ORIGINAL RESEARCH

An evolutionary approach for segmentation of noisy speech signals for efficient voice activity detection

Farook Sattar*[1], Frank Rudzicz[1,2], Moe Pwint[3]

[1] Toronto Rehabilitation Institute, University of Toronto, Toronto, Canada
[2] Department of Computer Science, University of Toronto, Toronto, Canada
[3] Computer University, Pyay, Myanmar

Received: August 21, 2015 **Accepted:** October 14, 2015 **Online Published:** October 26, 2015
DOI: 10.5430/air.v5n1p56 **URL:** http://dx.doi.org/10.5430/air.v5n1p56

Abstract

This paper presents a new approach to automatically segmenting speech signals in noisy environments. Segmentation of speech signals is formulated as an optimization problem and the boundaries of the speech segments are detected using a genetic algorithm (GA). The number of segments present in a signal is initially estimated from the reconstructed sequence of the original signal using the minimal number of Walsh basis functions. A multi-population GA is then employed to determine the locations of segment boundaries. The segmentation results are improved through the generations by introducing a new evaluation function which is based on the sample entropy and a heterogeneity measure. Experimental results show that the proposed approach can accurately detect the noisy speech segments as well as noise-only segments under various noisy conditions.

Key Words: Segmentation, Speech signals, Genetic algorithm, Sample entropy, Voice activity detection

1. Introduction

Speech segmentation is the process of partitioning a speech signal into homogeneous, contiguous units (*e.g.*, words, phonemes). Automatic segmentation is a crucial part of many applications of speech and audio processing including automatic speech recognition (ASR),[1] transcription,[2] classification of audio-visual data,[3] indexing,[4] segmenting of broadcast news[5] and speech/music discrimination.[6] Manual segmentation can be tedious, time consuming, and therefore unrealistic for applications with increasingly large databases.

Noise is one of the most prominent causes of error in speech processing applications such as ASR.[7] In this paper, we introduce a method to partition spoken words of continous speech signals which are corrupted by unknown levels and types of noise. The segmentation method described here is based on a genetic algorithm, which is a stochastic global search method. Genetic algorithms (GAs) were originally developed by Holland[8] and are relatively robust against local optima due to their parallel exploration of the search space. GAs are numerical optimization algorithms inspired by both natural selection and natural genetics.[9] GAs operate on a population of "chromosomes", which are defined here as a group of potential solutions to a problem. In this work, a chromosome is a real-valued vector representing the ordered locations of the candidate segment boundaries. To evaluate the alternative simultaneous solutions within the population, the "fitness" of each solution is calculated according to a provided evaluation function. At each generation, a new set

[*] **Correspondence:** Farook Sattar; Email: farook_sattar@yahoo.com.sg; Address: Toronto Rehabilitation Institute, University of Toronto, Toronto, Canada.

of solutions are produced by selecting the fittest chromosomes in the domain and through the application of "genetic operators" such as crossover and mutation, described below.

Unlike our previously proposed single-stage approach,[10] here we propose a two-stage GA-based segmentation approach. In the first stage, the original signal is reconstructed into a modified sequence using minimal binary Walsh basis functions. The Walsh functions form an ordered set of rectangular waveforms taking only two amplitude values +1 and -1 defined over a limited time interval.[11] In most cases, a set of Walsh functions is arranged in ascending order by the number of zero-crossings. Obtaining the local and global variations from the reconstructed signal, the number of segments present in the segmenting signal is determined using the *mean difference measure*.[12]

The second stage applies a genetic algorithm to segment the noisy speech signal into homogeneous regions according to the speech and non-speech conditions. The evaluation function here measures the regularity and homogeneity of speech segments using a metric called *sample entropy*[13] and *heterogeneity*.[14] Locations of segment boundaries are optimized through multiple generations of GA. The performance of the proposed GA-based speech segmentation method is evaluated on signals from the TIDIGITS database[15] with a sampling frequency of 8 kHz. The experimental results show that the proposed method can detect the boundaries of speech and non-speech events with high accuracy in various noise conditions, based on work described in Ref.[16]

2. Related work

Sohn *et al.*(1999)[17] proposed a voice activity detector (VAD) based on statistical models, which employ the decision-directed parameter estimation method for the likelihood ratio test together with a hang-over scheme using HMMs. Ramirez *et al.*(2004)[18] presented a VAD algorithm that measured the long-term spectral divergence (LTSD) between speech and noise, while making the speech/non-speech decision by comparing the long-term spectral envelope to the average noise spectrum. Ramirez *et al.*(2006),[19] presented a segmentation method based on speech endpoint detection using contextual feature vectors and SVM classifiers. The contextual feature vector consists of subband SNRs calculated from the long-term spectral envelopes using surrounding frames. Ramirez *et al.*(2007)[20] derived a revised contextual likelihood ratio test (LRT) defined over a multiple observation window and applied for VAD with a range of SNRs. Fujimoto *et al.*(2007),[21] proposed a noise robust VAD technique by integrating the periodic-to-aperiodic component ratio (PAR) and switching Kalman filter (SKF). Fujimoto *et al.*(2008)[22] presented another statistical VAD, where the estimate of the noise mean vector and the calculation of the likelihood are based on a parallel Kalman smoother and a backward probability estimation. Ishizuka *et al.*(2010),[23] used power ratios of the periodic and aperiodic components of observed signals for VAD, where a sum of the powers of harmonic components and the average power over the whole frequency range are used to calculate the power ratio. Deng *et al.*(2013)[24] proposed a statistical VAD method based on sparse representations over the learned dictionary for which the non-zero elements in the sparse representation is modeled as Gaussian distribution, and the decision rule is derived based on Bayesian framework. Aneeja *et al.*(2015)[25] presented a single-frequency filtering approach for VAD based on the weighted envelop from the output of a filter with fixed frequency. Zhang *et al.*(2013)[26] introduced a multiple-feature fusion method using deep-belief networks (DBNs) for VAD by exploiting the deep model to combine multiple features nonlinearly. Zou *et al.*(2014)[27] exploited the MFCC features and SVM classifier for VAD and Tu *et al.*(2014)[28] proposed VAD methods using discriminative acoustic features from computational auditory scene analysis.

3. Methods – determining the number of segments

In this paper, the number of segments present in an input speech signal is estimated from its modified output sequence. In order to modify the original signal, an analysis and synthesis scheme and a set of basis functions are employed. Here, binary Walsh basis functions are selected for modification, since they are computationally simple. The number of segments present in a given signal is obtained approximately from this modified signal by applying the *mean difference measure*.[12] This is performed by finding the number of local maxima as calculated from the difference of the local means by gliding two adjacent windows of equal length over the modified signal followed by thresholding.

3.1 Minimal Walsh basis functions

In order to capture differences between speech dynamics and non-speech segments, appropriate basis functions must be selected. Here, we empirically determine the minimal number of Walsh basis functions, as these functions essentially compress the signal in a way so that the modified signal could capture the separability of speech and non-speech segments. For this purpose, we use an algorithm that selects the global natural scale in the discrete wavelet transform[29] when the method adaptively detects the optimal scale using singular value decomposition (SVD).

The Walsh transform matrix constitutes complete orthogonal functions with only two possible values, +1 and -1, over their definition interval. A Walsh function of order M can be represented as

$$g(x,u)=\frac{1}{M}\prod_{i=0}^{q-1}(-1)^{b_i(x)b_{q-1-i}(u)} \tag{1}$$

where $u=0,1,\cdots,M-1$, $M=2^q$ and $b_i(x)$ is the i^{th} bit value of vector x. The Walsh functions are arrayed into *sequency* order to obtain a set of basis functions, W. *Sequency* is the total count of zero crossing occurrences of the Walsh function over the definition interval, which increases with the order of those basis functions. While the 0^{th}-order basis function ϕ_0 has only one interval, first order basis function ϕ_1 has two equal sub-intervals and ϕ_{M-1} has 2^q equal sub-intervals.

$$W=[\phi_0,\ \ \phi_1,\ \ \cdots\ \ ,\phi_{M-1}] \tag{2}$$

The number of basis functions is determined analytically, by evaluating the probability distributions of these orders as a function of signal-to-noise (SNR) ratios. Here, we add the traditional “white” Gaussian noise is with SNR levels 20 dB, 10 dB, 5 dB, and 0 dB where

$$\mathrm{SNR}=10\log_{10}\frac{\sum_{n=1}^{N_s}s^2(n)}{\sum_{n=1}^{N_v}v^2(n)} \tag{3}$$

n is the time index, and N_s and N_v are the lengths of speech s and noise v, respectively.

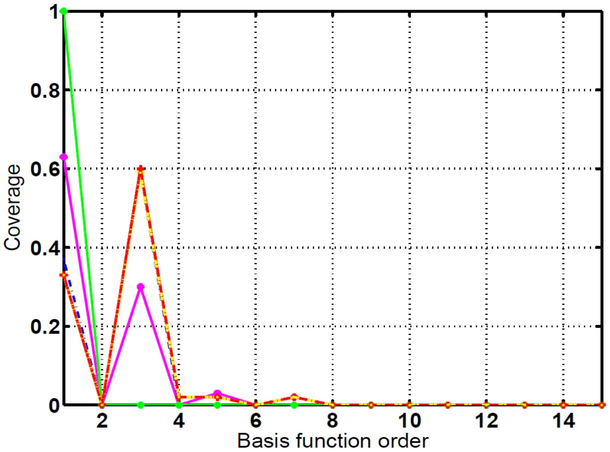

Figure 1. The distribution of the order of basis functions for the signals from clean to 0 dB (red: Clean, yellow: 20 dB, blue: 10 dB, magenta: 5 dB, green: 0 dB)

Figure 1 illustrates probability distributions over basis function orders, called *coverage*, with respect to SNR. Dominant eigenvalues are only located at the first few basis functions. This is especially pronounced for the noisier signals at 5 dB and 0 dB SNR, where the minimal order is one. Signals with relatively high SNRs (*i.e.*, 10 dB and above) have dominant eigenvalues with third-order basis functions. In our SVD analysis, lower-order basis functions of Walsh transform matrices appear useful. Therefore, only the first few orders of basis functions are selected for the modification process with a proper implementation.[30]

Practically, it is not always possible to expect prior knowledge about noise level or type. We therefore propose that the minimal order of basis functions is three, which is the median of the corresponding range[1,5] (see Figure 1). In the original algorithm, Quddus and Gabbouj[29] defined the optimal scale as an average over the first level to the natural scale. Unfortunately, averaging of this type may introduce clipping effects for signals with low SNR.[31] In response, we first apply a shifting operator which swaps the left and right halves of the basis function coefficients. This makes the basis functions conjugate symmetric. The estimated binary Walsh basis function at the dominant eigenvalue is therefroe defined as

$$\phi_m=\frac{\phi_0-\sum_{i=1}^{N_{min}}CS(\phi_i)}{\max\{|\phi_0-\sum_{i=1}^{N_{min}}CS(\phi_i)|\}} \tag{4}$$

where $N_{min}=3$ is the largest order of basis function with the most prominent eigenvalues and $CS(\cdot)$ is the shifting operator as it performs a N/2-circular shift to avoid clipping out the information at both ends.

3.2 Signal modification

The noisy input signal is reconstructed as a modified sequence based on an analysis/synthesis scheme described in Ref.[32] At the analysis stage, the signal is represented using fast Fourier transforms (FFTs). These representations are then modified by Walsh basis functions before reconstruction at the synthesis stage. First, the input signal $x(n)$ is multiplied by a Hanning window to get the successive windowed segments $x_s(n)$ which are then converted to the spectral domain by FFT. In this manner, a time-varying spectrum $X_s(n,k)=|X_s(n,k)|e^{j\varphi(n,k)}$ with $n=0,1,\cdots,N-1$ and $k=0,1,\cdots,N-1$ for each windowed segment is computed. Here, $X_s(n,k)$ denotes the spectral component of the noisy input signal at frequency index k and time index n, where $|X_s(n,k)|$ represents the magnitude and $\varphi(n,k)$ is the phase of the time varying spectrum. Before synthesis, each s^{th} windowed segment is modified as the weighted sum of the magnitude $|X_s(n,k)|$ using binary Walsh basis functions. The modified sequence, $y_s(n)$, for each windowed segment is computed as

$$y_s(n) = \sum_{k=0}^{N-1} |X_s(n,k)|.\phi_m(k) \tag{5}$$

We then concatenate each of the adjusted S segments, which results in the output signal $y(n)$:

$$y(n) = \sum_{s=0}^{S-1} y_s(n - sN) \tag{6}$$

The effectiveness of the proposed modification scheme is demonstrated in Figure 2. A noisy signal in the white Gaussian noise at 5 dB SNR is shown in Figure 2(a). For comparison, the corresponding clean signal is displayed in Figure 2(b). Provided that sharper representation and higher discriminating features in the modified sequence, the noisy input signal has been reconstructed by applying the optimal basis function ϕ_m into Eqs. 5 and 6. The resultant signal is presented in Figure 2(c) with very low values at the locations of noise-only intervals.

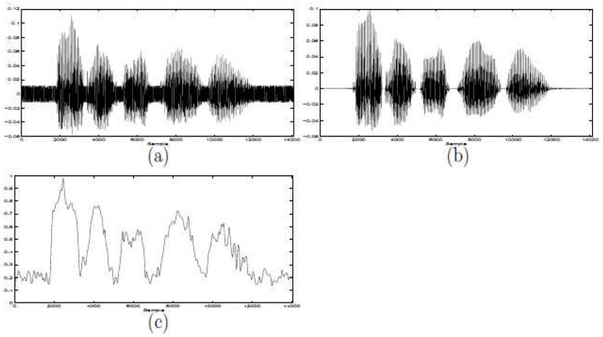

Figure 2. (a) The noisy signal, (b) the corresponding clean signal, (c) the modified signal using the basis function ϕ_m

For illustration, we have compared the results of the modified signal with that of the minimum mean-square error (mmse) filter[33] used for speech enhancement (see Figure 3). As we see in Figure 3(b), the signal-to-noise contrast of the modified signal is higher (by approximately double) than that of the envelope for the output of mmse filter.

3.3 Mean difference measure

The input to GA (*i.e.*, the number of segments) is determined using the modified sequence obtained in the previous section and the mean difference measure described in Ref.[12] In this approach, the two adjacent windows of equal length are moved through the modified signal; at each position, the magnitude of the difference of the means within each window is calculated. This sequence of mean differences is thresholded to constrain the local maxima. Local maxima having a width greater than a specified value are taken as the segments. All the other maxima which do not meet these conditions are discarded. Finally, the total number of significant local maxima satisfying both the requirements are assumed to correspond to the number of segments present in the segmenting signal. Here, the width of the window is 62.5 ms.

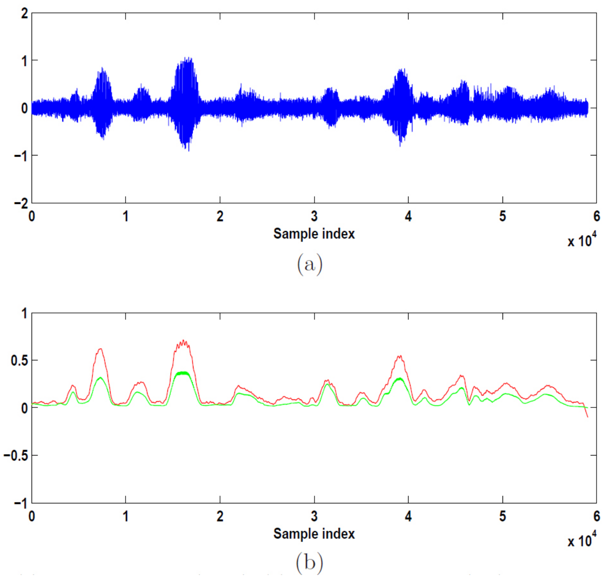

Figure 3. (a) The noisy signal (5 dB), (b) the modified signal (red) and envelope of the output of mmse filter (green)

4. Methods – Segmentation by genetic algorithm

As the second stage, the locations of the segment boundaries are detected using GA. Depending on the total number of segments estimated from the first stage as discussed in the previous section, an initial population is randomly generated with uniform distribution where the number of random variables are twice the total number of segments. To guide the search space of GA, a new evaluation function measures the *homogeneity* and *heterogeneity* of the candidate segments as the locations of segment boundaries are optimized through the generations.

4.1 Sample entropy

In this segmentation method, a similarity measure of the time series (sample entropy) is employed to determine the boundaries of speech segments. The origin of sample entropy ($SampEn$)[13] is the *approximate entropy* ($ApEn$), which is introduced in Ref.[34] to measure the regularity in the time series. $ApEn(m,r,N)$ is defined as the negative natural logarithm of the conditional probability that a data set of length N, having repeated itself (*i.e.* iterated) within a tolerance r for m points, will also repeat itself for $(m+1)$

points. Small values of $ApEn$ indicate a high regularity in time series while large values of $ApEn$ implies that the time series is irregular.

Recently, $ApEn$ has been applied to analyze the time series of clinical cardiovascular data.[35] Since the $ApEn$ algorithm counts self-matching, it is (i) lacking relative consistency and (ii) heavily dependent on the signal length. To reduce the bias and inconsistent results caused by self-matching, sample entropy ($SampEn$), which does not count self-matches, is developed in Ref.[13] $SampEn(m, r, N)$ is defined as 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, without allowing self-matches. Thus, a low value of *SampEn* reflects a high degree of self-similarity in a time series. In Ref.,[36] the dynamics of neonatal heart rate variability was investigated using $SampEn$.

Figure 4(b) illustrates a sequence of sample entropy measurements calculated for the clean speech shown in Figure 4(a). For non-speech portions, the value of $SampEn$ is minimal. $SampEn$ increases for speech segments and decreases during the appearance of non-speech regions. Figure 5 also shows the analysis of *SampEn* for a speech signal in a noisy background. The plots of the clean signal and the noisy signal at 5 dB SNR are shown in Figures 5(a) and (b). The sample entropy sequence calculated for the noisy signal is shown in Figure 5(c). In this example, the *SampEn* of the speech signal often increases before noisy speech components and decreases quickly thereafter.

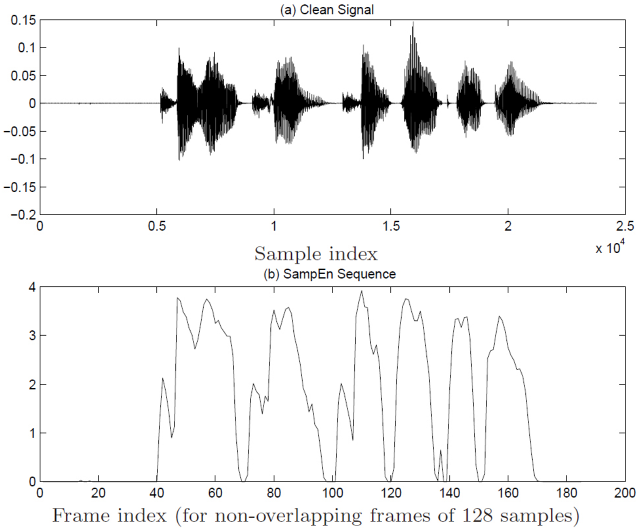

Figure 4. The sample entropy of a clean speech signal

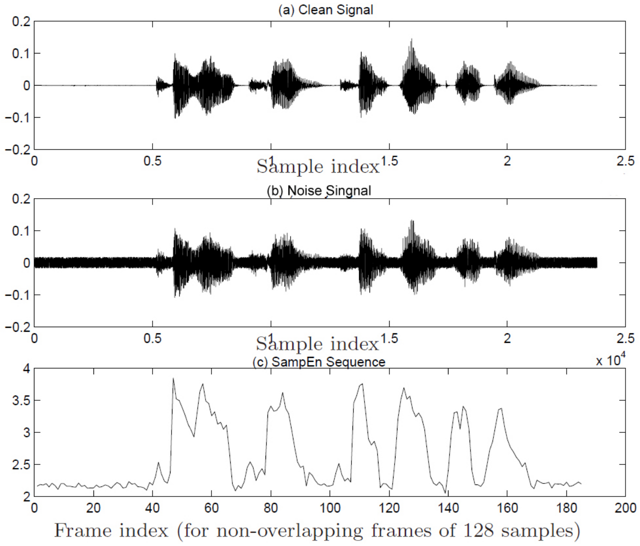

Figure 5. The sample entropy of the noisy speech signal

4.2 Real-valued genetic algorithm

Real-valued GA uses selection, crossover, and mutation operators to generate the offspring of the existing population. Specifically,

Selection Our implementation of the real-valued GA incorporates stochastic universal sampling (SUS) and the "roulette wheel" method. SUS is a kind of fitness proportionate selection which exhibits no bias and minimal spread.[37]

Crossover Once a pair of chromosomes has been selected for crossover, these vectors are randomly split in one or more positions and recombined to generate new members of the population.

Mutation This determines how a chromosome should be mutated in the next generation. In this study, a non-uniform mutation method is applied in which the k^{th} parameter is set to $LB_k + r(UB_k - LB_k)$ where r is a random number taken from Gaussian $N(0, \sigma)$, and LB_k and UB_k denote the lower and upper bound at location k, respectively.

4.2.1 *Initial population*

The number of potential local minima determined in Section 3 is defined as the number of segments present in a given signal. In order to detect both start and end locations of each segment, a population is generated with chromosomes whose lengths are twice the total number of segments obtained above. Although binary-coded GAs are the most commonly used representation of chromosome,[38] a real-valued representation is used in this system to increase the efficiency of GA. Using the real-valued chromosomes, there is no need to convert chromosomes to phenotypes to evaluate their fitness.[38]

4.2.2 *Evaluation function*

In order to obtain accurate boundaries of each segment, an evaluation function is designed using the heterogeneity measure and sample entropy. This function simultaneously maximizes the homogeneity within the segments and heterogeneity among different segments using sample entropy. In this context, $SampEn$ of the original segmenting signal is calculated first, capturing the dynamics on each data set of length $N = 80$ within a tolerance $r = 0.1 \times \sigma$ for 1 point (*i.e.* $m = 1$) where σ is the standard deviation of the data set. If H_w is the total within-segment homogeneity and H_b is the total between-segment heterogeneity, our segmentation evaluation function is defined as

$$H = \frac{H_b + 1}{H_b + H_w + 1} \tag{7}$$

where total within-heterogeneity H_w is defined as

$$H_w = \frac{\sum\limits_{i=1}^{S} L_i \sigma_i^2}{L} \tag{8}$$

and 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 the 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 i^{th} and j^{th} segments.

$$H_b = \frac{\sum\limits_{(i,j) \in X | X = \{i,j\}} \|\mu_i - \mu_j\|^2}{ns} \tag{9}$$

where ns is the total number of the adjacent segments in the segmented signal, μ_i and μ_j are the mean values of the sample entropy of the i^{th} and j^{th} segments respectively. $H = 1$ when the speech unit internals (*e.g.* the variances of all segmented speech) are completely homogeneous.

4.2.3 *Evolution procedure*

One problem with simple or sequential GA is its premature convergence to a suboptimal solution. In order to effectively search the solution space, and to take advantage of the parallelism of GAs, the proposed algorithm applies the multiple subpopulations approach provided by Ref.[39] where multiple subpopulations evolve independently toward different optima. More diverse subpopulations can be maintained by exchanging genetic materials between subpopulations, mitigating premature convergence. Subsets from subpopulations migrate to others according to a migration interval (*i.e.*, the number of generations between such migrations) and the migration rate (*i.e.*, the number of individuals to be migrated). The initial population is created using 16 subpopulations containing 120 individuals each providing least segmentation error (see section 5.1). At each generation, 90% of the individuals (determined empirically) with the highest fitness within each subpopulation are selected for breeding using *stochastic universal sampling*. By applying *discrete recombination crossover*, a uniform crossover for real-valued representations, the new offspring within each subpopulation are produced.

In this segmentation method, offspring are inserted into the appropriate subpopulations depending on *fitness-based reinsertion* with a rate of 0.3, meaning that 30% of the offspring are inserted based on the fitness. Here, migration of individuals between subpopulations is performed at every 20 generations with a migration rate (*i.e.*, the migration probability of the individuals) of 0.4. Optimization stops after 80 iterations due fast converstion given the small population size.[40] The individual with the maximum fitness represents the optimized solution for the boundaries of the segments for the segmented signal. The values of the GA parameters such as population size, reinsertion rate, migration rate mentioned above, are chosen empirically as demonstrated in section 5.1.

5. Experiments

To evaluate the proposed GA-based segmentation, experiments are carried out on the speech of 10 speakers (5 male and 5 female) from the TIDIGITS database, each producing 20 utterances of 10-digit strings with length of each string varies between 3 to 7 digits. White noise, babble noise, car noise, street noise, and train noise from the NOISEX-92 database[41] and the database of typical background noise used to simulate real-world conditions in the evaluation of AURORA[42] are then added to obtain the corrupted signals at different SNRs.

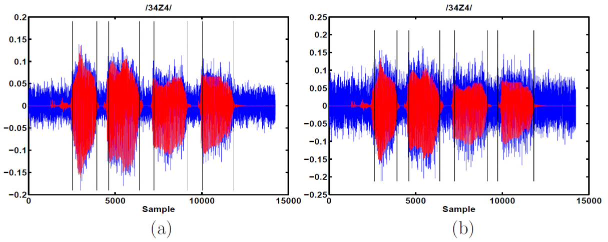

Figure 6. (a) Illustrative results for white Gaussian noise at (a) 5 dB SNR and (b) 0 dB SNR. Clean signals are superimposed in red and the detected boundaries of each segment are shown by vertical lines

Five levels of SNRs are considered (20 dB, 15 dB, 10 dB, 5 dB, and 0 dB). Therefore, the proposed method is applied to a total of 1,000 signals sampled at 8 kHz, downsampled to match the data in Ref.[42] Figures 6(a) and 6(b) demonstrate the segmentation results of an input noisy speech signal. Illustrative results for white Gaussian noise at 5 dB and 0 dB SNR are depicted in Figures 6(a) and 6(b), respectively.

5.1 Results - relative error

The relative error of a segment is RE=$\frac{|AD-ED|}{AD}$ where AD is the true duration of the segment, ED is duration of the segment estimated by the proposed method, and $|\cdot|$ is the absolute value. Table 1 shows RE across SNR levels and type of noise. Table 1 shows that nearly 96% of the segments are detected within 25 ms of manually determined boundaries. The highest segmentation errors (between 21% and 25%) occur at 0 dB SNR. White noise results in the lowest segmentation error at all levels of SNRs (except 15 dB), while the babble noise has the highest relative error for all instances (except 10 dB). Table 2 shows the results of an ANOVA indicating that the proposed method is statistically significant ($p << .05$).

Table 1. Relative Error (RE[%]) for the TIDIGITS database

Item			**Noise Types**		
SNR	White	Car	Babble	Street	Train
Clean	8.35	8.35	8.35	8.35	8.35
20 dB	9.39	9.52	11.50	10.50	10.52
15 dB	9.96	9.62	12.14	11.64	11.62
10 dB	11.44	11.82	14.85	15.40	13.05
5 dB	17.60	18.50	19.23	18.84	18.61
0 dB	21.37	23.45	24.25	24.14	23.17

Table 2. ANOVA table for the relative error

Source of Variation	**SS**	**df**	**MS**	**F**	***p***
Between Groups	811.374	5	162.275	124.92	2.22045×10^{-16}
Within Groups	31.176	24	1.299	-	-
Total	842.550	29	-	-	-

Figure 7 shows the variations of the relative error with respect to the number of generations. This convergence is generated by testing on the signals over all speakers in white Gaussian noise.

Figure 8 shows the effects of using different GA parameters. The corresponding results are obtained for speech signals with 10 dB SNR and white Gaussian noise using the TIDIGITS database. As shown in Figure 8(a), the migration rate of 0.4 is chosen since it provides the lowest relative error. In Figures 8(b) and 8(c), the RE with respect to the reinsertion rate and the number of subpopulation are depicted where 0.3 and 16 are selected in terms of the lowest relative error. The other important parameter is the total number of individuals per subpopulation, N, which should be large enough (*e.g.*, $N = 120$ as used here) to give higher level of genetic diversity within the populations and perform the GA operations properly.

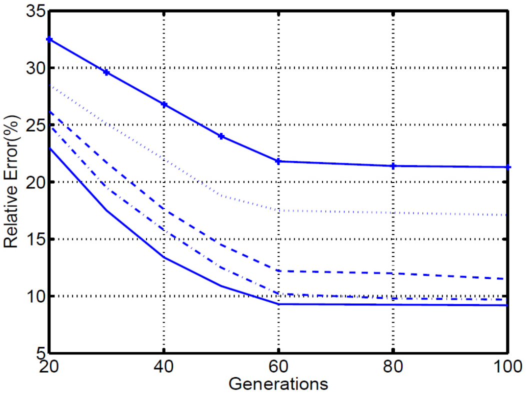

Figure 7. Variation of the relative error with the number of generations ("-": 20 dB, "-": 15 dB, "- -": 10 dB, ":": 5 dB, "-+": 0 dB)

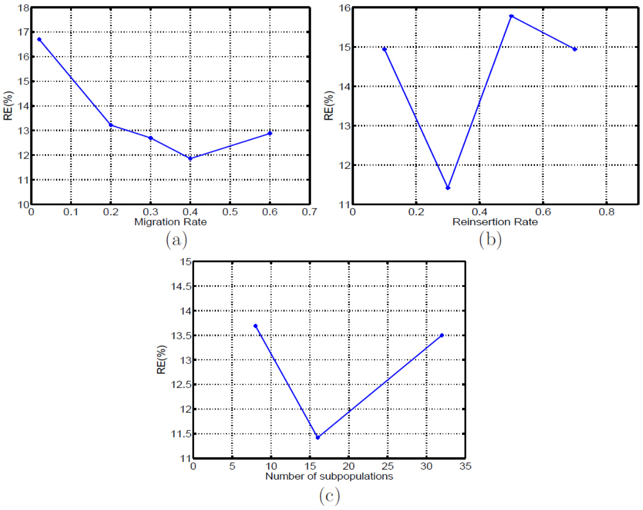

Figure 8. The effects of GA parameters (a) RE *vs.* migration rate, (b) RE *vs.* reinsertion rate, (c) RE *vs.* number of subpopulations

5.2 Results - precision, recall, and F-measure

Segmentation errors can be categorized into insertion errors and deletion errors. An insertion error occurs when a detected segment boundary does not correspond to one in the manual transcription, *i.e.*, when a detected boundary is outside of the search region which in turn is defined as over-segmentation or when there is more than 1 boundary within a single reference boundary region. Deletion errors occur if a detected segment has a segmentation error greater than 25 ms.[43] Deletion and insertion errors are used to compute precision (*PRC*) and recall (*RCL*).[44] The F-measure combines precision and recall as

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \tag{10}$$

The performance of the proposed method at different noise conditions is shown across Tables 3 and 4. The present method can maintain high recall rates, high precision rates and provide high F-measures at different SNRs.

Table 3. Recall and precision for different types of noise

Item	Recall					Precision				
Noise	20 dB	15 dB	10 dB	5 dB	0 dB	20 dB	15 dB	10 dB	5 dB	0 dB
White	0.90	0.85	0.79	0.67	0.56	0.79	0.75	0.69	0.59	0.50
Car	0.90	0.86	0.78	0.67	0.54	0.84	0.79	0.72	0.62	0.50
Babble	0.87	0.82	0.74	0.64	0.52	0.82	0.78	0.71	0.62	0.50
Street	0.89	0.83	0.76	0.64	0.53	0.84	0.79	0.72	0.60	0.53
Train	0.88	0.82	0.76	0.64	0.52	0.84	0.79	0.73	0.62	0.53

Table 4. F-measure for different types of noise

Item	F-measure				
Noise	20 dB	15 dB	10 dB	5 dB	0 dB
White	0.83	0.79	0.73	0.62	0.51
Car	0.86	0.82	0.75	0.64	0.50
Babble	0.84	0.80	0.73	0.62	0.51
Street	0.86	0.81	0.73	0.62	0.51
Train	0.85	0.80	0.74	0.63	0.52

5.3 Results – Speech/non-speech classification

Table 5 compares the performance of the proposed method with those of Ref.[45] and the baseline hidden Markov model (HMM) method, using speech signals in white Gaussian noise of the TIDIGITS database. The method in Ref.[45] converts the input signal into the frequency domain then estimates voiced activity based on a channel energy estimator, channel SNR estimator, spectral deviation estimator, background noise estimator, and a peak-to-average-ratio module. The HMM method uses the first 12 Mel-frequency cepstral coefficients and 1 mixture component per state. Here, D_s is the ratio of correct speech decisions to the total number of manually marked speech frames and D_n is the ratio of correctly detected non-speech decisions to manually determined non-speech frames, respectively. E_r is the ratio of total false decisions to total frames. In this comparison, frames are 10 ms in length. Under different levels of SNR, our method achieves better performance in terms of speech/non-speech classification with small error rates, and significantly different than the alternatives, given an ANOVA analyses on accuracy (see Table 6).

Table 5. Classification accuracy of speech and non-speech frames

Item	Proposed			AMR2 [45]			HMM [46]		
SNR	E_r (%)	D_n (%)	D_s (%)	E_r (%)	D_n (%)	D_s (%)	E_r (%)	D_n (%)	D_s (%)
20 dB	6.13	96.10	91.55	30.87	68.77	69.06	7.67	92.05	90.66
15 dB	7.47	95.53	89.66	31.82	71.08	65.76	9.17	91.47	89.76
10 dB	9.29	95.00	86.72	33.34	75.11	60.38	35.95	79.06	69.60
5 dB	11.49	93.60	83.67	38.10	85.97	44.70	46.81	62.23	49.97
0 dB	13.55	92.21	81.18	47.20	95.02	23.39	50.94	54.36	50.55

Table 6. ANOVA table for the classification accuracy by the proposed method

Source of Variation	SS	df	MS	F	*p*
Between Groups	21,992.7	2	10,996.4	1,122.04	2.26485 ×10^{-14}
Within Groups	117.6	12	9.8	-	-
Total	22,110.3	14	-	-	-

5.4 Illustrative performance with TIMIT

Some illustrative results are obtained for the non-digit TIMIT database,[47] specifically the 10 sentences spoken by each of 35 speakers with the New York dialect. For the actual segment duration *AD*, we have taken the manually annotated labels of the word segment boundaries[48] as the number of segments are determined in terms of words. The relative error of the proposed method is shown in Table 7 for various noise levels and noise types (as described in section 5). The performance of the proposed method has been compared with the baseline HMM method in terms of speech/non-speech classification as shown in Table 8. Also, ANOVA tests are performed for the results obtained by the proposed method using TIMIT database. The corresponding results are presented in Tables 9 and 10 indicating that the results of the proposed method are significantly more accurate.

Table 7. Relative Error (RE[%]) at different SNRs for the TIMIT database

Item	Noise Types				
SNR	White	Car	Babble	Street	Train
Clean	4.83	4.83	4.83	4.83	4.83
20 dB	12.63	15.05	14.67	14.76	14.60
15 dB	14.75	15.16	15.93	15.76	15.04
10 dB	15.59	15.42	16.44	15.96	16.32
5 dB	22.01	22.77	24.60	22.93	23.78
0 dB	25.76	26.30	27.71	28.03	26.71

Table 8. Classification accuracy of speech and non-speech frames for the TIMIT database

Item	Proposed			HMM [46]		
SNR	E_r (%)	D_n (%)	D_s (%)	E_r (%)	D_n (%)	D_s (%)
20 dB	8.23	90.53	90.45	33.36	90.01	90.18
15 dB	9.47	92.01	88.66	34.89	77.97	80.36
10 dB	11.29	94.00	86.72	35.60	75.62	69.12
5 dB	12.94	96.10	83.67	43.13	67.34	63.99
0 dB	26.95	96.80	75.92	50.03	49.96	50.04

Table 9. ANOVA table for the relative error (TIMIT database)

Source of Variation	SS	df	MS	F	*p*
Between Groups	1,477.79	5	295.558	538.51	0
Within Groups	13.17	24	0.549	-	-
Total	1,490.96	29	-	-	-

Table 10. ANOVA table for the classification accuracy by the proposed method (TIMIT database)

Source of Variation	SS	df	MS	F	*p*
Between Groups	19,300.5	2	9,650.23	298.34	5.87165×10^{-11}
Within Groups	388.2	12	32.35	-	-
Total	19,688.6	14	-	-	-

5.5 Comparison with a conventional VAD

We have compared the results with a conventional VAD based on change-point detection over speech segments. Here, it is assumed that speech signals often change around the beginning or end of each speech/non-speech segment. Several steps must be performed to detect the change-points. First, the time-domain features (*i.e.*, mean and standard deviation) of all frames are extracted to construct feature vectors. The cosine similarity is then used to measure the distance between the feature vectors of the two consecutive frames and the distances are calculated between all consecutive frames. Finally, the distance values are compared with an empirical threshold to detect the change-points in order to determine speech/non-speech segments as well as their start- and end-timestamps. Note that the frame size and the threshold value used are 10 ms and mean of the mean distances over frames, respectively whereas the frames are overlapped by 50%.

Tables 11 and 12 show the corresponding RE across SNR levels and type of noise for the TIDIGITS and TIMITS database. It can be inferred from the experimental results that the RE obtained for the conventional VAD is much higher than using the proposed method as shown in Tables 1 and 7.

Table 11. Relative Error (RE[%]) for the TIDIGITS database using a conventional VAD

Item	Noise Types				
SNR	White	Car	Babble	Street	Train
Clean	18.96	18.96	18.96	18.96	18.96
20 dB	19.01	19.07	19.30	18.98	19.04
15 dB	19.02	19.20	19.30	19.01	19.24
10 dB	20.02	20.19	20.34	20.01	20.26
5 dB	24.05	24.23	24.42	24.05	24.28
0 dB	29.08	29.34	29.45	29.08	29.34

Table 12. Relative Error (RE[%]) for the TIMIT database using a conventional VAD

Item	Noise Types				
SNR	White	Car	Babble	Street	Train
Clean	19.01	19.01	19.01	19.01	19.01
20 dB	19.49	20.11	19.91	19.94	20.12
15 dB	20.36	20.02	19.93	19.92	20.12
10 dB	20.04	20.55	20.88	19.93	20.15
5 dB	34.15	34.86	34.89	34.92	34.17
0 dB	39.33	39.78	39.85	38.04	38.55

6. Discussion

This paper presents a novel evolutionary scheme for segmenting speech signals in different noisy conditions. The challenge of speech segmentation is formulated as an optimization problem where the start- and end-points of segments are determined by a genetic algorithm using a measure of "fitness" combining sample entropy, a regularity measure, and a heterogeneity measure. The proposed method can accurately detect the speech and noise-only segments – for 96% of the segments determined by the presented method, the deviation of the detected boundaries from manually determined ones is less than 25 ms. In addition, the classification accuracy of speech and non-speech frames are high even at low SNRs and significantly higher than the baseline methods. Future work includes implementing an adaptation scheme for adjusting the hyperparameters of GA and using additional acoustic features to improve the performance.

REFERENCES

[1] Rybach D, Gollan C, Schluter R, *et al*. Audio segmentation for speech recognition using segment features. ICASSP'09; 2009.

[2] Sprugnoli R, Moretti G, Fuoli M, *et al*. Comparing two methods for crowdsourcing speech transcription. ICASSP'13; 2013. pp. 8116–20.

[3] Subashini K, Palanival S, Ramaligam V. Audio-video based segmentation and classification using SVM. Int. Conf. Comp., Comm. and Networking Tech (ICCCNT); 2012 pp. 1-6.

[4] Kiranyaz S, Qureshi AF, Gabbouj M. A generic audio classification and segmentation approach for multimedia indexing and retrieval. IEEE Trans. Audio, Speech and Language Proc. 2006; 14(3).

[5] Lu P, Yan YH. Audio segmentation and classification in a broadcast news task. Journal of Electronics and Information Tech. 2006: 2292-5.

[6] Ajmera J, McCowan I, Bourlard H. Speech/music segmentation using entropy and dynamic features in a HMM classification framwork. Speech Communication. 2003; 40.

[7] Juang BH, Rabiner LR. Fundamentals of speech recognition. Prentice-Hall; 1992.

[8] Holland JH. Adaptation in natural and artificial systems. University of Michigan Press: USA; 1975.

[9] Coley DA. An introduction to genetic algorithms for scientists and engineers. World Scientific; 2001.

[10] Pwint M, Sattar F. Speech/non-speech detection using minimal Walsh basis functions. EURASIP Journal on Audio, Speech and Music Processing. 2007; 39546.

[11] Beauchamp KG. Applications of Walsh and related functions. Academic Press; 1984.

[12] MacDougall S, Nandi AK, Chapman R. Multiresolution and hybrid Bayesian algorithms for automatic detection of change points. Academic Press. 1998; 145(4): 280-6.

[13] Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol. 2000; 278(6): 2039-49.

[14] Jin X, Davis CH. A genetic image segmentation algorithm with a fuzzy-based evaluation function. IEEE Proc. Fuzzy Systems. 2003; 2: 938-43.

[15] Leonard R. A database for speaker-independent digital recognition. Proc. ICASSP. 1984; 9: 328-31.

[16] Pwint M, Sattar F. A segmentation method for noisy speech using genetic algorithm. ICASSP'05; 2005.

[17] Sohn J, Kim NS, Sung W. A statistical model-based voice activity detection. IEEE Signal Process Letters. 1999; 6(1): 1-3.

[18] Ramirez J, Segura JC, Benitez C, *et al*. Efficient voice activity detection algorithm using long-term speech information. Speech Communication. 2004; 42: 271-87. http://dx.doi.org/10.1016/j.specom.2003.10.002

[19] Ramirez J, Yelamos P, Gorriz JM, *et al*. SVM-based speech endpoint detection using contextual speech features. IEEE Electronics Letters. 2006; 42(7).

[20] Ramirez J, Segura JC, Gorriz JM. Revised contextual LRT for voice activity detection. Proc. of ICASSP'07; 2007. pp. 801-4.

[21] Fujimoto M, Ishizuka K. Noise robust voice activity detection based on switching Kalman filtering. Proc. of Interspeech'07; 2007. pp. 2933-6.

[22] Fujimoto M, Ishizuka K, Nakatani T. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. Proc. of ICASSP'08; 2008. pp. 4441-4.

[23] Ishizuka K, Nakatani T, Fujimoto M, *et al*. Noise robust voice activity detection based on periodic to aperiodic component ratio. Speech Communication. 2010; 52(1): 41-60. http://dx.doi.org/10.1016/j.specom.2009.08.003

[24] Deng SW, Han JQ. Statistical voice activity detection based on sparse representation over learned dictionary. Digital Signal Processing. 2013; 23: 1228-32. http://dx.doi.org/10.1016/j.dsp.2013.03.005

[25] Aneeja G, Yegnanarayana B. Single frequency filtering approach for discriminating speech and nonspeech. IEEE/ACM Trans Audio. Speech, and Language Proc. 2015; 23(4): 705-17.

[26] Zhang XL, Wu J. Deep belief networks based voice activity detection. IEEE/ACM Trans Audio, Speech, and Language Proc. 2013; 21(4): 697-710. http://dx.doi.org/10.1109/TASL.2012.2229986

[27] Zou YX, Zheng WQ, Shi W, *et al*. Improved voice activity detection based on support vector machine with high separable speech feature vectors. Proc. of Int. Conf. Digital Signal Processing. 2014: 763-7.

[28] Tu M, Xie X, Na X. Computational auditory scene analysis based voice activity detection. Proc. of Int. Conf. Pattern Recognition; 2014. pp. 797-802.

[29] Quddus A, Gabbouj M. Wavelet-based corner detection technique using optimal scale. Pattern Recognition Letter. 2002; 23: 215-20. http://dx.doi.org/10.1016/S0167-8655(01)00090-3

[30] Adjouadi M, Candocia F, Riley J. Exploiting Walsh-based attributes to stereo vision. IEEE Trans. Signal Processing. 1996; 44(2): 409-20.

[31] Braun S. The effects of clipping in some signal extraction processors. 1973; 18(2): 215-20.

[32] Arfib D, Keiler F, Zoler U. DAFX–digital audio effects; 2002.

[33] Benesty J, Jensen JR, Christensen MG, *et al*. Speech enhancement: A signal subspace perspective. Academic Press; 2014.

[34] Pincus SM. Approximate entropy (ApEn) as a complexity measure. Chaos. 1995; 5: 100-17.

[35] Sparacino G, Bardi F, Cobelli C. Approximate entropy studies of hormone pulsatility from plasma concentration time series: influence of the kinetics assessed by simulation. Ann Biomed Eng. 2000; 28(6): 665–76. http://dx.doi.org/10.1114/1.1306344

[36] Lake DE, Richman JS, Griffin MP, *et al*. Sample entropy analysis of neonatal heart rate variability. Am J Physiol. 2002; 283(3): 787–97.

[37] Baker JE. Reducing bias and inefficiency in the selection algorithm. Proc. Int. Conf. on Genetic Algorithms and their Application; 1987.

[38] Tang KS, Man KF, Kwong S, *et al*. Genetic algorithms and their applications. IEEE Signal Processing Magazine. 1996; 13(6): 22–7. http://dx.doi.org/10.1109/79.543973

[39] Chipperfield A, Fleming P, Pohlheim H, *et al*. Genetic algorithm toolbox. Dept. of Automatic Control and Systems Eng, University of Sheffield; 1995.

[40] Gotshall S, Rylander B. Optimal population size and the genetic algorithm. WSEAS Conf.; 2002.

[41] Verga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX-92:: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication. 1993; 12(3): 247-51. http://dx.doi.org/10.1016/0167-6393(93)90095-3

[42] Ellis D. Sound examples: Noise. Available from: http://www.ee.columbia.edu/dpwe/sounds/noise

[43] Young S. Large vocabulary continuous speech recognition: a review. IEEE Signal Processing Magazine. 1996; 13(5): 45-57. http://dx.doi.org/10.1109/79.536824

[44] Olson DL, Delen D. Advanced data mining techniques. Springer Verlag; 2008.

[45] ETSI GSM 06.94. Voice activity detectors (VAD) for adaptive multi-rate (AMR) speech traffic channels. European Telecommunications Standards Institute; 1999.

[46] Rabiner LR. A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE. 1989; 77: 257-86.

[47] Garofolo J, Lamel L, Fisher W, *et al*. The DARPA TIMIT acoustic-phonetic continuous speech corpus CD ROM. National Institute of Standards and Technology; 1990.

[48] Mporas I, Ganchev T, Fakotakis N. Speech segmentation using regression fusion of boundary predictions. Computer, Speech, and Language Process. 2010; 27: 273-88.