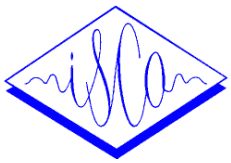

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

GROUP DELAY BASED SEGMENTATION OF SPONTANEOUS 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, Chennai. 600 036.
hema@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 corpus and the error in segmentation is found to be utmost 40msec for 85% of the syllable segments, in addition to 5.25% insertions and 7.10% deletions.

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.

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 the local variations in the energy function. To overcome these problems, instead of directly using the short-term energy function, in our laboratory, we have proposed a method [4] 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. To demonstrate the power of the additive property of group delay spectrum, three different systems are chosen, 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 (Figs.1(b), 1(e) and 1(h)), it is observed that even though the peaks in Fig.1(b) and Fig.1(e) are clearly visible, in a system where these two poles are combined together, the peaks are not resolved well as shown in Fig.1(h). This is due to the multiplicative property of the magnitude spectra. But from Fig.1(c), Fig.1(f) and Fig.1(i), it is evident that the group delay spectrum obtained by combining the poles together, the peaks are well resolved as shown Fig.1(i). 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.

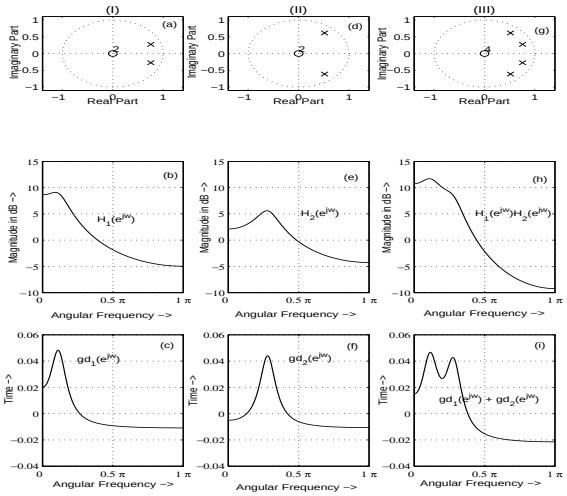

Fig. 1. Resolving power of group delay spectrum: z-plane, magnitude spectrum and group delay spectrum of the cases when I) a pole inside the unit circle at $(0.8, \pi/8)$, II) a pole inside the unit circle at $(0.8, \pi/4)$ and III) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$, inside the unit circle.

2. 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

$$|X(e^{j\omega})|^2 = X(e^{j\omega})X^*(e^{j\omega})$$

$$= X(z)X^*(1/z^*)|_{z=e^{j\omega}} \tag{5}$$

Let

$$C(z) = X(z)X^*(1/z^*)$$

$$C(z) = \frac{1}{\Pi_{i=1}^{N}(1 - a_i e^{jw_i} z)(1 - a_i e^{jw_i} z^{-1})} \tag{6}$$

From eqn.6, 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 [5]. Since Fourier transform of eqn.6 exists, inverse Z-transform of eqn.6 leads to :

$$c(n) = \sum_{i=1}^{N} \left[A_i (a_i e^{jw_i})^{-n} u(-n-1) + B_i (a_i e^{jw_i})^n u(n) \right] \tag{7}$$

If we consider only the causal portion of $c(n)$, then,

$$c_m(n) = \sum_{i=1}^{N} B_i (a_i e^{jw_i})^n u(n) \tag{8}$$

From eqn.8, 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 [6, 7] 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, but with local variations. 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.

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

[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.

- 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 $\tilde{e}(m)$ ([7, 6]). Let this sequence be $\tilde{E}_{gd}(K)$.
- 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.2 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.2, it can be noticed that most of the syllable boundaries detected by our approach nearly coincide with the manually marked boundaries. But it fails considerably in non-speech regions, for example the first segmentation point, whose duration is quite high and at fricative speech segments.

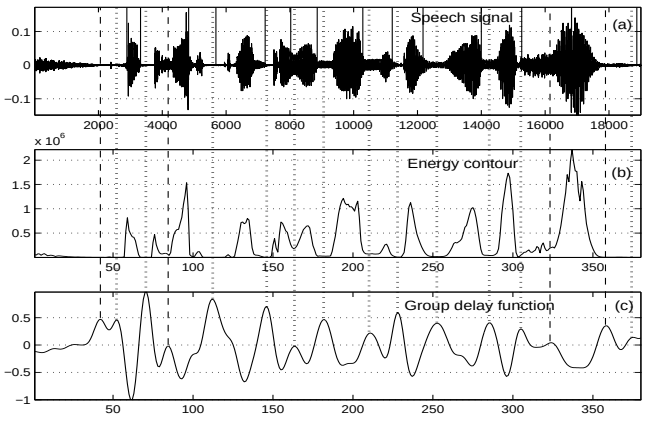

Fig. 2. Group delay based segmentation - an example.

3.2. Sub-band based system

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 this problem, on advice from Steven Greenberg at ICSI [8], a sub-band based approach to syllable segmentation is attempted. 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 sub-bands 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. 2).

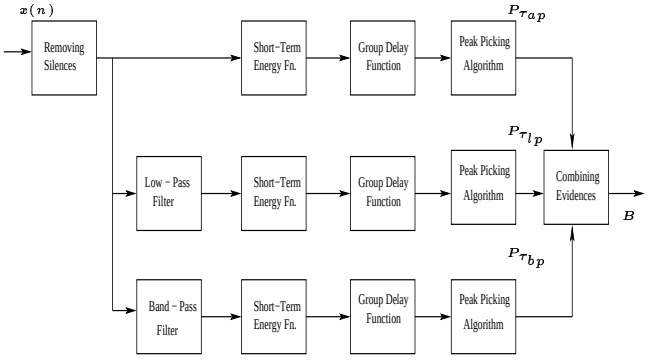

Fig. 3. Block-diagram of sub-band based approach.

(a) Silence removal : In the silence region of speech, we are supposed to get one peak in the middle. But since the poles are placed consecutively without any zeros in between, we may get two peaks or even multiple peaks, which is undesirable. To avoid this problem, the silence segments present in the speech signal should be removed. Based on the knowledge derived from the energy, zero-crossing rate and the spectral flatness of a frame the decision is made. If the silence is present in more than two frames of the signal, then that particular segments are removed from the original speech signal.

(b) Filtering : 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. To avoid this, the signal, $x(n)$ is low-pass filtered to remove the high frequency fricatives. 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. If we apply a band-pass filter to the original signal, since the energy of the semivow-

els are concentrated at low formant frequencies alone, 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.

(c) Energy function Computation : For all the three types of signals, the short-time energy function is calculated with frame-size of 25msec and frame-shift of 5msec. The time resolution in the short-time energy function depends on the frame-shift. Therefore, the average minimum error introduced in segmentation is equal to $frameshift/2$.

(d) Group-delay function Computation : From the short-time energy functions, the group-delay function is computed using the same algorithm explained in Section.(3).

(e) Peak-picking algorithm : From the group delay functions derived from the different short-time energy functions, the positive peaks alone picked using a simple peak-picking algorithm. The peaks derived from all-pass, low-pass and band-pass filtered signals are denoted as, $P_{\tau ap}$, $P_{\tau lp}$ and $P_{\tau bp}$ respectively.

(f) Combining Evidences : The peaks derived from the different group-delay functions are combined using the following logic.

$$P_{\tau al} = (P^i_{\tau ap} \wedge P^i_{\tau lp}) \quad (9)$$

if $(P_{\tau ap} \sim P_{\tau lp}) \leq$ 20msec $\forall$ i where 'i' is the ith peak in the group delay function.

$$P_{\tau alb} = P_{\tau al} \vee (P^i_{\tau al} \wedge P^i_{\tau bp}) \quad (10)$$

if 50msec $\leq (P^i_{\tau al} \vee P^i_{\tau bp}) \leq$ 100msec, $\forall$ i.

3.3. Performance

The Switchboard corpus is 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. The performance of the base line system and the sub-band based approach are given in Table.1.

error (in msec)	A	B
<25	52.00	66.93
25 - 40	19.66	18.46
40 - 60	13.98	12.54
60 - 80	14.35	2.05
insertion	10.75	5.25
deletion	9.50	7.10

Table 1. Performance (in %) of the group delay based segmentation approach (A) Base line system. (B) Sub-band based system.

4. CONCLUSIONS

In this paper, a novel approach for segmenting the speech signal into syllabic units is presented. The performance of the minimum phase group delay function based segmentation approach is tested on Switchboard corpus and it is found to be quite satisfactory. It is further shown that the performance can be significantly improved by a sub-band based approach. 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.

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] 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.

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

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

[7] 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.

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