Segmentation of speech into syllable-like units

T.Nagarajan, Hema A. Murthy and Rajesh M. Hegde

Department of Computer Science and Engineering
Indian Institute of Technology, Madras
raju@lantana.iitm.ernet.in

Abstract

In the development of a syllable-centric ASR system, segmentation of the acoustic signal into syllabic units is an important stage. This paper presents a minimum phase group delay based approach to segment spontaneous speech into syllable-like units. Here, three different minimum phase signals are derived from the short term energy functions of three sub-bands of speech signals, as if it were a magnitude spectrum. The experiments are carried out on Switchboard and OGI-MLTS corpus and the error in segmentation is found to be utmost 40msec for 85% of the syllable segments.

1. Introduction

One of the major reasons for considering syllable as a basic unit for 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 the effective minimal unit in the time-domain. Since then, several ASR systems have been developed for different languages, most recently for Indian languages [3]. In [3], it is demonstrated that segmentation at syllable-like units followed by isolated style recognition of continuous speech performs well.

Most of the existing speech segmentation techniques are based on modeling each of the sub-word unit, usually by HMMs. The subword unit considered in those techniques is mostly phonemes. Only very few works have been done on segmenting the speech into syllabic units. In [4], for Japanese language, 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 sequence. In [5], a temporal flow model (TFM) network has been developed to extract syllable boundary information from continuous speech, where the TFM captures the time varying properties of the speech.

The simple candidate for segmenting speech is the short-term energy function of the speech signal. But, the basic problems with the short-term energy function based segmentation are thresholding and local energy fluctuations due to the presence of consonants. To overcome these problems, instead of directly using the short-term energy function, in our laboratory, we have proposed a method [6] for segmenting the acoustic signal into syllable-like units, in which we derive a minimum phase signal from the short term energy function as if it were a magnitude spectrum. We have found that the group delay function of this minimum phase signal is a better representative of the short term energy function to perform segmentation.

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

$$H(\omega) = H_1(\omega).H_2(\omega) \quad and, \tag{1}$$

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

$$\tau_h(\omega) = -\partial(arg(H(\omega)))/\partial\omega \tag{2}$$

$$= \tau_{h1}(\omega) + \tau_{h2}(\omega) \tag{3}$$

From eqns. 1 and 3, we see that a multiplication in the spectral domain becomes an addition in the group delay domain. 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 magnitude spectrum), we can show 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 at syllable-like entities.

In Section 2, we briefly discuss the properties of the signal derived from the magnitude spectrum. In section 3, we discuss the proposed group delay based approach for segmenting the speech signal.

2. 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 z-transform of the square of the magnitude of the system frequency response is given by

$$C(z) = \frac{1}{\Pi_{i=1}^{N}(1 - a_i e^{jw_i} z)(1 - a_i e^{jw_i} z^{-1})} \tag{5}$$

From eqn.5, we can infer that, for each pole of X(z), there is a pole of C(z) at a_i and $\frac{1}{a_i^*}$. Consequently, if one element of each pair is outside the unit circle, then the conjugate reciprocal will be inside the unit circle [7]. Since Fourier transform of eqn.5 exists, the causal protion of the inverse Z-transform of eqn.5 is given by,

$$c_m(n) = \sum_{i=1}^{N} B_i (a_i e^{jw_i})^n u(n) \tag{6}$$

From eqn.6, 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 '$\frac{1}{a_i}$' , with $|a_i| < 1$, will have a root at 'a_i', ie, the resultant signal will always be a minimum phase

signal. But since the duration of the causal portion of the root cepstrum is finite, the z-transform of the signal will have spurious zeros. These zeros may affect the positions of the actual zeros present in the signal. Hence to overcome this problem, the squared magnitude spectrum can be inverted $(1/(|X(e^{j\omega})|^2)$ and another minimum phase signal 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]. Then, $|X(e^{j\omega})|$ 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 $y(n)$.

3. Group delay based segmentation of speech

3.1. Base line system

In [8, 9] it is shown that if the signal is minimum phase, the group delay function resolves the peaks and valleys of the spectrum well. If the short-term energy function is thought of as a magnitude spectrum, an equivalent minimum phase signal can be derived, as explained in Section. 2. The peaks and valleys of group delay function of this signal will now correspond to the peaks and valleys in the short-term energy function. In general, the number of syllables present in a speech signal is equal to number of voiced segments. In the short-time energy function of any syllable as a segment, the energy is quite high in the voiced region and tapers down at both the ends, where a consonant may be present, which results in local energy fluctuations. If these local variations are smoothed, then the valley points at both the ends of a voiced region can be cosidered as syllabe boundary. The algorithm for segmentation of continuous speech using this approach is given below, which essentially smoothens the energy contour and removes the local energy fluctuations.

- Let $x(n)$ be the given digitized speech signal of a continuous speech utterance.
- Compute the short term energy function $E(n)$, using overlapped windows. Since this is viewed as an arbitrary magnitude spectrum, let it be denoted as $E(K)$
- Construct the symmetric part of the sequence by producing a lateral inversion of this sequence about the Y-axis. Let this sequence be $\tilde{E}(K)$.
- Invert the sequence $\tilde{E}(K)$, since our interest is in the valleys, which are supposed to be the syllable boundaries. Let the resultant sequence be $\tilde{E}^i(K)$.
- Compute the inverse DFT of the sequence $\tilde{E}^i(K)$. This resultant sequence $\tilde{e}(m)$, is the root cepstrum and the causal portion of it resembles the properties of the minimum phase signal.
- Compute the minimum phase group delay function of the windowed causal sequence of $\tilde{e}(m)$ ([9, 8]). Let this sequence be $\tilde{E}_{gd}(K)$. Here the window size(N_c) applied in the root cepstrum is chosen as,

$$N_c = \frac{Length\ of\ short\ term\ energy\ sequence}{Window\ scale\ factor(WSF)} \quad (7)$$

- The location of the peaks in the minimum phase group delay function $\tilde{E}_{gd}(K)$ approximately correspond to the sub-word/syllable boundaries.

Fig.1 demonstrates the segmentation of a spontaneous speech signal at syllable boundaries. The manually marked boundaries are indicated by solid lines, while the group delay boundaries are indicated by dotted lines. The spurious boundaries are indicated by dashed lines. From Fig.1, it can be noticed that most of the syllable boundaries detected by our approach nearly coincide with the manually marked boundaries.

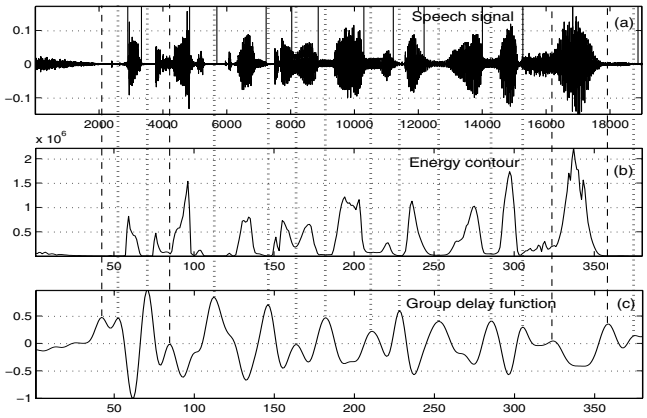

Figure 1: *Group-delay based segmentation - an example*

3.2. Problems and their remedies

The group-delay function resolves even very closely spaced poles well when they are separated by a zero, provided the radius of the zero is also close to radii of the adjacent poles. In other cases, there may be some degradation in the performance. Three possible places where failure may occur are, (i) at the silence region, where the duration of the silence is considerable. (ii) at the fricative segments, where the energy of the fricative is quite high and (iii) at semivowels, when it comes in the middle of any word. To overcome these problems, on advice from Steven Greenberg at ICSI [10], a sub-band based approach to syllable segmentation is attempted.

3.2.1. Presence of long-silences

In our 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, the presence of long-silence in between, say more than about 30msec., this rule may not apply, instead we may get more than one peak in the group delay domain, depending upon the resolution (Fig.2b). This is due to the presence of more than one consecutive poles. To avoid this problem, the silence segment present in the continuous speech whose duration is more than 30msec 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 a silence or speech. If the duration of the silence is more than 30msec, that particular segment is removed from the signal and then processed. The resultant peaks in the group delay spectrum are the proper segment boundaries. This process will reduce the spurious segment boundaries (Fig.2d)

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

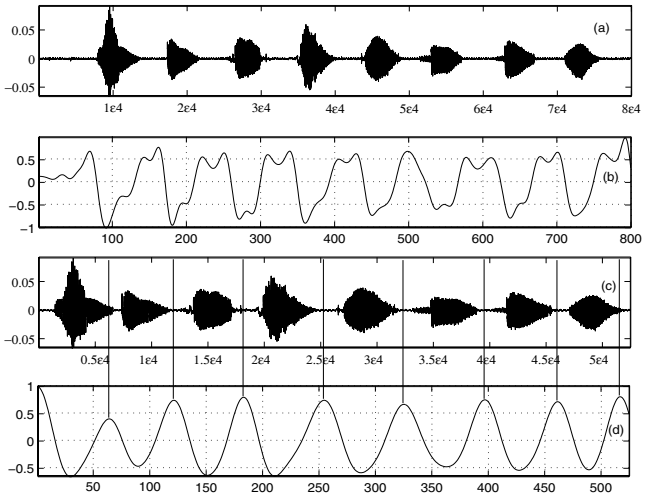

Figure 2: (a) Speech signal with silence (b) Group delay function derived from signal given in (a) (c) Speech signal after removing long silences (d) Group-delay function derived from signal given in (c).

3.2.2. Presence of fricatives

In the speech signal, $x(n)$ if a fricative is present, when we compute the energy function, we may get a peak at those segments. This will be manifested in the group-delay domain also, which is a spurious peak (Fig.3b). To avoid this, the signal, $x(n)$ is low-pass filtered to remove the high frequency fricatives.

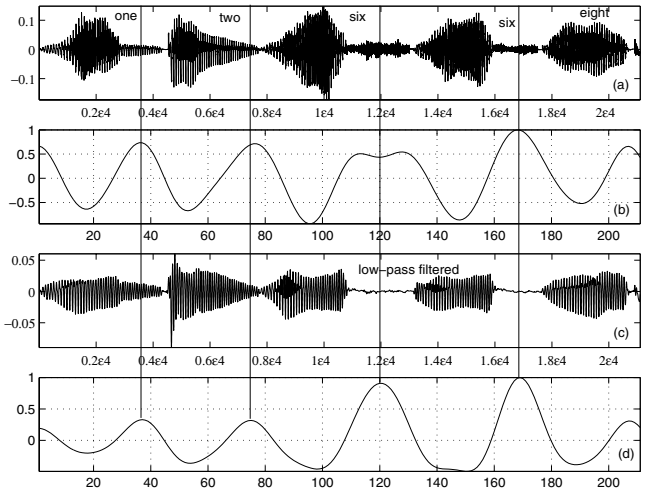

Figure 3: (a) Speech signal (b) Group-delay function derived from the signal given in (a) (c) Low-pass filtered signal given in (a) (d) Group-delay function derived from signal given in (c).

Because of this, the segment boundary may be slightly shifted. So the group delay function derived from this should not be considered as the reference. But it is used to ensure whether the peak present in the group delay function derived from the original signal is because of the fricative or not (Fig.3d).

3.2.3. Presence of a semivowel

The semivowels are very similar to vowels in that they have periodic, intense waveforms with most energy in the low formants. Eventhough they are slightly weaker than vowels, if they come in the middle of a word in the continuous speech, in may places, there may not be a visible energy reduction. Because of this, in the group delay spectrum also, we may not get a peak in between two vowels when they are separated by a semivowel (Fig.4b). If we apply a suitable band-pass filter to the original signal, since the energy of the semivowels are concentrated at low formants, the semivowels will be attenuated severely without affecting the vowel regions much. This will ensure that a peak will be present at semivowel segment also (Fig.4d).

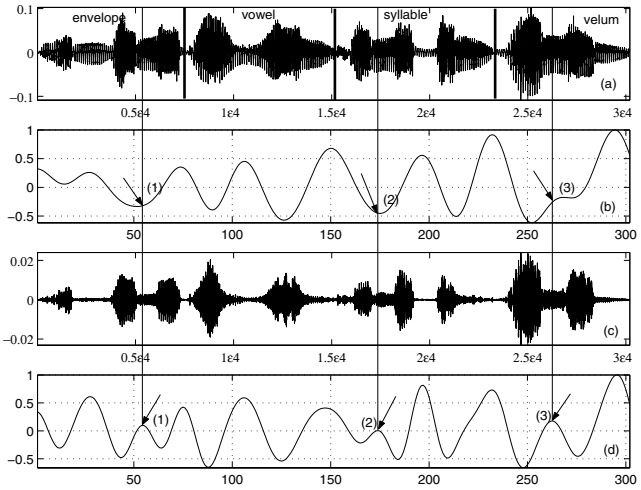

Figure 4: (a) Speech signal (b) Group-delay function derived from the signal given in (a) (c) Band-pass filtered signal given in (a) (d) Group-delay function derived from signal given in (c).

3.2.4. Refining segment boundaries

The boundaries derived from the sub-band based algorithm may have slight deviation from the actual boundaries, mostly in the nasal consonant regions (Fig.5b). This is due to the lower resolution. If the resolution of the group delay spectrum is increased by increasing the window size applied in the cepstral domain, we will get a spurious segment at the begining of a nasal consonant. But, when the resolution is more, the error in the segment boundary is less (Fig.5c). 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 peak.

3.2.5. Combining evidences

Instead of using the group delay function derived from the short term energy function of the original signal alone, here, three group delay functions are derived, each derived from three subbands of the original speech signal. The basic steps involved in this approach for segmenting the speech signal at syllable-like units is given in the block diagram (Fig.6). The peaks derived

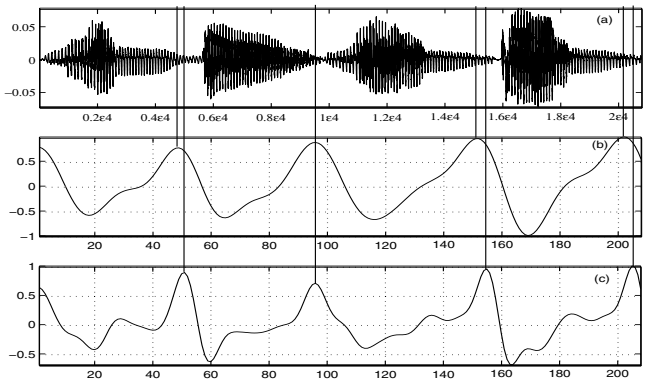

Figure 5: (a) Speech signal (b) Group-delay function derived from signal given in (a) with lower resolution (d) Group-delay function derived from signal given in (a) with higher resolution.

from the different group-delay functions are combined using the following logic.

$$P_{\tau_{al}} = P^{i}_{\tau_{ap}} \tag{8}$$

if $(P^{i}_{\tau_{ap}} \sim P^{j}_{\tau_{lp}}) \leq$ 20msec, for each peak 'i' in $P_{\tau_{ap}}$ and for each peak 'j' in $P_{\tau_{lp}}$.

$$P_T = P^{j}_{\tau_{bp}} \tag{9}$$

if 50msec $\leq (P^{i}_{\tau_{al}} \sim P^{j}_{\tau_{bp}}) \leq$ 100msec, for each peak 'i' in $P_{\tau_{al}}$ and for each peak 'j' in $P_{\tau_{bp}}$.

$$P_{\tau_{alb}} = P_{\tau_{al}} \vee P_T \tag{10}$$

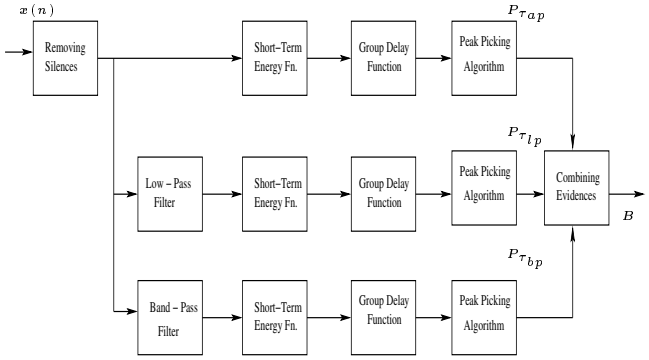

Figure 6: Block-diagram of sub-band based approach.

3.3. Performance

The Switchboard corpus and OGI-MLTS corpus are used for analyzing the performance of our system. Switchboard is a corpus of several hundred informal speech dialogs recorded over the telephone. For our analysis, we have considered over 5000 speech dialogs. The duration of the speech signals varies from 0.5sec to 25sec. In OGI-MLTS, only Tamil language 45sec utterances are manually segmented into syllabic units and is used to check the performance of our segmentation approach (see Table.1).

4. Conclusions

In this paper, a novel approach for segmenting the speech signal into syllabic units is presented. Several refinements are also suggested for improving the segmentation performance. The performance of the minimum phase group delay function based segmentation approach is tested on Switchboard and OGI corpus and it is found to be quite satisfactory. 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 post-processing can be done to build systems for specific task.

error (in msec)	A	B
<25	66.93	76.58
25 - 40	18.46	9.62
40 - 60	12.54	7.86
60 - 80	2.05	5.94
insertion	5.25	5.02
deletion	7.10	4.38

Table 1: Performance (in %) of the group delay based segmentation approach (A) Switchboard Corpus. (B) OGI-MLTS

5. References

[1] Su_Lin Wu, Brian E. D. Kingsbury, Nelson Morgan and Steven Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Seatle, W A, May 1998, pp. 721–724.

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

[3] Kamakshi V. Prasad, *Segmentation and Recognition of Continuous Speech*, PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, 2002.

[4] Seiichi Nakagawa and Yasuhide Hashimoto, "A method for continuous speech segmentation using hmm," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1988, pp. 960–962.

[5] Logendra shastri, Shuangyu chang, and Steven Greenberg, "Syllable detection and segmentation using temporal flow neural networks," in *ICPhS*, Sanfrancisco, 1999, pp. 1721–1724.

[6] T.Nagarajan, V.Kamakshi Prasad and Hema A.Murthy, "The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation," in *Sixth Biennial Conference of Signal Processing and Communications*, July 2001.

[7] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-time signal processing*, Prentice Hall, 2000.

[8] Hema A. Murthy and B Yegnanarayana, "Formant extraction from minimum phase group delay function," in *Speech Comm.*, August 1991, vol. 10, pp. 209–221.

[9] Hema A. Murthy, "The real root cepstrum and its applications to speech processing," in *National Conference on Communication*, IIT Madras, Chennai, India, January 1997, pp. 180–183.

[10] Steven Greenberg, "Private communication," May - July 2002.