Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering
Shaw Conference Center, Edmonton, Alberta, Canada May 9-12 1999

AUTOMATIC DETECTION OF ACOUSTIC SUB-WORD BOUNDARIES FOR SINGLE DIGIT RECOGNITION

Maen M. Artimy, William Robertson, William J. Phillips

Dalhousie University, DalTech
artimym@is2.dal.ca, bill.robertson@is.dal.ca, william.phillips@is.dal.ca

Abstract- **This paper investigates the use of a spectral variation function to automatically detect acoustic subword boundaries in single digits. The developed algorithm generates subwords for a single-digit recognition system using an RBF neural network.**

1. Introduction

Many speech recognition systems based on pattern recognition approaches use frame-level acoustical information to model speech. Those systems usually have to time-align frame vectors belong to the same sounds in both test and template words before comparison. Traditionally, a dynamic time-warping (DTW) algorithm is needed to perform this task. Use of higher-level segments or subwords can eliminate the alignment problem by presenting acoustically similar parts of speech as a single unit. This approach eliminates any complex computations associated with the DTW algorithm.

Automatic decomposition of speech into linguistically defined subwords (phonemes, syllables, etc.) is not always reliable because of the mismatch between the acoustic properties of these subwords and their linguistic descriptions [4]. Alternatively, using acoustically defined subwords can lead to more consistent results [4,10].

In this paper we investigate the use of an automatic procedure to decompose spoken digits into subwords based on acoustical information contained within the speech signal. The procedure locates the boundaries between subwords by finding the peaks in a function representing the spectral change between consecutive speech frames. Energy parameters derived from a filter bank of Mel-frequency scale [6,10] are used to model the spectral characteristics of the speech. The results are enhanced further by using zero-crossing rate parameters and some heuristic rules to eliminate excessive segmentation.

Vectors representing the Mel-band parameters in each subword are then presented to a Radial Basis Function (RBF) neural network to perform the recognition [3,10].

In Section 2, we describe all stages of the boundaries detection process. In Section 3, we list our experiments followed by the results in Section 4. A discussion of the results is in Section 5.

2. Boundaries Detection

The automatic detection procedure starts by isolating a sampled speech signal from the surrounding background noise. We used a modified version of the algorithm described by [5] to detect the end points of the spoken digits. This step is very important to reduce the amount of processing time required in subsequent stages. The rest of the algorithm consists of four stages:

2.1 Feature Analysis

At this stage, the isolated speech signal is divided into frames of 10ms each. Samples in each frame are then multiplied by a Hamming window of 20ms length. The larger window is chosen to ensure overlapping between frames [6,7]. The speech frames are then analyzed with 512-point FFT. A 20-channel Mel-band filter bank is simulated by averaging spectrum coefficients in each frequency band to cover a range of 0-4KHz. All energy parameters are transformed into a decibel scale of 0-60dB. We refer to the 20-element vector of energy values as a frame vector F. The speech signal S has N frame vectors.

$$F_n = [F(0), F(1), \dots, F(19)]^T \quad 0 \le n \le N-1 \text{ ....(1)}$$

$$S = [F_0, F_1, \dots, F_{N-2}, F_{N-1}] \text{ ....(2)}$$

2.2 Spectral Variation Function

The spectral variation values are indications of the degree of distortion between the associated speech signals [4]. High values of a variation function *v(n)* exhibit a rapid change in the spectra between adjacent frames while low values indicate similar spectrum characteristics. Therefore, edges between different sounds can be located by picking the peaks in the spectral variation function.

The spectral variation function *v(n)* is created by measuring the Euclidean distance [1,7] between consecutive frame vectors F as shown in Fig.1-b.

$$v(n)=\sqrt{\sum_{k=0}^{19}\left(F_n(k)-F_{n-1}(k)\right)^2} \quad 1\le n\le N-1 \qquad ......(3)$$

$$v(0)=0$$

The function is then smoothed by splines polynomials to eliminate spurious peaks that may cause false boundaries to be detected. The result, *vs(n)*, is shown in Fig.1-c.

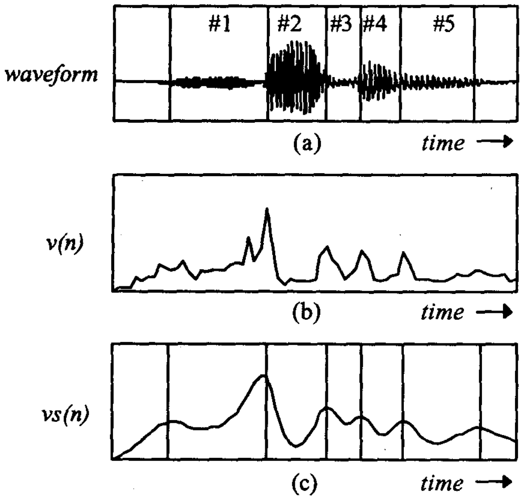

Figure 1 (a) A waveform of digit "seven" divided into five subwords, (b) A spectral variation function, (c) The smoothed function and detected boundaries.

A procedure is implemented to search for local maxima in *vs(n)* and return their time indices in a vector *B*.

$$B=\left[b_0,b_1,\dots,b_{M-1}\right] \qquad ....(4)$$
$$b_0=0,\ b_{M-1}=N-1,\ 1\le b_i\le N-2$$

The vector *B* contains *M* frame numbers that lie at the boundaries of *M-1* subwords.

2.3 Boundaries Reduction

The optimal number of subwords *D* differs in each digit. Nevertheless, by assuming the number of subwords cannot exceed the number of phonemes in a digit, we set the maximum allowed number of subwords to be five [8,10]. At this stage, some knowledge of the phonetic composition of the vocabulary will be used to enhance the results by merging some subwords and eliminating others to create the final set of subwords. This process employs two facts, (1) eight digits ("2" to"8", and "0") have one noise-like consonant at either or both ends, (2) nine digits ("1" to "6", "8", "9", and "oh") are single syllable words. Noise-like sounds can be easily detected using the zero-crossing (ZC) rate of the signal [6,7,8]. Therefore, the rate is measured in all frames of the speech signal. The ZC values are then averaged in all frames within each subword. A subword is considered a noise-like sound if the zero-crossing rate value exceeds a dynamically-set threshold. Since no two noise-like sounds can be adjacent in a single digit, any two or more adjacent noise-like subwords are merged to form one subword. Fig.2-a shows a false boundary that divides the fricative /s/ into two subwords. Fig.2-b shows that the ZC rate in these subwords exceeds the threshold; thus, both subwords can be merged.

A syllable consists of vowel nucleus surrounded by an optional consonant [7]. Vowels have relatively higher energy levels than other sounds such fricatives or nasals. Therefore, the syllable nucleus in a digit can be easily located by finding the subword of the highest energy. The number of subwords in a digit can be reduced to optimal value by limiting the number of subwords surrounding the vowel in the digit.

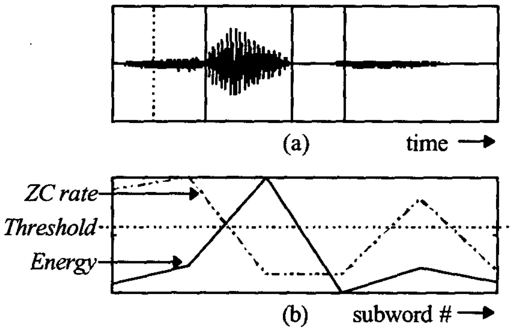

Figure 2 (a) A waveform of digit "six" with a false boundary marked by a dotted line. (b) The energy and zero-crossing rate values at each subword and the dynamically set threshold.

These rules are not strictly applicable in all cases; however, implementing them can remove most false boundaries in a spoken digit.

2.4 Subword Vectors

The last stage of this algorithm averages energy coefficients in the same filter bank channel across all frame vectors between two consecutive boundaries. The result is a set of subword vectors *SV*.

$$SV_d=\left[SV(0),SV(1),\dots,SV(19)\right]^T \qquad ....(5)$$
$$0\le d\le D-1$$

$$SV_d(k)=\frac{1}{b_{d+1}-b_d}\cdot\sum_{i=b_d}^{b_{d+1}-1}F_i(k) \quad 0\le k\le 19 \qquad ....(6)$$

To create a common representation for all digits, a digit vector is created by concatenating subword vectors into a 100-coefficient vector *DV*. Zero vectors *Z* are added to digits with less than five subwords.

$$DV=\left[SV_0,SV_1,\dots,SV_{D-1},Z,\dots,Z\right]^T \qquad ....(7)$$

3. Experiments

The digit vectors were used to train an RBF neural network for recognition experiments. The RBF network is a two-layer structure that can be trained for pattern classification using a clustering technique. An extensive discussion of using RBF network for digit recognition is given by [2,3,10].

A database of spoken digits' corpus (TIDIGITS) [9] was used to evaluate the performance of the boundaries detection algorithms.

For performance comparison, we recreated a set of experiments used by an early work [2,10] in which digit vectors were created manually. The set includes 17 speakers (8 males, 9 females) from two American English dialects (Rochester and Pittsburgh). Each speaker has two tokens of each of the eleven digits ('0'-'9', and 'oh'). The total of 374 samples were processed to generate a boundaries' vector and a digit vectors for each digit.

A total of 27 experiments were performed. In each, a pair of different speaker sets were chosen arbitrarily. Every pair consists of training and test sets. The sets were chosen to include variable number of speakers, dialects, and genders.

Digit	*% Set #1*	*% Set #2*	*% Other*
One	61	30	9
Two	68	18	14
Three	52	26	22
Four	59	23	18
Five	68	23	9
Six	86	0	14
Seven	72	14	14
Eight	36	32	18*, 14
Nine	50	27	23
Zero	68	27	5
Oh	54	41	5

*Set #3

Table 1 Distribution of Digits Among Boundary Sets

To further evaluate the performance of the algorithm, a different set of experiments were performed. This set has 43 speakers (21 males, 22 females) from four dialects (Boston, Philadelphia, Rochester, and Pittsburgh) with total of 946 digits. The set were divided into 36 experiments. As in the previous set, the training sets were chosen to represent a number of speakers ranges from 2 to 10. The test sets contain either 15 or 21 speakers.

4. Results

A visual inspection indicates that most boundaries fall into two sets of locations per digit. One set is usually a subset of the other with one boundary missing. Some digits have uncommon set of boundaries due to either under-segmentation or over-segmentation. Table 1 shows the percentage of digits that fall in each set of boundaries. Table 2 shows the average recognition rate of all experiments with common size of a training set along with the corresponding results obtained from the manual system under the same conditions.

Table 3 lists all the recognition rates obtained from the second set of experiments with the size of both the training and test sets of speakers.

Size of Training Set	*Recognition %*	
	Manual System	*Automatic System*
2	98.48	60.10
3	97.84	73.05
4	99.07	79.02
5	99.11	82.74
8	99.44	87.60

Table 2 Recognition Rates from Set #1

Fig. 3 shows a plot of the recognition rate versus number of speakers in the training set in both sets of experiments.

5. Discussion

The comparison between the recognition rates of this algorithm with those obtained by manual means favors the latter by an average of 19%. However, we notice from Fig.3 that the recognition rate improves considerably when the size of the training set rises to five speakers or more. In this case the recognition rate exceeds the 82% up to 89.7%. These results are reasonable considering inherited inability in any automatic procedure to achieve the precision of the manual segmentation. The automatic procedure may falsely add some boundaries or miss others resulting a wrong number of subwords.

The boundaries detection algorithm does not seem to be affected by the number of speakers or the dialects included since it achieved consistent results in both sets of experiments.

Considering these results, we suggest the following:

1. Enhancing the set of rules used to eliminate excessive segmentation by removing false boundaries. Current rules are chosen to ensure the simplicity of the algorithm.
2. Using the information included in Table 1 to design training sets that statistically represent the distribution of boundaries; thus, enhancing the recognition rate.

6. Summary

In this paper, we proposed an algorithm for automatic detection of acoustic subword boundaries in single digits. The algorithm uses a spectral variation function to locate the boundaries between subwords then enhances the results further by implementing simple rules to remove false boundaries. An RBF neural network was used to test the efficiency of this algorithm in digit recognition. The experiments included a database of 43 speakers from both genders representing four American English dialects.

#	*Size Train. Set*	*Size Test Set*	*Rec. %*	#	*Size Train. Set*	*Size Test Set*	*Rec. %*
1	2	15	72.42	19	7	15	84.55
2	2	21	71.65	20	7	21	85.06
3	2	15	70.00	21	7	15	83.94
4	2	21	71.43	22	7	21	85.93
5	2	15	64.55	23	7	15	83.03
6	2	21	68.18	24	7	21	83.77
7	3	15	78.18	25	8	15	83.33
8	3	21	76.62	26	8	21	85.06
9	3	15	70.61	27	8	15	86.06
10	3	21	72.29	28	8	21	86.58
11	3	15	72.42	29	8	15	84.55
12	3	21	73.16	30	8	21	86.15
13	5	15	83.03	31	10	15	88.79
14	5	21	83.98	32	10	21	87.45
15	5	15	82.42	33	10	15	89.70
16	5	21	82.25	34	10	21	89.39
17	5	15	82.73	35	10	15	86.97
18	5	21	82.03	36	10	21	87.66

Table 3 Recognition Rates from Experiments Set #2

7. References

[1] A. H. Gray Jr., J. D. Markel, "Distance Measures for Speech Processing", *IEEE Transactions on ASSP*, 24(5):380-391, October 1976.

[2] C. Tosuner, "Single Spoken Digit Recognition Using RBF Networks", *M.A.Sc Thesis, Technical University of Nova Scotia*, 1996.

[3] D. R. Hush, B. G. Horne, "Progress in Supervised Neural Networks: What's New Since Lippmann?", *IEEE Signal Processing Magazine*, 10(1):8-39, January 1993.

[4] G. Wilpon, B. H. Juang, L. R. Rabiner, "An Investigation on the Use of Acoustic Sub-word Units for Automatic Speech Recognition", *Proceedings ICASSP 87*, 821-824, April 1987.

[5] E. S. Dermats, N. D. Fakotakis, G. K. Kokkinakis, "Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment", *Proceedings ICASSP 91*, 733-736, 1991.

[6] J. W. Picone, "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, 81(9):1215-1247, September 1993.

[7] L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[8] M. R. Sambur, L. R. Rabiner, "A Speaker-Independent Digit-Recognition System", *The Bell System Technical Journal*, (54)1 81-103, January 1975.

[9] R. G. Leonard, "A Database for Speaker-Independent Digit Recognition", *Proceedings IEEE ICASSP 84*, 42.11, March 1984.

[10] W. J. Phillips, C. Tosuner, W. Robertson, "Speech Recognition Techniques Using RBF Networks" *Proceedings IEEE WESCANEX 95*, 185-190.

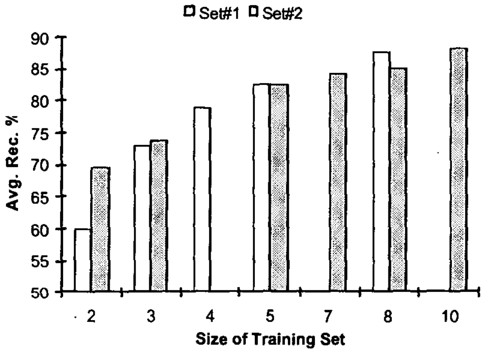

Figure 3 Size of Training Set vs. Average Recognition Rate.