, R., 2010. In: Theory and Applications of Digital Speech Processing. Prentice Hall, USA.

Santhanam, B., Maragos, P., 2000. Multicomponent AM–FM demodulation via periodicity-based algebraic separation and energy-based demodulation. IEEE Trans. Commun. 48 (3), 473–490. http://dx.doi.org/10.1109/26.837050.

Sjolander, K., Beskow, J., 2000. Wavesurfer – An open source speech tool. In: Proceedings of the International Conference on Spoken Language Processing, Beijing, China, pp. 464–467.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251 <http://www.speech.cs.cmu.edu/comp.-speech/Section1/Data/noisex.html>.

Vargas, J., McLaughlin, S., 2008. Cascade prediction filters with adaptive zeros to track the time-varying resonances of the vocal tract. IEEE Trans. Audio Speech Lang. Process. 16 (1), 1–7. http://dx.doi.org/10.1109/TASL.2007.907573.

Welling, L., Ney, H., 1998. Formant estimation for speech recognition. IEEE Trans. Speech Audio Process. 6 (1), 36–48.

Yegnanarayana, B., 1978. Formant extraction from linear prediction phase spectra. J. Acoust. Soc. Am. 63 (5), 1638–1640.

Yegnanarayana, B., Murthy, H.A., 1992. Significance of group delay functions in spectrum estimation. IEEE Trans. Signal Process. 40 (9), 2281–2289.

Yegnanarayana, B., Murty, K.S.R., 2009. Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Process. 17 (4), 614–624. http://dx.doi.org/10.1109/TASL.2008.2012194.

Yegnanarayana, B., Veldhuis, R., 1998. Extraction of vocal-tract system characteristics from speech signals. IEEE Trans. Speech Audio Process. 6 (4), 313–327.