EURASIP Journal on Applied Signal Processing 2004:17, 2614–2625

Subband-Based Group Delay Segmentation of Spontaneous Speech into Syllable-Like Units

T. Nagarajan
Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India
Email: raju@lantana.iitm.ernet.in

H. A. Murthy
Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India
Email: hema@lantana.tenet.res.in

Received 16 January 2004; Revised 17 June 2004; Recommended for Publication by Chin-Hui Lee

In the development of a syllable-centric automatic speech recognition (ASR) system, segmentation of the acoustic signal into syllabic units is an important stage. Although the short-term energy (STE) function contains useful information about syllable segment boundaries, it has to be processed before segment boundaries can be extracted. This paper presents a subband-based group delay approach to segment spontaneous speech into syllable-like units. This technique exploits the additive property of the Fourier transform phase and the deconvolution property of the cepstrum to smooth the STE function of the speech signal and make it suitable for syllable boundary detection. By treating the STE function as a magnitude spectrum of an arbitrary signal, a minimum-phase group delay function is derived. This group delay function is found to be a better representative of the STE function for syllable boundary detection. Although the group delay function derived from the STE function of the speech signal contains segment boundaries, the boundaries are difficult to determine in the context of long silences, semivowels, and fricatives. In this paper, these issues are specifically addressed and algorithms are developed to improve the segmentation performance. The speech signal is first passed through a bank of three filters, corresponding to three different spectral bands. The STE functions of these signals are computed. Using these three STE functions, three minimum-phase group delay functions are derived. By combining the evidence derived from these group delay functions, the syllable boundaries are detected. Further, a multiresolution-based technique is presented to overcome the problem of shift in segment boundaries during smoothing. Experiments carried out on the Switchboard and OGI-MLTS corpora show that the error in segmentation is at most 25 milliseconds for 67% and 76.6% of the syllable segments, respectively.

Keywords and phrases: group delay, minimum-phase signal, syllable, subband-based segmentation.

1. INTRODUCTION

One of the major reasons for considering the syllable as a basic unit for automatic speech recognition (ASR) systems is its better representational and durational stability relative to the phoneme [1]. The syllable was proposed as a unit for ASR as early as 1975 [2], in which irregularities in phonetic manifestations of phonemes were discussed. It was argued that the syllable will serve as an effective minimal unit in the time domain. In [3], it is demonstrated that segmentation at syllable-like units followed by isolated style recognition of continuous speech performs well.

Researchers have tried different ways of segmenting the speech signal either at the phoneme level or at the syllable level, with or without the use of phonetic transcription. These segmentation methods can further be classified into two categories, namely, time-domain-based methods, where short-term energy (STE) function, zero-crossing rate, and so forth are used, and frequency-domain-based methods, where short-term spectral features are used.

In [4], a loudness function, defined as the time-smoothed and frequency-weighted summation of the signal spectrum, is used for segmenting speech into syllabic units. Syllable boundaries are placed at local minima in the loudness function, subject to various conditions.

A syllabification procedure developed in [5] for German makes an initial estimate of syllable boundaries based on voicing, energy level, and place of articulation and then locates syllables based on a more detailed acoustic analysis.

In [6], for Japanese, a syllable-level segmentation technique is proposed, which is based on a common syllable model. The segment boundaries are detected by finding the optimal HMM state sequence.

In [7], a multilayered neural network structure for continuous speech recognition, based on isolation and identification of syllables, is presented. The syllable boundaries are detected at the first layer of a neural network, which is an adaptation of Kohonen's phonotopic feature map trained by unsupervised learning.

An STE-based method for detecting syllable nuclei is presented in [8]. In this work, the speech is first bandpass filtered and then the short-term magnitude function is computed. To suppress the ripples caused by f_0 or transient phonemes, the short-term magnitude function is further lowpass filtered at approximately 10 Hz. The peaks of the resulting energy contour are declared as the syllable nuclei.

In [9], a temporal flow model (TFM) network has been developed to extract syllable boundary information from continuous speech, where TFM captures the time-varying properties of the speech signal.

The syllable is structurally divisible into three parts, the onset, nucleus, and coda [10]. Although many syllables contain all three elements, a significant number contain either one or two. With rare exceptions, when a single component is present, it is the nucleus. Generally, the nucleus is vocalic, while the onset and coda are usually consonantal in form. In terms of STE function, the syllable can be viewed as an energy peak in the nucleus region and it tapers off at both ends of the nucleus where a consonant may be present, which results in local energy fluctuations. If these local energy fluctuations are smoothed out, then the valleys at both ends of the syllable nucleus can be considered as syllable boundaries.

Many languages of the world possess a relatively simple syllable structure consisting of several canonical forms [10]. Most of the syllables in such languages contain just two phonetic segments, typically of CV type (e.g., Japanese language). The remaining syllabic forms are generally of V or VC variety. In contrast, English and German possess a more highly heterogeneous syllable structure. In such forms, the onset and/or coda constituents often contain two or more consonants. But a salient property shared by stress- and syllable-timed languages is the preference of CV syllabic forms in *spontaneous speech*. Nearly half of the forms in English and over 70% of the syllables in Japanese are of this variety. There is also a substantial proportion of CVC syllables in the spontaneous speech of both the languages [10]. The analysis done on the Switchboard corpus shows that nearly 88% of the syllables are of simple structure and only 12% of the syllables are of a more complex structure with consonant clusters [10]. This shows that even for the languages which are not syllable-timed, the syllable can be defined using a simple structure. Further, the definition of syllable in terms of STE function is suitable for almost all the languages, in the case of spontaneous speech. Keeping this fact in mind, in this paper, a time-domain-based speech segmentation procedure is described, which segments the speech signal into syllable-like units, without the knowledge of the phonetic transcription. This approach is somewhat similar to *homomorphic filtering*, which essentially smoothes the magnitude spectrum of the windowed speech signal.

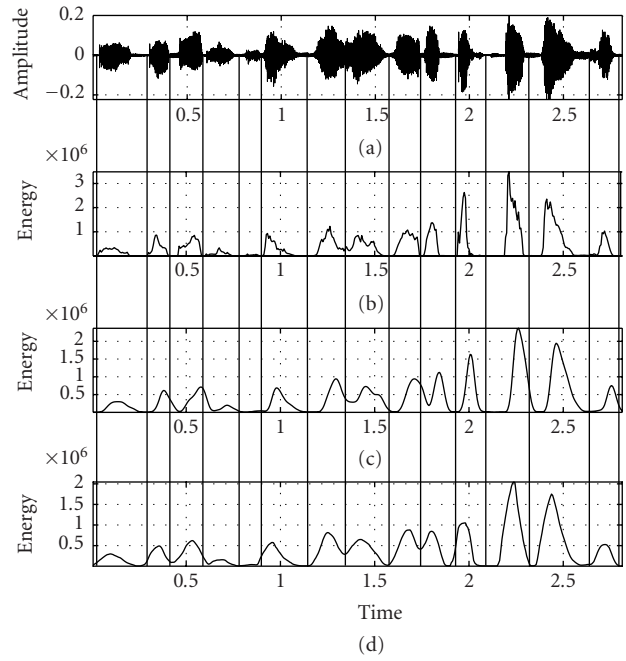

Figure 1: (a) Speech signal. (b) Corresponding STE function. (c) Lowpass filtered STE function. (d) Mean-smoothed STE function.

Earlier, a method was proposed [11] for segmenting the acoustic signal into syllable-like units, in which a minimum-phase signal is derived from the STE function as if it were a magnitude spectrum. It is observed that the group delay function of this minimum-phase signal is a better representative of the STE function to perform segmentation. Later, several refinements have been made to improve the performance of the baseline segmentation algorithm [12]. In this paper, we specifically discuss the refinements made on the system described in [11].

2. SHORT-TERM ENERGY-BASED SEGMENTATION

A simple candidate for segmenting speech is the STE function of the speech signal. The high-energy regions in the STE function correspond to syllable nuclei, and the valleys at both ends of the syllable nuclei are approximately the syllable boundaries. But the raw STE function cannot be directly used to perform segmentation due to significant local energy fluctuations. This is due to the presence of transient consonants and f_0 (see Figure 1b). Techniques like fixed thresholding can be used but suffer when energy variation across the signal is quite high. For continuous speech, especially for spontaneous speech, the energy is quite high at the beginning of a phrase and tapers off towards the end of the phrase. An adaptive thresholding can be used to address this problem but the threshold value 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.

To overcome the problems due to local energy fluctuations, the STE function should be smoothed. Smoothing the STE function can be performed in several ways. Firstly, the STE function can be computed with increased window size, but with the consequence of shift in boundary locations. The STE function is normally mean smoothed with a narrow window size (see Figure 1d). In this case, the order of mean smoothing is crucial. If the order is large, it will result in significant shift in boundaries or even miss detection of boundaries altogether, while if the order is small, it will not properly serve the purpose of smoothing. In [13], it is mentioned that the syllable duration can be conceptualized in terms of *modulation frequency*. For example, a syllable duration of 200 milliseconds is equivalent to a modulation frequency of 5 Hz. Further, the syllable duration analysis [10] performed on the *Switchboard corpus* [14] shows that the duration of syllables mostly varies from 100 milliseconds to 300 milliseconds with a mean of 200 milliseconds. In terms of modulation frequency, it varies from 3 to 10 Hz, with a mean of 5 Hz. Using this approach, in [8], a lowpass filter with cutoff frequency of 10 Hz is applied on the logarithmic STE amplitude to suppress the ripples caused by f_0 or transient consonants. This forces the system to oscillate at syllable frequencies (see Figure 1c). The selection of cutoff frequency is crucial; it should be different for different speech rates.

In this paper, an attempt is made to overcome these issues. The STE function is a nonzero, positive function. But the magnitude spectrum of any real signal has the symmetry property, that is,

$$|X(\omega)| = |X(-\omega)|. \tag{1}$$

If the STE function is symmetrized, it will have the properties similar to that of the magnitude spectrum. Therefore, techniques applied for processing the magnitude spectrum can be applied to the energy function. The inverse DFT (IDFT) of this assumed magnitude spectrum will be a two-sided signal (the real cepstrum). If the causal portion of this signal alone is considered, it is a perfect minimum-phase signal since it is derived from the magnitude spectrum alone. Now, smoothing of this assumed magnitude spectrum can be performed using one of the following techniques.

(1) *Cepstrum-based smoothing*. It is well established that high-frequency ripples can be removed by applying a lifter in the cepstral domain, thereby retaining the low-frequency ripples alone [15]. Using the same analogy, in our work, the symmetrized STE function is treated as if it were a magnitude spectrum of an arbitrary signal. The low-frequency oscillations in the STE function correspond to the syllable rate and the high-frequency oscillations or ripples correspond to the presence of transient consonants and f_0. The high-frequency ripples in the STE function can be removed as is done in *homomorphic filtering*.

(2) *LP-cepstrum-based smoothing*. By choosing a proper order, which is based on the number of syllables present in the speech signal, the cepstrum can be modeled.

(3) *Root-cepstrum-based smoothing*. In [16], it is shown that spectral root homomorphic deconvolution performance is similar to, or even better than, the log homomorphic deconvolution, where the root-cepstrum is defined by the IDFT $|X(\omega)|^{\gamma}$, with $0 < \gamma \ll 1$.

It has been well established in the literature that minimum-phase group delay functions are very useful in formant extraction [17]. In the present work also, instead of deriving a new magnitude spectrum from the cepstrum, group delay functions are derived as explained in the following section.

3. GROUP-DELAY-BASED SEGMENTATION OF SPEECH

The negative derivative of the Fourier transform phase is defined as *group delay*. The group delay function exhibits additive properties. If

$$H(\omega) = H_1(\omega) \cdot H_2(\omega), \tag{2}$$

then the group delay function $\tau_h(\omega)$ can be written as

$$\begin{aligned} \tau_h(\omega) &= -\frac{\partial(\arg(H(\omega)))}{\partial\omega} \\ &= \tau_{h_1}(\omega) + \tau_{h_2}(\omega). \end{aligned} \tag{3}$$

From (2) and (3), observe that a 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 (Figure 2a, 2b, and 2c), where the first system consists of a complex conjugate pole pair at an angular frequency ω_1, the second system with a complex conjugate pole pair at an angular frequency ω_2, and the third with two complex conjugate pole pairs, one at ω_1 and the other at ω_2. From the magnitude spectra of these three systems (Figures 2d, 2e, and 2f), it is observed that even though the peaks in Figures 2d and 2e are clearly visible, in a system where these two poles are combined together, the peaks are not resolved well as shown in Figure 2f. This is due to the multiplicative property of the magnitude spectra. But from Figures 2g, 2h, and 2i, it is evident that the group delay spectrum obtained by combining the poles together, the peaks are well resolved as shown in Figure 2i. Further, in the group delay spectrum of any signal, the peaks (poles) and valleys (zeros) will be resolved properly only when the signal is a minimum-phase signal. In our work, since the signal is derived from the positive function (which is similar to the magnitude spectrum), it can be shown that the resultant signal is a minimum-phase signal. We have exploited the minimum-phase property of the signal derived from any positive function and the additive property of the group delay function to segment the speech into syllable-like entities.

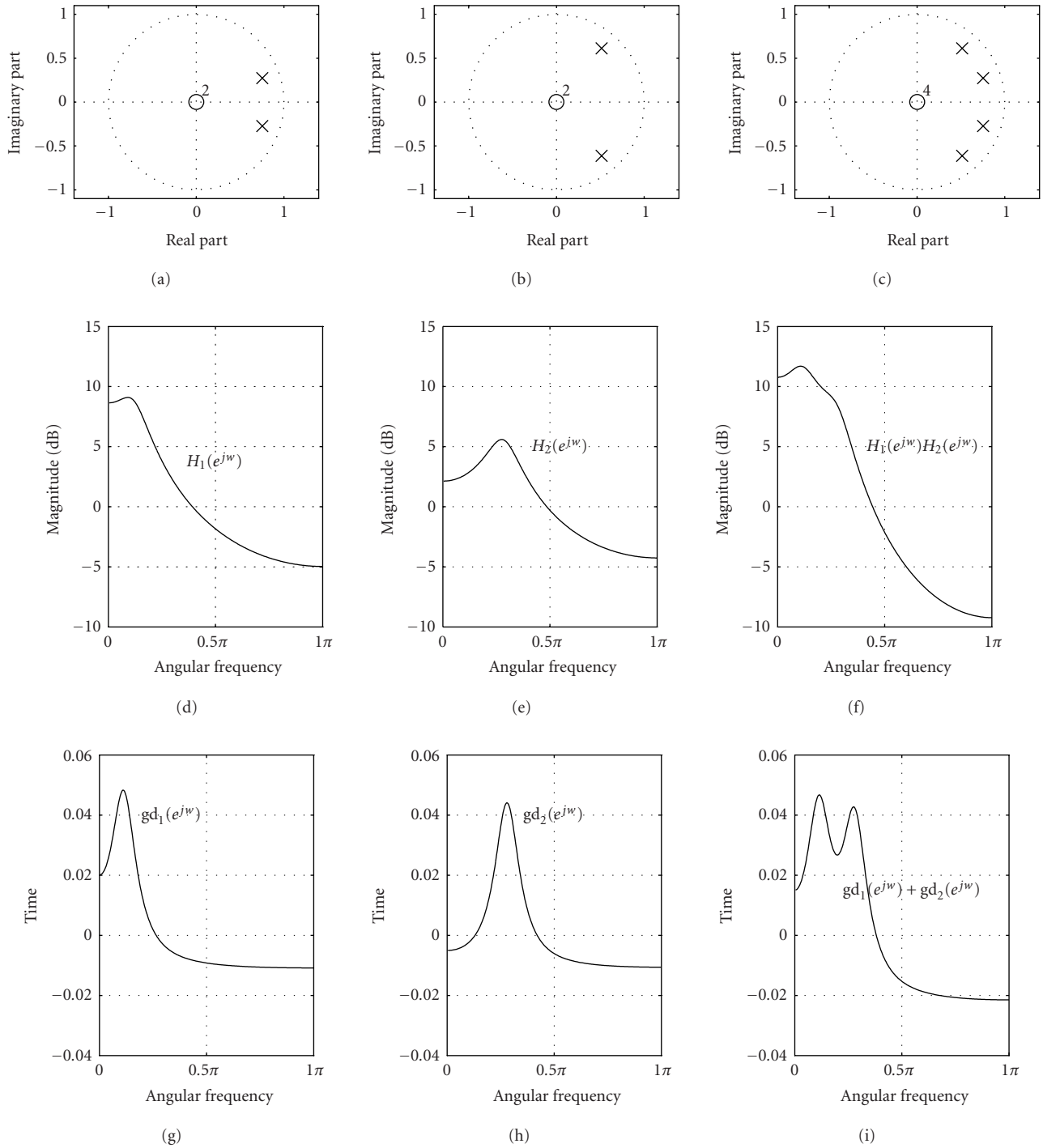

Figure 2: Resolving power of group delay spectrum: z-plane, magnitude spectrum, and group delay spectrum of the cases of ((a), (d), and (g)) a pole inside the unit circle at $(0.8, \pi/8)$, ((b), (e), and (h)) a pole inside the unit circle at $(0.8, \pi/4)$, and ((c), (f), and (i)) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$ inside the unit circle.

3.1. The minimum-phase property of the magnitude spectrum

Consider a system function $X(z)$ given below:

$$X(z) = \frac{1}{\Pi_{i=1}^{N}(1 - a_i e^{jw_i} z)}. \tag{4}$$

The square of the magnitude of the system frequency response is given by

$$|X(e^{j\omega})|^2 = X(e^{j\omega})X^c(e^{j\omega}) = X(z)X^c\left(\frac{1}{z^c}\right)\Big|_{z=e^{j\omega}}. \tag{5}$$

Let

$$C(z) = X(z)X^c\left(\frac{1}{z^c}\right), \tag{6}$$

where "c" denotes complex conjugation;

$$C(z) = \frac{1}{\Pi_{i=1}^{N}(1 - a_i e^{jw_i} z)(1 - a_i e^{jw_i} z^{-1})}. \tag{7}$$

From (7), we can infer that, for every pole in $X(z)$, there is a pole in $C(z)$ at a_i and $1/a_i^c$. Consequently, if one element of each pair is outside the unit circle, then the conjugate reciprocal will lie inside the unit circle [18]. Since the Fourier transform of (7) exists, inverse z-transform of (7) leads to

$$c(n) = \sum_{i=1}^{N} \left[A_i (a_i e^{jw_i})^{-n} u(-n-1) + B_i (a_i e^{jw_i})^{n} u(n)\right], \tag{8}$$

where $u(n)$ is the unit-step function.

If only the causal portion of $c(n)$ is considered, then

$$c_m(n) = \sum_{i=1}^{N} B_i (a_i e^{jw_i})^{n} u(n). \tag{9}$$

From (9), we conclude that the causal portion of the inverse Fourier transform of the squared magnitude spectrum of a signal whose root is at "a_i" or "$1/a_i$," with $|a_i| < 1$, will have a root at "a_i," that is, the resultant signal will always be a minimum-phase signal. But, since a window is applied in the cepstral domain, the root-cepstrum is of finite length. Because of this, the z-transform of the signal will have spurious zeros. These zeros may affect the positions of the actual zeros present in the signal. To overcome this problem, the squared magnitude spectrum can be inverted ($1/(|X(e^{j\omega})|^2)$) and another minimum-phase signal can be derived using the same algorithm, if zeros are of interest.

Instead of taking the squared magnitude spectrum, in fact, we can take $|X(e^{j\omega})|^{\gamma}$, where γ can be any value.[1] If the signal $x(n)$ is an energy bounded signal, from the Akhiezer-Krein and Fejer-Riesz theorems [19], it can be shown that

$$\begin{aligned} F^{-1}\left(|X(e^{j\omega})|^{\gamma}\right) &= F^{-1}\left(|X(e^{j\omega})|^{0.5\gamma} |X(e^{j\omega})|^{0.5\gamma}\right) \\ &= F^{-1}\left(Y(e^{j\omega}) Y^c(e^{j\omega})\right) \\ &= y(n) * y(-n), \end{aligned} \tag{10}$$

where c and $*$ denote complex conjugation and convolution operations, respectively. Thus $|X(e^{j\omega})|$ can be expressed as the Fourier transform of the autocorrelation of some sequence $y(n)$. Basically, the root-cepstrum of any signal $x(n)$ can be thought of as the autocorrelation of some other sequence $y(n)$.

[1] Other values of γ, say $\gamma < 1$, are especially useful in formant and antiformant extraction from the speech signal when the dynamic range is very high.

3.2. Algorithm for segmentation

In [17, 20], it is shown that if the signal is of minimum phase, the group delay function resolves the peaks and valleys of the spectrum well. If the STE function is thought of as a magnitude spectrum, an equivalent minimum-phase signal can be derived, as explained in Section 3.1. The peaks and valleys of the group delay function of this signal will now correspond to the peaks and valleys in the STE function. In the STE function of any syllable, the energy is quite high in the voiced region and tapers off at both ends, where a consonant may be present, which results in local energy fluctuations. If these local variations are smoothed, then the minima at both ends of a voiced region correspond to syllable boundaries. The algorithm for segmentation of continuous speech using this approach is given below, which essentially smoothes the energy contour and removes the local energy fluctuations.

(i) Let $x(n)$ be the given digitized speech signal (Figure 3a) of a continuous speech utterance.

(ii) Compute the STE function $E(m)$, where $m = 1, 2, \ldots, M$ (Figure 3b), using overlapped windows. Let the minimum value of the STE function be $E_{\min}$.

(iii) Compute the order N of FFT as given below:

$$N = 2^{\lceil \log(2M)/\log(2) \rceil}. \tag{11}$$

(iv) Invert the function $E(m)^{\gamma}$ (where $\gamma = 0.001$) after appending $(N/2 - M)$ number of $E_{\min}$ to the sequence $E(m)$. Let the resultant function be $E^i(m)$ (Figure 3c).

(v) Construct the symmetric part of the sequence by producing a lateral inversion of this sequence about the Y-axis. Let this sequence be $E(K)$ (Figure 3d). Here, the sequence $E(K)$ is treated as the magnitude spectrum of some arbitrary signal. In this time-frequency substitution, N is replaced by 2π irrespective of the value of N.

(vi) Compute the IDFT of the sequence $E(K)$. This resultant sequence $e(n')$ is the root-cepstrum. The causal portion of $e(n')$ has the properties of a minimum-phase signal.

(vii) Compute the minimum-phase group delay function of the windowed causal sequence of $e(n')$ (see [17, 20]). Let this sequence be $E_{\text{gd}}(K)$. Let the size of the window applied on this causal sequence, that is, the size of the cepstral lifter, be N_c.

(viii) Detect the positive peaks in the minimum-phase group delay function ($E_{\text{gd}}(K)$) as given below. If $E_{\text{gd}}(K)$ is positive and if

$$E_{\text{gd}}(K-1) < E_{\text{gd}}(K) < E_{\text{gd}}(K+1), \tag{12}$$

then $E_{\text{gd}}(K)$ is considered as a peak. These peaks approximately correspond to the syllable boundaries.

As explained in Section 2, for a given speech signal $x(n)$ (Figure 4a), a group delay function may be derived in three different ways. The group delay function shown in Figure 4b

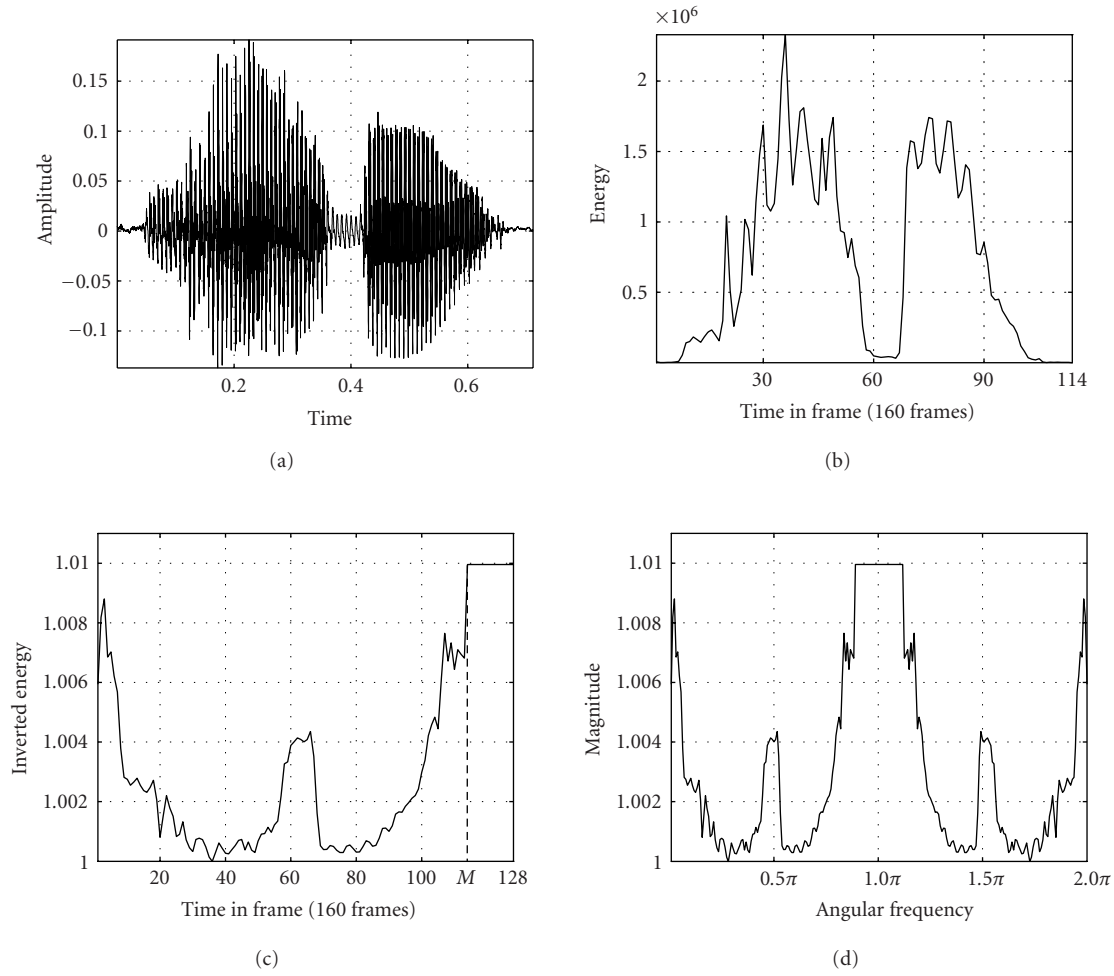

Figure 3: (a) Speech signal. (b) Corresponding STE function. (c) Inverted STE function. (d) Inverted and symmetrized STE function.

is derived using the root-cepstrum-based approach. The group delay functions derived using the other two methods, that is, cepstrum- and LP-cepstrum-based smoothing methods, are also given in Figures 4c and 4d, along with the group delay function derived using root-cepstrum-based smoothing. Interestingly, all the three group delay functions are almost similar, except for slight shifts in boundary locations in the case of LP-cepstrum based smoothing. But, each method has its own advantages and disadvantages. In the cepstrum- and root-cepstrum-based smoothing, the group delay functions are exactly similar in shape. But the computation of the conventional cepstrum requires a *log* operation. The common problem with these two methods is the choice of the cepstral lifter size N_c. Appropriate choices for this parameter are discussed in the next section. If we use LP-cepstrum-based method, the cepstral lifter size is not crucial and in fact, the whole causal portion of the cepstrum can be considered for prediction. Even though this seems to be very attractive, this method suffers from the fact that the choice of the predictor order is related to the number of boundaries.

3.3. Choice of N_c

The frequency resolution in the magnitude spectrum as well as in the group delay spectrum depends on the size of the cepstral lifter N_c applied in the root-cepstrum. Here, N_c is defined as

$$N_c = \frac{\text{Length of STE function}}{\text{WSF}}. \tag{13}$$

In (13), the length of the STE function corresponds to the number of samples in the STE function and the window scale factor (WSF) represents a scaling factor which is used to truncate the cepstrum. In this context, the value of WSF is always greater than 1. If N_c is high, the resolution will also be high, that is, it can resolve two closely spaced boundaries. If N_c is chosen to be high, a boundary will appear between CV/CVC at the CV transition. For syllable segmentation, this is undesirable. On the other hand, if the resolution is too low, even syllables will not be resolved, which is also not desirable. To choose N_c appropriately, durational analysis was performed. For this analysis, about 5000 speech dialogs of

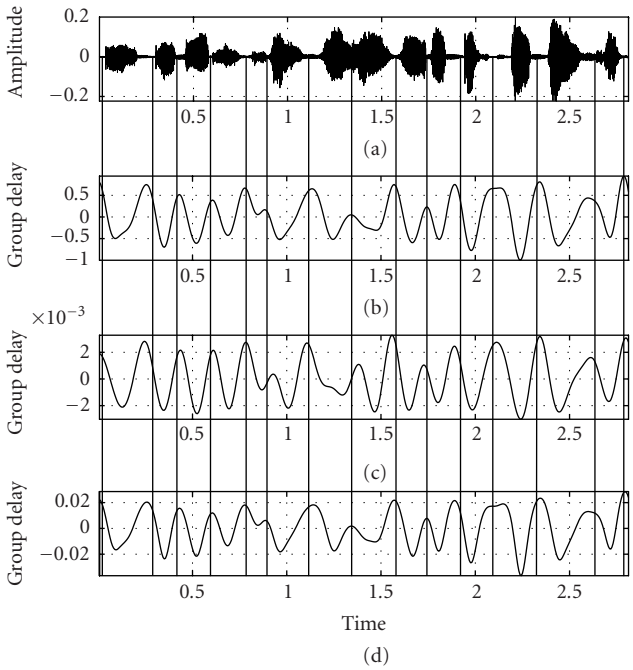

Figure 4: The utterance "group-delay-based segmentation" (an example) (a) Speech signal. (b) Group delay function derived from root-cepstrum. (c) Group delay function derived from LP-cepstrum. (d) Group delay function derived from conventional cepstrum.

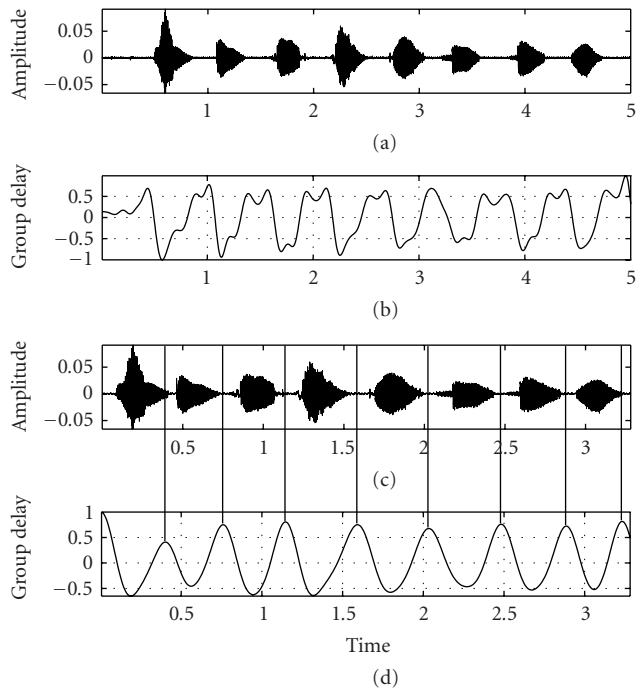

Figure 5: (a) Speech signal for the alphanumeric string "1258abdg" with silence. (b) Group delay function derived from the signal given in (a). (c) Speech signal after removing long silences. (d) Group delay function derived from the signal given in (c).

Table 1: Durational properties of syllables in Switchboard corpus (subset).

Duration (ms)	Syllables (%)
< 50	1.92
50–100	13.10
100–200	42.78
200–300	25.37
300–400	11.11
400–500	4.19
500–600	1.52

the Switchboard data [14] were considered. Table 1 gives the durations of a subset of syllables in Switchboard data. From this table, observe that the lengths of approximately 70% of the syllables vary from 100 milliseconds to 300 milliseconds. The mean duration of a subset of Switchboard syllable data is 201.2 milliseconds. For these different durations of syllables, the experiments show that the values for N_c can be within a fixed range. For example, a carefully uttered speech signal with different syllable durations from 75 milliseconds to 325 milliseconds is considered. The WSF is varied from 2 to 12. The analysis shows that when the WSF is varied from 4 to 10, the number of syllable boundaries detected is equal to the number of actual boundaries. Based on this experiment, the WSF in the computation of N_c can be set between 4 and 10. The number of samples in the STE function is directly related to the number of syllables present in the speech signal. In a few instances, the syllable duration may be more than 300 milliseconds or less than 100 milliseconds. If the syllable duration is more than 300 milliseconds, then that particular segment may be split into two segments. Similarly, if the syllable duration is less than 100 milliseconds, there is a chance that the syllable boundary is not resolved. But most importantly, other syllable boundaries remain unaffected.

4. SILENCES, FRICATIVES, AND SEMIVOWELS

The group delay function resolves even very closely spaced poles well when they are separated by a zero, provided the zero is located at approximately the same radius as that of the poles. In other cases, there may be some degradation in performance. Three possible places where failure may occur are (i) at the silence region, where the duration of the silence is considerable, (ii) at fricative segments, where the energy of the fricative is quite high, and (iii) at the semivowels, when it comes in the middle of any word. To overcome these problems, on advice from Greenberg at ICSI, a subband-based approach to syllable segmentation is attempted.

4.1. Presence of long silences

In this approach, since the symmetrized energy contour is inverted, any drastic energy reduction in between two syllables is considered as a pole in the z-domain and a positive peak in the group delay domain. But, for a long silence (see Figure 5a) in between two syllables (say more than

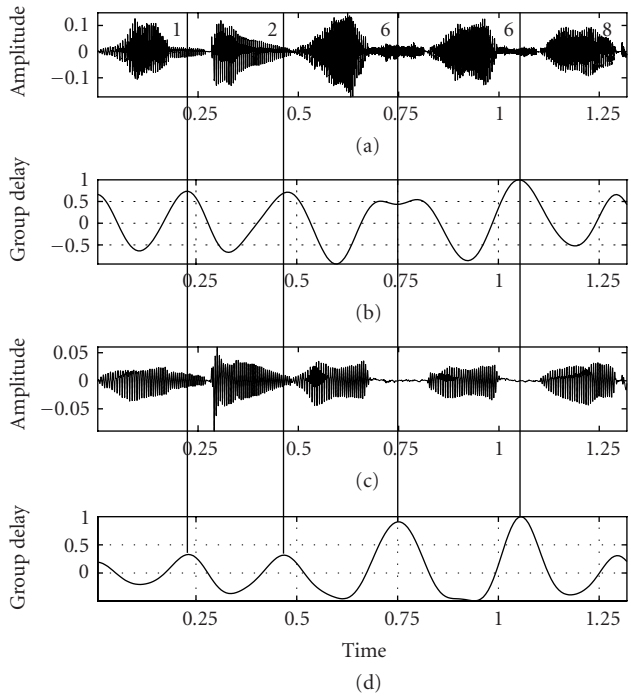

Figure 6: (a) Speech signal. (b) Group delay function derived from the signal given in (a). (c) Lowpass filtered (f_c = 500 Hz) signal given in (a). (d) Group delay function derived from the signal given in (c).

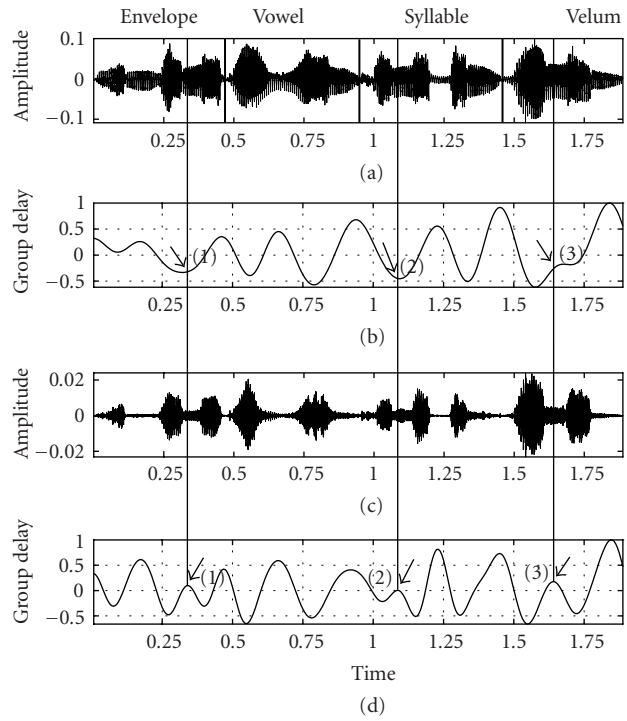

Figure 7: (a) Speech signal. (b) Group delay function derived from the signal given in (a). (c) Bandpass filtered (f_l = 500 Hz and f_u = 1500 Hz) signal given in (a). (d) Group delay function derived from the signal given in (c).

about 30 milliseconds), this rule may not apply. Instead we may get more than one boundary in the group delay domain, depending upon the resolution (Figure 5b). Syllable boundaries correspond to poles in the group delay domain. The long silence is equivalent to having two or more consecutive poles with identical radii. To overcome this problem, the silence segments present in the continuous speech, whose duration is high, namely, about 30 milliseconds, should be removed. Based on the knowledge derived from the energy, zero-crossing rate, and spectral flatness of a frame, the decision is made whether that frame of signal is silence or speech. If the duration of the silence is more than 30 milliseconds, that particular segment is removed from the signal (see Figure 5c) and then processed. The resultant peaks in the group delay spectrum now correspond to correct segment boundaries. This process reduces the spurious segment boundaries (Figure 5d).

4.2. Presence of fricatives

In the speech signal $x(n)$, if a fricative is present (Figure 6a), when we compute the energy function, a boundary will be generated at the middle of a fricative. This will be manifested in the group delay domain also, which is a spurious peak (see the 3rd and 4th peaks in Figure 6b). To avoid this, the signal, $x(n)$ is lowpass filtered to remove the high-frequency fricatives. Observe that the energy of the signal in the fricative regions is significantly reduced (Figure 6c). Consequently, in the group delay spectrum too, the spurious peak/boundary is removed (Figure 6d). This results in the segment boundary being slightly shifted. So the group delay function derived from this should not be considered as the reference. Nevertheless, it can be used to remove peaks due to fricatives in the original group delay (Figure 6d).

4.3. Presence of a semivowel

The semivowels are very similar to vowels in that they have periodic, intense waveforms with most of the energy in the low formants. Even though they are slightly weaker than vowels, if they come in the middle of a word in continuous speech, in most cases, a visible energy reduction may not be perceived (see Figure 7a). Because of this, in the group delay spectrum too, we may not get a boundary in between two vowels when they are separated by a semivowel (see the three vertical lines drawn in Figure 7 and the intersecting points (1), (2), and (3) in Figure 7b). For example, in the word *envelope*, since there is no significant energy reduction in between the syllables /ve/ and /lope/, in the group delay spectrum too, the peak is not present (see the intersecting point (1) in Figure 7b). If a suitable bandpass filter is applied to the original signal, since the energy of the semivowels are concentrated at low formants, the semivowels will be attenuated severely (see Figure 7c) without affecting the vowel regions much. This will ensure that a boundary will be present at the semivowel segment also (see the points/peaks (1), (2), and (3) in Figure 7d).

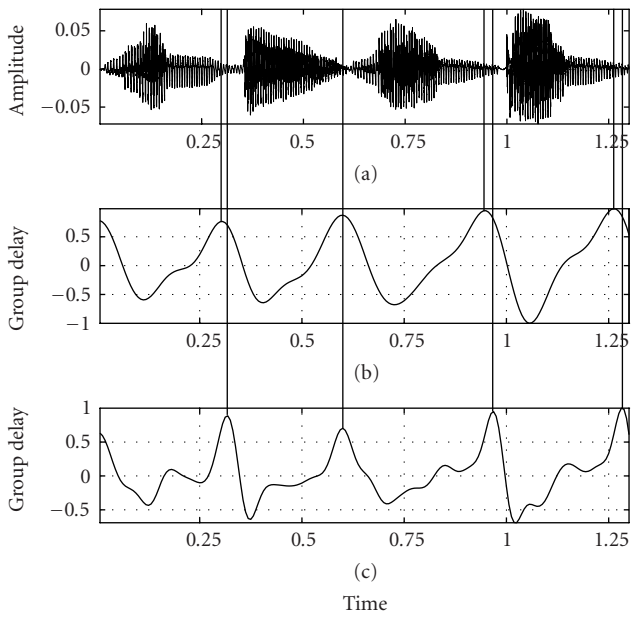

FIGURE 8: (a) Speech signal for the utterance of the digit string "1 2 9 10." (b) Group delay function derived from signal given in (a) with lower resolution (WSF = 2.5). (c) Group delay function derived from the signal given in (a) with higher resolution (WSF = 1.2).

4.4. Refining segment boundaries

The boundaries derived from the group-delay-based algorithm may have slight deviations from the actual boundaries, for example, in the nasal-consonant regions (Figures 8a and 8b). This is due to lower resolution. If the resolution of the group delay spectrum is increased by increasing the cepstral lifter size N_c applied in the cepstral domain, a spurious segment is observed at the beginning of a nasal consonant. Nevertheless, when the resolution is increased, the error in the segment boundary is small (Figure 8c). Each boundary location in the lower-resolution group delay spectrum is compared with all the peaks in the higher-resolution group delay spectrum and the nearest peak is considered as the actual segment boundary.

4.5. Combining evidence

Instead of using the group delay function derived from the STE function of the original signal alone, here, the speech signal is passed through a bank of three filters. The group delay function of the outputs of each of these three filters is computed. The basic steps involved in this approach for segmenting the speech signal at syllable-like units is given in the block diagram (Figure 9). The boundaries derived from the different group delay functions are combined using the following logic:

$$P_{\tau_{\text{al}}} = P^{i}_{\tau_{\text{ap}}} \tag{14}$$

if $(P^{i}_{\tau_{\text{ap}}} \sim P^{j}_{\tau_{\text{lp}}}) \leq 20$ milliseconds for each peak "i" in $P_{\tau_{\text{ap}}}$ and for each peak "j" in $P_{\tau_{\text{lp}}}$;

$$P_{\text{temp}} = P^{j}_{\tau_{\text{bp}}} \tag{15}$$

if $50 \leq (P^{i}_{\tau_{\text{al}}} \sim P^{j}_{\tau_{\text{bp}}}) \leq 100$ milliseconds for each peak "i" in $P_{\tau_{\text{al}}}$ and for each peak "j" in $P_{\tau_{\text{bp}}}$;

$$P_{\tau_{\text{alb}}} = P_{\tau_{\text{al}}} \vee P_{\text{temp}}, \tag{16}$$

$$P_{\tau_{\text{albm}}} = P_{\tau_{\text{m}}} \tag{17}$$

if $(P_{\tau_{\text{alb}}} \sim P_{\tau_{\text{m}}}) \leq 30$ milliseconds, where

(i) $\vee$ represents "OR" operation and $\sim$ represents the difference operation (i.e., only magnitude of the time difference is considered);

(ii) $P_{\tau_{\text{ap}}}$—boundaries derived from the allpass signal;

(iii) $P_{\tau_{\text{lp}}}$—boundaries derived from the lowpass filtered signal;

(iv) $P_{\tau_{\text{bp}}}$—boundaries derived from the bandpass filtered signal;

(v) $P_{\tau_{\text{m}}}$—boundaries derived from the higher-resolution group delay function;

(vi) $P_{\tau_{\text{al}}}$—boundaries derived from allpass and lowpass filtered signals after combining;

(vii) $P_{\tau_{\text{alb}}}$—boundaries derived from allpass, lowpass, and bandpass filtered signals after combining;

(viii) $P_{\tau_{\text{albm}}}$—boundaries derived from allpass, lowpass, and bandpass filtered signals and from the higher-resolution group delay function after combining.

For example, the speech signal for the utterance "group-delay-based segmentation" (Figure 10a) is considered to describe the method of combining evidence. First, the silences in between the syllables, if any, are removed. In Figure 10, the solid vertical lines drawn between Figures 10b and 10c denote the segment boundaries detected after combining the evidence from the group delay functions of allpass and lowpass filtered speech signals using (14). The dashed line between Figures 10b and 10c (labeled as "1") denotes the spurious boundary, which is removed after combining. The solid vertical line drawn between Figures 10b and 10d denotes the new boundary detected after combining the evidence from the group delay functions of allpass and bandpass filtered speech signals using (15) and (16). The dotted vertical lines drawn from Figure 10e to Figure 10a denote the boundaries detected after refinement (see (17)) using the higher-resolution group delay function derived from the allpass filtered signal. Observe that a spurious segment boundary produced at the fricative region is removed after lowpass filtering the signal and a new boundary is detected (as indicated by the solid vertical line with label "2") in between the syllables /de/ and /lay/ because of bandpass filtering the signal.

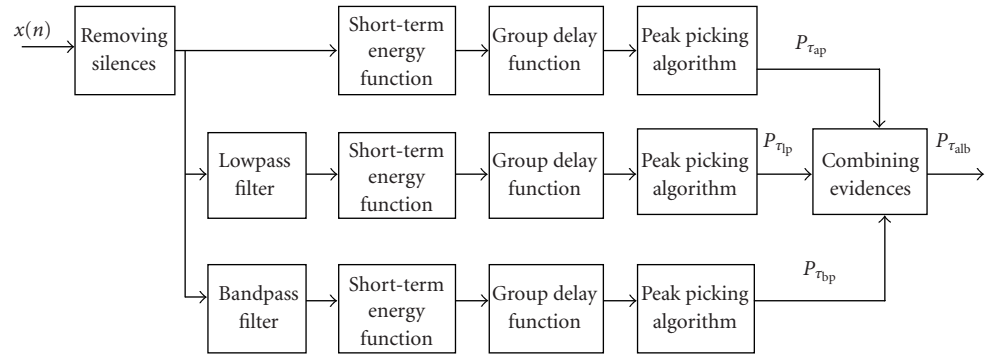

Figure 9: Block diagram of subband-based approach.

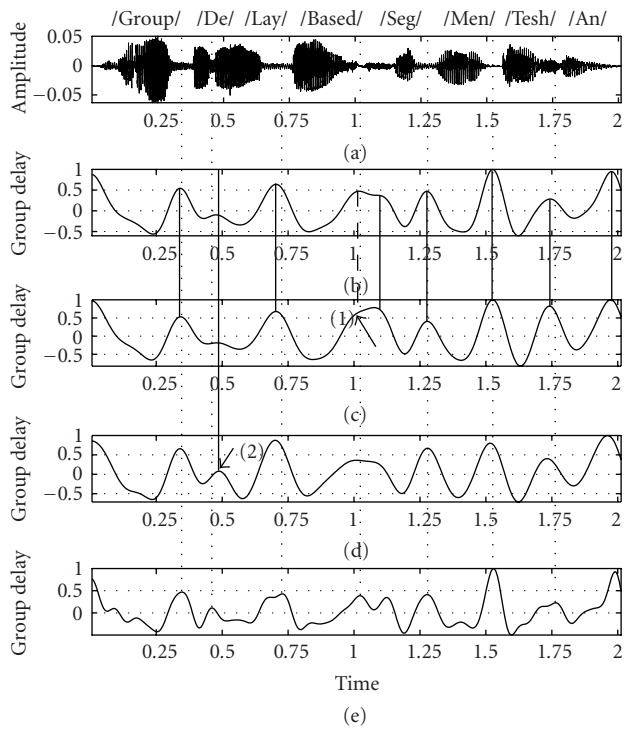

Figure 10: (a) Speech signal. (b) Group delay function derived from the allpass filtered signal. (c) Group delay function derived from lowpass filtered (f_c = 500 Hz) signal. (d) Group delay function derived from the bandpass filtered (f_l = 500 Hz and f_u = 1500 Hz) signal. (e) Group delay function with higher resolution derived from the allpass filtered signal.

5. EXPERIMENTS AND RESULTS

5.1. Speech corpora

The Switchboard corpus [14] and OGI-MLTS corpus [21] are used for analyzing the performance of our system. Switchboard is a corpus of several thousands of informal speech dialogs recorded over the telephone. For our analysis, a portion of the corpus, which has syllable-level transcription, is considered. For these speech dialogs, syllable-level transcription [13] is also provided in this corpus. The duration of the speech signals varies from 0.5 seconds to 25 seconds. In OGI-MLTS, 40 speech files uttered by 40 different speakers are considered for the analysis. In this subset, each file is of 45 seconds duration. These files are manually segmented into syllabic units and used as a reference to verify the performance of our segmentation approach.

5.2. Experimental setup

Prior to automatic segmentation, the speech signals are first preprocessed by removing the long silences (if any) as explained in Section 4.1. For the computation of STE function, overlapped rectangular windows are used, where the window length is of duration 20 milliseconds and the overlap is of 10 milliseconds duration. Further, the value of γ in $E(m)^{\gamma}$ is set to 0.001 to reduce the dynamic range of the STE function, irrespective of the speech corpus considered. In fact, any value of $\gamma < 0.01$ has been found to be appropriate. As defined in Section 3.3, the WSF used to compute the size of the Hanning window (cepstral lifter size N_c) is set to 4.0. Since the value of the WSF is fixed, the length of the root-cepstrum is proportional to the length of the STE function. Three different group delay functions are computed from (a) the original speech signal (allpass filtered), (b) lowpass filtered (f_c = 500 Hz) speech signal, and (c) bandpass filtered (f_l = 500 Hz and f_u = 1500 Hz) speech signal. The evidence derived from these group delay functions are combined as explained in Section 4.5. In order to see the effect of each of the group delay function in the performance of the final system, four different experiments are carried out separately on the Switchboard corpus and the results are tabulated (see Table 2). In all these experiments, a boundary is said to be detected if the error between an automatic segmentation boundary and manual segmentation boundary is less than 80 milliseconds. Based on the error, four different categories are observed (see 1st column of Table 2). In each of these four categories, the performance is calculated by computing the ratio between the number of boundaries in each category and the total number of automatically detected boundaries.

From Table 2, observe that, for the baseline system ($P_{\tau_{ap}}$ alone), the number of insertions and deletions are very high (see 5th and 6th rows of 2nd column). The number of insertions is considerably reduced when $P_{\tau_{ap}}$ and $P_{\tau_{lp}}$ are

Table 2: Performance (%) of different experiments.

Error (ms)	$P_{\tau_{ap}}$	$P_{\tau_{al}}$	$P_{\tau_{alb}}$	$P_{\tau_{albm}}$
< 25	64.08	63.8	63.39	66.81
25–40	12.82	12.66	13.35	18.10
40–60	7.49	7.69	7.75	11.21
60–80	15.6	15.82	15.5	3.88
Insertion	8.14	5.64	5.80	5.96
Deletion	10.72	10.88	6.45	6.53

Table 3: Performance (%) of the group-delay-based segmentation approach. (A) Switchboard corpus. (B) OGI-MLTS.

Error (ms)	A	B
< 25	66.81	76.58
25–40	18.10	9.62
40–60	11.21	7.86
60–80	3.88	5.94
Insertion	5.96	5.02
Deletion	6.53	4.38

combined (see 3rd column) and the number of deletions is also reduced when $P_{\tau_{ap}}$, $P_{\tau_{lp}}$, and $P_{\tau_{bp}}$ are combined (see 4th column of Table 2). The error in segmentation boundaries are found to be greatly reduced when $P_{\tau_{alb}}$ are combined with P_{τ_m} (see 5th column). The performance of the final system on Switchboard data is compared with the performance on OGI data (see Table 3). The performance of the final system on Tamil data in OGI corpus is found to be better than that of Switchboard corpus. The better performance for the language Tamil may be due to its simple syllable structure.

6. DISCUSSION

After several refinements, the performance of the segmentation algorithm is reasonably better than that of the baseline system described in Section 3. But still, there are some issues which are yet to be addressed. For example, the knowledge derived from the durational analysis can be incorporated into the system for reducing the number of insertions and deletions of syllable boundaries, which is yet to be done. The major problem in our approach is with syllables whose durations (D_s) are out of range, that is, when $D_s < 100$ milliseconds or $D_s > 300$ milliseconds. For this particular case, even the durational knowledge will not be of any help. This problem can be handled if the phonetic transcription is available or at least if the number of syllables present in the signal is known a priori. In our approach, the silence regions are detected and removed from the signal in the preprocessing stage itself. The STE function, zero-crossing rate, and spectral flatness measure are used with proper thresholds for silence detection. This method may suffer when analyzing signals with very low SNR.

As such, this segmentation approach is successfully used in two different tasks, namely, spoken language identification and automatic speech transcription, as described below. A syllable-level, unsupervised, and incremental clustering procedure is proposed in [22] for spoken language identification. In this work, for each language, a syllable inventory is created by first segmenting the speech signal into syllable-like units. Similar syllable segments are then clustered using an incremental approach which results in a set of syllable models for each language. These language-dependent syllable models are then used for language identification. Further, in [23], the same approach is extended for the speech transcription task. This segmentation algorithm can also be used in a real-time speech recognition system by considering a few syllables (say a phrase) at a time.

7. CONCLUSIONS

In this paper, a novel approach for segmenting the speech signal into syllable-like units is presented. Several refinements are suggested for improving the segmentation performance. The performance of the minimum-phase group-delay-function-based segmentation approach, before and after refinements, is tested on Switchboard and OGI corpora. When compared with the performance of the baseline system, there is a considerable reduction in segmentation errors and the number of insertions and deletions. The advantage of segmentation prior to labeling in speech is that it can be independent of the task. Simple isolated syllable models can be built from the segmented data. Once syllable sequences are available, appropriate postprocessing can be done to build systems for specific tasks.

ACKNOWLEDGMENTS

The authors are grateful to Steven Greenberg at ICSI, Berkeley, for making available a portion of the Switchboard corpus with syllable-level transcription. To overcome the problems with the baseline group-delay-based segmentation technique, Steven Greenberg suggested the subband-based approach described in Section 4.

REFERENCES

[1] S.-L. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 2, pp. 721–724, Seattle, Wash, USA, May 1998.

[2] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 82–87, 1975.

[3] V. K. Prasad, *Segmentation and recognition of continuous speech*, Ph.D. dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India, 2002.

[4] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[5] O. Schmidbauer, "Syllable-based segment-hypotheses generation in fluently spoken speech using gross articulatory features," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '87)*, vol. 12, pp. 391–394, Dallas, Tex, USA, April 1987.

[6] S. Nakagawa and Y. Hashimoto, "A method for continuous speech segmentation using HMM," in *Proc. IEEE 9th International Conference on Pattern Recognition*, vol. 2, pp. 960–962, Rome, Italy, November 1988.

[7] A. Noetzel, "Robust syllable segmentation of continuous speech using neural networks," in *Electro International*, pp. 580–585, New York, NY, USA, April 1991.

[8] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. 4th International Conf. on Spoken Language Processing (ICSLP '96)*, vol. 2, pp. 1261–1264, Philadelphia, Pa, USA, October 1996.

[9] L. Shastri, S. Chang, and S. Greenberg, "Syllable detection and segmentation using temporal flow neural networks," in *Proc. 14th International Congress of Phonetic Science (ICPhS '99)*, pp. 1721–1724, San Francisco, Calif, USA, August 1999.

[10] S. Greenberg, "Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2–4, pp. 159–176, 1999.

[11] V. K. Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, no. 3-4, pp. 429–446, 2004.

[12] T. Nagarajan, H. A. Murthy, and R. M. Hegde, "Segmentation of speech into syllable-like units," in *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 2893–2896, Geneva, Switzerland, September 2003.

[13] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proc. 4th International Conf. on Spoken Language Processing (ICSLP '96)*, pp. 24–27, Philadelphia, Pa, USA, October 1996.

[14] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '92)*, vol. 1, pp. 517–520, San Francisco, Calif, USA, March 1992.

[15] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[16] J. S. Lim, "Spectral root homomorphic deconvolution system," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.

[17] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Communication*, vol. 10, no. 3, pp. 209–221, 1991.

[18] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[19] A. Papoulis, *Signal Analysis*, McGraw-Hill International Editions, Singapore, 1984.

[20] H. A. Murthy, "The real root cepstrum and its application to speech processing," in *Proc. National Conference on Communications*, pp. 180–183, IIT Madras, Chennai, India, January 1997.

[21] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. International Conf. on Spoken Language Processing (ICSLP '92)*, pp. 895–898, Banff, Alberta, Canada, October 1992.

[22] T. Nagarajan and H. A. Murthy, "Language identification using parallel syllable-like unit recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '04)*, vol. 1, pp. 401–404, Montreal, Quebec, Canada, May 2004.

[23] T. Nagarajan and H. A. Murthy, "An approach to segmentation and labeling of continuous speech without bootstrapping," in *Proc. National Conference on Communications*, pp. 508–512, IISc, Bangalore, India, January 2004.

T. Nagarajan received the B.S. and M.S. degrees in physics from Bharathidasan University, Tamil Nadu, India, in 1985 and 1987, respectively, the A.M.I.E. degree in electronics and communication from the Institution of Engineers, India, in 1988, the M.E. degree in microwave and optical engineering from Madurai Kamaraj University, Tamil Nadu, India, in 1991, and the Ph.D. degree from the Indian Institute of Technology, Madras, India, in 2004. From 1991 to 1999, he held various positions in Periyar Maniammai College of Technology for Women. Since January 2004, he has been with TeNet Group, Department of Electrical Engineering, Indian Institute of Technology, Madras, as a Senior Project Officer. His special interest is in the area of speech processing.

H. A. Murthy received her B.E. degree in electronics and communications engineering from Osmania University, Hyderabad, India, in 1980, an M.E. degree in electrical and computer engineering from McMaster University, Canada, in 1986, and a Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras (IIT-M), India, in 1992. From 1980 through 1983, she was a Scientific Officer at the Speech and Digital Systems Group, Tata Institute of Fundamental Research, Bombay, India. In 1988, she joined the faculty of the Department of Computer Science and Engineering, IIT-M, India. During the year 1995–1996, she was a Postdoctoral Fellow at the Speech Technology and Research Laboratory, SRI International.