Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Speech Communication 42 (2004) 429–446

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic segmentation of continuous speech using minimum phase group delay functions

V. Kamakshi Prasad, T. Nagarajan *, Hema A. Murthy

Department of Computer Science and Engineering, Indian Institute of Technology, Madras, IIT Campus, Chennai, Tamil Nadu 600036, India

Received 9 October 2003

Abstract

In this paper, we present a new algorithm to automatically segment a continuous speech signal into syllable-like segments. The algorithm for segmentation is based on processing the short-term energy function of the continuous speech signal. The short-term energy function is a positive function and can therefore be processed in a manner similar to that of the magnitude spectrum. In this paper, we employ an algorithm, based on group delay processing of the magnitude spectrum to determine segment boundaries in the speech signal. The experiments have been carried out on TIMIT and TIDIGITS databases. The error in segment boundary is $\leqslant 20\%$ of syllable duration for 70% of the syllables. In addition to true segments, an overall 5% insertions and deletions have also been observed.

Keywords: Minimum phase group delay functions; Root cepstrum; Speech segmentation

1. Introduction

Segmenting the continuous speech signal according to the phonetic transcription is a fundamental task in any voice activated system. Manual segmentation is tedious, time consuming and error prone. Further, it is almost impossible to reproduce the manual segmentation results due to the variability in human visual and acoustic perception. It is also difficult to arrive at a common labeling strategy across different researchers. Automatic segmentation is not faultless, but it is inherently consistent and results are reproducible. Ideally, one would like to have an automatic segmentation and labeling system which is capable of handling language and speaker independent speech.

In general, there are two broad categories of speech segmentation algorithms. One class of algorithms perform the segmentation when the underlying sequence of phonemes is assumed known (Rabiner et al., 1982). Another class of algorithms use no knowledge of the underlying phoneme sequence contained within the speech waveform, instead the segment boundaries are identified at time instants, where there is a high degree of change in the acoustic properties of the waveform (Wilpon et al., 1987). There is yet another class of procedures which combine explicit information about the speech with frame to frame spectral change (van Hemert, 1991).

* Corresponding author. Tel.: +91-44-2257-8342; fax: +91-44-2257-8352.

E-mail addresses: raju@lantana.iitm.ernet.in (T. Nagarajan), hema@lantana.tenet.res.in (H.A. Murthy).

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2003.12.002

Nomenclature

τ	group delay function
$r_{xx}(l)$	autocorrelation of the signal in time domain
$R_{xx}(z)$	z-transform of autocorrelated signal $r_{xx}(l)$
$x_{\mathrm{nm}}(n)$	non-minimum phase signal
$x_{\mathrm{mp}}(n)$	minimum phase correspondent signal of the signal $x_{\mathrm{nm}}(n)$

The proposed approach for segmenting the speech signal is based on processing the short-term energy function of the speech signal. This approach only uses the information about the approximate number of voiced segments present in the given utterance. No information about the phonetic content of the speech signal is used. Such algorithms are well suited for tasks such as language independent segmentation of multilingual speech. The motivation for this is that, whatever the target language, the sentences in a language are made up of a sequence of linguistic units which correspond to one or more sequences of acoustic units, namely, phoneme, syllable, word and sentence.

The co-articulation effects present at the phoneme level, make segmentation at phoneme boundaries an impossible task. Further, large portions of phonemes either change their identity or are altogether missing in action (Greenberg, 1999). Hence, finding a direct correspondence between a speech segment and a phoneme is a difficult task. Therefore a higher level of linguistic organization, namely, syllable, is a better linguistic unit for segmentation. Syllable seems to be an intuitive unit for representation as the variation observed is more systematic at the level of the syllable than at the level of the phoneme (Greenberg, 1999). The significance of syllable units for improving performance of continuous speech recognition systems is demonstrated in (Ganapathiraju et al., 2001).

In automatic segmentation of speech, there are two issues to be addressed namely, the presence of background noise and local energy variations. Frequency domain approaches may not be suitable for handling noisy speech signals as the frequency components caused by noise affect the entire spectrum and corrupt the spectral envelope of the original speech signal. For segmenting the speech signal at syllable boundaries, time domain approaches such as energy based methods are good. Because, the segment structure is preserved in the short-term energy, in spite of noise. One time domain approach for segmenting the speech signal at syllable like units uses the loudness function. This is computed by weighting the short time power spectrum (Mermelstein, 1975). The difference between the convex hull and loudness is computed and the point of maximal difference between the loudness function and the convex hull is identified as a potential syllable boundary. Other approaches include measurement of peak to peak amplitude and root mean square intensity (Sargent et al., 1974).

The high energy regions in the short-term energy function correspond to syllable centres. The short-term energy function cannot be used directly to perform segmentation due to significant local variations that could often result in misidentified boundaries. Techniques like fixed thresholding can be used but when energy variations across the signal is high, they suffer. For continuous speech, energy is generally high at the beginning of a sentence and tapers off towards the end of a sentence. An adaptive threshold can be used to address this problem but the value of the threshold used will have to be learnt continuously from the speech signal. Further, the region over which the adaptive threshold is computed will become crucial: too large a region will miss boundaries, while too short a region will generate spurious boundaries. Fig. 1(a) shows a speech signal corresponding to the digit string '77'. Solid lines indicate manually segmented boundaries. Fig. 1(b) and (c) demonstrate the use of an adaptive threshold to

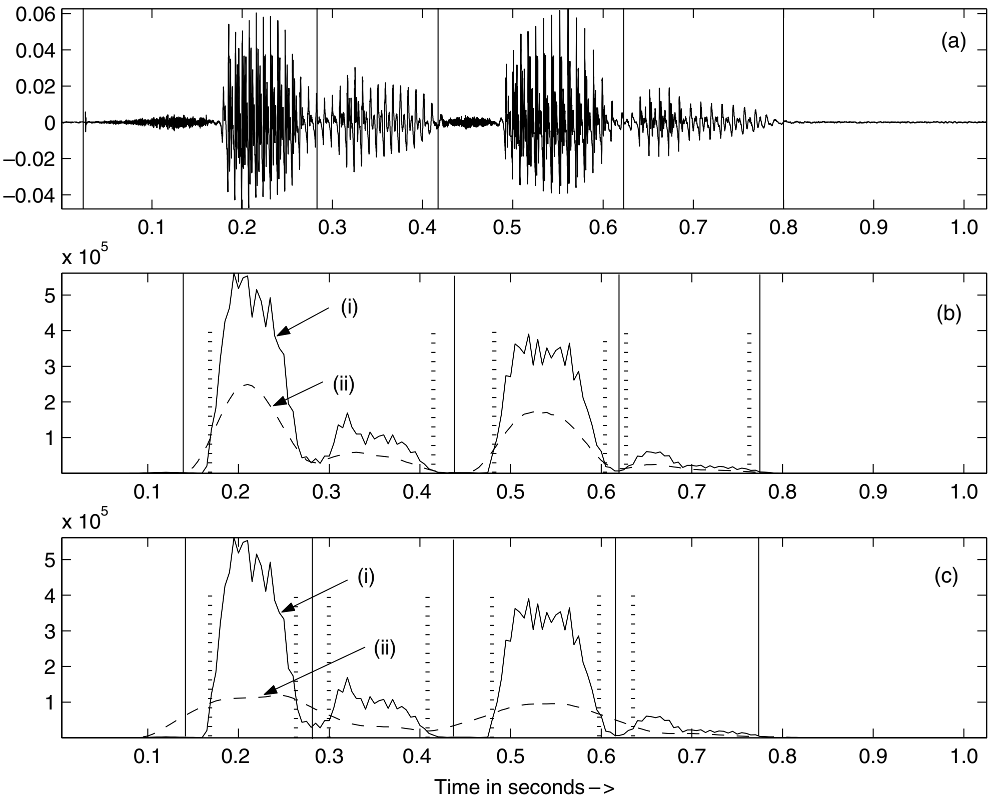

Fig. 1. Segmentation using adaptive thresholding technique: (a) Speech signal for the utterance of digit string 77. (b,c) Illustration of adaptive thresholding (dotted curve (ii)) on short-term energy function (solid curve (i)) with mean-smoothing order 25 and 50, respectively.

segment the speech signal. The threshold is applied on the short-term energy function. Two threshold functions are computed using the average energy over two different window lengths on the energy function: 25 and 50 samples. The points of intersection of the threshold function and the energy contour are denoted by short vertical lines. Energy minima between two consecutive short vertical lines are assumed to be segment boundaries. Observe that the boundary at 0.3 s is missed in Fig. 1(b), while it is detected in Fig. 1(c). Clearly, the choice of region size over which the adaptive threshold is computed, affects the performance of the system.

It has been well established that minimum phase group delay functions are very successful in formant/anti-formant extraction (Hema A. Murthy and Yegnanarayana, 1991) and spectrum estimation (Yegnanarayana and Hema A. Murthy, 1992). In this work, we propose an algorithm for processing the short-term energy function using the group delay approach to spectral smoothing. In the proposed technique, we process the short-term energy function as if it were a magnitude spectrum. In the context of segmentation, the valleys in the energy function approximately correspond to syllable boundaries. The group delay spectrum resolves the peaks and valleys properly, only when it is derived from a minimum phase signal (Nagarajan et al., 2001). Therefore it is necessary to derive a minimum phase signal corresponding to that of the short-term energy function.

In Section 2, we review some of the properties of the minimum phase group delay function. In Section 3, we detail the root cepstrum based minimum phase group delay algorithm for segmenting continuous speech. In Section 4, we evaluate the segmentation performance of the proposed algorithm on two different speech databases namely, the TIMIT (Fisher et al., 1986) and TIDIGITS (Leonard, 1984).

2. Properties of the minimum phase group delay function

It has been empirically shown that the causal portion of the inverse Fourier transform of the magnitude spectrum of the speech signal behaves like a minimum phase signal (Hema A. Murthy, 1992). It has also been well established that the group delay function of the of the minimum phase signal can be used for spectrum estimation (Yegnanarayana and Hema A. Murthy, 1992).

The theory of minimum phase signals has been developed extensively in the past (Berkhout, 1973, 1974). In particular, the properties of the minimum phase and zero phase time functions, have received considerable attention (Berkhout, 1973). In this section, we review the properties of the minimum phase group delay function.

2.1. Minimum phase signal

In terms of poles and zeroes, $x(n)$ is a minimum phase signal if and only if all the poles and zeroes of the z-transform of $x(n)$ (denoted as $X(z)$) lie within the unit circle. Symbolically,

$$X(z) = \frac{b_0 \cdot \prod_{i=1}^{m}(1 - b_i z^{-1})}{a_0 \cdot \prod_{i=1}^{n}(1 - a_i z^{-1})}, \tag{1}$$

where,

$$\forall i \quad [(b_i < 1) \ \wedge \ (a_i < 1)]$$

and $X(z) \cdot X^{-1}(z) = 1$.

From the roots of any energy bounded non-minimum phase signal, a minimum phase equivalent signal can be derived by replacing the roots, which are outside the unit circle, at their reciprocal locations. Although, there are efficient methods available to estimate the roots, these methods are model based. Any model-based estimator of roots requires *a priori* knowledge of the number of roots.

We present a non-model, root cepstrum based approach, to derive a minimum phase signal $x_{\mathrm{mp}}(n)$ from any signal $x(n)$ under the constraint that it is derived from the magnitude spectrum of $x(n)$, i.e., $|X(\mathrm{e}^{\mathrm{j}\omega})|$. The reason for this constraint is that the magnitude spectrum of a given root inside the unit circle (at a radial distance 'α' from the origin of the unit circle) is the same as that of a root outside the unit circle (at a distance '$1/\alpha$' at the same angular frequency). In general, if a system function has 'N' roots, then there are 2^N possible pole/zero configurations that will yield the same magnitude spectrum. Therefore, it is not possible to determine whether a given signal is minimum phase or non-minimum phase from the magnitude spectrum alone.

2.2. Properties of the group delay function

The negative derivative of the Fourier transform phase is defined as *group delay*. The group delay function exhibits an additive property. Let

$$H(\mathrm{e}^{\mathrm{j}\omega}) = H_1(\mathrm{e}^{\mathrm{j}\omega}) \cdot H_2(\mathrm{e}^{\mathrm{j}\omega}) \tag{2}$$

and,

$$|H(\mathrm{e}^{\mathrm{j}\omega})| = |H_1(\mathrm{e}^{\mathrm{j}\omega})| \cdot |H_2(\mathrm{e}^{\mathrm{j}\omega})|, \tag{3}$$

$$\arg(H(\mathrm{e}^{\mathrm{j}\omega})) = \arg(H_1(\mathrm{e}^{\mathrm{j}\omega})) + \arg(H_2(\mathrm{e}^{\mathrm{j}\omega})). \tag{4}$$

Then the group delay function, which is defined as the negative derivative of phase is given by

$$\begin{aligned} \tau_h(\mathrm{e}^{\mathrm{j}\omega}) &= -\frac{\partial(\arg(H(\mathrm{e}^{\mathrm{j}\omega})))}{\partial\omega} \\ &= -\frac{\partial(\arg(H_1(\mathrm{e}^{\mathrm{j}\omega})))}{\partial\omega} - \frac{\partial(\arg(H_2(\mathrm{e}^{\mathrm{j}\omega})))}{\partial\omega}, \end{aligned}$$

$$\tau_h(\mathrm{e}^{\mathrm{j}\omega}) = \tau_{h_1}(\mathrm{e}^{\mathrm{j}\omega}) + \tau_{h_2}(\mathrm{e}^{\mathrm{j}\omega}), \tag{5}$$

where, $\tau_{h_1}(\mathrm{e}^{\mathrm{j}\omega})$ and $\tau_{h_2}(\mathrm{e}^{\mathrm{j}\omega})$ correspond to the group delay function of $H_1(\mathrm{e}^{\mathrm{j}\omega})$ and $H_2(\mathrm{e}^{\mathrm{j}\omega})$, respectively.

From Eqs. (2) and (5), we see that multiplication in the spectral domain becomes an addition in the group delay domain. To demonstrate the power of the additive property of the group delay spectrum, three different systems are chosen, (i) a complex conjugate pole pair at an angular frequency ω_1, (ii) a complex conjugate pole pair at an angular frequency ω_2 and (iii) two complex conjugate pole pairs one at ω_1, and, the other at ω_2. From the magnitude spectra of these three systems (Fig. 2(b), (e) and (h)), it is observed that even though the peaks in Fig. 2(b) and (e) are resolved well, in a system consisting of these two poles, the peaks are not resolved well (see Fig. 2(h)). This is due to the multiplicative property of magnitude spectra. From Fig. 2(c), (f) and (i), it is evident that in the group delay spectrum obtained by combining the poles together, the peaks are well resolved as shown Fig. 2(i).

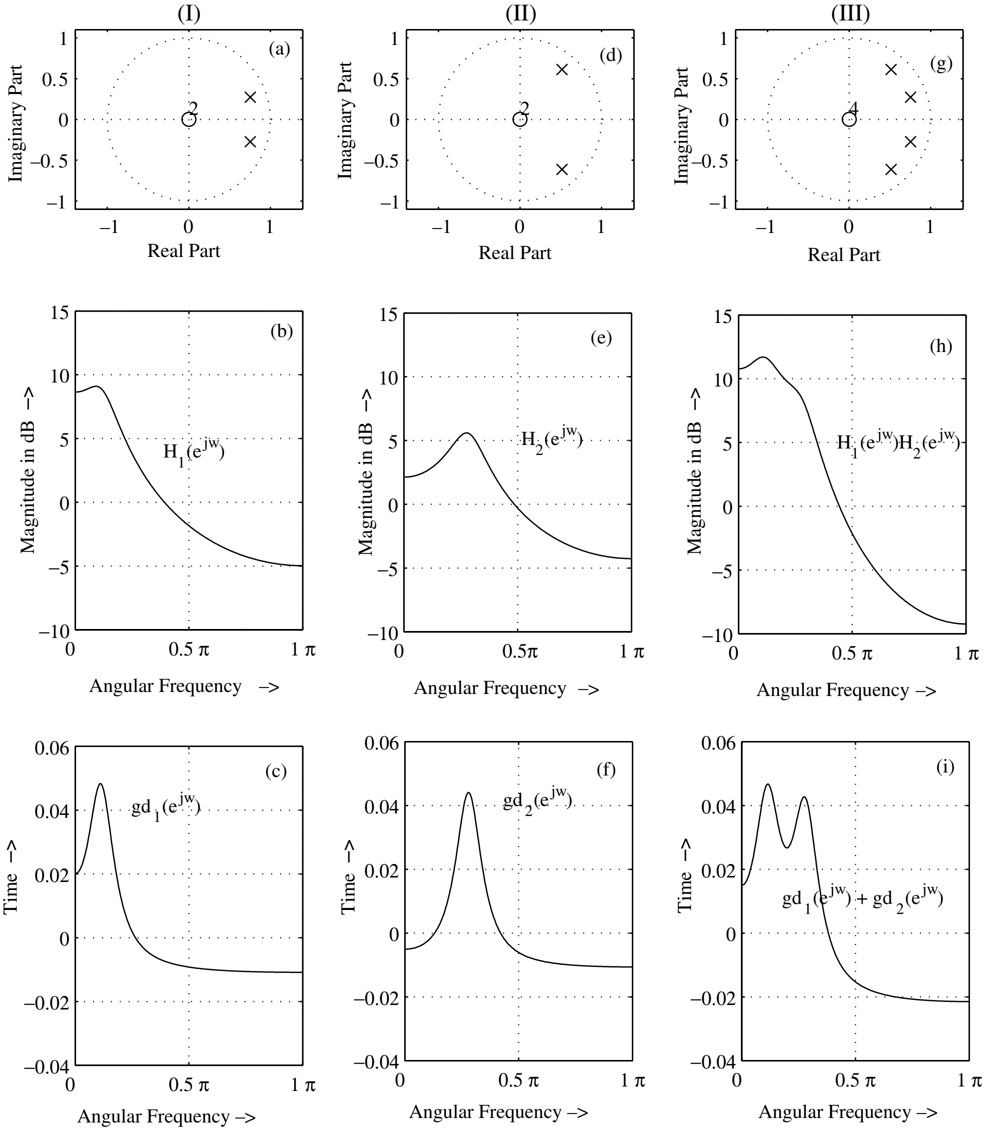

Fig. 2. Resolving power of group delay spectrum: z-plane, magnitude spectrum and group delay spectrum (I) a pole inside the unit circle at $(0.8, \pi/8)$, (II) a pole inside the unit circle at $(0.8, \pi/4)$ and (III) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$, inside the unit circle.

2.3. Properties of minimum phase group delay function

The group delay function derived from the minimum phase signal is called a *minimum phase group delay function*. In the minimum phase group delay function, poles and zeroes can be distinguished easily; peaks correspond to poles while valleys correspond to zeroes. Non-minimum phase signals do not possess this property. This is illustrated with an example in Fig. 3. For analysis, we have chosen the roots of minimum phase and non-minimum phase signals in Fig. 3, such that the magnitude spectrum of all the three different signals are identical. Further, the signals are all chosen to be real and stable and the roots come in

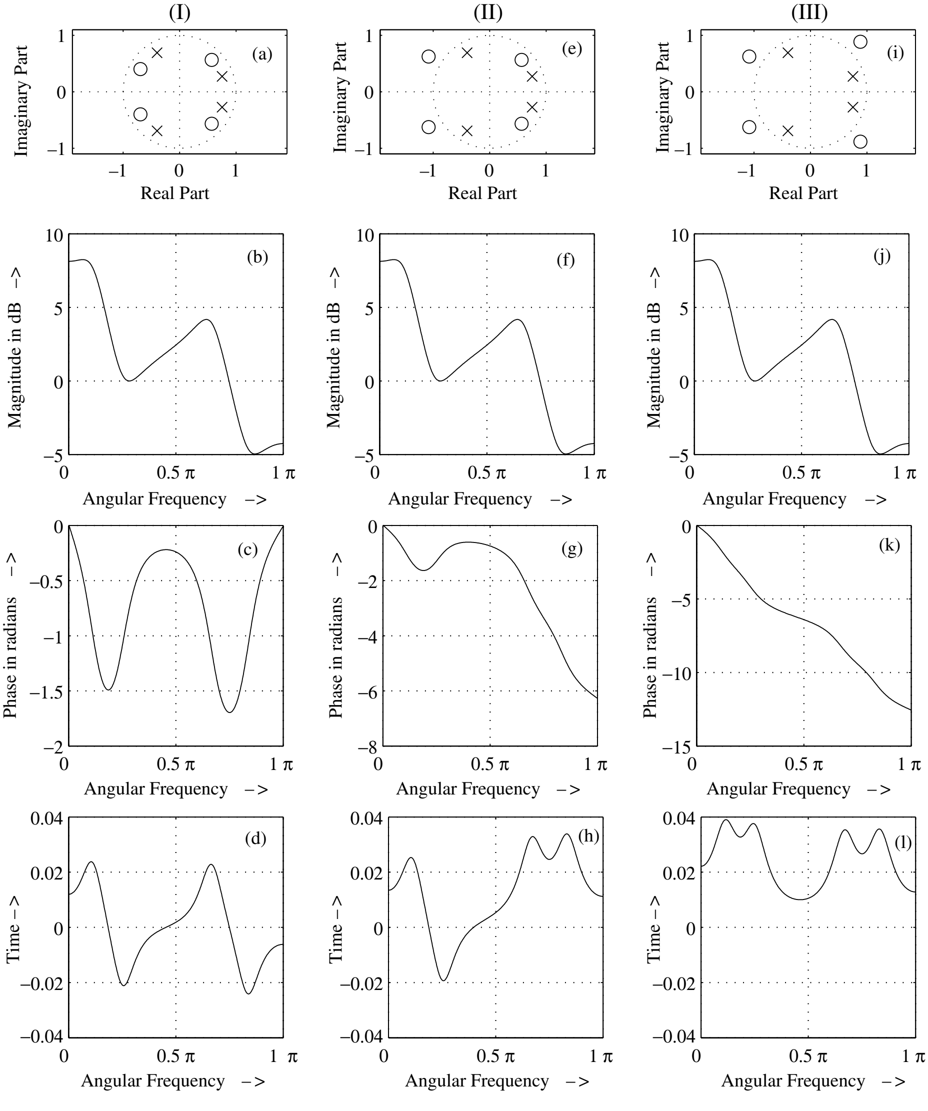

Fig. 3. Group delay property of different types of signals: the *z*-plane, the magnitude spectrum, the phase spectrum, and the group delay spectrum for (I) minimum phase, (II) non-minimum phase—type (1) and (III) non-minimum phase—type (2) systems.

complex conjugate pairs. The corresponding system function $H(z)$ is given by,

$$H(z) = \frac{(z - b_1) \cdot (z - b_1^*) \cdot (z + b_2) \cdot (z + b_2^*)}{(z - a_1) \cdot (z - a_1^*) \cdot (z + a_2) \cdot (z + a_2^*)}, \qquad (6)$$

where $|a_i| < 1$ for $i = 1, 2$ for all types of signals; $|b_i| < 1$ for $i = 1, 2$ for minimum phase signal; $|b_1| < 1$ and $|b_2| > 1$ for type (1) signal; $|b_i| > 1$ for $i = 1, 2$ for type (2) signal.

For the system function given in Eq. (6), magnitude, phase, and, group delay spectra are computed (see Fig. 3). From Fig. 3, we observe that

(a) For all three types of systems, the magnitude spectra are identical in shape (Fig. 3(b), (f) and (j)).

(b) For the minimum phase system (Fig. 3(a)), the net phase change from 0 to π radians, $(\arg(H(\pi)) - \arg(H(0)))$ is negligible (Fig. 3(c)). For non-minimum phase systems (Fig. 3(e) and (i)), the net phase change is proportional to the number of zeroes outside the unit circle (Fig. 3(g) and (k)).

In summary, for minimum phase system, the net phase change is negligible, while for type (2) system, the net phase change is significant and greater than that of the type (1) system (Fig. 3(c), (g) and (k)).

(c) In the group delay spectrum, for the minimum phase system, both the peaks and valleys are resolved correctly (Fig. 3(d)), where peaks correspond to poles and valleys correspond to zeroes. In the case of non-minimum phase systems, the zeroes which are outside the unit circle are not resolved properly as shown in Fig. 3(h) and (l). The zeroes outside the unit circle, instead of showing up as valleys, appear as peaks at the corresponding angular frequencies. It is therefore, difficult to distinguish between poles and zeroes (when the zeroes are outside the unit circle) in the group delay spectrum.

From the above example and extensive earlier studies (Yegnanarayana et al., 1984), we observe that the group delay function resolves the zeroes and poles better than the magnitude and phase spectra when the signal is minimum phase. This is the primary motivation for converting a non-minimum phase signal to a minimum phase signal.

3. The root cepstrum approach to segment continuous speech

As observed from the results of the previous section, the magnitude spectra are identical in shape for minimum phase and non-minimum phase signals (Fig. 3(b), (f) and (j)), when the roots are located at reciprocal locations. Clearly, from the magnitude spectrum alone, one cannot identify whether the signal is minimum phase, type (1) or type (2). In this section, we first present an approach based on the root cepstrum to derive a minimum phase signal from any arbitrary magnitude spectrum. Next, we apply this technique to process the short-term energy function. We exploit the property that the short-term energy function is a positive function and can therefore be processed in a manner similar to that of magnitude spectrum.

3.1. Derivation of a minimum phase signal from the magnitude spectrum

To derive the minimum phase signal from any magnitude spectrum $|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|$, the following algorithm is proposed:

1. Compute the squared magnitude spectrum $|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|^2$ from $|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|$.
2. Compute the IDFT $(|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|^2)$. Let this be $x_{\mathrm{c}}(n)$.
3. The causal portion of $x_{\mathrm{c}}(n)$ is a minimum phase signal whose poles correspond to the peaks in the original magnitude spectrum $|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|$.

3.2. The minimum phase property of the root cepstrum

Consider a non-minimum phase signal $x_{\mathrm{nm}}(n)$ which is generated by a system $X_{\mathrm{nm}}(z)$ with one pole outside the unit circle at a distance $1/a$, where $|a| < 1$, i.e.,

$$X_{\mathrm{nm}}(z) = \frac{1}{1-az}. \tag{7}$$

The squared magnitude spectrum of $x_{\mathrm{nm}}(n)$ is

$$\begin{aligned}|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|^2 &= X(z)X^*(1/z^*)|_{z=\mathrm{e}^{\mathrm{j}\omega}} \\ &= \left.\frac{1}{1-a(z+z^{-1})+a^2}\right|_{z=\mathrm{e}^{\mathrm{j}\omega}} \\ &= R_{xx}(z)|_{z=\mathrm{e}^{\mathrm{j}\omega}}. \end{aligned} \tag{8}$$

From Eq. (8), we can infer that the squared magnitude spectrum has two poles, one inside and the other outside the unit circle. This is equivalent to the Fourier transform of the autocorrelation of the original signal $x_{\mathrm{nm}}(n)$. Now,

$$\begin{aligned}Z^{-1}(R_{xx}(z)) &= \frac{1}{1-a^2}a^{|l|} \quad -\infty < l < +\infty \\ &= r_{xx}(l), \end{aligned} \tag{9}$$

If we consider only the causal portion of the $r_{xx}(l)$, say $y(l)$, then

$$y(l) = \frac{1}{1-a^2}a^l \quad 0 \leqslant l < \infty. \tag{10}$$

The z-transform of $y(l)$ is given by

$$Y(z) = \frac{1}{1-a^2}\left(\frac{1}{1-az^{-1}}\right), \tag{11}$$

where $|a| < 1$. Using partial fractions, this result can be extended to any number of poles (Nagarajan et al., 2003). From Eq. (11), it can be concluded that the causal portion of the inverse Fourier transform of the squared magnitude spectrum of any type of signal is a minimum phase correspondent of the original signal in that the pole is located at the conjugate reciprocal location inside the unit circle. By the same token, theoretically, if the Fourier transform of a non-minimum phase signal exists, then the corresponding minimum phase signal can be derived using the power spectrum of the signal.

We can choose a value for 'γ' in $|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|^\gamma$ (step 1 in Section 3.1) such that $0 < \gamma \leqslant +2$ for poles and $0 > \gamma \geqslant -2$ for zeroes.

As long as γ is real, the causal portion of the root cepstrum derived from any magnitude spectrum exhibits the properties of a minimum phase signal (Nagarajan et al., 2001). This is because the root cepstrum can be represented as the convolution of some sequence $y(n)$ and $y(-n)$. For a Fourier transform to exist, $y(-n)$ and $y(n)$ must be bounded signals. If the system is stable, then $y(-n)$ must be a non-causal sequence while $y(n)$ must be a causal sequence. Hence, the causal portion of $y(n) * y(-n)$ is a decaying sequence. In general, the root cepstrum derived from $|X_{\mathrm{nm}}(\mathrm{e}^{\mathrm{j}\omega})|^\gamma$ has the following properties:

- The roots of the causal portion of the signal derived from the magnitude spectrum are all inside the unit circle (Eq. (11)).
- The angular frequencies of the poles are not disturbed.
- Since the duration of the causal portion of the root cepstrum is finite, the z-transform of that signal will have spurious zeroes. These zeroes affect the positions of the actual zeroes present in the signal. To overcome this problem, the spectrum is inverted $(1/(|X(\mathrm{e}^{\mathrm{j}\omega})|)^\gamma)$ and the minimum phase signal is derived using the algorithm given in Section. 3.1.

This clearly shows that, the root cepstrum method places the roots inside the unit circle and so, any non-minimum phase signal $x_{\mathrm{nm}}(n)$ can be converted to a minimum phase signal. What is crucial to this approach is that the angular frequency of the pole is not altered. This is an important feature, particularly, in the context of estimation of formants and anti-formants in speech processing (Hema A. Murthy, 1997). In this paper, we have developed this property of minimum group delay functions for detecting transitions between falls and rises in any kind of signal, as long as the signal can be represented by a positive function. In Section 2.3, it is mentioned that in the group delay spectrum, both the peaks and valleys are resolved correctly only for the minimum phase signal. Further in Section 3.1, it is established that a minimum phase signal can be derived from a given magnitude spectrum. Any arbitrary positive function symmetrized along the Y-axis (Fig. 4(a)), can be considered as a magnitude spectrum and a minimum phase signal can be derived from the same. To demonstrate this, an

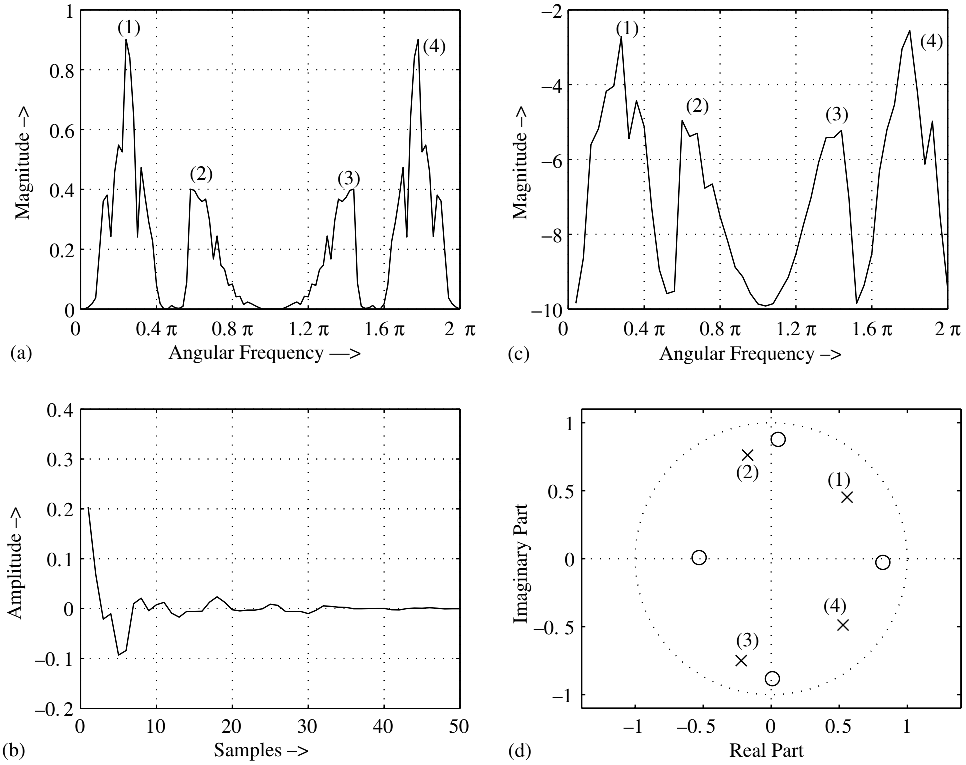

Fig. 4. Conversion of an arbitrary positive function to a minimum phase signal: (a) arbitrary positive function symmetrized about the *y*-axis; (b) the causal portion of the IDFT of the symmetrized energy contour shown in (a); (c) the magnitude spectrum of the signal shown in (b); (d) the *z*-plane with roots estimated from the magnitude spectrum shown in (c).

arbitrary symmetric positive function has been taken and the root cepstrum approach, explained in Section 3.1, has been applied. It is found that for the resultant signal (Fig. 4(b)), all the poles and zeroes (using a least square approach to estimate an ARMA model) [1] are inside the unit circle as shown in Fig. 4(d) and the angular frequencies of poles of the minimum phase signal (Fig. 4(b)) are same as the angular frequencies of the peaks of its power spectrum (Fig. 4(a)). But, there is a slight variation in the angular frequencies of zeroes which correspond to valleys of the power spectrum. This problem is addressed in the next section.

3.3. Minimum phase group delay based segmentation of speech

In Section 3.2, it was shown that significant events, namely, location of peaks/valleys for any arbitrary positive function can be obtained using the group delay function derived from the root cepstrum. In Section 1, it was shown that the short-term energy function is a good candidate for segmentation of continuous speech, but the issue is primarily the choice of an appropriate threshold. Since the short-term energy function is a positive function of time, it can be processed in a manner similar to that of processing an arbitrary magnitude spectrum (Fig. 4). The valleys

[1] The ARMA model based estimator is used only to confirm the fact that the causal portion of root cepstrum is indeed minimum phase.

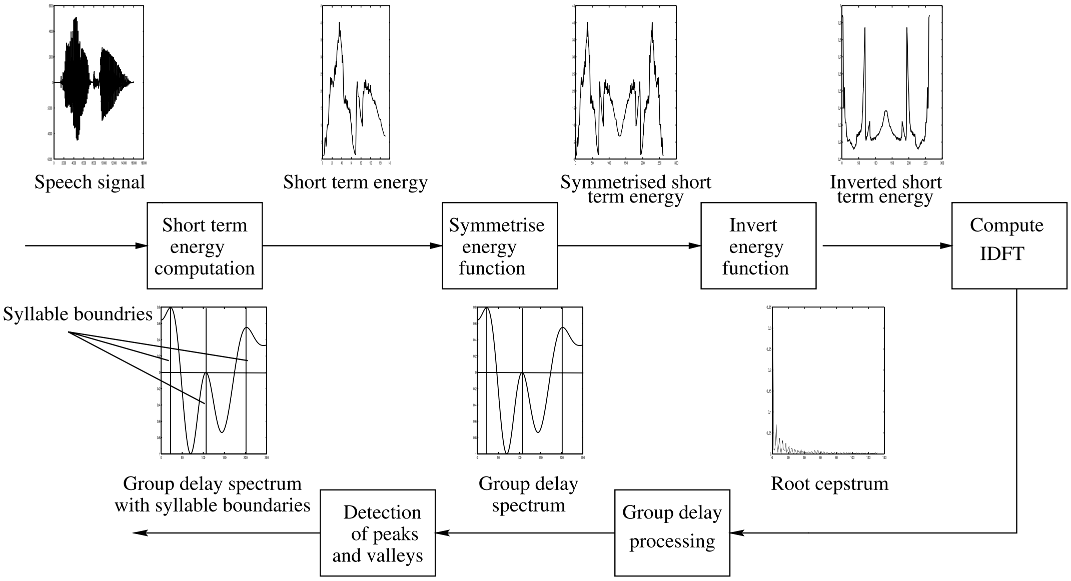
Speech signal
Short term energy
Symmetrised short term energy
Inverted short term energy
Short term energy computation
Symmetrise energy function
Invert energy function
Compute IDFT
Syllable boundries
Group delay spectrum with syllable boundaries
Detection of peaks and valleys
Group delay spectrum
Group delay processing
Root cepstrum

Fig. 5. Steps involved in finding syllable boundaries.

correspond to the location of segment boundaries. In the context of segmentation, we have observed that the duration of syllable segments does not vary very significantly. This ensures that equal emphasis is given to all sub-word units.

Truncation of the signal in the root cepstral domain can cause spurious valleys due to windowing effects. These valleys affect the position of valleys which correspond to actual segment boundaries in the speech signal. To overcome this problem, the short-term energy function is inverted. The positive peaks in the inverted energy function now correspond to the original segment boundaries. The steps involved in the segmentation of a continuous speech signal are as follows (see also Fig. 5):

- Let $x(n)$ be a given speech signal.
- Compute the short-term energy function $E(n)$, using overlapped windows.
- Construct the symmetric part of the sequence by producing a lateral inversion of this sequence about the *Y-axis*. This new sequence is viewed as an arbitrary magnitude spectrum and denoted by $E(k)$.
- Compute $(E(k))^{\gamma}$ where γ is $0 < \gamma \leqslant 2$. (Specifically, the value of γ has been optimized to 0.01.)
- Invert the function $(E(k))^{\gamma}$. Let the resultant function be $\widetilde{E}^{i}(k)$.
- Compute the inverse DFT of the function $\widetilde{E}^{i}(k)$. The resultant sequence $\tilde{c}(n)$, is the root cepstrum and the causal portion of it has minimum phase properties.
- Compute the minimum phase group delay function of the windowed [2] causalsequence $c(n)$ of $\tilde{c}(n)$ (Hema A. Murthy and Yegnanarayana, 1991; Hema A. Murthy, 1997) which follows the steps mentioned below.
  - ○ Compute $\phi(k)$, the phase spectrum of $c(n)$.
  - ○ Compute the group delay function as the forward difference of the phase function, i.e., $\underset{\sim}{\phi}'(k) = \phi(k) - \phi(k-1)$. Let this function be $\underset{\sim}{E}_{\text{gd}}(k)$.

[2] The size of the window (N_{c}) applied on this causal sequence is proportional to the length of the short-term energy function and is defined as

$$N_{\text{c}} = \frac{\text{Short-term energy function size}}{\text{Window scale factor (WSF)}}. \tag{12}$$

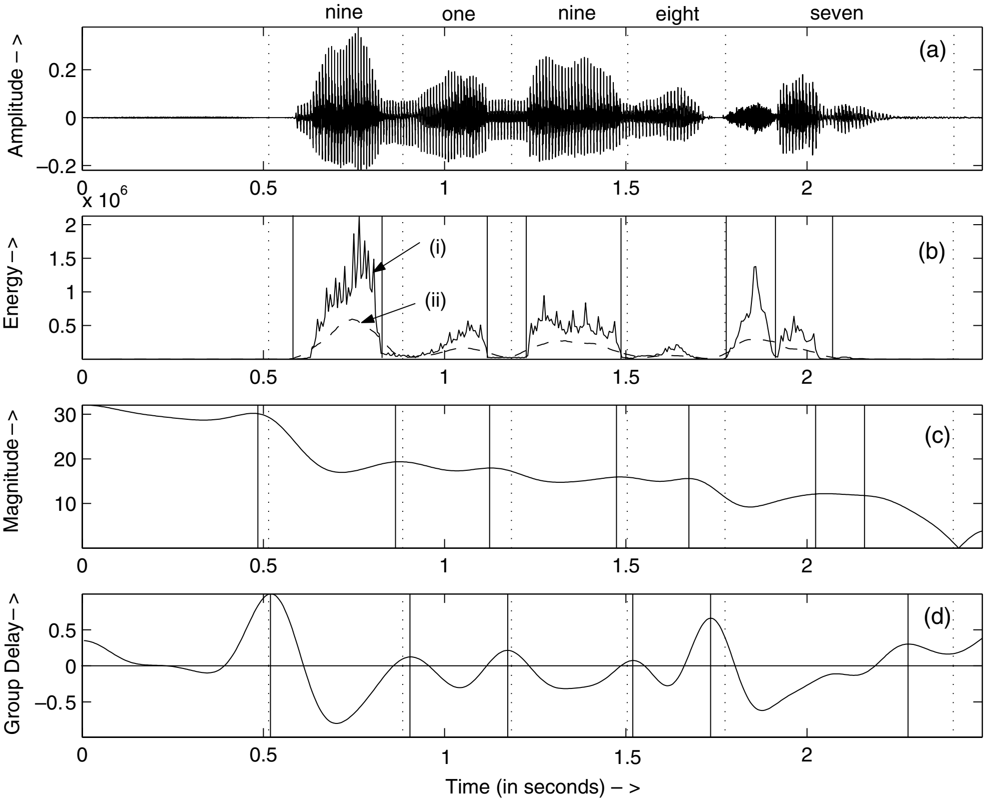

Fig. 6. Comparison of group delay function based segmentation with other techniques: (a) speech signal for the utterance of the digit string 91987. (b) Illustration of adaptive thresholding (dotted curve (ii)) on short-term energy function (solid curve (i)) with mean-smoothing order 25. (c) Cepstral smoothing, (d) minimum phase group delay function. In (b)–(d), the solid vertical lines denote segment boundaries obtained. The dotted vertical lines denote manually identified boundaries.

- The positive [3] peaks in the minimum phase group delay function $\widetilde{E}_{\mathrm{gd}}(k)$ approximately correspond to sub-word/syllable boundaries.

To demonstrate the effectiveness of the minimum phase group delay based speech segmentation algorithm, a comparison has been made with adaptive thresholding and the traditional cepstrum applied to a connected digit speech signal. This is illustrated in Fig. 6. The threshold for the adaptive thresholding based approach is computed over a 25 sample window on $E(n)$. If the minima between the two successive intersections of the energy function with the threshold function is less than the energy values at the intersection points, then that minimum is viewed as a valid syllable boundary. Fig. 6(b) shows the short-term energy function for the speech signal shown in Fig. 6(a), with the adaptive thresholding superimposed on it. It is found that there are spurious segments. Observe the spurious boundary in Fig. 6(b) between 1 and 1.5 s.

By viewing the short-term energy function as an arbitrary magnitude spectrum, conventional cepstrum based smoothing is applied. A one-sided Hanning window is applied on the traditional cepstrum. Simple peak picking algorithm is used on the spectrum (derived from the cepstrum), to detect the segment boundaries (Fig. 6(c)). It is found that in the resultant spectrum, the errors in

[3] Only positive peaks are chosen, as negative peaks are primarily caused by two consecutive valleys.

segmentation are quite high. For example, observe the erroneous segment boundaries corresponding to that of 'one' and 'eight'. But in the segmentation based on group delay function, as shown in Fig. 6(d), the peaks corresponding to segment boundaries are more accurate.

4. Performance evaluation

To evaluate the performance of the proposed segmentation algorithm, two different types of databases are used, namely the TIMIT (Fisher et al., 1986) and TIDIGITS (Leonard, 1984). In both the databases, the speech signals are not corrupted by background noise. To remove DC offsets in the speech signal, the signal is pre-emphasized. If there are any long inter-word silences present, these are removed before segmentation by using a coarse voiced–unvoiced detection algorithm based on zero-crossing rate. The short-term energy function is computed using overlapped rectangular windows, where the window length is of duration 12.5 ms and the overlap is of 5 ms duration. As explained in Section 3.3, the root cepstrum is computed on the short-term energy function and a one-sided Hanning window is used to truncate the cepstrum. The length of the window applied to the root cepstrum is tuned iteratively so that the number of peaks in the group delay function is equal to the number of voiced units present in the input speech signal. As explained in Section 3.3, to pick the valleys properly, the spectrum is inverted. The positive peaks in the group delay function correspond to segment boundaries. To overcome the problem of overflow when the short-term energy function is zero, zero values are replaced by the smallest non-zero value. Further the γ value in $(1/(|X(e^{j\omega})|)^{\gamma})$ is set to 0.01 to reduce the dynamic range of the short-term energy function.

4.1. Continuous speech segmentation

Since the number of syllables present in the speech signal is equal to the number of voiced units, the length of the Hanning window applied to the causal portion of the root cepstrum is adjusted iteratively. Initially, the window applied on the causal portion of the root cepstrum is chosen as 50 samples and the window size is iteratively adjusted so that the number of peaks in the group delay function is equal to the number of voiced units in the speech signal. Tuning is done separately for each continuous speech utterance in the database. The tuning process is demonstrated in Fig. 7. Fig. 7(a) is the speech signal and Fig. 7(b) denotes its short-term energy function. The group delay function derived from the energy function is shown in Fig. 7(c) which identifies only four segments. Further, When the window size is increased iteratively, the missed peak is also identified as shown in Fig. 7(d).

Performance of the proposed segmentation algorithm is evaluated on the sentence *she had your dark suit in greasy wash water all year* from the TIMIT (Fisher et al., 1986) database. For all monosyllabic words, the word boundaries nearly coincide with the syllable boundaries. The bisyllabic words are split further at syllable boundaries. Although the phrase *suit in* consists of two words *suit* and *in*, acoustically it is represented as two syllables, *su* and *tin*. Hence the word sequence *suit in* is viewed as a syllable sequence *su* and *tin*. Fig. 8 demonstrates the segmentation of the given continuous speech signal at syllable boundaries. Fig. 8(a) shows the continuous speech utterance, and, Fig. 8(b) is its short-term energy function. The location of peaks in the minimum phase group delay plot correspond to syllable boundaries which are represented by solid lines in Fig. 8(c), and, the manually found syllable boundaries are represented by dotted vertical lines. The proposed method is applied on all the 462 utterances of the sentence *she had your dark suit in greasy wash water all year* from the TIMIT database. The error observed, in addition to an overall 5% insertions and 5% deletions, is shown in Table 1. Given that the average syllable duration is 250 ms, the error in segmentation for the worst case is 50 ms which is 20% of the syllable duration. Post-processing of segment boundaries can be taken up as future research to revise the segment boundaries.

Fig. 9 demonstrates the consistency in the proposed segmentation approach. If the number of segments generated by this segmentation approach is not equal to the number of syllables present in

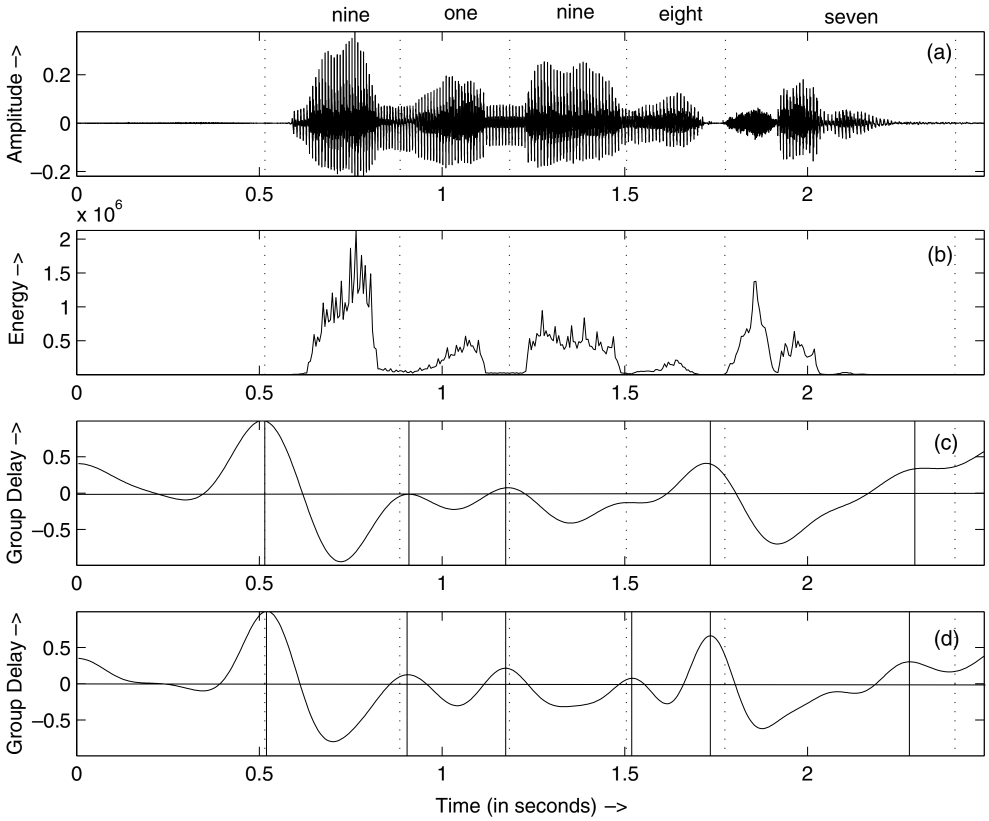

Fig. 7. Iterative adjustment of group delay function parameter: (a) speech utterance of the digit string '91987', (b) short-term energy function of the signal, (c) initial group delay spectrum (d) group delay spectrum obtained after tuning the parameters. Solid vertical lines in (c) and (d) denote the segment boundaries. The dotted vertical lines denote manually identified boundaries.

the speech signal, it does not result in altering the actual segment boundaries. In Fig. 9, the manually marked boundaries are indicated by dotted vertical lines, while the group delay boundaries are indicated by solid vertical lines. When the number of segments are less than the number of syllables present, as shown in Fig. 9(a), the group delay peaks near 1.9 and 2.5 s are missed, because their amplitudes are negative, but boundaries on either side are not misplaced. When the root cepstral window size is increased, the amplitude of the group delay peak near 2.5 s becomes positive, and a spurious segment boundary is introduced (Fig. 9(b)). Further increase of the window size (N_c) results in an additional spurious segment boundary near 0.4 s as shown in Fig. 9(c). In either case, there is no significant displacement in other segment boundaries.

4.2. Segmentation of connected digit speech

Segmentation performance of the proposed algorithm is also evaluated on the male speaker TIDIGITS (Leonard, 1984) database. The tuning procedure applied on the root cepstral window is same as that of continuous speech segmentation except that the number of digits present in the connected digit utterance is considered in place of voicing units. The vocabulary of TIDIGITS database consists of 11 digits (1 to 9, *zero* and *oh*). Among the eleven digits, eight digits (1, 2, 3, 4, 5, 8, 9 and *oh*) consist of only one syllable unit. Other digits (6, 7 and *zero*) consist of two sub-word units; the digit 6 contains of a sub-word unit which does not consist of voicing, whereas the digits 7 and *zero* consist of two sub-word units which correspond to two syllables. To demonstrate the

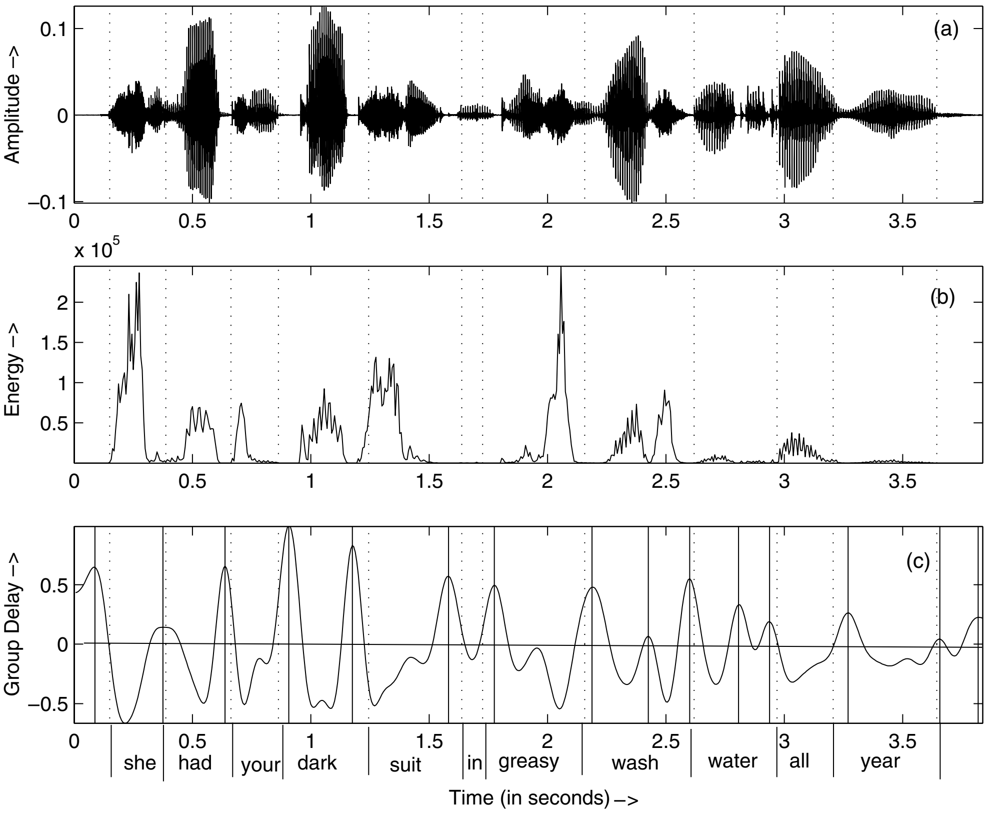

Fig. 8. An example for segmenting the continuous speech signal using minimum phase group delay function: (a) continuous speech signal, (b) short-term energy function and (c) minimum phase group delay function, for the utterance *she had your dark suit in greasy wash water all year* from the TIMIT database.

Table 1
Segmentation performance of continuous speech utterance *she had your dark suit in greasy wash water all year* from the TIMIT database

Error range (in ms)	Coverage (in %)
⩽ 10	29.6
10–20	45.0
21–30	15.9
31–40	4.0
41–50	1.8
⩾ 51	3.1

segmentation performance in different cases, the digit strings of lengths varying from 2 digits to 7 digits have been considered.

When there is a significant intra-digit energy variation, the proposed algorithm may split digits with two sub-word units into two segments. To address this problem, durational information of digits is used. The entire male speaker database from TIDIGITS is manually segmented. The mean and standard deviation of digit durations are estimated from the segmented database. It is found that the mean value is 390 ms and the standard deviation is 60 ms. The durational information for the entire male speakers' database for all the digits is shown in Fig. 10.

Any segment of duration not within the range '$\mu \pm 3\sigma$' is treated separately. If the duration of a segment is more than '$\mu + 3\sigma$', this segment is processed further using the same segmentation algorithm to determine whether further segmentation is possible. If the duration of segment is less than '$\mu - 3\sigma$', it is treated as a syllabic fragment and moderate post processing is done

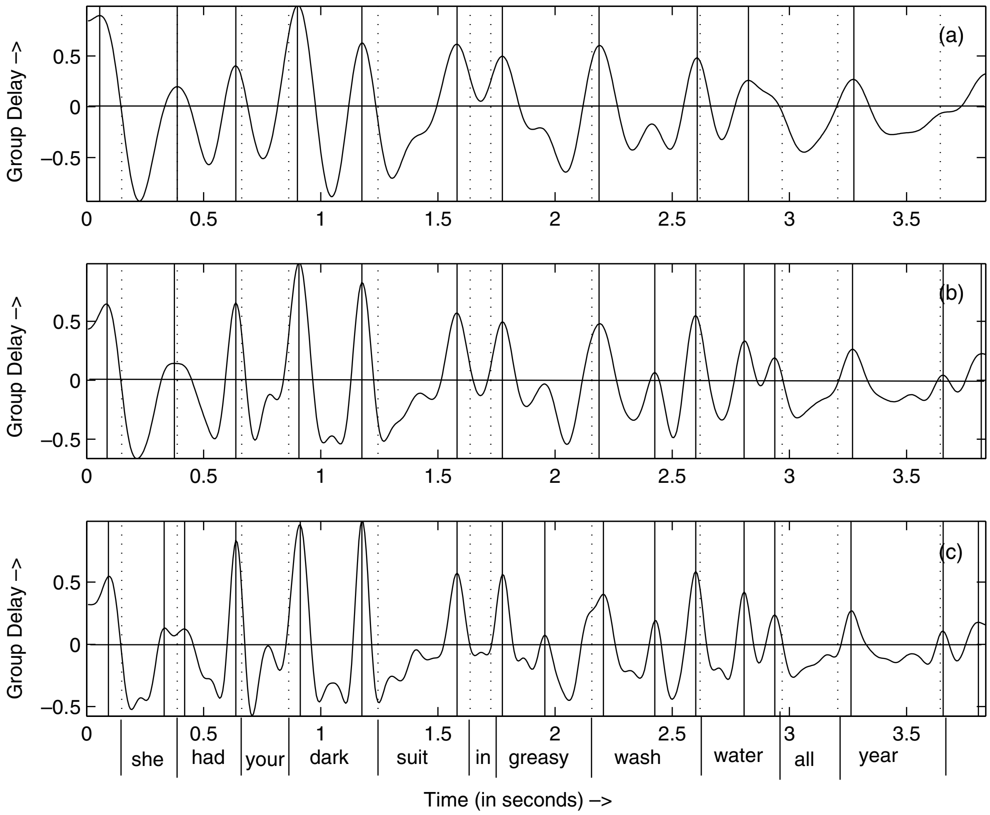

Fig. 9. Consistency in the proposed segmentation approach. (a)–(c) show the minimum phase group delay functions correspond to the windows applied on the causal portion of the root cepstrum, in the increasing order of window size (0.96, 1.28 and 1.92 s, respectively).

to detect whether the fragment is a fricative or not. Fricative segments are characterized by high zero crossing rate, high spectral flatness and low energy. If the segment is found to be a fricative, it is merged with one of the neighbouring segments, that is shorter in duration.

Fricatives are generally not tightly bound to the syllabic units with which they are associated but are frequently separated from them by a short interval of weak voicing or even silence. As a result, fricative sounds on either side of the utterance *six* are sometimes treated as separate segments by the proposed algorithm. These segments are processed and merged with one of the neighbours in a manner similar to the one explained earlier.

The error in segmentation using the proposed algorithm is computed as follows:

$$\text{Relative error} = \frac{|(\text{Actual duration} - \text{Estimated duration})|}{\text{Actual duration}}. \tag{13}$$

Fig. 11 demonstrates the distribution of the error relative to the average duration of all digit segments. In about 90% of the instances, the error in segmentation is less than 20% of the duration of the digit utterance.

Segmentation performance is also assessed with respect to transition from one digit to another. Segmentation performances for different permutations of digit transitions are shown in Table 2. In Table 2, the row corresponding to digit ‘six’, corresponding to the transition from digit ‘six’ to any other digit, shows large errors. In the utterance ‘six’, the fricative sounds on either side is not

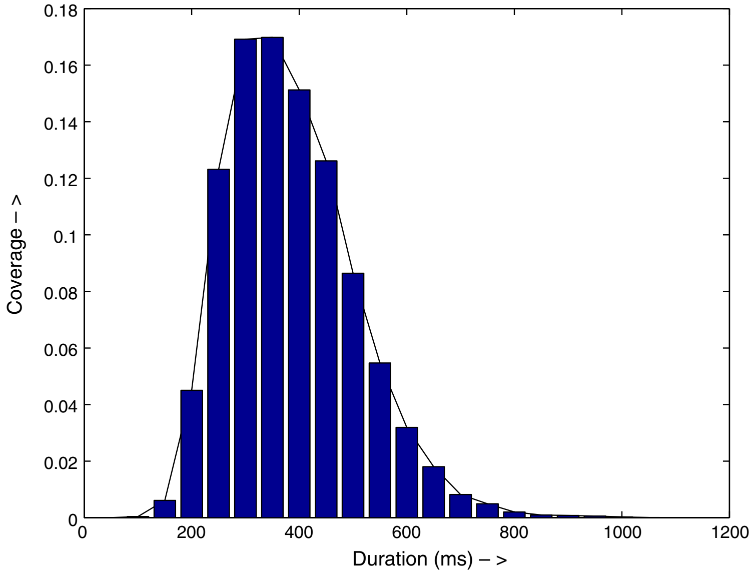

Fig. 10. Durational distribution of all digit segments from TIDIGITS male speakers database.

tightly bound with the rest of the utterance, resulting in low energy regions in the short-term energy function. This characteristic results in large errors.

To evaluate the segmentation performance in terms of insertions and deletions, the database is grouped into three classes. The first class consists of connected digit utterances, where each digit in the digit string contains one syllable. The second class consists of connected digit utterances, where one or more occurrences of digit 6, contains an unvoiced sub-word unit, along with digits with one syllable. The third class consists of connected digit utterances where one or more digits consists of digits with two sub-word units (6, 7 and *zero*) along with one sub-word unit digit strings. The performance for different digit string lengths is presented in Table 3.

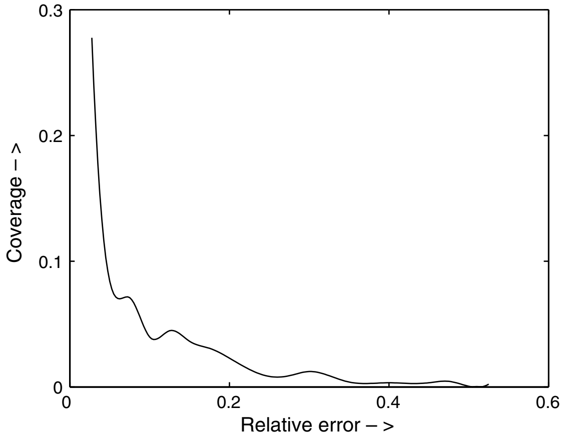

Fig. 11. The distribution of the relative error for all 11 digits from the male speakers in the TIDIGITS database.

From Table 3, we observe that, as the number of digits in the digit string increases, the percentage of insertions/deletions also increases for all the three classes of digit strings. In particular, for the second and third classes, the percentage of insertions/deletions are slightly more when compared with the same in the first class. This is because of the occurrence of digit 6 in the digit string. In the digits 7 and *zero*, the sub-word units are relatively close to each other compared to the neighbouring digit segments. Hence, when the group delay function is tuned to obtain segments equal to the number of digits present, it is likely that sub-word units belonging to the same digit are merged and identified as one unit. Due to this behaviour of the group delay function, segmentation performance degrades gracefully.

Table 2
The averaged segmentation error for the transition between different digits (in ms)

Digit class transition	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero	Oh
One	25	44	65	65	52	10	15	33	52	32	63
One	(19)	(13)	(18)	(20)	(24)	(21)	(23)	(12)	(24)	(18)	(17)
Two	45	64	75	87	123	38	63	30	28	46	19
Two	(18)	(23)	(15)	(15)	(22)	(13)	(19)	(16)	(26)	(23)	(15)
Three	36	39	62	84	64	49	42	69	26	59	75
Three	(13)	(18)	(17)	(19)	(15)	(24)	(18)	(9)	(19)	(18)	(15)
Four	51	75	67	67	66	66	110	38	26	83	47
Four	(13)	(22)	(18)	(25)	(18)	(17)	(15)	(13)	(21)	(14)	(14)
Five	22	57	75	71	37	14	15	20	21	33	22
Five	(21)	(20)	(16)	(16)	(11)	(24)	(24)	(12)	(19)	(22)	(20)
Six	54	48	65	53	62	101	118	82	66	130	133
Six	(18)	(19)	(23)	(20)	(17)	(15)	(10)	(14)	(16)	(26)	(19)
Seven	62	23	31	46	39	33	33	64	29	63	69
Seven	(20)	(16)	(21)	(16)	(24)	(16)	(14)	(17)	(23)	(15)	(25)
Eight	61	43	93	81	71	44	46	22	27	42	53
Eight	(13)	(20)	(14)	(19)	(13)	(20)	(18)	(17)	(27)	(21)	(17)
Nine	25	26	44	70	35	10	19	32	19	16	45
Nine	(26)	(26)	(30)	(16)	(22)	(15)	(22)	(20)	(22)	(15)	(23)
Zero	53	66	72	94	81	30	30	73	26	32	–
Zero	(26)	(17)	(16)	(21)	(18)	(18)	(23)	(16)	(17)	(36)	–
Oh	48	45	67	66	59	21	93	102	34	–	27
Oh	(23)	(14)	(18)	(18)	(24)	(22)	(17)	(19)	(24)	–	(30)

The value in brackets denote the number of occurrences of digit pairs.

Table 3
Segmentation errors in terms of insertions and deletions using the proposed approach

No. of digits in the utterance	Digit strings with digits of		
	one syllable (%)	one syllable with one or more occurrences of digit 6 (%)	one syllable with one or more occurrences of digit 6, 7 and zero (%)
2	0.6	1.2	1.8
3	1.5	3.9	3.7
4	3.6	4.9	3.9
5	5.2	8.2	7.7
7	7.6	9.6	7.8

5. Conclusions

In this paper, we have proposed a novel approach for segmenting the speech signal into syllable-like units. Although, the raw short-term energy function of the speech signal contains information about the syllable segment boundaries by means of energy minima, we have shown that a simple adaptive thresholding technique is of limited use for extracting boundaries. The major reason for this is the presence of local energy fluctuations in the raw short-term energy function.

As an alternative to adaptive thresholding, we propose a group delay based approach to processing the short-term energy for determining segment boundaries. The performance of this technique is tested on both continuous speech utterances and connected digit sequences. It is shown that the segmentation performance is quite satisfactory. The error in segment boundary is $\leqslant 20\%$ of syllable duration for 70% of the syllables. In addition to true segments, an overall 5% insertions and deletions have also been observed. Our results illustrate that segmentation prior to labelling speech can be performed with the group delay approach, at least for the two types of read speech that were studied in this investigation.

Acknowledgements

The authors would like to thank the reviewers for very fruitful comments. In particular, they would like to thank one of anonymous reviewers who helped (i) in making significant changes to the presentation and (ii) the English. The authors would also like to thank Dr. V. Bharathi, TeNet group, for editing the final draft.

References

Berkhout, A.J., 1973. On the minimum length property of one-sided signals. Geophysics 38 (4), 657–672.

Berkhout, A.J., 1974. Related properties of minimum phase and zero phase time functions. Geophys. Prospect. (22), 683–709.

Fisher, W.M., Doddington, G.R., Goudie-Marshal, K.M., 1986. The darpa speech recognition research database: specifications and status. In: Proc. DARPA Workshop on Speech Recognition. pp. 93–99.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G.R., 2001. Syllable-based large vocabulary continuous speech recognition. IEEE Trans. Speech, Audio Process. 9 (4), 358–366.

Greenberg, S., 1999. Speaking in short hand—a syllable-centric perspective for understanding pronunciation variation. Speech Comm. 29, 159–176.

Hema A. Murthy, 1992. Algorithms for processing fourier transform phase of signals. PhD dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India.

Hema A. Murthy, 1997. The real root cepstrum and its applications to speech processing. In: National Conf. on Communication. 180–183.

Hema A. Murthy, Yegnanarayana, B., 1991. Formant extraction from minimum phase group delay function. Speech Comm. 10, 209–221.

Leonard, R.G., 1984. A database for speaker independent digit recognition. In: Proc. IEEE Internat. Conf. on Acoust., Speech, and Signal Processing, Vol. 3. pp. 42–45.

Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Amer. 58 (4), 880–883.

Nagarajan, T., Kamakshi Prasad, V., Hema A. Murthy, 2001. Minimum phase signal derived from the magnitude spectrum and its application to speech segmentation. In: 6th Biennial Conf. Proc. on Signal Processing and Communications. IISc, Bangalore, India, pp. 95–101.

Nagarajan, T., Kamakshi Prasad, V., Hema A. Murthy, 2003. Minimum phase signal derived from root cepstrum. IEE Electron. Lett. 39 (12), 941–942.

Rabiner, L.R., Rosenberg, A.E., Wilpon, J.G., Zampini, T.M., 1982. A bootstrapping training technique for obtaining demisyllabic reference patterns. J. Acoust. Soc. Amer. 71, 1588–1595.

Sargent, D.C., Li, K.P., Fu, K.S., 1974. Syllabic detection in continuous speech. J. Acoust. Soc. Amer. 45 (4), 880–883.

van Hemert, J.P., 1991. Automatic segmentation of speech. IEEE Trans. Signal Process. 39 (4), 1008–1012.

Wilpon, J.G., Juang, B.H., Rabiner, L.R., 1987. An Investigation on the use of acoustic sub-word units for automatic speech recognition. In: Proc. of IEEE Internat. Conf. on Acoust., Speech, and Signal Processing. Dallas, TX, pp. 821–824.

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

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), 610–622.