A SEGMENTATION METHOD FOR NOISY SPEECH USING GENETIC ALGORITHM

Moe Pwint, Student Member, IEEE and Farook Sattar, Member, IEEE

School of Electrical and Electronic Engineering, Nanyang Technological University
Nanyang Avenue, Singapore 639798
emails: moepwint@pmail.ntu.edu.sg; efsattar@ntu.edu.sg

ABSTRACT

This paper presents a technique to automatically segment a speech signal in noisy environments. The speech segmentation is formulated as an optimization problem and boundaries of the speech segments are detected using genetic algorithm (GA). The initial number of segments is estimated from the modified version of the signal using the minimal number of binary Walsh basis functions. The segmentation results are improved through the generations of GA by introducing a new evaluation function, which is based on the sample entropy and a heterogeneity measure. The results of the experiments, which have been carried out on TIDIGITS database and different types and levels of noise, show the efficiency of the proposed genetic segmentation algorithm.

1. INTRODUCTION

The detection of speech segments which are corrupted by unknown type and level of noise is considered. The segmentation method described in this paper is based on the genetic algorithm, which is a stochastic global search method. In GAs, the search directions are influenced only by the evaluation function and its corresponding fitness value without requiring any derivative information or other auxiliary knowledge. GAs are also able to utilize parallel exploration of the search space by reducing the possibility of being stuck in local optima. This motivates to use GA for detecting accurate boundaries of noisy segments in the proposed segmentation technique.

The proposed noisy speech segmentation algorithm can be divided into two stages. In the first stage, the number of segments are estimated from the modified version of the segmenting signal using the minimal number of binary Walsh basis functions together with the mean difference measure described in [1]. The second stage is the segment boundaries determination. In that stage, the start and end points of segments are detected using a multi-population genetic algorithm. To guide the search space of GA, an evaluation function is introduced by the combination of sample entropy [2] and a heterogeneity measure [3].

2. SAMPLE ENTROPY

The origin of sample entropy is the approximate entropy ($ApEn$), which is originally introduced in [4] to measure the regularity in time series. $ApEn(m, r, N)$ is the negative natural logarithm of the conditional probability that a data set of length N, having repeated itself within a tolerance r for m points, will also repeat itself for $m+1$ points. However, $ApEn$ is lack of relative consistency and heavily dependent on the record length and is uniformly lower than expected for short records. Therefore, sample entropy ($SampEn$) that does not count self-matches is developed in [5] to reduce those biases.

For an input signal of length N, $\{u(j) : 1 \leq j \leq N\}$ forms the $N-m+1$ vectors $x_m(i)$ for $\{i|1 \leq i \leq N-m+1\}$, where $x_m(i) = \{u(i+k) : 0 \leq k \leq m-1\}$ is the vector of m data points from $u(i)$ to $u(i+m-1)$. Let $B^m(r)$ is the probability that two sequences will match for m points and $A^m(r)$ is the probability that two sequences will match for $m+1$ points. $B_i^m(r)$ is defined as $(N-m-1)^{-1}$ times the numbers of vectors $x_m(j)$ within r of $x_m(i)$, where $1 \leq j \leq N-m$, and $j \neq i$ to exclude self-matches. Then $B^m(r)$ is defined as

$$B^m(r) = (N-m)^{-1} \sum_{i=1}^{N-1} B_i^m(r) \tag{1}$$

Similarly, $A_i^m(r)$ is defined as $(N-m-1)^{-1}$ times the numbers of vectors $x_{m+1}(j)$ within r of $x_{m+1}(i)$, where $1 \leq j \leq N-m$ and $j \neq i$. Then set $A^m(r)$ as

$$A^m(r) = (N-m)^{-1} \sum_{i=1}^{N-1} A_i^m(r) \tag{2}$$

Finally, sample entropy ($SampEn$) is calculated by

$$SampEn(m, r, N) = -\ln \frac{A^m(r)}{B^m(r)} \tag{3}$$

3. INITIAL SEGMENT ESTIMATION

For a given input signal, the number of input segments to be detected is estimated firstly from its modified version.

Boundaries of speech segments are then detected by the GA routine.

3.1. Analysis and Synthesis Scheme

Modification of the noisy input signal is performed by using an analysis and synthesis scheme described in [6]. At the analysis part, the input signal $x(n)$ is multiplied by a Hann window to obtain windowed segments. Then a time varying spectrum $X_s(n,k) = |X_s(n,k)|e^{j\varphi(n,k)}$ with $n = 0, 1, ..., N-1$ and $k = 0, 1, ..., N-1$ for each s^{th} windowed segment is computed by transforming into spectral domain using FFTs. Here, $X_s(n,k)$ denotes the spectral component of the noisy input signal at frequency index k and time index n.

At the synthesis part, the magnitude $|X_s(n,k)|$, of each s^{th} windowed segment is processed to reconstruct a modified sequence of $y_s(n)$ as the weighted sum of the magnitudes using binary Walsh basis functions. Walsh basis functions, $\phi_0, \phi_1, ..., \phi_{N-1}$ are the kernel of Walsh transform which are arranged into ascending order of zero-crossings.

$$W = [\phi_0, \ \phi_1, \ \cdots \ , \phi_{N-1}]. \tag{4}$$

3.2. Selection of Minimal Basis Functions

A technique for selecting the global natural scale in discrete wavelet transform [7] is employed to determine the required minimum number of basis functions. Using these basis functions, a modified signal is reconstructed to capture both the global characteristics and local details of the segmenting signal. This method adaptively detects the optimal scale using singular value decomposition (SVD), while decomposition is being carried out. Let $y_d(n)$ be the modified sequence developed by using the basis function of order d, then modified sequences $\{y_d(n)\}_{d=0}^{D-1}$ can be represented in a matrix P of dimension $D \times N$.

To detect the order of basis functions with dominant eigenvalues, the SVD of the matrix P is computed adaptively starting with the first two orders (i.e. ϕ_0 and ϕ_1) while adding the higher orders. The probability distributions of the order of basis function as a function of noise levels is studied using a number of speech signals, spoken by male and female speakers from TIDIGITS database. It is observed that the dominant eigenvalue is found at the order of one in highly noisy cases (5 dB and 0 dB). On average, the dominant eigenvalue is found at the order of three in 10 dB, 20 dB and clean signal. Thus the prominent order of basis function is chosen as three throughout the experiments since the *a priori* information of the SNR or noise type cannot be obtained in practice. A good estimate of the Walsh basis function at dominant order is then defined as

$$\phi_m = \frac{\phi_0 - \sum_{i=1}^{M} CS(\phi_i)}{\max\{|\phi_0 - \sum_{i=1}^{M} CS(\phi_i)|\}} \tag{5}$$

where M=3 is the largest order with the most prominent eigenvalue and $CS(\cdot)$ is the shifting operator which swaps the left and right halves of the coefficients of basis function. This time domain shifting is similar to the phase shifting while keeping the details of the signal. This new basis function ϕ_m provides sharper representation and higher discriminating features in the modified sequence, such as noisy speech periods and noise only intervals. Figure 1 demonstrates how a noisy speech waveform is reconstructed after exploiting the proposed modification procedure as discussed above.

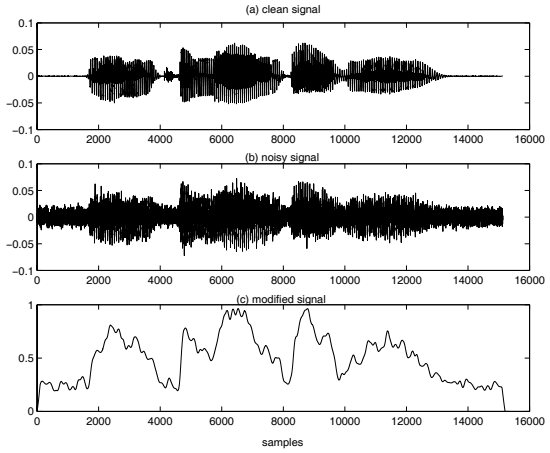

Fig. 1. (a) Clean speech signal; (b) Noisy speech signal; (c) The modified signal.

3.3. Mean Difference Measure

The input to GA (i.e. number of segments) is then determined from this modified signal using mean difference measure [1]. In this approach, the two adjacent windows of equal length are moved through the modified signal. At each position, magnitude of the difference of the means within each window of length 500 is calculated. This mean differences sequence is thresholded to determine local maxima of a certain value. Local maxima having a width greater than a specified value is taken as segments. All the other maxima which do not meet these conditions are discarded.

4. SEGMENTING BY THE GENETIC ALGORITHM

As the second step, the precise locations of the segment boundaries are detected using genetic algorithm. Number of segments determined from the initial segmentation is considered as the input to the GA.

4.1. Genetic Algorithm

The GA is a searching process based on the laws of natural selection and genetics [8]. The procedure of a simple GA can be described as follows, where the population of candidate solutions at time t is represented by P(t):

```
begin
    t = 0;
    initialize P(t);
    while not termination criteria do
    begin
        t = t + 1;
        select P(t) from P(t-1);
        reproduce pairs in P(t);
        evaluate P(t);
    end
end
```

4.2. Initial Population

Depending on the total number of segments estimated from the initial segmentation, an initial population is randomly generated. In order to detect the start and end locations of each segment, the length of an individual is defined as two times total number of segments. To increase the efficiency of GA, a real-valued representation is used in the proposed method as there is no need to convert chromosomes to phenotypes before fitness function evaluation.

4.3. Evaluation Function

In order to obtain the accurate boundaries of each segment, evaluation function or fitness function is designed using the heterogeneity measure proposed in [3] and sample entropy. This function simultaneously maximize the homogeneity within the segments and heterogeneity among different segments using sample entropy. In this context, $SampEn$ of the original segmenting signal is calculated on each data set of length 80 (i.e. N=80) within a tolerance r of $0.1 \times SD$ for 1 point (i.e. m=1). Here SD is the standard deviation of the data set. A segmentation evaluation function is defined as

$$H = \frac{H_b + 1}{H_b + H_w + 1} \tag{6}$$

where total within-heterogeneity H_w is defined as

$$H_w = \frac{\sum_{i=1}^{S} L_i \sigma_i^2}{L} \tag{7}$$

where L is the total length of the segmented signal, L_i is the length of i^{th} segment, σ_i^2 is the variance of the sample entropy of i^{th} segment and S is the number of segments in the segmented signal. The between-segment heterogeneity, H_b, is defined as the average Euclidean distance between the mean value of the sample entropy of any two adjacent segments.

$$H_b = \frac{\sum_{(i,j) \in adjacent, i \neq j} \|\mu_i - \mu_j\|^2}{ns} \tag{8}$$

where ns is the total number of the adjacent segments in the segmented signal, μ_i and μ_j are the mean value of the sample entropy of the i^{th} and j^{th} segments.

4.4. Evolution Procedure

In the proposed algorithm, the multiple subpopulations approach provided by [9] is applied for the evolutionary process. It is implemented through the use of high-level genetic operator functions and exchanging individuals between subpopulations. Over the generations, each subpopulation is evolved as in traditional simple genetic algorithm (SGA) using the basic operators *crossover* and *mutation*. The initial population is created using 8 subpopulations containing 20 individuals each. At each generation, 90% of the individuals with higher fitness values within each subpopulation are selected for breeding using a *stochastic universal sampling* function.

New offsprings within each subpopulation are produced by *discrete recombination crossover*, which is a uniform crossover for real-valued representation. The offspring are then mutated with a mutation rate of $1/nvar$, where $nvar$ is the length of an individual. Offspring may now be inserted into the appropriate subpopulations depending on fitness-based reinsertion with a rate 0.9. In this multi-population GAs, migration of individuals between subpopulations is performed at every 20 generations with a migration rate of 0.2. After GA iterates for *maxgen* times (here *maxgen*=80), the evolution of this GA stops. The best individual with the maximum fitness value presents the optimized solution for the boundaries of the segments of the segmented signal.

5. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed GA based segmentation, the experiments are carried out on the speech signals of 20 speakers (10 male and 10 female) from the TIDIGITS database. 10 digit strings of lengths varying form 3 to 7 digits have been considered from each speaker. If there exist long inter-word silences, they are extracted first. For reference decision the boundaries of each segment are manually determined. White noise, babble noise and car noise are then mixed to obtain the corrupted signals at different SNRs (20 dB, 15 dB, 10 dB, 5 dB and 0 dB). Therefore a total of 3000 signals are applied to the proposed algorithm. Fig. 2 demonstrates the segmentation of a given

Table 1. Relative Error at Different SNRs.

	Noise Types		
SNR	White	Car	Babble
20 dB	0.0858	0.0983	0.1077
15 dB	0.1088	0.1194	0.1314
10 dB	0.1378	0.1512	0.1737
5 dB	0.1753	0.1883	0.1990
0 dB	0.2092	0.2332	0.2553

noisy speech signal. Fig. 2(a) shows a noisy speech signal at 0 dB SNR and Fig. 2(b) shows its corresponding clean signal. The detected boundaries of each segment are also shown by the vertical dashed line in Fig. 2(b).

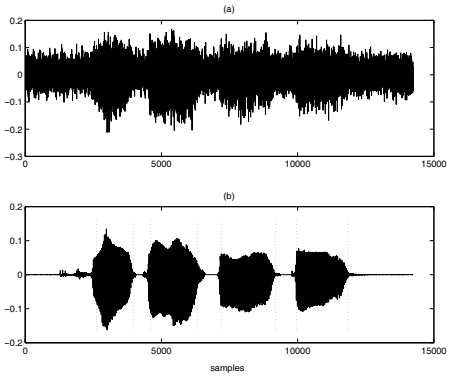

Fig. 2. (a) Noisy input signal; (b) Segmentation results shown in dashed line together with the clean speech signal.

When there is no inter-word silence between consecutive digits, the proposed method detect them as one segment only. To tackle this problem, durational information of digits studied in [10] is applied. When a detected segment has the duration greater than 390 ms, the mean duration of digits spoken by male speakers, further segmentation is performed to find another segment within this interval. Similarly, if the duration of a segment is less than 150 ms, this is merged to one of its neighboring segments having shorter duration. As a performance measure, relative error of a segment RE is calculated as follows:

$$RE = \frac{|AD - ED|}{AD} \tag{9}$$

where AD is the duration of the segment which is determined manually and ED is duration of the segment estimated by the proposed method. In Table 1, the results for the performance measure of the proposed segmentation algorithm are given. For white noise it is found to have the lowest segmentation errors at all levels of SNRs while for the babble noise it has the highest relative error.

6. CONCLUSION

A segmentation method for noisy speech is presented. The problem of segment boundary detection is formulated as an optimization problem. Based on the genetic algorithm (GA) and sample entropy, the start and end points of the segments are determined. Experimental results show that the proposed method can detect the speech segments as well as noise only segments precisely. To our best knowledge, GA based noisy speech segmentation has not been published earlier. Thus the comparisons with other types of method will be done in future.

7. REFERENCES

[1] S. MacDougall, A. K. Nandi and R. Chapman, "Multiresolution and Hybrid Bayesian Algorithms for Automatic Detection of Change Points," *IEE Proc. Vision, Image and Signal Processing,* vol. 145, no. 4, pp. 280-286, 1998.

[2] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman, "Sample Entropy Analysis of Neonatal Heart Rate Variability,", *Am J Physiol* vol. 283, no.(3), R. 789-797, 2002.

[3] X. Jin and C. H. Davis, "A Genetic Image Segmentation Algorith with a Fuzzy-Based Evaluation Function," *IEEE Proc. Fuzzy Systems*, vol. 2, pp. 938-943, 2003.

[4] S. M. Pincus, "Approximate Entropy(ApEn) as a Complexity Measure," *Chaos*, vol. 5, pp. 110-117, 1995.

[5] J. S. Richman and J. R. Moorman, "Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy," *Am J Physiol Heart Circ Physiol.,* vol. 278, no. (6), H2039-H2049, 2000.

[6] D.Arfib, F. Keiler, and U. Zöler, *DAFX - Digital Audio Effects*, John Wiley Publisher, pp. 237-277, 2002.

[7] A. Quddus and M. Gabbouj,"Wavelet-based Corner Detection Technique Using Optimal Scale," *Pattern Recognition Letters*, vol. 23, pp. 215-220, 2002.

[8] K. S. Tang, K. F. Man, S. Kwong and Q. He, "Genetic Algorithms and their Applications,", *Signal Processing Magazine, IEEE.,* vol.13, no. (6), pp. 22-37, 1996.

[9] A. Chipperfield, P. Fleming, H. Pohlheim and C. Fonseca, "Genetic Algorithm Toolbox," *Department of Automatic Control and Systems Engineering, University of Sheffield.,* 1995.

[10] V. K. Prasad, T. Nagarajan and H. A. Murthy, "Automatic Segmentaion of Continuous Speech Using Minimum Phase Group Delay Functions,", *Speech Communication.,* vol. 42, no(3-4), pp. 429-446, 2004.