AN AUTOMATIC ALGORITHM FOR SEGMENTING AND LABELLING A CONNECTED DIGIT SEQUENCE

V. Kamakshi Prasad and Hema A. Murthy

Department of Computer Science and Engineering,
Indian Institute of Technology, Madras, Chennai - 600 036
e-mail : hema@lantana.iitm.ernet.in

ABSTRACT

Group delay functions provide an alternative representation of signal information. The main features of group delay functions are the additive and high resolution properties. The Fourier transform (FT) phase is generally featureless due to random polority and wrapping. But the group delay function which is defined as the negative derivative of phase, can be processed to derive significant information such as peaks and valleys in the spectral envelope. In this paper, we show an application of group delay function to solve the segmentation problem in speech. In the proposed method a new signal is generated by symmetrising the short term energy function. The minimum phase group delay function of this signal is computed, the valleys of which correspond to segment boundaries. The proposed technique was tested on manually segmented digit utterances of the TI-DIGITS database. The overall correct segmentation performance is 77.8%. Digitwise recognition performance on the correctly segmented database is 87.1%

1. INTRODUCTION

Unlike printed text of a given language, in which words are separated by blanks, speech does not contain any reliable demarcation. As a result, developing a system for automatic segmentation of the speech signal at appropriate phonetic units is a non trivial task. The segmentation of the speech signal is a pre-requisite for many real time tasks like speech recognition, speaker recognition and language identification. In particular, in the developement of connected digit recognition systems if the speech signal is segmented at digit boundaries, a simple HMM based isolated digit recognition system can be used to recognise each digit. The advantage of such an implementation is, it overcomes the computational complexity and error propagation which are found in the approaches like level building and two level dynamic programming[1, 2].

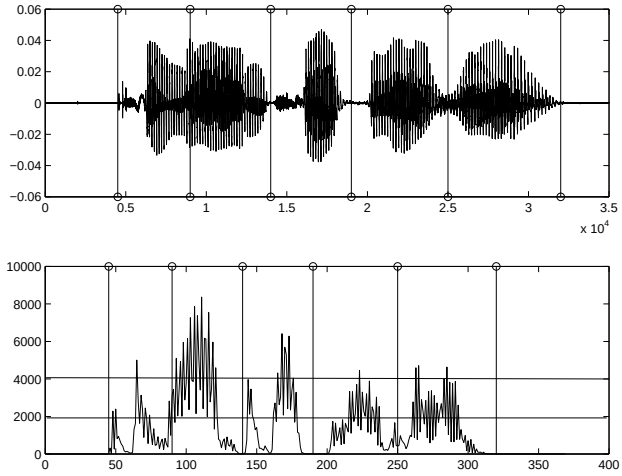

Figure 1: Segmentation of the connected digit utterance '2934z' from TI-DIGITS database using a simple thresholding. Fig.1(a) shows the original CD utterance, Fig.1(b) shows the short term energy function of the signal with a threshold applied.

Energy based approaches are considered to be appropriate to detect the boundaries [3, 4]. The short term energy function of the speech signal is not smooth due to significant local variations in the amplitude values and the co-articulation effects between successive units. Although visible to the naked eye, it is difficult to extract the syllable boundaries from the energy function. Applying a threshold on short term energy function or power spectrum may be used for detecting segment boundaries [4]. But finding an appropriate threshold is a difficult task and thresholding techniques may not all the times give consistent results as shown in Fig.1. In this paper, we propose an energy based algorithm where the minimum phase group delay function is used for processing the short term energy to detect the digit boundaries.

In the technique discussed in this paper, we exploit the additive and high resolution properties of group

delay functions [5]. It has already been well established that group delay processing of speech signals yield more consistent estimates of parameters than the conventional approaches based on homomorphic signal processing and linear prediction analysis [5, 6].

In Section 2, we describe the algorithm based on minimum phase group delay functions to identify the digit boundaries for fixed length connected digit speech utterance. In Section 3, we show why group delay function approach is more successful than other approaches. In Section 4, we evaluate the performance of the segmentation of the connected digit utterances from TI-DIGITS [7] database. Application of segmentation to continuous speech is given in Section 5 and in Section 6, we present the conclusions.

2. ALGORITHM FOR IDENTIFICATION OF DIGIT BOUNDARIES

In connected digit recognition, determining the digit boundaries directly from the energy function is difficult because each digit can affect the succeeding/preceding digit boundaries. Generally the boundaries merge for the case of connected digit utterances of length 4 and more. In addition, the energy function will have a large number of spurious peaks.

The energy function has some interesting properties. The energy function is a nonzero positive function of time and it faithfully follow the amplitude variations in the speech signal. So it can be said that the energy function behaves like a magnitude spectrum. Due to this property, algorithms for smoothing the magnitude spectrum can be explored for smoothing the energy function.

LPC based smoothing algorithm model the poles well but do not model the valleys [8]. Cepstrum based smoothing of the magnitude function can be used to detect the segment boundaries [5]. To employ cepstral based smoothing, the energy function has to be preprocessed. We first symmetrise the energy function by reflecting the same about the y-axis. We now treat this signal as the magnitude spectrum and cepstrally smooth it. Fig.2(b) shows the cepstrally smoothed energy function. The valleys of this function should correspond to digit boundaries. But it is clear that the boundaries are not identified correctly. Minimum phase group delay functions have been successfully used to smooth the magnitude spectra [6] of the speech signal for determining the location of the formants. A similar minimum phase group delay approach is used on the energy function to identify the digit boundaries. As shown in Fig.2(c) the valley points are clearly resolved, which gives the digit boundaries in the connected digit speech signal.

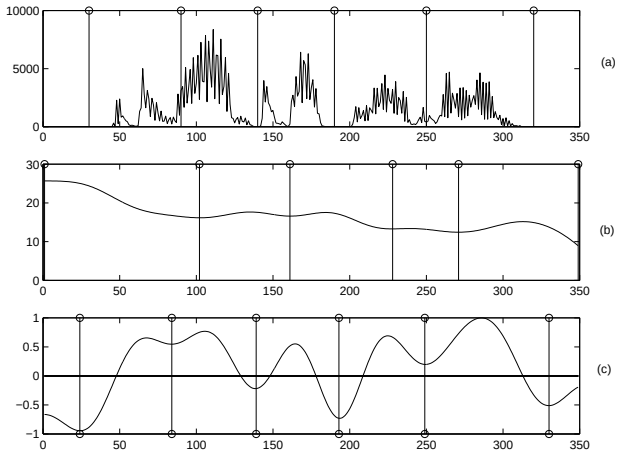

Figure 2: Segmentation of the connected digit utterance '2934z' from TI-DIGITS database. Fig.2(a) shows the short term energy function of the signal, Fig.2(b) shows the cepstrum smoothing of the speech signal and Fig.2(c) shows the group delay based smoothing of this signal

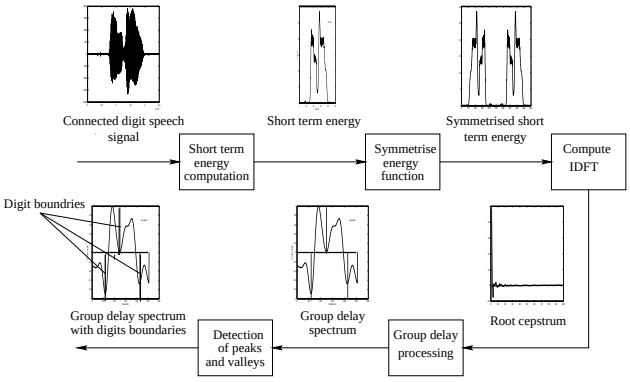

Figure 3: Steps involved in finding digit boundaries.

Steps involved in the algorithm for identification of digit boundaries are as follows [Fig.3]:

- Let $x\{n\}$ be the given digitized speech signal of a connected digit utterance.
- Compute $E\{n\}$, the energy in each segment of 25.6ms of speech using overlapped windows. Since this sequence does not have any negative values, it behaves like the short-term spectrum of a given signal.
- Construct the symmetric part of the sequence by producing a lateral inversion of this sequence about the Y-axis. Let this sequence be $\tilde{E}\{n\}$.
- Compute the inverse DFT of the sequence $\tilde{E}\{n\}$. This sequence behaves like the root cepstrum [9]. Let this sequence be $\tilde{e}(n)$.

- Compute the minimum phase group delay function [6, 9] of the causal portion of the sequence $\tilde{e}(n)$.
- The locations of the valleys approximately correspond to the digit boundaries, and the number of the peaks corresponds to the number of digits in the sequence.

Since the phase function is additive, the group delay function which is defined as the negative derivative of the phase is also additive. Since no model is assumed in this approach, this technique should capture the underlying characteristics of the energy function well.

3. WHY DOES THE GROUP DELAY FUNCTION WORK ?

In the minimum phase group delay processing of the energy function, the energy function is treated like the magnitude spectrum and the width of the digit in the energy function is being treated like band width. The average duration of the digit segments in the five digit speech utterance is found to vary from 311 ms to 460 ms for short duration utterances like 'oh' and long duration utterance like 'six' respectively. Since the duration does not vary significantly, the group delay based approach which is primarily dependent on band width, performs very well. Since the group delay function has the property that the height of the peak is inversely proportional to the band width, equal emphasis is given to all the digits in the sequence. Due to this property the information about a given digit is concentrated around the digit center. This is referred to as the additive and high resolution property of the group delay function in the literature [6, 10].

4. PERFORMANCE OF THE SYSTEM

We first evaluated the performance of the segmentation process. Five digit male speaker database from TI-DIGITS database has been manually segmented and tested against the segmentation results obtained using the proposed algorithm.

If the number of digits in the uttered connected digit are known apriori, the parameters of the algorithm are tuned to give the segments equal to that of number of digits. The performance of the segmentation is as shown in Table 1.

The speech signal is divided into single digit utterances by segmenting at digit boundaries. The Mel-frequency cepstral coefficients (MFCC) feature vectors of the single digit utterances are now used to train an isolated digit recognition system based on HMMs. As the TI-DIGITS database consists of the digits '1' through '9','oh' and 'zero', 11 HMMs have been built for the digits.

Table 1: Digit-wise segmentation performance of the 5 digit utterance.

Digit	Total no.of digit segments	No.of segments correctly identified	Percentage of correct segments
1	276	216	78.3
2	283	234	82.7
3	291	237	81.2
4	251	216	86.1
5	271	233	86.0
6	299	175	58.5
7	280	190	67.9
8	275	197	71.6
9	269	221	82.2
zero	271	208	76.8
oh	258	208	80.6
average			77.8

Only the correctly identified digit segments are tested against the isolated digit HMM models. The feature vector sequence of each segment is computed as in training phase. The recognition performance on the segmented data is shown in the Table 2.

Table 2: Digit-wise recognition performance of the correctly identified segments.

Digit	Total no.of digit segments	No.of digits correctly recognised	Percentage of correct recognition
1	216	186	86.1
2	234	198	84.6
3	237	221	93.2
4	216	203	94.0
5	233	201	86.3
6	175	155	88.5
7	190	165	86.8
8	197	160	81.2
9	221	191	86.4
zero	208	196	94.2
oh	208	159	76.4
average			87.1

5. APPLICATION OF SEGMENTATION ALGORITHM TO CONTINUOUS SPEECH SIGNAL

The algorithm given in Section 2, was applied to continuous speech. Fig.4(a) shows a continuous speech

utterance (from the TIMIT database) and the corresponding short term energy function Fig.4(b) and the group delay function Fig.4(c). The valleys in the group delay plot correspond to syllable boundaries. The dotted vertical lines in the Fig.4(c) denote the manually identified syllable boundaries of the speech signal and the thick lines denote the syllable boundaries detected using the group delay function. It is noticed that most of the syllable boundaries are identified correctly. The performance of the segmentation was evaluated on 76 utterances and it was found that about 80% of the syllable boundaries were identified with an error of atmost 20 ms.

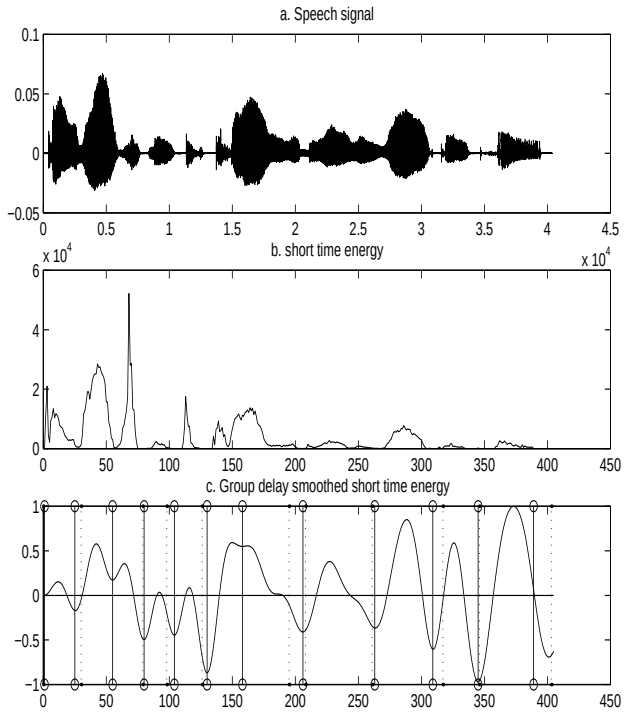

Figure 4: Segmentation of the continuous speech utterance "Don't ask me to carry an oily rag like that" from the TIMIT database

6. CONCLUSIONS

In this paper, we have proposed a new approach to segment the connected digit utterance. A method based on minimum phase group delay function for smoothing the magnitude spectrum is used to obtain a smooth envelope of the short term energy function. Our results on segmentation show that although there are some errors in the detection of boundaries, these errors primarily correspond to merged boundaries or splitting of a given segment. Never has the group delay function misplaced the boundaries. Even when a segment is split, we find that in the energy function there is a valley, although this may not correspond to an actual boundary. We are currently exploring the use of knowledge based on prosody to correct these errors.

7. REFERENCES

[1] L.R.Rabiner and B.H.Juang, *Fundamentals of Speech Recognition*. New Jersy: Prentice Hall, 1993.

[2] C.S.Myers and L.R.Rabiner, " Connected digit recognition using level-building DTW algorithm," in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, pp. 351–363, June 1981.

[3] Su_Lin Wu, MicaelL.Shire, Steven Greenberg, Nelson Morgan, "Integrating syllable boundary information into speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Munich, Germany), pp. 987–990, May 1997.

[4] Paul Mermestein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Amer.*, vol. 58, pp. 880–883, October 1975.

[5] Hema A. Murthy, *Algorithms for Processing Fourier Transform Phase of Signals*. PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, 1992.

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

[7] R.G.Leonard, "A database for speaker-independent digit recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 3, pp. 42–45, 1984.

[8] J.Makhoul, "Linear prediction: A tutorial review," *Proceedings of IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[9] Hema A. Murthy, "The real root cepstrum and its application to speech processing," in *National Conference on Communication*, (Chennai, India), pp. 180–183, Jan 1997.

[10] B. Yegnanarayana, "Formant extraction from linear prediction phase spectram," in *J. Acoust. Soc. Amer.*, pp. 1638–1640, 1978.