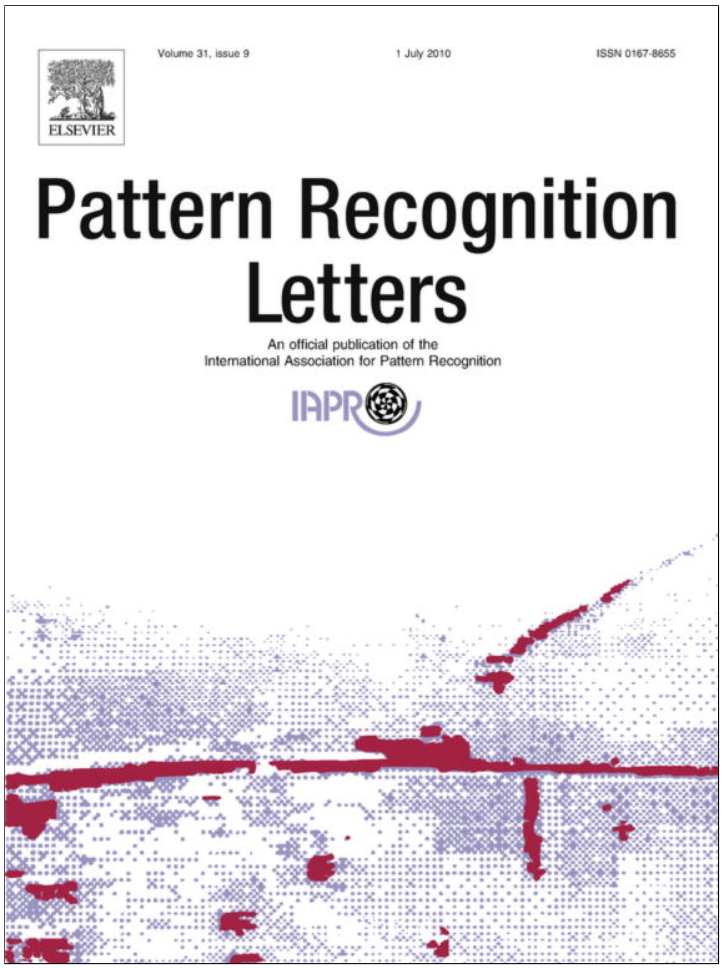

Pattern Recognition Letters 31 (2010) 936–948

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A two-stage speech activity detection system considering fractal aspects of prosody

Soheil Shafiee [a], Farshad Almasganj [a], Bahram Vazirnezhad [a,b,*], Ayyoob Jafari [a]

[a] Biomedical Engineering Department, Amirkabir University of Technology (Polytechnic of Tehran), Iran
[b] Iran Electronics Research Institute, Iran

ARTICLE INFO

Article history:
Received 11 August 2008
Received in revised form 6 November 2009
Available online 4 January 2010

Communicated by O. Siohan

Keywords:
Speech activity detection
Prosody
Fractal dimension

ABSTRACT

Speech Activity Detectors (SADs) are essential in the noisy environments to provide an acceptable performance in the speech applications, such as speech recognition tasks. In this paper, a two-stage speech activity detection system is presented which at first takes advantage of a voice activity detector to discard pause segments out of the audio signals; this is done even in presence of stationary background noises. In the second stage, the remained segments are classified into speech or non-speech. To find the best feature set in speech/non-speech classification, a large set of robust features are introduced; the optimal subset of these features are chosen by applying a Genetic Algorithm (GA) to the initial feature set. It has been discovered that fractal dimensions of numeric series of prosodic features are the most speech/non-speech differentiating features. Models of the system are trained over a Farsi database, FARSDAT, however, test experiments on the TIMIT English database have been also conducted. Employing the SAD system in conjunction with an ASR system, has been resulted in a relative Word Error Rate (WER) reduction of as high as 28.3%.

1. Introduction

Automatic speech activity detection has many applications especially in the noisy environments. For example in a speech recognition task, the efficiency in automatic detection of speech boundaries considerably influences the accuracy of the recognition engine. Even a minor improvement in speech boundary detection improves the overall system performance in long run. The most common methods for speech endpoint detection are based on short-time spectral energy (Savoji, 1989; Mauuary, 1994; Martin et al., 2000; Ramirez et al., 2005). In (Ramirez et al., 2004), the information of long-term spectral energy in sub-bands is the main idea of speech detection. These approaches work well over clean signal, but the performance is not acceptable in presence of various kinds of noise. Typically an adaptive threshold based on the features of the energy profile is employed to differentiate speech segments from silence/background voices (Martin and Mauuary, 2006; Prasad et al., 2002). In (Martin and Mauuary, 2006), a dynamic approach is employed to detect speech and non-speech segments. A finite state machine (FSM) with five states decides whether an input frame is speech or not, in a manner that the state of a frame depends on the state of its previous frame. Energy and some duration constraints are scrutinized in order to make decision for every input frame. In this approach, some non-speech noises specified by high energy in a short duration can be rejected, which is an advantage. Moreover, low length pauses between natural speech will not be classified as non-speech. In order to increase the algorithm robustness in presence of slowly variable background noise, in each step, the energy threshold is updated based on the statistics of the last non-speech frames. Lack of complex features differentiating speech from non-speech voice and noise is the main drawback of the approach introduced in (Martin and Mauuary, 2006). Using "Teager" energy is another useful technique in speech/non-speech separation. This operator, depends both on the signal energy and the basic frequency of the signal (Ying et al., 1993; Chen et al., 2005). Teager energy is robust in noisy environments, however, the main drawback is its weakness in dealing with signals mixed with certain types of non-speech voices.

Entropy as a measure of order and similarity is used in many SAD systems. This parameter that was firstly used by Shannon in information technology shows the degree of organization or the uncertainty level of the signal (Wu and Wang, 2005). Entropy can be used in both time (Waheed et al., 2002) and frequency domain (called spectral entropy) (Shen et al., 1998; Basu, 2003). Entropy is a powerful feature in speech and noise discrimination; however, the main drawback is the inability to discard certain types of non-speech voices produced by speakers such as cough and breath.

MFCCs (Mel Frequency Cepstral Coefficients) and their derivatives are often used in speech processing applications. The main

* Corresponding author. Address: Biomedical Engineering Department, Amirkabir University of Technology (Polytechnic of Tehran), Iran. Fax: +98 21 6649 5655.
E-mail addresses: sshafiee@aut.ac.ir (S. Shafiee), almas@aut.ac.ir (F. Almasganj), bvazirnezhad@aut.ac.ir (B. Vazirnezhad), ajafari20@aut.ac.ir (A. Jafari).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.12.014

idea of calculating these coefficients comes from hearing mechanism in human. In (Martin and Mauuary, 2006), after Linear Discriminative Analysis (LDA) of MFCC parameters, they are used in a FSM to determine weather a frame is speech or not. The problem of using just MFCCs is that they are not completely capable to differentiate between speech and non-speech signals generated by speaker, such as cough and breath, as experiments in this paper confirm. In (Padrell et al., 2005) MFCCs are used besides other features like short-time energy, sub-band energy and zero crossing rate (ZCR) as discriminative features and finally a Support Vector Machine (SVM) is employed to classify speech and non-speech signals. Cepstral variability is another approach used in (Skorik and Berthommier, 2000). These complementary features significantly improve the classification task, as these kind of features have important information on entropy and the order of signal.

The spectral distributions of speech and non-speech signals are not alike; wavelet transform would be a suitable technique to show the difference, i.e. the sub-band energies could be compared in wavelet domain. This approach is mainly used in Voice Activity Detection (VAD) applications (Shaojun et al., 2004; Chen et al., 2005). Their efficiency in SAD systems is not comprehensively studied.

Some researches have been employed different forms of garbage models to detect and reject non-speech parts (Villarrubia and Acero, 1993). In these approaches, to have good estimations of parameters in non-speech models, a large amount of training data is required.

In this work, we have considered features with high discrimination power and robustness against various background non-speech/noise. Fractal dimension of short time prosodic features have been proposed in this paper. Prosodic patterns differ extensively in different languages; however, all languages, obey some inherently fixed rules and specifications because of anatomical similarities of vocal tracts of humans (Hori and Furui, 2003; Koumpis and Renals, 2001). Autocorrelation is a well-known technique to compute the pitch frequency. Pitch frequency is mainly limited in a special range and does not have abrupt changes along signal. In (Tian et al., 2002), the smoothness of the variations of the autocorrelation function is used as the main feature for speech/non-speech discrimination. In this work, two criteria are used to estimate the smoothness of the autocorrelation function, which are called Mean Crossing Ratio and Energy Distribution Coefficient.

Different classifiers have been used to classify audio signals into speech/non-speech. Neural networks as a supervised technique have been widely used in classification problems. In (Hussain et al., 2000) two kinds of neural networks have been used for this purpose. An ADAptive LINEar network (ADALINE) with Widrow–Hoff training rule and a two layer Perceptron network. The disadvantage for neural networks is costly training procedure. Cepstral Linear Predictive Coding (LPC) is the main feature to train the neural networks. In (Beaufays et al., 2003), the features are extracted from signal frames and used to train a three layer feed forward network with 400 neurons in hidden layer to detect speech signals. To evaluate this speech detector, a speech recognition system is used with and without this stage and WER is calculated for each of the conditions; by using the speech detector, error rate shows 12% reduction.

Decision tree is another supervised classification technique used in SAD systems. In (Padrell et al., 2005), a 31-dimensional feature vector is used as discriminative feature vector and a decision tree is used as speech/non-speech classifier.

SVM is a powerful method in classifying an input space into two classes and the convergence speed in the training phase is faster than other classification techniques. However, in comparison with neural networks, SVM is not efficient in multi-class classification tasks. SVM has been extensively used in speech/non-speech classification problem (Abdulla et al., 2003; Enqing et al., 2002). In (Xianbo and Guangshu, 2005), an adaptive SVM algorithm has been used, in which, the support vectors of SVM are updated in each step. The feature vector in this work consists of MFCCs, short-time energy, sub-band energy and ZCR.

In this paper, which is an extension report of our previous work presented in (Shafiee et al., 2008), a two-stage approach is proposed for the SAD system. First, a segmentation block detects high-energy segments in signal, which can potentially be speech. These segments will be subject to further analysis in the following speech/non-speech classifier. It should be clarified here that in large part of the literature the term VAD is used in place of SAD; however, in this paper SAD is supposed to mean speech detector while VAD is simply a classification system to detect any sound, including speech and non-speech segments as cough, laugh, breath, etc. The conventional MFCC-based features are the main features to segment auditory signals to speech/non-speech parts. Moreover, prosodic features have been employed directly and indirectly, as duration, energy and fractal calculation of the autocorrelation numeric series to increase the performance of the classifier. Since, there may be a large amount of redundancy in some of these features; GA is applied to find an optimal subset of the feature vector to be employed in the classification task. Finally, a SVM classifier is exploited to discriminate speech from non-speech segments.

The remaining parts of the paper are organized as follows; Section 2 describes the framework of the proposed system. In Section 3, databases and evaluation methods are introduced. Section 4 details the first stage of the system. In Section 5, features to be used in speech/non-speech classification have been discussed. Feature selection using GA is also presented in this section. In Section 6, the SVM classifier will be discussed. Experimental results are presented in Section 7. Sections 8 and 9 are discussion and conclusion, respectively.

2. The overall framework of the system

The proposed SAD system consists of two basic stages, as shown in Fig. 1. The first stage is a "segmentation block" which keeps potentially speech parts and rejects other parts which are certainly non-speech, considering energy and duration aspects simultaneously. Segmentation block utilizes a FSM which will be explained in details later. The main idea of the FSM is from (Martin and Mauuary, 2006), which relies on the fact that speech parts of an audio signal exist among segments with certain energy and duration. The second stage is a "speech/non-speech discriminator" which tags the remaining segments into speech and non-speech. For this purpose, a large set of features are computed, and finally, a SVM classifier is used to decide whether the segment is speech or not. The features are mainly divided into two main groups which are MFCC based features and prosodic and autocorrelation-based features. The features will be discussed in detail in Section 5.

The two-stage SAD system has an excellent performance in the presence of both stationary and non-stationary noises. In the first stage, the segmentation block is applied to the audio signal and consequently in the next stage, speech/non-speech classification is made for each segment. In this manner, the system is able to reject transient noise segments which comprise high level energies.

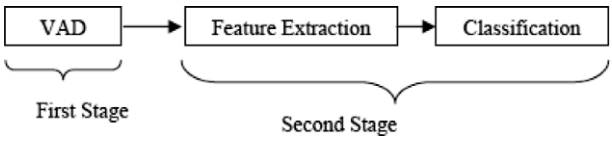

Fig. 1. The block diagram of the proposed SAD system.

The SAD system is not assigned to eliminate background noises overlapped with speech segments. In such a situation, the solution is the application of a post-processing unit just after the SAD system to enhance the input speech segments prior to applying to an ASR engine.

To detail the approach, after introducing the databases in Section 3, segmentation block and the second stage of the system will be described in detail in Sections 4 and 5, respectively.

3. Materials and methods, corpus and empirical workbench

3.1. Speech databases

We have used different databases to train and test the proposed SAD system. The first used speech database is FARSDAT (Bijankhan and Sheikhzadegan, 1994). FARSDAT consists of 6000 sentences uttered by 304 Farsi speakers. Each speaker has uttered 20 sentences in the acoustic booth of the Linguistics Laboratory of the University of Tehran. The utterances are recorded with a sampling rate of 22050 Hz and are phonetically transcribed. A SONY cardioids dynamic microphone with a frequency response within 80–16 kHz range has been used. The distance between microphone and lip position of the speaker was about 12 cm. The speech was collected using sound blaster hardware cards, installed in four 80,486 IBM microcomputers and sampled at 22,050 Hz samples per second at 16 bit resolution. Signal to noise ratio of the FARSDAT is about 31 db.

In this work, these speech signals are fed into the segmentation block, introduced in Section 4, and 8579 segments detected as speech are collected. These speech segments are used to train and evaluate different parts of the proposed SAD. to test the performance of the system over a language other than Farsi, "TIMIT" database, which is a well-known English speech corpus has been used In order to evaluate the language independency of the proposed system. Moreover, the performance of an ASR system has been examined with and without using SAD system at the front-end speech recognition task. The ASR system had been trained by the train set of TIMIT and is evaluated by its test set from 24 speakers distorted by non-speech segments manually added in the pause intervals between speech segments.

3.2. Non-speech databases

FARSDAT does not contain sufficient non-speech segments. Therefore, non-speech signals were collected or recorded in the following three ways. Totally, 10,892 non-speech segments have been used in this study.

3.2.1. Non-speech parts of the T-FARSDAT database

T-FARSDAT is a Farsi corpus consisting of phonetically hand-labeled telephony conversations (Bijankhan et al., 2003). This database consists of conversation files from 64 Farsi speakers recorded in a sampling rate of 11,025 Hz. In this database, non-speech parts of conversations, such as breath and laugh, are labeled. These parts have been used as a source for non-speech segment collection.

3.2.2. Non-speech signals from internet available sources

Some internet sources, such as FREESOUND, include non-speech sounds recorded in different formats and sampling rates. Sampling rate of all of these samples has been changed into 22,050 Hz to be unified for this work.

3.2.3. Recorded non-speech signals

In order to increase the variety of non-speech samples, some frequent non-speech samples have been recorded especially for this work. These samples have been uttered by 10 speakers and include cough, laugh, sneeze, crying, hiccup, scream, breath, etc.

3.3. Noise database

In order to evaluate the overall system in the presence of background noises, four different additive background noises, including white, pink, babble and factory from NOISEX-92 have been examined. The sampling frequency of these signals has been changed from 19,980 Hz to 22,050 Hz in order to be consistent with the other speech and non-speech samples.

3.4. Evaluation parameters

The proposed system, as shown in Fig. 1, consists of two main stages. In this paper, these two stages have been evaluated separately, and the overall performance is also evaluated by conducting a speech recognition experiment with/without SAD system.

The first stage of the SAD is a segmentation block to detect audio segments. This block of the system is evaluated using a signal consisting of 20 utterances. The audio segments have been manually labeled. Two parameters, HR0 and HR1 have been used to evaluate the segmentation block.

$$\text{tHR0} = (\text{The total duration of correctly detected silence})/(\text{Total duration of silence in the signal}) \quad (1)$$

$$\text{tHR1} = (\text{The total duration of correctly detected audio segments})/(\text{Total duration of audio segments in the signal}) \quad (2)$$

The second stage of the SAD system classifies audio segments into speech and non-speech. In order to evaluate this stage, two criteria have been proposed: "non-speech segment rejection ratio" or "hit rate for non-speech" and "speech segment false rejection ratio" or "hit rate for non-speech" (Górriz et al., 2005). These criteria are called HR0 (Hit rate for non-speech) and HR1 (Hit rate for speech), respectively. These parameters are defined as follows:

$$\text{HR0} = (\text{The number of non-speech segments detected as non-speech segments})/(\text{Total number of non-speech segments}) \quad (3)$$

$$\text{HR1} = (\text{The number of speech segments detected as speech segments})/(\text{Total number of speech segments}) \quad (4)$$

In addition, to evaluate the ASR system performance with/without the SAD system, Word Recognition Rate (WRR) has been measured for the ASR system.

$$WRR = \frac{TW - SW - DW - IW}{TW} \quad (5)$$

where *TW* is the total number of the words,*SW* sits for the number of substituted words, *DW* is the number of deleted words and IW is the number of false insertions.

4. Segmentation module, the first stage of the proposed SAD

The proposed SAD system, at the first stage, detects those parts of the signal which are likely to be speech. These are often high-energy segments of the signal which have a reasonable duration to be speech. Errors in this section may cause a false rejection of speech segments. These errors degrade the overall performance of the SAD system. FSM is a dynamic technique used to extract likely to be speech segments from the input audio signal similar to the approach in (Martin and Mauuary, 2006). The FSM proposed in this work, applies duration constraints besides energy constraints, to

reject non-speech segments, especially those ones with high energy levels and small durations. The inputs of the FSM are frames of the audio signal. Energy level across the signal is compared with an adaptive threshold. The energy threshold is updated periodically based on the statistics driven from the last non-speech frames. The adaptive threshold increases the robustness of the system to semi-stationary background noises with variable amplitude. Fig. 2 shows the proposed FSM.

The FSM shown in Fig. 2 has five states: (1) low-energy, (2) high-energy presumption, (3) high-energy, (4) brief energy increasing and (5) high-energy continuation. The transition regime checks energy and duration conditions to determine the present state. Constraints and actions are shown in Fig. 2 with letters C and A, respectively. In each transition, some actions take place. Constraints and actions are defined in Table 1.

The energy threshold update regime is based on the statistics of the few last non-speech frames similar to (Prasad et al., 2002). For every frame of the signal, the updated threshold value is calculated by the following equation.

$$T_{new} = (1-p)T_{old} + p\overline{T_{silence}} \quad (6)$$

where T_{old} is the previous energy threshold, $\overline{T_{silence}}$ is the mean energy of the last eight non-speech frames and $0 < p < 1$ is determined based on the variation rate of energy for non-speech frames. The frames here contain 512 samples with a sampling frequency of 22,050 Hz. The energies of the last eight non-speech frames are always stored in $T_{silence}$ array. The variation rate of background noise is calculated as $\delta_{new}/\delta_{old}$ ratio, where δ_{new} and δ_{old} are noise energies calculated from updated and previous energy arrays, respectively. Table 2 shows the P value determination scheme.

In the proposed FSM, frame sequences with more than 70 ms with certain energy level, are labeled as likely to be speech and low energy segments with a duration of more than 116 ms will be rejected. The algorithm checks the duration, when a low to high energy transition is detected; in this case if the energy of the input frames remains higher than the threshold, at least for 70 ms, it would be labeled as a likely to be speech segment. In contrary, when the energy level remains at least 116 ms below the threshold, the frames will be rejected. The energy threshold is adaptively updated in each step. In this scheme, a small duration high energy noise or a brief silence in the middle of a speech segment would not switch the state. There are five different states for the frames of the input signal. The five states of FSM are then used to make a final decision on whether the frame sequence is likely to be speech or non-speech. The original five states and the final labels FSM are shown in Fig. 3, for a sample audio signal. In this figure, five states of the FSM, from 1 to 5, are shown on the vertical axis with rational numbers 0.2, 0.4, 0.6, 0.8 and 1, respectively. The final

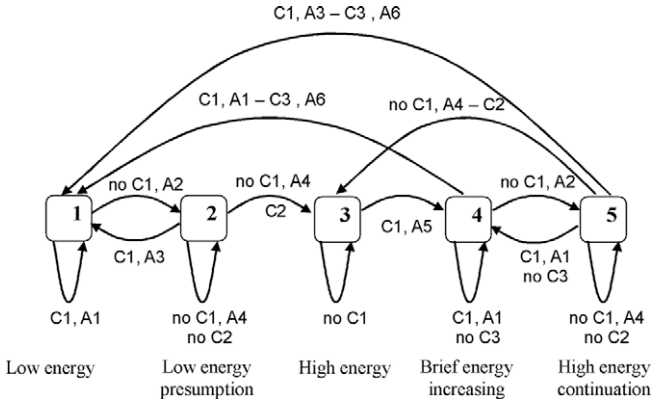

Fig. 2. Schematic of the segmentation block.

Table 1
Actions and constraints in the state transition regime FSM.

Actions	Constraints
A1: LD = LD + 1	C1: Frame Energy < Energy Threshold
A2: HD = 1	
A3: LD = LD + HD	C2: High Energy Duration (HD) > 70 ms
A4: HD = HD + 1	
A5: LD = 1	C3: Low Energy Duration (LD) > 116 ms
A6: LD = HD = 0	

Table 2
"p" value determination, based on the $\delta_{new}/\delta_{old}$ ratio.

$\delta_{new}/\delta_{old}$	P value
$\delta_{new}/\delta_{old} \geqslant 1.25$	0.25
$1.25 \geqslant \delta_{new}/\delta_{old} \geqslant 1.1$	0.2
$1.1 \geqslant \delta_{new}/\delta_{old} \geqslant 1$	0.15
$1 \geqslant \delta_{new}/\delta_{old}$	0.10

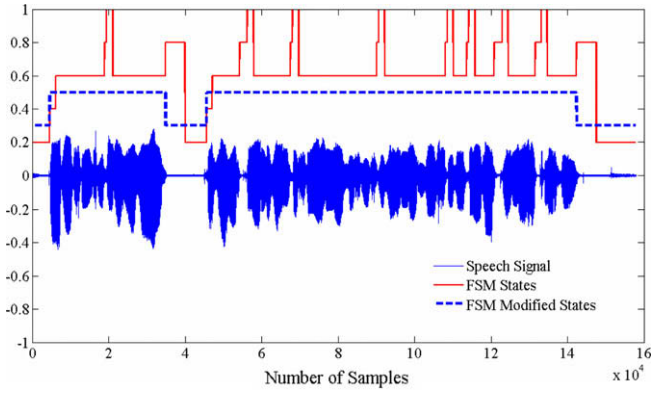

Fig. 3. The original 5 states of the FSM (red solid line), and the corresponding final labels (blue dashed line).

labels are shown with the dashed line with the higher level for speech segments.

In the next stage, after applying FSM, the remaining segments are subject to further processing. Sections 5 and 6 explain features and the proposed SVM used for speech/non-speech classification task in the next stage of SAD system.

5. Feature extraction and selection

As mentioned earlier, the second stage of the proposed system, classifies the remaining segments, which are likely to be speech segments, into speech/non-speech. For this purpose, a segment is divided into a number of overlapping frames with the same lengths. Here, a frame length of 256 samples (for sampling frequency of 22,050 it leads to 11.6 ms frames), with overlap of 128 samples (5.8 ms) between adjacent frames is considered. Next to the framing process, a number of MFCC-based features are calculated distinctively for each of the frames. The average value for features is calculated across frames to have an averaged unit MFCC-based feature vector. Some extra features extracted from the same segment will be added to the feature vector, and finally, a SVM binary classifier will decide on whether or not the segment is speech. In the following subsections, feature extraction and selection methods will be explained.

5.1. Features

One of the main goals of this work is to find the most discriminating features for speech/non-speech classification of audio seg-

ments. The most differentiating features out of a crude set of features have been discovered using GA. The features investigated here can be mainly divided into two groups: (1) MFCC-based features and (2) prosodic features. In the following subsections, features will be discussed in more detail.

5.1.1. MFCC-based features

Mel Frequency Cepstral Coefficients or MFCCs and their derivatives are known to be the most common features in speech processing applications (Padrell et al., 2005; Skorik, and Berthommier, 2000). To compute these features from signal, a filter bank is applied to the signal spectrum. The center frequencies of these filters are distributed in Mel scale. MFCCs are cosine transform of the logarithm of filter banks' outputs. In this work, the first 13 MFCCs and their derivatives are exploited. As an audio segment consists of a number of frames, the mean values of MFCCs can be calculated over frames of a segment. As a result, a feature vector of 13 components called mean-MFCC can be calculated. at the second step, first derivatives of MFCCs, for the consecutive frames in a segment, are calculated by the following equation.

$$\Delta MFCC_i[n] = C_i[n+1] - C_i[n], \qquad 1 \leqslant n \leqslant N-1, \qquad 1 \leqslant i \leqslant 13 \tag{7}$$

where $C_i(n)$ is the ith MFCC from the nth frame in a segment, and N is the total number of frames in a segment. Moreover, a new feature vector consisting of maximum values of $\Delta MFCC[n]$ is obtained for a segment as follows.

$$D_i = \max_n(\Delta MFCC_i[n]), 1 \leqslant n \leqslant N-1, 1 \leqslant i \leqslant 13 \tag{8}$$

where D_i is the maximum value of the ith component in $\Delta MFCC$ vector across the segment. This vector is called max-diff-MFCC. The discrimination ability of various components of mean-MFCC vectors is shown in Fig. 4, which shows the histograms over a large number of segments.

As shown in Fig. 4, coefficients with indices 1, 3, 10 and 12 seem to be better in discrimination speech and non-speech. Fig. 5 shows the histograms of max-diff-MFCC coefficients over speech/non-speech segments, distinctively. The coefficients with indices 1, 2, 12 and 13 seem to have the maximum discriminative power.

5.1.2. Autocorrelation-based features

Autocorrelation function has a tight relation to the pitch trajectory in speech signals, and one of the most well-known techniques in pitch frequency calculation of speech signals. The pitch frequency of a speech signal could vary in a limited range. Autocorrelation function can be used to extract features which are useful in speech/non-speech discrimination. Autocorrelation function for the frame l, with N samples of sequence s, is given by the following equation.

$$\Phi_l(m) = \sum_{n=1}^{N-m} s_l(n)s_l(n+m) \tag{9}$$

The maximum output of the equation always occurs at $m = 0$. Φ_l can be normalized by dividing $\Phi_l(m)$ by $\Phi_l(0)$. This value is called "Normalized AutoCorrelation Function" (NACF) similar to (Tian et al., 2002). NACFs are calculated for all of the frames taken from the segment. Index and magnitude of the first peak of NACF function (the peak at $m = 0$ is ignored) are called PI and PM, respectively. The frames of a segment have a series of PI and PM values called PIS and PMS, respectively. The length of these series depends on the length of a segment, or the number of frames in the segment. An example of PIS and PMS extracted from a speech and a non-speech segment is shown in Fig. 6.

As shown in Fig. 6, PIS and PMS of a speech segment have less variation than a non-speech one. So, smoothness of PIS and PMS sequences extracted from audio segments is an important measure to distinguish between speech and non-speech segments. Here, three criteria are introduced to measure the smoothness of PIS and PMS, as brought in the following subsections.

5.1.2.1. Mean crossing ratio of PIS and PMS. Mean Crossing Ratio (MCR) of PIS and PMS sequences are employed to discriminate speech/non-speech segments. These features are defined as the number of times that PIS and PMS series cross their mean values. This can be formalized as follows, in which $x(l)$ is PIS or PMS.

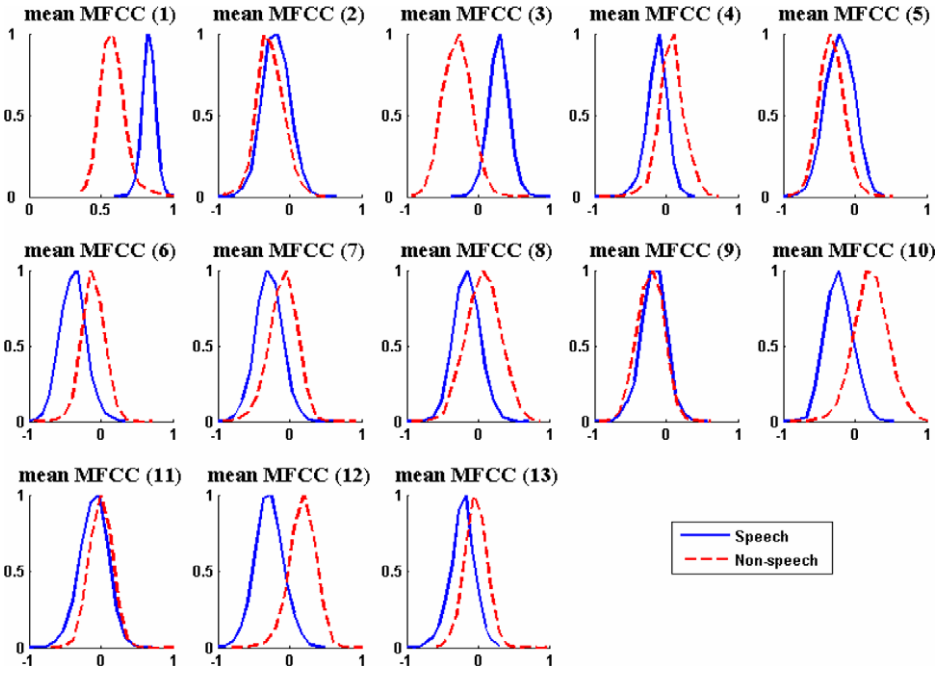

Fig. 4. Distributions of various components of mean-MFCC vector for speech and non-speech segments.

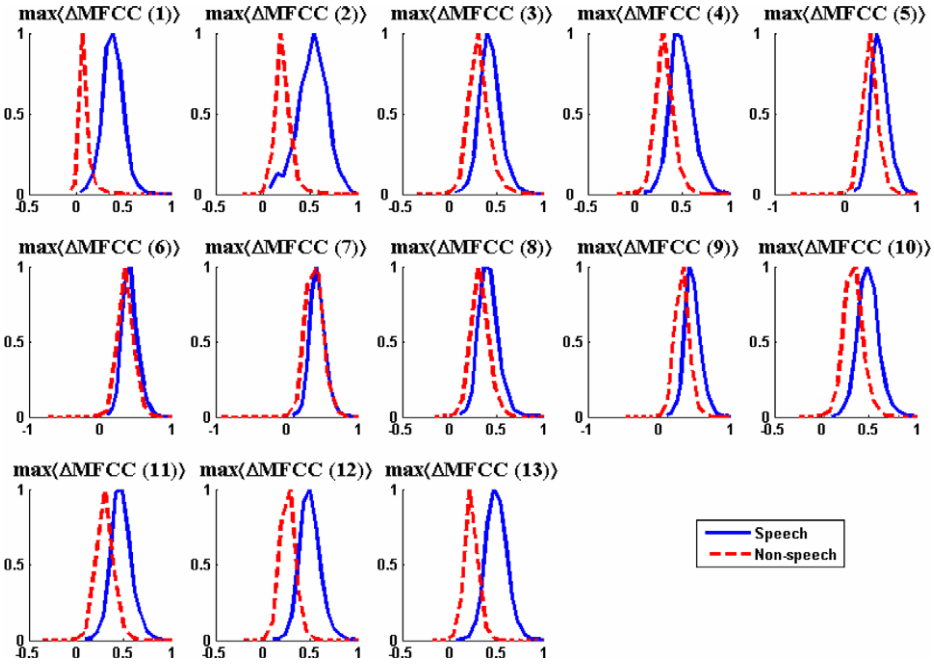

Fig. 5. Distributions of "max-diff-MFCC" coefficients for speech and non-speech segments.

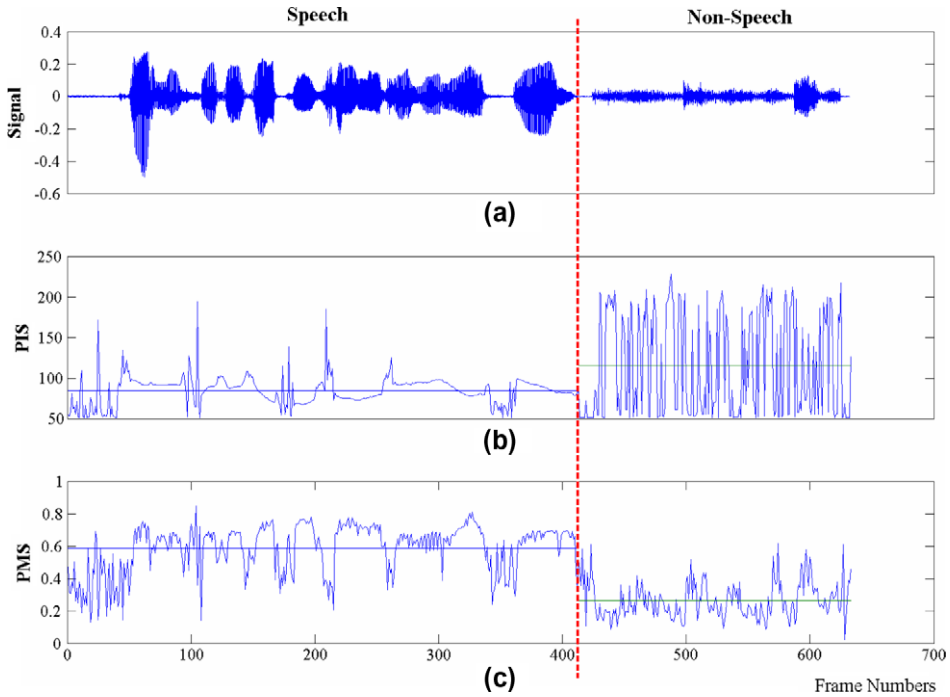

Fig. 6. (a) An audio signal, (b) Extracted PIS and (c) Extracted PMS

$$\mathrm{MCR} = \frac{1}{L}\sum_{l=2}^{L}\{|sgn[x(l) - M] - sgn[x(l-1) - M]|\} \tag{10}$$

where x is PIS or PMS, l is the index component in x, $M = \left(\frac{1}{L}\right)\sum_{l=1}^{L} x(l)$ and L is the length of PIS or PMS (i.e. the total number of frames in the audio segment). These features are called PIS–MCR and PMS–MCR.

Histograms of these two features are shown in Fig. 7(a) and (b). It could be seen in figures that they are just able to partially separate speech and non-speech segments and some more robust features are then needed. Although the MCR distribution for speech and non-speech are almost similar in Fig. 7, some kinds of non-speech, like breath, have a very high MCR which can make them distinguishable from speech. Fig. 6 shows such a case for a typical breath and a speech signal.

5.1.2.2. Energy distribution of PIS and PMS. Energy Distribution Coefficient (EDC) of PIS and PMS sequences are two other features applied in this work. EDC is originally inspired from the idea that non-smooth signals have greater frequency components than smooth ones. This means that for non-smooth signals, the frequency components distribution is mainly located in higher fre-

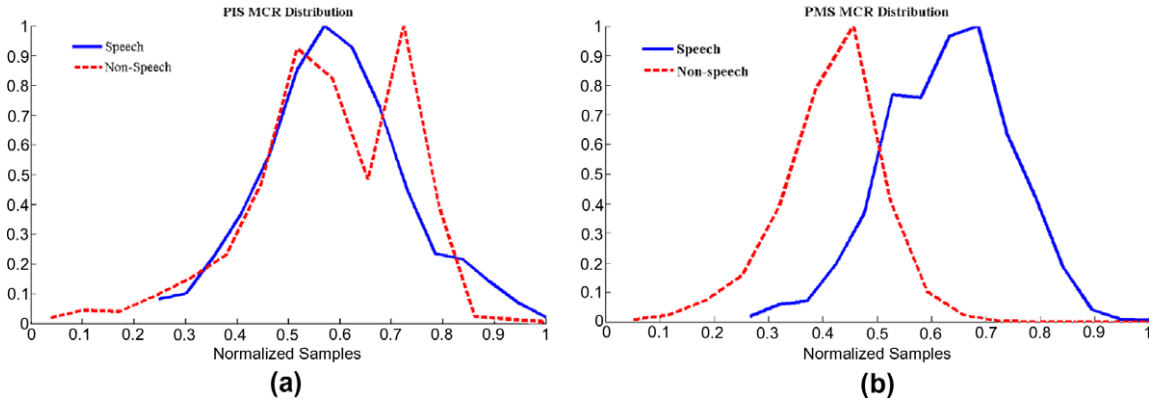

Fig. 7. (a) PIS-MCR distributions and (b) PMS-MCR distributions.

quencies. Correspondingly, the energy distributions of PIS and PMS extracted from speech segments are mostly expected to be located in lower frequencies, and vice versa for non-speech segments. EDC parameter for an audio segment is calculated as follows.

$$\mathrm{EDC} = \frac{\left(\sum_{l=1}^{L} l\omega_l\right)}{\left(L\sum_{l=1}^{L}\omega_l\right)} \tag{11}$$

where ω_l is the spectral amplitude of the lth sample of the DFT of PIS or PMS sequences with L samples. By applying Eq. (11) to PIS and PMS sequences, two parameters called PIS–EDC and PMS–EDC would be obtained, which are used as two other features in this work. These two features have been calculated for the speech/non-speech segments. The histograms are shown in Fig. 8(a) and (b).

5.1.2.3. Fractal dimension of PIS and PMS. Recently, many researchers have treated speech as a chaotic, complex and unpredictable phenomenon, and many nonlinear processes have been successfully exploited in speech processing (Kokkinos and Maragos, 2005; Banbrook and McLaughlin, 1994). Fractal dimension is one of the conventional methods used to describe the geometry and dimension of state space attractors of chaotic phenomena. Of course, in this paper, we have no intention to discuss chaotic nature of speech, but it is just intended to use fractal dimension of prosodic numeric series as a measure of irregularity. It is believed that irregular signals have greater fractal dimension than the regular ones. As mentioned previously, PIS and PMS extracted from non-speech segments are usually more irregular than the ones extracted from speech segments. Petrosian which is a well-known method to calculate fractal dimension is used in this work (Esteller et al., 2001; Hilborn, 2001). In this method, fractal dimension is calculated as follows.

$$D = \frac{\log_{10} N}{\log_{10} N + \log_{10}\left(\frac{N}{N+0.4N_\Delta}\right)} \tag{12}$$

where N is the length of the sequence and N_Δ is the number of times that the direction of variations between consecutive samples has been changed in the sequence. The method for calculation of N_Δ constructs a binary sequence by subtracting consecutive samples. From the sequence of subtractions, a binary sequence of + 1 and −1 is created by assigning on whether the result of the subtraction is positive or negative, respectively. Sequence, here, can be PIS or PMS described previously in Section 5.1.2. Applying this function to PIS and PMS, two features can be extracted for audio segments. We have computed these two features for speech/non-speech segments. The histograms are shown in Fig. 9(a) and (b). From the figure, the proposed features seem to be powerful features for separating speech from non-speech.

Fractal features are basically defined on the self-similarity specification of a signal. For a speech segment, especially for a vowel, this self-similarity must be more realized than for a non-speech one. The low-level overlapping histograms in Fig. 9 and the feature selection study discussed in Section 5.2, verify the high capability of these two new introduced features in discriminating speech from non-speech.

5.1.3. Duration

Duration of a segment is another feature discussed in this work. Some non-speech events such as breath, cough, sneeze, etc. gener-

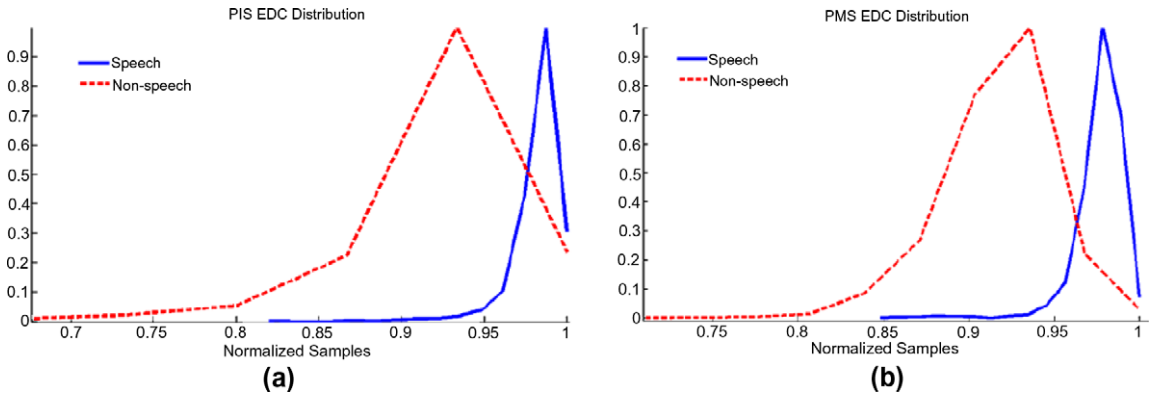

Fig. 8. (a) PIS-EDC distributions and (b) PMS-EDC distributions.

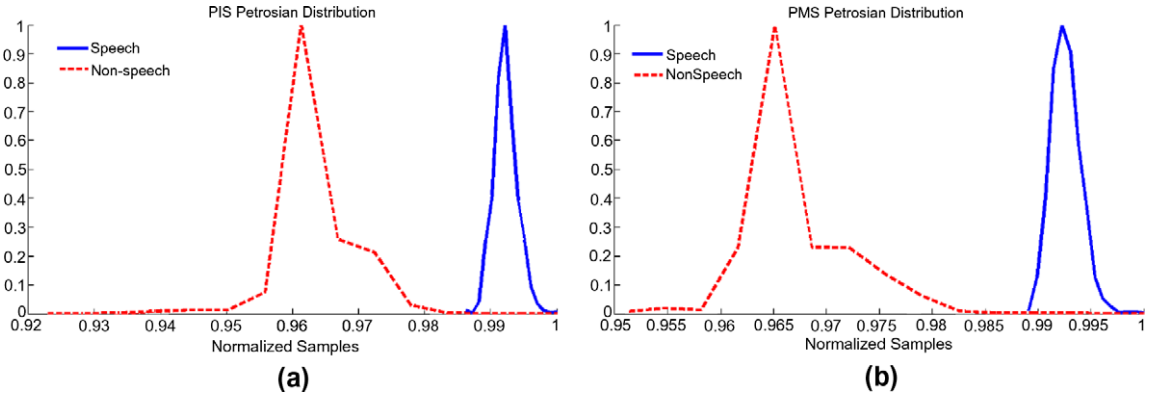

Fig. 9. (a) PIS-Petrosian distributions and (b) PMS-Petrosian distributions.

ally have short durations. For such non-speech segments, duration is a constructive feature within the set of features for discrimination. Histograms of this feature for speech and non-speech segments could be compared in Fig. 10. There is a considerable overlap between these two distributions which shows the limitation of this feature in speech-non-speech classification.

5.1.4. Energy gradient

Energy variation in a segment is another prosodic feature which has been used in this study. To calculate this feature, next to breaking the audio segment to sub-frames, frames energy is calculated distinctively and energy gradient is given by

$$\Delta E[n] = abs(E[n+1] - E[n]), 1 \leqslant n \leqslant N-1 \tag{13}$$

where N is the total number of frames in the segment. Mean of $\Delta E[n]$ is computed to have a unique value of this parameter for an audio segment. Normalized histograms of this feature for speech and non-speech segments, as shown in Fig. 11, have a wide-spread overlap region.

5.2. Feature selection

There are 34 features introduced in Section 5.1, which are considered to be fed to a classifier to classify audio segments. Many features may contain overlapped information and could be pruned with a minimal loss of information. It must be notified that the most common classifiers are mostly suffering from different limitations. An increase in the feature vector dimension, generally not only increases the complexity of the classification process, but also may sometimes decrease the performance of the classification task. To overcome this problem, a reduced version of the original feature

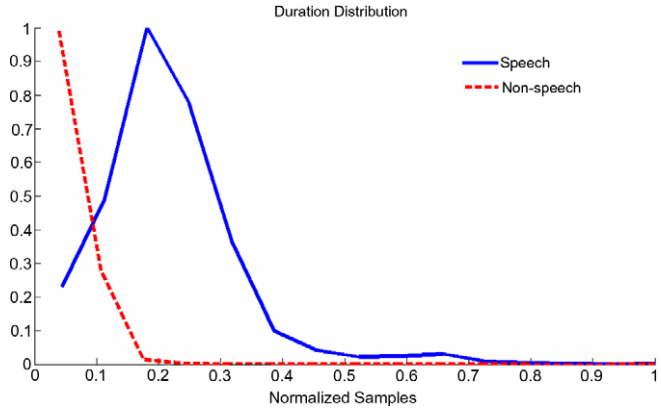

Fig. 10. Distribution of duration over the speech and non-speech segments.

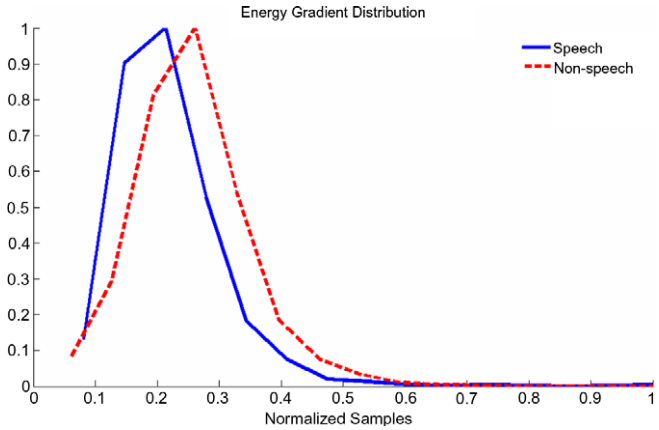

Fig. 11. Distribution of energy gradient over the speech and non-speech segments.

set, can be used. For instance, in the speech/non-speech classification of audio segments, PIS-Petrosian and PMS-Petrosian are the best discriminators (as will be discussed in Section 5.2.1), but they have considerable repeated information; so, it is sensible to employ only one of them in the classification process.

To follow this approach, we applied F-Ratio value (Cohen, 1986). By running GA, which uses F-Ratio as its fitness function, we get forward to find an optimally pruned version of the original crude feature set.

5.2.1. F-Ratio

Fisher discriminator analysis for a two-class problem, leads us to the Eq. (14) as the Fisher's measure or F-Ratio (Cohen, 1986):

$$F = \frac{B}{W} \tag{14}$$

where B matrix, the "between class scatter matrix", is given by

$$B = (\underline{\hat{\mu}}_1 - \underline{\hat{\mu}}_2)(\underline{\hat{\mu}}_1 - \underline{\hat{\mu}}_2)^T \tag{15}$$

where $\underline{\hat{\mu}}_i$ is the mean of the N_i samples of class w_i in the n dimensional space. The W matrix is the sum of scatter matrices for two classes and is given by

$$W = W_1 + W_2 \tag{16}$$

$$W_i = \sum_{\beta \in w_i} (\underline{\beta} - \underline{\hat{\mu}}_i)(\underline{\beta} - \underline{\hat{\mu}}_i)^T, i = 1, 2 \tag{17}$$

It is expected that a feature vector with a higher F-Ratio has more discrimination power, and is a good measure to compare extracted features in the involved task.

To have some experiments in this field, first, F-Ratio has been individually computed the introduced features over the two class dataset consisting of the speech/non-speech segments. Fig. 12 shows the F-Ratio values calculated for each of the 34 features in a left to right descending order.

These values are brought in Table 3, in a different manner.

As can be seen in the table, fractal dimension-based features have the highest F-Ratio value. By considering F-Ratio criterion, the best discriminator feature is known to be PMS-Petrosian.

5.2.2. Feature selection using GA

In the previous section, 34 distinct features, extracted from audio segments, were compared to each other using the F-Ratio factor. Employing all of these features in the classification process will not necessarily lead to the best results. On the other hand, some features with high F-Ratio values have repeated information and are redundant. For example, from Table 3 and Fig. 12, the best discriminator features are two fractal features, PMS-Petrosian and PIS-Petrosian. But, these features have almost similar information. This kind of problems can significantly reduce the performance of most classifiers. To avoid these problems, GA, which is an adaptive feature selection procedure, can be employed in order to choose the best feature set. There is not a problem with the high complexity and low speed of the GA, since, the feature selection procedure needs to be done only one time, and the found optimum feature set will remain fixed to be employed in classification process afterwards.

It has been shown that, the GA is an efficient approach for large scale optimal subset selection problems (Xu and Chan, 2002; Selouani and O'Shaughnessy, 2004). The search algorithm in GA does not follow the steepest path of the error gradient, to find the minimum value for the error function; in contrary it examines various points in the search space based on a random approach which considers the laws of evolution, to minimize the error function. Therefore, it is particularly appealing when such information is unavailable or costly to obtain. GA, unlike the gradient-based training algorithms, can handle the global search problem better in a vast, complex, multi-modal and non-differentiable surface. Moreover, GA is generally much less sensitive to initial conditions of the training data. GA always globally searches for an optimal solution, while a gradient descent algorithm can only look for a local optimum in a neighborhood of the initial condition.

GA is basically inspired by the laws of natural evolution and genetics. Its principal idea is to search for the optimal solution in a large population. It usually uses a fixed length continuous or discrete string, called chromosome, to represent a possible solution. Usually, a simple GA consists of three main operations called parent selection, crossover and mutation. The population comprises a group of chromosomes from which the candidates can be selected, as the possible solutions of a given problem. The initial population often consists of a group of individuals whose chromosomes are randomly selected. During the optimization task, the appropriateness of the individuals, in the current generation, is evaluated using a user-defined fitness function.

In this work, each individual's chromosome is made of a binary string which represents a set of active or inactive genes. Active genes' indexes represent those components in the feature vector which participate in the classification task.

To optimize the speech/non-speech classification task, the optimum number of features to be used as well as their combination should be determined. In this work, parent selection in GA works based on a stochastic uniform selection algorithm. The cross over function employs a so-called scattered function to generate children. This function creates a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent, and combines the genes to form the child. The mutation function, adds a random number taken from a Gaussian distribution with mean 0 to each entry of the parent vector. In generation replacement, 2 individuals are guaranteed to be survived in the next generation. The fitness function used in GA was F-Ratio discussed in Section 5.2.1. GA searches for the feature vector which represents the highest F-Ratio. To find the best feature vector, GA searches for the best 1, 2, 3, ..., 34-dimentional feature sets, distinctively. This is done by running GA for 34 times, each run is for a certain size of the feature set. As an example, for a known feature vector size, e.g. 15, GA has been applied and it resulted in a vector with 15 features which represented the highest F-Ratio value among other selections of 15 dimensional vectors. This procedure has been applied for other vector sizes and finally, the F-Ratio values of the optimum feature vectors were compared. These values are shown in Fig. 13.

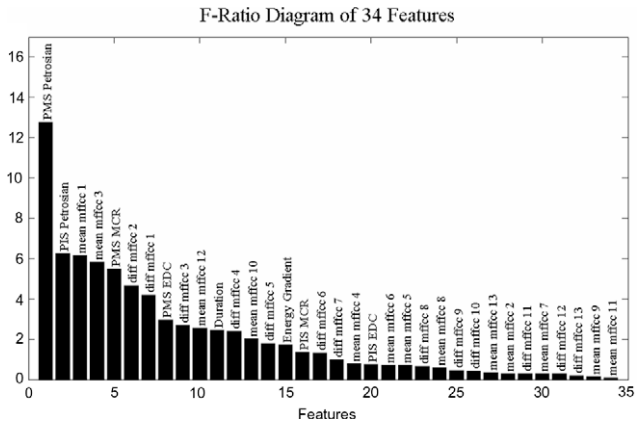

Fig. 12. F-Ratio values calculated individually for the 34 features extracted from the two-class audio segments.

Table 3
The exact F-Ratio values calculated individually for the 34 features extracted from the two-class audio segments.

Feature	F-Ratio
PMS-Petrosian	12.78
PIS-Petrosian	6.27
Mean-MFCC 1	6.16
Mean-MFCC 3	5.83
PMS-MCR	5.51
Diff-MFCC 2	4.67
Diff-MFCC 1	4.18
PMS-EDC	2.94
Diff-MFCC 3	2.68
Mean-MFCC 12	2.54
Diff-MFCC 4	2.39
Mean-MFCC 10	2.02
Diff-MFCC 5	1.77
PIS-MCR	1.35
Diff-MFCC 6	1.32
Diff-MFCC 7	0.99
Mean-MFCC 4	0.8
PIS-EDC	0.73
Mean-MFCC 6	0.71
Mean-MFCC 5	0.71
Diff-MFCC 8	0.65
Mean-MFCC 8	0.58
Diff-MFCC 9	0.45
Diff-MFCC 10	0.43
Mean-MFCC 13	0.33
Mean-MFCC 2	0.28
Diff-MFCC 11	0.28
Mean-MFCC 7	0.28
Diff-MFCC 12	0.27
Diff-MFCC 13	0.18
Mean-MFCC 9	0.14
Mean-MFCC 11	0.08

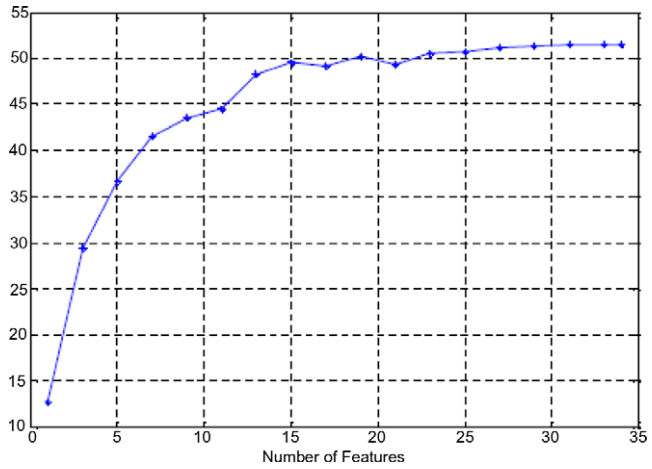

Fig. 13. The best F-Ratio values obtained by GA for different numbers of feature set dimensions.

It can be seen that F-Ratio value has a considerable growth rate along the feature sizes of 1–15, and after that, its growth decreases significantly. As, it is needed to keep the feature vector size small as possible, this point has been chosen as the optimum point for the mentioned problem. Using 15 features (instead of 34) in the classification task reduces the computational complexity of the whole SAD system significantly and makes it faster and more time efficient. The best 15-dimensional feature set selected by GA contains PMS-Petrosian, PMS-MCR, PMS-EDC, PIS-EDC, indices 2–5 of max-diff-MFCC and indices 1, 2, 3, 5, 7, 8 and 9 of Mean-MFCC. These features have been extracted from audio segments and used as inputs for the speech/non-speech classifier introduced in the next section.

6. SVM, the speech/non-speech classifier

In order to discriminate speech segments from non-speech ones, SVM is employed which is known as a useful technique in the field of statistical learning theory (Burges, 1998; Vapinik, 1998). This technique has been widely used in different signal processing, pattern recognition and classification applications (Justino et al., 2005; Pal and Mather, 2004). Because of its good performance, especially in the case of two-class discrimination problems, this technique has been selected as the classification method in this work. Here is a brief description of the algorithm; assume that the training data with k number of samples is represented by $\{xi, yi\}$, $i = 1, \ldots, k$, where $x \in R^n$ is an n dimensional vector and $y \in \{-1, +1\}$ is the class label as shown in Fig. 14. The aim is to find a hyper-plane that separates the data with the minimum error. This depends on finding w and b such that

$$y_i(w \cdot x_i + b) + \xi_i \geqslant 1 \tag{18}$$

where ξ_i is the error value for ith sample of the training dataset.

The optimal separating hyper-plane is achieved by minimization of the following criterion.

$$\min_{w,b,\xi_i} \left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right] \tag{19}$$

The first term in this criterion is used to set the learning capacity, and the second term controls the number of misclassified points. The regularization constant ($C > 0$) also determines the trade-off between the empirical error and the complexity term. The parameter C is chosen intuitively by the user. A large value for C means a higher penalty of errors.

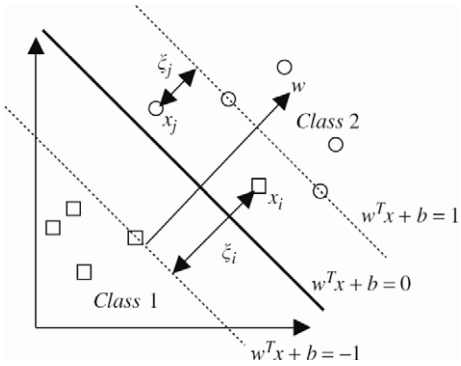

Fig. 14. SVM aim is to find a hyper-plane that separates the data classes with the minimum error.

When it is not possible to have a linear hyper-plane, which separates the training data, the technique can be extended to nonlinear decision surfaces. Such a procedure can be considered as the mapping of input data to a higher-dimensional feature space, using some non-linear functions. The transformation into a higher-dimensional space can be achieved by a kernel and makes outlier data points to be classified linearly. There are several kernel functions introduced in the literature such as polynomial, Radial Basis Function (RBF), Multi-layer Perceptron (MLP), spline and Fourier series. In this work, two kinds of kernels i.e. linear and polynomial are used in order to classify the speech and non-speech segments. Polynomial kernel of SVM is defined as follows.

$$K(x, x') = (Gamma\langle x, x'\rangle + Coeff)^d \tag{20}$$

where "*Gamma*" and "*Coeff*" are two constant values and d is the polynomial degree. In this equation, the two dimensional input vector (x, x') is mapped into a higher dimensional feature space, where two classes can be separated linearly.

7. Experimental results

7.1. Experiments on the segmentation block stage

Experiments have been conducted on the Farsi and English speech signals separately. Farsi test set consists of about 30% of FARSDAT speech corpus described in Section 3.1. Two criteria, tHR0 and tHR1 (described in Section 3.4), are exploited to evaluate the performance of the segmentation block. These values for the clean test data are calculated 96.64% and 93.11%, respectively. This test is repeated for the test signals with additive white noise in different levels of SNR. The results are brought in Table 4.

Similar experiments, using the system trained only over Farsi database, are conducted on the English utterances. The specification of the English test set was described in Section 3.1. Here, the core test of 24 speakers of TIMIT database in the clean and noisy conditions has been used. The results are brought in Table 5.

7.2. Experiments on the speech/non-speech classifier stage

When an audio signal is fed into the first-stage (segmentation block) of the SAD system, the output will be a number of audio segments which must be then classified into speech/non-speech segments. To implement the training process, 1325 speech segments extracted from FARSDAT database, besides 667 non-speech segments of different kinds described in Section 3.2 have been used. The test set consists of 537 speech and 282 non-speech segments (without any overlap with the train set). The 15-dimensional fea-

Table 4
tHR0 and tHR1 calculated for the segmentation block of the SAD system, applied to the Farsi test signal, for different levels of additive noise.

Noise type	SNR (dB)	tHR0 (%)	tHRI (%)
Clean signal	–	96.64	93.11
White	10	91.32	94.67
	5	85.20	93.89
	0	76.47	90.54
Pink	10	87.20	93.60
	5	81.37	91.52
	0	73.99	88.31
Babble	10	84.03	93.83
	5	77.81	79.79
	0	74.28	68.20
Factory	10	87.80	94.14
	5	78.75	93.22
	0	74.45	83.86

Table 5
tHR0 and tHR1 calculated for the segmentation block of the SAD system, applied to English utterances, for different levels of additive noise.

Noise type	SNR (dB)	tHR0 (%)	tHRI (%)
Clean signal	–	94	96
White	10	91	93.2
	5	89.2	88.3
	0	81.4	83
Pink	10	92	93.4
	5	88.2	83
	0	82.4	79.8
Babble	10	91	92.4
	5	86.1	86.3
	0	77.3	80.4
Factory	10	90	92
	5	84	84.1
	0	73.4	77.3

ture vectors, based on the optimized feature set discussed in Section 5.2.2, are extracted for the segments.

For the linear SVM, constant C is selected experimentally equal to 50. To find a proper value for this constant, we examined HR0 and HR1 values over different values of C. By changing C from 1 to 50, HR0 and HR1 values increased, but they decreased for C values more than 50 (this may be due to over fitting).

First, SVM classifier has been trained by feature vectors extracted from the clean segments. Due to applying this classifier to the clean test set, HR0 and HR1 values are obtained 100% and 99.92%, respectively. This performance is degraded by adding background noise to the test set. For additive white noise with SNR values of 10, 5 and 0 dB, HR1 reduces to 66%, 55% and 47%, respectively. In these conditions, no considerable change in HR0 value is seen.

In order to increase the efficiency of the classifier in presence of background noise, linear SVM is trained with a noisy train set. For this purpose, different noise types including white, pink, babble and factory are added to the training dataset, with different SNR values of 10, 5 and 0 dB. This noisy database is then added to the clean train dataset. In this manner, a new mixed train set is prepared. By renewing the train phase of the SVM classifier, new test results obtained for HR0 and HR1, which are brought in Table 6.

To examine a more complex SVM over our favorite task, a SVM with polynomial kernel is trained by the mixed noisy train set. The polynomial parameters of d, *Gamma* and *Coeff* values (in Eq. (20)) are experimentally selected 4, 2 and 25, respectively. The test results are brought in Table 7.

Table 6
HR0 and HR1 for the linear SVM classifier (train and test on Farsi).

Noise type	SNR (dB)	tHR0 (%)	tHRI (%)
Clean signal	–	99.47	95.39
White	10	95	86
	5	94	83
	0	91	83
Pink	10	96	84
	5	96	77
	0	96	71
Babble	10	98	76
	5	97	70
	0	97	62
Factory	10	97	90
	5	97	83
	0	97	83

To show the language independency of the approach, SVM classifier (trained only by Farsi train set) is examined over an English database consisting of 2541 speech segments from TIMIT database (taken from the 24 speaker test core discussed in Section 3.1), plus 950 non-speech segments (from those introduced in Section 3.2). The obtained results are brought in Table 8.

7.3. Speech recognition experiments

To evaluate the performance of the proposed SAD in a speech recognition task, we have measured the WRR in an ASR system with and without using the proposed SAD. For this purpose, HTK toolbox is employed; which could be configured as a HMM based ASR system. The baseline system doesn't have any word language model. The feature set used in the ASR system consists of 39 MFCC-based parameters (MFCCs plus their delta and delta-delta coefficients). 6-state, 16-mixture left-to-right diagonal covariance HMMs have been employed to model traditional phonemes, plus a 3-state silence model and a single-state short pause model. The frame rate is 10 ms, with a frame length of about 23 ms. The test set is the 24 speaker core test of TIMIT which 950 non-speech segments are added inside the silences intervals. Of course, different stationary noises could be added to this mixed signal, to have more general distorted conditions for the test set. Table 9 presents the WRR values obtained by applying the clean and some noisy versions of the test set to the ASR system.

As shown in Table 9, in all cases, there is about 5–9% improvement in speech recognition accuracy. For clean test set, WER has been decreased from 32% to 22.8% which shows a relative WER reduction improvement of as high as 28%.

Table 7
HR0 and HR1 for the SVM classifier with polynomial kernel (train and test on Farsi).

Noise type	SNR (dB)	tHR0 (%)	tHRI (%)
Clean signal	–	99.47	95.39
White	10	95	87
	5	94	86
	0	92	83
Pink	10	96	87
	5	95	84
	0	93	79
Babble	10	95	79
	5	96	77
	0	95	74
Factory	10	96	91
	5	97	90
	0	97	90

Table 8
HR0 and HR1 results for the SVM classifier with polynomial kernel (train by Farsi and test with English).

Noise type	SNR (dB)	tHR0 (%)	tHRI (%)
Clean signal	–	99.6	95.3
White	10	96.8	88.2
	5	93.6	85.3
	0	92.4	84.5
Pink	10	95.7	86
	5	93.6	85.9
	0	94	81.3
Babble	10	92.2	78.2
	5	93	75.7
	0	91.7	75
Factory	10	98.2	92
	5	96	89.1
	0	93.5	86.9

Table 9
WRR values using SAD at the input of the ASR system.

Signal	WRR with SAD	WRR without
Clean	77.2	68
White-10 dB	71.4	63.2
Pink-10 dB	66.8	60.4
Babble-10 dB	65.6	58.7
Factory-10 dB	64.3	59.5
Volvo-10 dB	66.3	58.9

8. Discussions

In this paper, a two-stage SAD is introduced. It is shown that this approach can be considered as a language independent technique. As mentioned in the previous sections, A Farsi speech database is used to extract speech segments for determining the system parameters and training of the classifiers. The first-stage of the system is a segmentation block that rejects those parts of signal which can easily be indentified as non-speech, based on energy and duration of the audio segments. Experiments conducted on English speech signal, Table 5, have shown the language independency of the system. The first row of Table 5 shows the segmentation results which are almost identical to the first row of Table 4 for Farsi. This is also true for the noisy conditions as well. Outputs of the segmentation block are audio segments which must be classified as speech or non-speech. The final classification process is done by a SVM classifier, fed by an optimal subset of feature vectors. A SVM classifier has been trained over features extracted from Farsi speech and some non-speech segments of various kinds. The optimal reduced feature set consists of features, such as PMS-Petrosian, and some MFCC-based features. Of course, there are many common phones and sub-phones in Farsi and English and these languages are not inherently different in a sense of phonology. The comparison of the results shown in Tables 7 (for Farsi) and 8 (for English) show the efficiency of the system for both languages. So, we can proclaim that the overall developed system is language independent to some extent and could be used for other languages as well. The final test in Section 7.3 conducted on English utterances, shows excellent results and can boost up the performance of the ASR system.

In SAD systems, performance degradation in presence of background noise is an unresolved problem. To make the proposed SAD system robust against background noise, an adaptive energy threshold is employed in the first stage of the system Tables 4 and 5 show acceptable performance of the segmentation block, in presence of background noises. In the second stage (speech/non-speech classifier), some noisy speech segments have been added to the original train set, to make the classifier more robust against noisy signal. Tables 6–8 show the corresponding results of the noisy speech/non-speech segments classification, over Farsi. It is shown that in high energy background noise, the SVM classifier with a complicated kernel like "polynomial kernel" is more robust to background noise than the linear SVM. But, in high SNRs, linear SVM, with a lower complexity, could have an acceptable performance.

9. Conclusions

In this research, a two-stage speech activity detector is proposed consisting of two distinct cascaded systems. First, a segmentation block based on a FSM machine, which operates on the energy and duration of the audio segments, rejects pauses of the input signal and passes speech and non-speech audio segments to the next stage. This preprocessing unit benefits from an adaptive energy threshold to overcome background noises which vary slowly in time. The second stage of the system is consisted of a SVM classifier which considers a feature vector for every segment to identify whether or not the segment is speech. An extensive investigation had been conducted on two new fractal features which are shown to be excellent discriminators for speech/non-speech classification. A GA is employed to reduce the feature set from 34 to 15 in order to reduce complexity and redundancy. Finally, conducting different experiments on Farsi and English test sets, it has been shown that our approach is to some extent language independent. The robustness of the system against background noises was shown to be acceptable. The proposed SAD was employed in conjunction with an English HMM-based ASR system, which resulted in an improvement of 28% in relative WER reduction.

References

Abdulla, W.H., Kecman, V., Kasabov, N., 2003. Speech-background classification by using SVM technique. In: 13th Internat. Conf. on Artificial Neural Networks (ICANN'03), pp. 310–315.

Banbrook, M., McLaughlin, S., 1994. Is speech chaotic?: Invariant geometrical measures for speech data. In: IEE Colloquium on Exploiting Chaos in Signal Processing, p. 810.

Basu, S., 2003. A linked-HMM model for robust voicing and speech detection. In: Proc. IEEE Conf. on Acoustic, Speech and Signal Processing (ICASSP'03), vol. 1, pp. 816–819.

Beaufays, F., Boies, D., Weintraub, M., Zhu, Q., 2003. Using speech/non-speech detection to bias recognition search on noisy data. In: Proc. IEEE Conf. on Acoustic, Speech and Signal Processing (ICASSP'03), vol. 1, pp. 424–427.

Bijankhan, M., Sheikhzadegan, M.J., 1994. FARSDAT- the Farsi spoken language database. In: Proc. Fifth Australian Internat. Conf. on Speech Sciences and Technology (SST'94), vol. 2, pp. 826–829.

Bijankhan, M., Sheikhzadegan, M.J., Roohani, M.R., Zarrintare, R., Ghasemi, S.Z., Ghasedi, M.E., 2003. Tfarsdat – the telephone Farsi speech database. In: Proc. of the Eurospeech 2003, pp. 1525–1528.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discovery 2 (2), 121–167.

Chen, S.H., Wu, H.T., Chen, C.H., Ruan, J.C., Truong, T.K., 2005. Robust voice activity detection algorithm based on the perceptual wavelet packet transform. In: Proc. IEEE Internat. Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'05), pp. 45–48.

Cohen, A., 1986. Biomedical Signal Processing, vol. 2. CRC Press.

Enqing, N., Guizhong, L., Yatong, Z., Xiaodi, Z., 2002. Applying support vector machines to voice activity detection. In: Proc. Internat. Conf. Signal Processing (ICSP'02), vol. 2, pp. 1124–1127.

Esteller, R., Vachtsevanos, G., Echauz, J., Litt, B., 2001. A comparison of waveform fractal dimension algorithms. IEEE Trans. Circuits Syst. 48 (2), 177–183.

FREESOUND. <http://freesound.iua.upf.edu/>.

Górriz, J.M., Puntonet, C.G., Ramírez, J., Segura, J.C., 2005. Statistical tests for voice activity detection. In: ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP'05), pp. 240–249.

Hilborn, R.C., 2001. Chaos and Nonlinear Dynamics, second ed. Oxford University.

Hori, C., Furui, S., 2003. A new approach to automatic speech summarization. IEEE Trans. Multimedia 5 (3), 368–378.

HTK toolbox, <http://htk.eng.cam.ac.uk/>.

Hussain, A., Abdul Samad, S., Fah, L.B., 2000. Endpoint detection of speech signal using neural network. In: Proc. IEEE TENCON, vol. 1, pp. 271–274.

Justino, E.J.R., Bortolozzi, F., Sabourin, R., 2005. A comparison of SVM and HMM classifiers in the off-line signature verification. Pattern Recognition Lett. 26 (9), 1377–1385.

Kokkinos, I., Maragos, P., 2005. Nonlinear speech analysis using models for chaotic systems. IEEE Trans. Speech Audio Process. 13, 1098–1109.

Koumpis, K., Renals, S., 2001. The role of prosody in a voicemail summarization system. In: Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ, pp. 87–92.

Martin, A., Karray, L., Gilloire, A., 2000. High order statistics for robust speech/non-speech detection. In: European Signal Processing Conference (EUPSICO 2000), pp. 469–472.

Martin, A., Mauuary, L., 2006. Robust speech/non-speech detection based on LDA-derived parameter and voicing parameter for speech recognition in noisy environments. Speech Comm. 48 (2), 191–206.

Mauuary, L., 1994. Improving the performances of interactive voice response services. Ph.D. thesis, University of Rennes (in French).

NOISEX-92. <http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html>.

Ramirez, J., Sagura, J.C., Benitez, C., de la Torre, A., Rubio, A., 2004. Efficient voice activity detection algorithms using long-term speech information. Speech Comm. 42, 271–287.

Ramirez, J., Sagura, J.C., Benitez, C., de la Torre, A., Rubio, A., 2005. An effective subband OSF-based VAD with noise reduction for robust speech recognition. IEEE Trans. Speech Audio Process. 13 (6), 1119–1129.

Padrell, J., Macho, D., Nadeu, C., 2005. Robust speech activity detection using LDA applied to FF parameters. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'05), vol. 1, pp. 557–560.

Pal, M., Mather, P.M., 2004. Assessment of the effectiveness of support vector machines for hyperspectral data. Future Generation Comput. Syst. 20 (7), 1215–1225.

Prasad, R.V., Sangwan, A., Jamadagani, H.S., Chiranth, M.C., Shah, R., Gaurav, V., 2002. Comparison of voice activity detection algorithms for VOIP. In: Proc. Seventh Internat. Symposium on Computers and Communications (ISCC'02), pp. 530–535.

Savoji, M.H., 1989. A robust algorithm for accurate endpointing of speech signals. Speech Comm. 8 (1), 46–60.

Selouani, S.A., O'Shaughnessy, D., 2004. Robustness of speech recognition using genetic algorithms and a mel-cepstral subspace approach. In: IEEE Internat. Conf. Acoustics Speech and Signal Processing (ICASSP'04), vol. 1, p. 201.

Shafiee, S., Almasganj, F., Jafari, A. 2008. Speech/non-speech segments detection based on chaotic and prosodic features. In: Proc. Interspeech 2008, pp. 111–114.

Shaojun, J., Haitao, G., Fuliang, Y., 2004. A New algorithm for VAD based on wavelet transform. In: Proc. Internat. Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 222–225.

Shen, J.L., Hung, J., Lee, L.S., 1998. Robust entropy based end point detection for speech recognition in noise. In: 5th Internat. Conf. on Spoken Language Processing (ICSLP'98), vol. 3, pp. 1015–1018.

Skorik, S., Berthommier, F., 2000. On a cepstrum-based speech detector robust to white noise. In: Proc. Internat. Conf. on Speech and Computer (Specom 2000).

Tian, Y., Wang, Z., Lu, D., 2002. Non-speech segment rejection based on prosodic information for robust speech recognition. IEEE Signal Process. Lett. 9 (11), 364–367.

Vapinik, V., 1998. Statistical Learning Theory. Wiley, NewYork.

Villarrubia, L., Acero, A., 1993. Rejection techniques for digit recognition in telecommunication applications. In: Proc. IEEE Conf. on Acoustic, Speech and Signal Processing (ICASSP'93), vol. 2, pp. 455–458.

Waheed, K., Weaver, F.M. Salam, 2002. A robust algorithm for detecting speech segments using an entropic contrast. In: The 45th Midwest Symposium on Circuits and Systems, vol. 3, pp. 328–331.

Wu, B.F., Wang, K.C., 2005. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. IEEE Trans. Speech Audio Process. 13 (5, Part 2), 762–777.

Xianbo, X., Guangshu, H., 2005. An incremental support vector machine based speech activity detection algorithm. In: Proc. 27th Annual Internat. Conf. of the Engineering in Medicine and Biology Society (IEEE-EMBS'05), pp. 4224–4226.

Xu, P., Chan, A.K., 2002. Optimal wavelet sub-band selection using genetic algorithm. In: IEEE Internat. Symposium on Geoscience and Remote Sensing (IGARSS'02), vol. 3, pp. 1441–1443.

Ying, G.S., Mitchell, C.D., Jamieson, L.H., 1993. Endpoint detection of isolated utterances based on a modified Teager energy measurement. In: Proc. IEEE Conf. on Acoustic, Speech and Signal Processing (ICASSP'93), vol. 2, pp. 732–735.