Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 3, 2022, 401-409.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Voice Activity Detection using Clustering-based Method in Spectro-Temporal Features Space

Nafiseh Esfandian[1*], Fatemeh Jahani bahnamiri[2] and Samira Mavaddati[3]

1. Department of Electrical Engineering, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran.
2. Department of Computer Engineering, Aryan Institute of Science and Technology, Babol, Iran.
3. Department of Electrical Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran.

Article Info

Article History:
Received 01 December 2021
Revised 06 January 2022
Accepted 17 February 2022

*DOI:*10.22044/jadm.2022.11439.2304

Keywords:
Spectro-temporal Features, Auditory Model, Gaussian Mixture Model, WK-means clustering, Voice Activity Detection.

**Corresponding author: na_esfandian@Qaemiau.ac.ir (N. Esfandian).*

Abstract

This paper proposes a novel method for voice activity detection based on clustering in the spectro-temporal domain. In the proposed algorithms, the auditory model is used in order to extract the spectro-temporal features. The Gaussian mixture model and the WK-means clustering methods are used to decrease the dimensions of the spectro-temporal space. Moreover, the energy and positions of the clusters are used for voice activity detection. Silence/speech is recognized using the attributes of clusters and the updated threshold value in each frame. Having a higher energy, the first cluster is used as the main speech section in computation. The efficiency of the proposed method is evaluated for silence/speech discrimination in different noisy conditions. Displacement of the clusters in the spectro-temporal domain is considered as the criterion to determine the robustness of the features. According to the results obtained, the proposed method improves the speech/non-speech segmentation rate in comparison to the temporal and spectral features in low signal to noise ratios (SNRs).

1. Introduction

In many speech processing applications, it is essential to use voice activity detection (VAD) in order to discriminate speech from silence [1-4]. In this paper, the spectro-temporal features were used for silence/speech detection. In fact, many VAD methods have been proposed [5-7]. In each method, the speech features such as short time energy and zero-crossing rate (ZCR), auto-correlation function analysis, cepstral peak, and Mel Frequency Cepstral Coefficients (MFCCs) have been used for speech segmentation [8-14]. In the recent years, VAD based on deep neural network (DNN) has got a great success. However, typically developing noise-robust and more generalized deep learning-based voice activity detection requires the collection of a great amount of annotated speech data [15-17]. In the proposed method, the spectro-temporal features were extracted using the auditory model in order to detect speech from silence. The Gaussian Mixture Model (GMM) and the WK-means clustering algorithms were employed to extract the main speech sections, and the attributes of clusters were used as the secondary features to detect silence and speech frames [18]. Three clusters were used in order to segment the spectro-temporal space. The clusters were ranked in energy; therefore, the first cluster has the highest energy, and includes the main information of speech signal. The first cluster's energy in each frame was compared with the threshold value to detect the speech frames. In the proposed method, the threshold was updated in each frame. If the first cluster's energy was higher than the threshold value, this frame was identified as speech; otherwise, it was identified as silence. In this work, the sentences of the TIMIT database were used for the efficiency evaluation of the proposed method [19]. The auditory model and clustering-based feature selection method in the spectro-temporal domain is discussed in Section 2. The proposed method for silence/speech detection using the spectro-temporal features is presented in Section 3. The experimental results and the performance evaluation are analyzed in Section 4. Finally, the research work is concluded in Section 5.

2. Spectro-Temporal Representation of Speech using Auditory Model

The auditory model is the simulated model of the inner ear and the first layer of the auditory cortex that is used in many speech processing applications [20, 21]. This model consists of two main sections [22, 23]. In the first section, the auditory spectrogram of speech signal is calculated. In the second section, this 2D representation is converted into the spectro-temporal features using a 2D filter bank [24]. In fact, the 2D wavelet transform of the auditory spectrogram is obtained at this stage. The spectro-temporal impulse response of 2D filters is called the spectro-temporal response field (STRF). The outputs of the cortical filters are computed using the convolution of STRFs with the auditory spectrogram. The cortical representation of speech has four dimensions including scale (Ω in cycles/octave), rate (ω in Hz), frequency (f–the number of the band-pass filter), and time (t–the frame number). Therefore, it is important to select the noise-robust features due to the feature space vastness in the spectro-temporal model. In the recent studies, the clustering-based feature selection method has been employed to decrease the features space dimensions [25]. The spectro-temporal features were used in the speech processing applications [26-29].

2.1. Selection of spectro-temporal feature using clustering methods

In this method, the auditory spectrogram of a speech signal was first computed for each frame [30, 31]. Then the spectro-temporal features were extracted using the auditory spectrogram and the auditory cortex model. Since the cortical output of the auditory model has four dimensions (scale, rate, frequency, and time), the clustering block was used to decrease the vast dimensions of the spectro-temporal feature space, and to select the valuable features. In this section, the feature space was segmented using the clustering methods. As a result, new feature vectors with smaller dimensions were extracted by determining the main speech clusters in each frame. In this method, the spatial information of each point in the feature space was considered in the initial feature vector. In this feature extraction mechanism, the mean and variance vectors of cluster centers were considered as the secondary features, $V = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$ by selecting three clusters in the spectro-temporal space and clustering the initial feature vectors. The energy measure was used to sort the mean and variance vectors of the cluster centers. μ_1 and σ_1 are the mean and variance vectors of the cluster centers with a higher weight. In fact, the mean vectors of the cluster centers were first sorted using energy measure. The variance vectors of the cluster centers were then sorted using the same measure to form the secondary feature vector in each frame. The main motivation of this work is to extract the attributes of these clusters as the secondary high level features for speech/silence classification.

3. Silence/Speech Detection using Clustering-based Feature Extraction Method in Spectro-Temporal Domain

The spectro-temporal features were used in the proposed method for silence/speech detection. The block diagram of the proposed method is shown in Figure 1. In the preprocessing stage, a 16-ms window was used for speech signal windowing. In the next stage, the 4-D cortical output was computed in each speech frame using the auditory model.

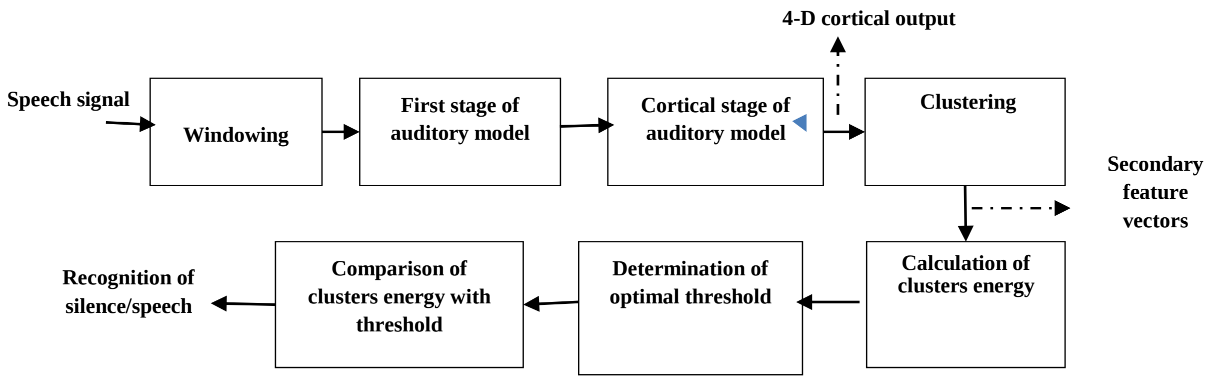

Figure 1. Block diagram of proposed method for silence/speech detection.

In this work, two clustering algorithms were used for spectro-temporal feature extraction. In the first method, the initial feature vectors had four dimensions; f_i denotes the frequency, s_i is the scale, r_i is the rate, and $|A_i|$ is the amplitude of points. The initial feature vectors, $v_i = (r_i, s_i, f_i, |A_i|)$, were clustered using GMM clustering. In the second method, the initial feature vectors, $v_i = (r_i, s_i, f_i)$, had three dimensions, and the amplitude of points, $w_i = |A_i|$, was considered as the weight vector. The initial feature vectors were clustered using WK-means clustering. The secondary feature vectors with smaller dimensions were computed by extracting the main clusters of speech in each frame. After that, the energy of the *k*th cluster in the *i*th frame was determined through the following equation:

$$W_{Ci,k} = \frac{1}{N}\sum_{n=1}^{N}|A_n|^2 \tag{1}$$

In this equation, N is the number of samples in the *k*th cluster, whereas A_n indicates the amplitude of the points belonging to the *k*th cluster. The clusters were ranked using energy measure; as a result, the first cluster had the highest energy. Therefore, the first cluster contains the valuable information of speech. Thus the first cluster's energy was compared with the threshold for speech/non-speech detection.

3.1. Determination of threshold value

The initial value of the threshold, which was updated along the VAD process, was determined using the mean energy of the first clusters in all frames.

$$T(0) = T_0 = \frac{1}{M}\sum_{i=1}^{M} W_{Ci,1} \tag{2}$$

In this equation, M shows the number of speech frames and $W_{Ci,1}$ indicates the first cluster's energy in the *i*th frame. If the first cluster's energy in the first frame was higher than T_0, the first frame was considered as the speech frame; otherwise, it was considered silence. In the proposed method, the threshold was updated in each frame. This initialization method was obtained empirically using the results of different simulations. In the next frames, the threshold was determined by applying the correction coefficients with respect to the type of the previous frame. The value of T was obtained from Equation (2). It was then updated for the next frames with respect to the frame type. If the *i*th frame included speech, the value of T was obtained from the following equation:

$$T(i+1) = T(i) - \alpha(i) \times \alpha_1 \times W_{C1}(i+1) \tag{3}$$

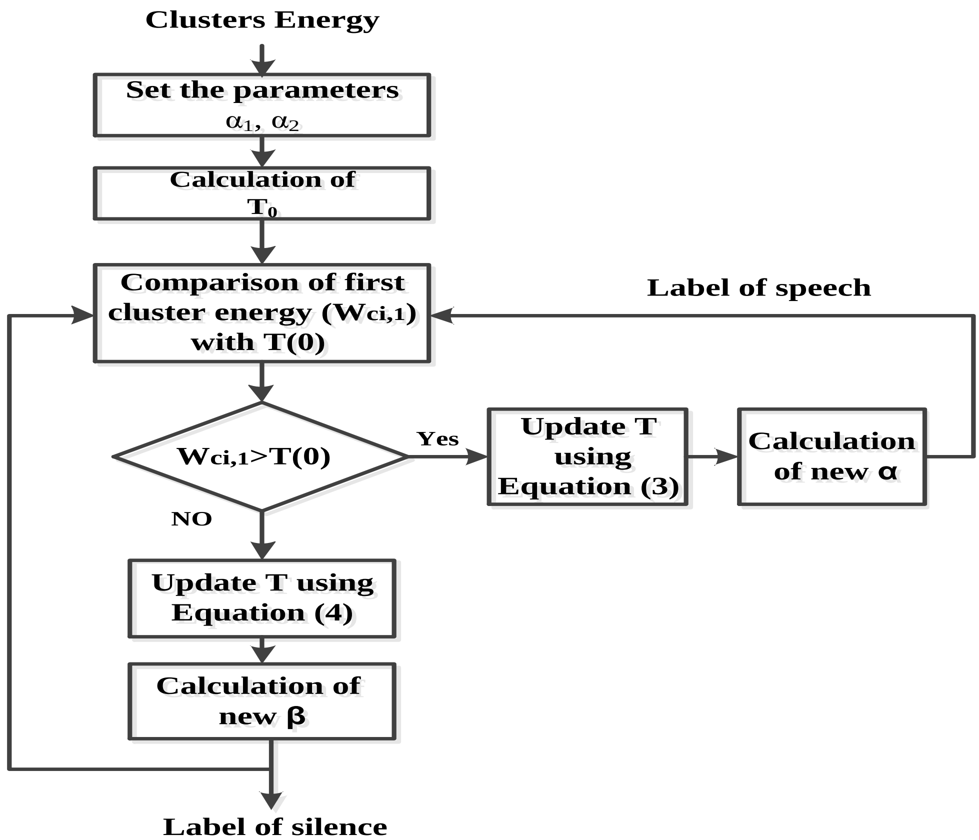

Figure 2. A flowchart of proposed method for silence/speech detection..

If the *i*th frame was silence, the value of T was calculated using Equation (4), in which α and β were considered to determine the optimal threshold value. A flowchart of the proposed method is shown in Figure 2.

$$T(i+1)=T(i)+\beta(i)\times\alpha_2\times W_{C1}(i+1) \quad (4)$$

Table 1 shows how to update the threshold value as the pseudo-code. The parameter α was considered so that the reduction rate, α_1, would not be the same in case a number of speech frames were put in a row. The larger number of consecutive speech frames, the lower the reduction slope of the threshold; otherwise, the non-speech frames might be identified as the speech frames. According to $\alpha(i+1)=\frac{\alpha(i)+1}{2}$, the definitive value of this parameter decreases if the number of consecutive speech frames decreases. It is put 1 until the detection of other speech frames starts. At the same time, the parameter β controls the increased threshold for non-speech frames. In this case, the threshold value increases at a lower rate for the large number of consecutive non-speech frames; otherwise, the speech frames might also be labeled as non-speech. According to $\beta(i+1)=\frac{\beta(i)+1}{2}$, the value of this parameter decreases if the number of consecutive non-speech frames increases.

Table 1. Pseudo-code for determining and updating threshold in proposed method.

```
Algorithm: Input: W_C1, α1, α2
Output: Flag
%parameter setting
α1 = 0.09
α2 = 0.06
T(0) = Mean(W_C1); % Equation (2)
α(0) = 0
β(0) = 0
%Threshold parameter adaptation
    For i=0: Number of frames
      IF W_C1(i+1) ≥ T(i)
       T(i+1)=T(i) − α(i) × α1 × W_C1(i+1)
        Flag(i) = 1;  % Speech frame
        α(i+1) = (α(i)+1)/2;
        β(i+1) = 0;
      else
       T(i+1)=T(i) + β(i) × α2 × W_C1(i+1)
        Flag(i) = 0;  % Silent frame
        β(i+1) = (β(i)+1)/2;
        α(i+1) = 0;
    end
  end
```

It is put 1 until the detection of another non-speech frame starts again.

3.1. Post-processing

After the input speech frames were labeled, the labels were post-processed to analyze the following conditions:

- If a zero label indicating non-speech frame is located between two frames with labels of one showing speech frame, it will be changed to 1.
- If a label of one indicating speech frame is located between two frames of zero labels showing non-speech frame, it will be changed to zero.

4. Experimental Results

Noisy speech was used for silence/speech detection in order to analyze the efficiency of the proposed algorithm. For this purpose, the proposed method was evaluated using 80 sentences; 40 female and 40 male speakers were randomly selected from the TIMIT database of each eight dialects. Different noises were used from noisex-92 [32].

4.1. Performance evaluation measure

In order to determine the accuracy, and evaluate the performance of the proposed method, the accuracy measure was used for silence/speech detection.

$$Accuracy\_Vocal=\frac{N_V}{N_T}\times 100 \quad (5)$$

$$Accuracy\_Vocal=\frac{N_V}{N_T}\times 100 \quad (6)$$

N_V and N_S respectively, are the number of correctly detected speech frames and the number of correctly detected silence frames. N_T is the number of total frames.

4.2. Speech/silence detection using empirical threshold

Figure 3 shows the male speaker's speech signal, whereas Figure 4 indicates the energy of the first, second, and third clusters extracted through GMM for a male speaker. Accordingly, the first cluster has a higher energy than the other clusters. It can be concluded that analysis of the first cluster's energy will have the greatest influence on the performance of the proposed method. Therefore, the energy of first cluster in each frame was used for detecting the input frame type (silence or speech).

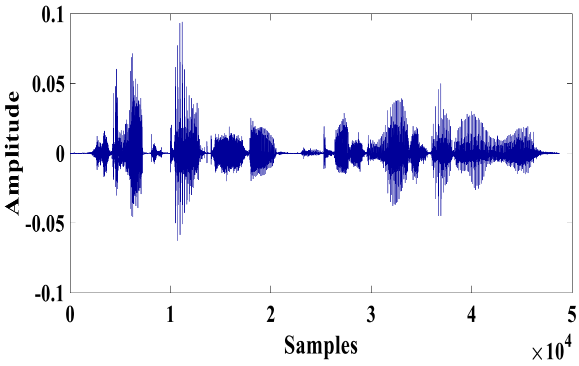

Figure 3. Male speaker's speech signal.

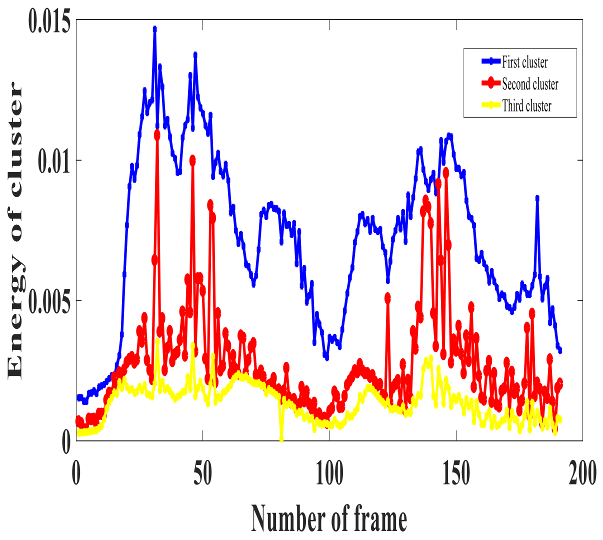

Figure 4. Energy of first, second, and third clusters for a male speaker.

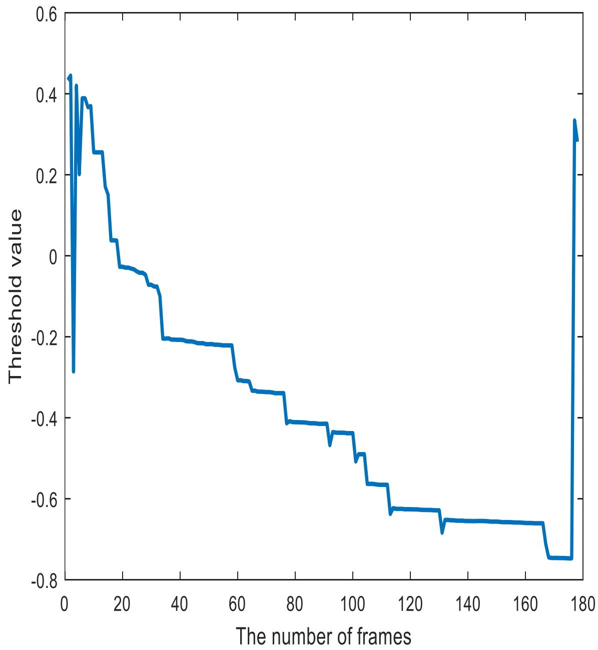

Figure 5. Updated threshold for a male speaker.

Accordingly, the threshold introduced by Equation (2) was updated based on the first cluster's energy in consecutive frames. The updating process was based on Equations (3) and (4). Figure 5 shows the updated threshold for a male speaker. Since most of the intermediate frames represent speech, the threshold for determining the input frame type decreases so that the intermediate speech frames can have appropriate labels. Table 2 presents the average accuracy rates of the proposed method for speech and silence detection using 80 sentences from the

Table 2. Evaluation of proposed method in different noises and various SNRs (%).

SNR (dB)	Noise					
	Features	**White**	**Street**	**Exhibition**	**Car**	**Babble**
20	Energy and ZCR	91.4	90.1	89.6	89.1	85.5
	WK-means Clustering	95.5	93.3	94.8	94.5	92.2
	GMM Clustering	98.7	96.9	97.1	98.9	94.9
15	Energy and ZCR	89.3	88.5	87.6	86.1	80.3
	WK-means Clustering	93.9	92.2	93.5	93.7	90.8
	GMM Clustering	97.1	94.8	96.3	96.4	92.5
10	Energy and ZCR	86.9	86.1	83.4	84.0	79.1
	WK-means Clustering	91.5	90.9	89.9	89.7	88.4
	GMM Clustering	94.9	92.5	93.3	92.8	89.6
5	Energy and ZCR	79.7	78.5	76.5	77.5	73.1
	WK-means Clustering	85.1	84.2	83.1	82.3	80.6
	GMM Clustering	86.2	85.9	85.5	84.1	80.3
0	Energy and ZCR	68.2	67.1	66.7	65.6	60.4
	WK-means Clustering	78.5	78.2	77.4	76.3	75.2
	GMM Clustering	79.0	78.5	77.8	77.6	76.0
-5	Energy and ZCR	54.5	53.8	52.3	51.5	49.3
	WK-means Clustering	64.3	64.1	62.1	61.9	59.2
	GMM Clustering	65.5	65.1	63.8	62.1	60.4

Table 3. Comparison of proposed features with conventional features for some sentences of timit

				Frame Error Rate (%)		
Dialect	**Sex**	**Speaker**	**Sentence**	**Sentence code**	**Energy and ZCR**	**Proposed Features**
DR1	Female	FDAW0	"Steve collects rare and novel coins"	SX326	14.9	12.3
DR1	Male	MCPM0	"She had your dark suit in greasy wash water all year"	SA1	15.1	14.9
DR2	Female	FAEM0	"Fill small hole in bowl with clay"	SI762	13.2	10.3
DR3	Male	MBEF0	"Far more frequently, overeating is a result of a psychological compulsion"	SI651	15.3	13.6
DR4	Female	FCAG0	"They all agree that the essay is barely intelligible"	SX243	14.2	10.1
DR8	Female	FBCG1	"Suburban housewives often suffer from the gab habit"	SX442	15.6	12.6

TIMIT database at different signal-to-noise ratio (SNR) levels. White, street, exhibition, car, and babble noises were added to clean speech. In this table, the proposed method using GMM and WK-means clustering techniques were compared with the combination of energy and ZCR features in various noisy conditions. According to the results, the accuracy rate of silence/speech detection was improved in low noise conditions. Moreover, the accuracy rate of GMM was higher than WK-means clustering since it was difficult to distinguish the fricative phoneme from the background noise. The frame error rate of the proposed algorithm in comparison to the conventional features for some sentences starting with the fricative phonemes is shown in Table 3. In this table, the results were obtained using white noise with SNR of 10dB. The frame error rate was decreased using the proposed features in the spectro-temporal domain. The main goal of this work was to compare the clustering-based features in the spectro-temporal domain with the conventional features used for voice activity detection. Therefore, in Figure 6, the proposed features were compared with the combined features such as zero crossing rate, short time energy, spectral entropy, and linear prediction error (LPE) [8, 33]. As it could be observed, the frame error rate was decreased using the spectro-temporal features in comparison to the conventional features.

In Figure 7, the frame error rate of the proposed method was compared with the unsupervised method using SNR of 15dB. In this method, the long-term features were computed using the fractal dimension estimation [34]. The results obtained show that frame error rate was improved using the proposed method.

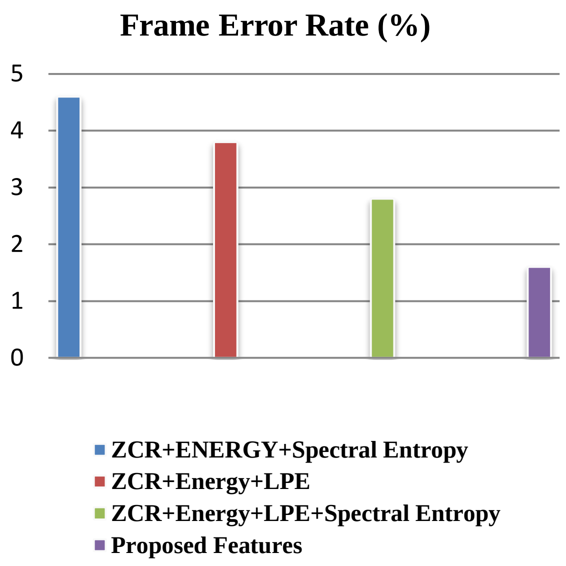

Figure 6. Frame error rate of proposed features in comparison to conventional features.

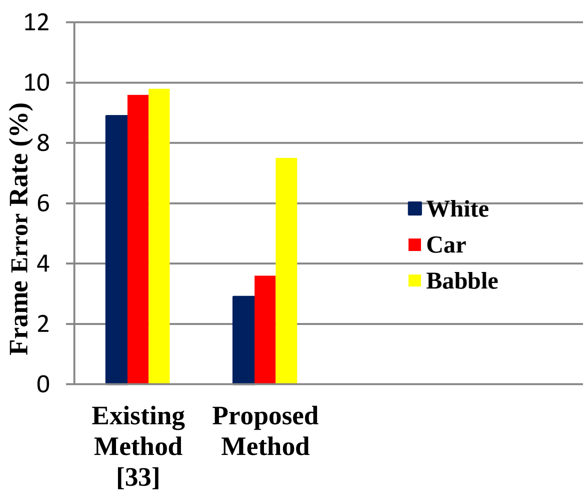

Figure 7. Frame error rate of proposed method in comparison to existing method.

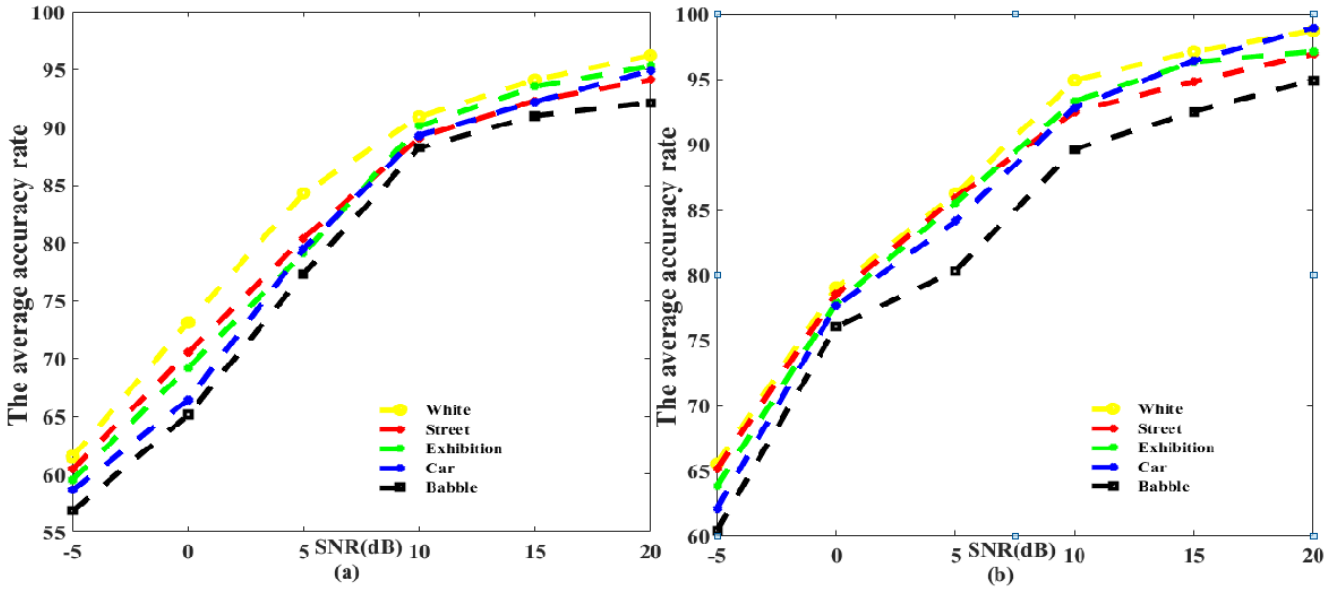

Figure 8. Evaluation of proposed method in comparison to previous technique for speech/silence detection. (a): Combination of auto-correlation function, ZCR, and cepstral peak. (b): Proposed method using GMM clustering.

Figure 8 indicates the results of evaluating the proposed method for speech/silence detection in comparison to the combination of auto-correlation function, ZCR, and cepstral peak in different noise and various SNRs. The mean accuracy of the proposed algorithm was improved through GMM clustering in comparison to the existing. Accordingly, the proposed method improved the accuracy rate of speech/silence detection in comparison to the temporal and spectral features.

4.3. Displacement of clusters in spectro-temporal domain

Displacement of cluster centers is one of the criteria to determine the robustness of features. $\bar{D}$, displacement of clean speech clusters relative to noisy speech clusters, is defined as:

$$\bar{D} = \frac{\sum_{j=1}^{m} \sqrt{D_j^{\ 2}}}{m} \tag{7}$$

$$D_j^{\ 2} = \sum_{i=1}^{n} \frac{(\overset{c}{\mu_{ij}} - \overset{n}{\mu_{ij}})^2}{\overset{c}{\mu_{ij}}^{\ 2}} \tag{8}$$

where m denotes the number of speech frames, and D_j^2 is the displacement of mean vector of each clean speech cluster relative to noisy speech in j[th] frame. In addition, n denote the number of features, $\overset{c}{\mu_{ij}}$ and $\overset{n}{\mu_{ij}}$ are the mean vector of clean speech cluster and mean vector of noisy speech cluster. $\bar{\Delta}$, change of variance of each clean speech cluster relative to noisy speech, is defined as:

$$\bar{\Delta} = \frac{\sum_{j=1}^{m} \sqrt{\Delta_j^{\ 2}}}{m} \tag{9}$$

$$\Delta_j^{\ 2} = \sum_{i=1}^{n} \frac{(\overset{c}{\sigma_{ij}} - \overset{n}{\sigma_{ij}})^2}{\overset{c}{\sigma_{ij}}^{\ 2}} \tag{10}$$

$\overset{c}{\sigma_{ij}}^{\ 2}$ and $\overset{n}{\sigma_{ij}}^{\ 2}$ are the variance of clean speech cluster and noisy speech cluster. Table 3 and Table 4 present $\bar{D}_k$ and $\bar{\Delta}_k$, displacement of mean and variance vectors of k[th] clean speech cluster relative to noisy speech cluster using WK-means and GMM clustering-based features. Distance of mean and variance vectors of clean and noisy clusters decreases in the high SNRs. As it can be observed, displacement of clusters using GMM clustering-based features is less than the WK-means clustering-based features. Therefore, the GMM features are more robust than the WK-means features.

Table 3. Displacement of clusters using WK-means clustering-based features.

	SNR (dB)					
	20	**15**	**10**	**5**	**0**	**-5**
$\bar{D}_1$	5.1	7.5	8.3	14.6	20.9	24.5
$\bar{D}_2$	4.5	8.1	8.9	9.5	12.2	15.6
$\bar{D}_3$	5.2	6.8	7.1	8.2	9.3	11.3
$\bar{\Delta}_1$	1.2	1.5	2.0	2.1	2.4	4.2
$\bar{\Delta}_2$	1.9	2.2	2.3	2.5	2.6	3.5
$\bar{\Delta}_3$	2.1	2.4	2.7	2.8	2.8	3.8

Table 4. Displacement of clusters using GMM clustering-based features.

	SNR (dB)					
	20	**15**	**10**	**5**	**0**	**-5**
$\overline{D}_1$	3.1	3.7	4.3	4.8	5.2	6.9
$\overline{D}_2$	3.0	3.5	5.8	6.2	6.6	8.3
$\overline{D}_3$	6.8	7.4	9.5	10.1	10.6	13.1
$\overline{\Delta}_1$	1.2	1.2	1/3	1.4	1.6	2.5
$\overline{\Delta}_2$	1.7	1.9	1/2	2.4	2.6	4.2
$\overline{\Delta}_3$	1.1	1.1	1.2	1.3	1.5	3.1

5. Conclusion

In this work, the spectro-temporal features were used for voice activity detection. The main energy concentration of speech was extracted in the spectro-temporal space through the GMM and WK-means clustering techniques. For this purpose, the spatial information and energy of points in the spectro-temporal space were used as the initial vectors in the clustering algorithm. Since the first cluster had the highest energy in the spectro-temporal space, its energy in each frame was compared with the threshold. The threshold value was updated in each speech frame to decrease the segmentation error by optimizing the determined parameters. The updated threshold was employed to detect the speech/silence frames in noisy condition. The results obtained indicated the higher efficiency of the proposed method in comparison to the existing techniques, although the frame error rate of speech/silence detection had been improved using the proposed features in comparison to the conventional features. However, the deep learning-based methods have shown to be more robust and accurate than the statistical methods and other existing approaches. Therefore, in the future research work, the deep learning-based methods will be used to develop the proposed method.

References

[1] Z. H. Tan and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech and Language,* Vol. 59, pp. 1-21, January 2020.

[2] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with ctc-based voice activity detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP).,* Barcelona, Spain., pp. 6999-7003, May 2020.

[3] J. Lee, Y. Jung, and H. Kim, "Dual Attention in Time and Frequency Domain for Voice Activity Detection," in *Proceedings of Interspeech 2020,* Shanghai, China, pp. 3670-3674, October 2020. Available: http://arxiv.org/abs/2003.12266.

[4] Y. G. Thimmaraja, B. Nagaraja, and H. Jayanna, "Speech enhancement and encoding by combining SS-VAD and LPC," *International Journal of Speech Technology,* Vol. 24, No. 1, pp. 165-172, 2021.

[5] R. Makowski and R. Hossa,"Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise," *Applied Acoustics,* Vol. 166:107344, September 2020.

[6] F. Liu and A. Demosthenous, "A Computation Efficient Voice Activity Detector for Low Signal-to-Noise Ratio in Hearing Aids," in *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS,),* Michigan, USA, pp. 524-528, August 2021.

[7] A. K. Alimuradov, "Enhancement of Speech Signal Segmentation using Teager Energy Operator," in *2021 23rd International Conference on Digital Signal Processing and its Applications (DSPA),* Moscow, Russia, pp. 1-7, March 2021.

[8] T. H. Zaw and N. War, "The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection," in *2017 20th International Conference of Computer and Information Technology (ICCIT),* Dhaka, Bangladesh, pp. 1-5, December 2021.

[9] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech2010),* Makuhari, Chiba, Japan, pp. 3118-3121, September 2010.

[10] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing,* Vol. 2015, No. 1, pp. 1-15, 2015.

[11] R. G. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education (ASEE),* pp. 1-7, 2008.

[12] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Ninth European Conference on Speech Communication and Technology,* Lisbon, Portugal, pp. 1-4, September 2005.

[13] S. Endah, R. Kusumaningrum, S. Adhy, and R. Ulfattah, "Automatic speech recognition by using local

adaptive thresholding in continuous speech segmentation," in *Journal of Physics: Conference Series*, Vol. 1943, pp. 1-8, 2021.

[14] S. Sharma, A. Sharma, R. Malhotra, and P. Rattan, "Voice Activity Detection using windowing and updated K-Means Clustering Algorithm," in *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom pp. 114-118, April 2021.

[15] H. Khalid, S. Tariq, T. Kim, J. H. Ko, and S. S. Woo, "ORVAE: One-Class Residual Variational Autoencoder for Voice Activity Detection in Noisy Environment," *Neural Processing Letters*, pp. 1-22, 2022.

[16] F. Jia, S. Majumdar, and B. Ginsburg, "MarbleNet: Deep 1D Time-Channel Separable Convolutional Neural Network for Voice Activity Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, pp. 6818-6822, June 2021.

[17] M. Asadolahzade Kermanshahi, and M. M. Homayounpour, "Improving Phoneme Sequence Recognition using Phoneme Duration Information in DNN-HSMM." *Journal of AI and Data Mining*, Vol. 7, No. 1, pp. 137-147, 2019.

[18] N. Esfandian, "Phoneme Classification using Temporal Tracking of Speech Clusters in Spectro-temporal Domain," *International Journal of Engineering*, Vol. 33, No. 1, pp. 105-111, 2020.

[19] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT Acoustic-phonetic continuous speech corpus documentation," *in Technical Report NISTIR 4930, National Institute of Standards and Technology*, 1993.

[20] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in neurosciences,* Vol. 34, pp. 114-123, 2011.

[21] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Mechanisms of noise robust representation of speech in primary auditory cortex," in *Proceedings of the National Academy of Sciences,* Vol. 111, pp. 6792-6797, 2014.

[22] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *The Journal of the Acoustical Society of America,* Vol. 123, pp. 899-909, 2008.

[23] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 14, pp. 920-930, 2006.

[24] N. Mesgarani, J. Fritz, and S. Shamma, "A computational model of rapid task-related plasticity of auditory cortical receptive fields," *computational neuroscience,* Vol. 28, pp. 19-27, 2010.

[25] N. Esfandian, F. Razzazi, and A. Behrad, "A clustering based feature selection method in spectro-temporal domain for speech recognition," *Engineering Applications of Artificial Intelligence,* Vol. 25, pp. 1194-1202, 2012.

[26] I. Zulfiqar, M. Moerel, and E. Formisano, "Spectro-temporal processing in a two-stream computational model of auditory cortex," *Frontiers in computational neuroscience,* Vol. 13, pp. 1-18, January 2020.

[27] D. R. Ruggles, A. N. Tausend, S. A. Shamma, and A. J. Oxenham, "Cortical markers of auditory stream segregation revealed for streaming based on tonotopy but not pitch," *The Acoustical Society of America,* Vol. 144, pp. 2424-2433, 2018.

[28] F. Z. Yen, M. C. Huang, and T. S. Chi, "A two-stage singing voice separation algorithm using spectro-temporal modulation features," in *Sixteenth Annual Conference of the International Speech Communication Association (Interspeech),* Dresden, Germany, pp. 3321-3324, September 2015.

[29] K. Lu, W. Liu, P. Zan, S. V. David, J. B. Fritz, and S. A. Shamma, "Implicit memory for complex sounds in higher auditory cortex of the ferret," *Neuroscience,* Vol. 38, pp. 9955-9966, 2018.

[30] N. Esfandian, F. Razzazi, and A. Behrad, "A feature extraction method for speech recognition based on temporal tracking of clusters in spectro-temporal domain," in *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012),* Shiraz, Iran, pp. 012-017, May 2012.

[31] N. Esfandian, F. Razzazi, A. Behrad, and S. Valipour, "A Feature selection method in spectro-temporal domain based on Gaussian mixture models," in *IEEE 10th International Conference on Signal Processing (ICSP)*, Beijing, China, pp. 522-525, October 2010.

[32] A. Varga and H. J. Steeneken, "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,* Vol. 12, pp. 247-251, 1993.

[33] B. H. Prasetio, E. R. Widasari, and H. Tamura, "Automatic Multiscale-based Peak Detection on Short Time Energy and Spectral Centroid Feature Extraction for Conversational Speech Segmentation," in *6th International Conference on Sustainable Information Engineering and Technology,* Indonesia, pp. 44-49, September 2021.

[34] Z. Ali and M. Talha, "Innovative method for unsupervised voice activity detection and classification of audio segments," *IEEE Access,* Vol. 6, pp. 15494-15504, 2018.

تشخیص نواحی فعال گفتار با استفاده از روش مبتنی بر خوشه بندی در فضای طیفی- زمانی

نفیسه اسفندیان[۱*]، فاطمه جهانی بهنمیری[۲] و سمیرا مودّتی[۳]

[۱] گروه مهندسی برق، واحد قائمشهر، دانشگاه آزاد اسلامی، قائمشهر، ایران.

[۲] گروه مهندسی کامپیوتر، موسسه آموزش عالی آریان، بابل، ایران.

[۳] دانشکده مهندسی و فناوری، دانشگاه مازندران، بابلسر، ایران .

ارسال ۲۰۲۱/۱۲/۰۱؛ بازنگری ۲۰۲۲/۰۱/۰۶؛ پذیرش ۲۰۲۲/۰۲/۱۷

چکیده:

این مقاله، یک روش جدید برای تشخیص نواحی فعال گفتار بر مبنای خوشه بندی در فضای طیفی- زمانی ارائـه مـی‌دهـد. در الگـوریتم پیشـنهادی، از مدل شنیداری برای استخراج ویژگی‌های طیفی- زمانی استفاده می‌شود. روش های خوشه بندی مدل مخلوط گوسی و K میانگین وزن دار برای کـاهش ابعاد فضای طیفی-زمانی بکار گرفته می‌شود. از انرژی و موقعیت مکانی خوشه ها برای تشخیص نواحی فعال گفتار استفاده می شود. بخش های سـکوت و گفتار با استفاده از ویژگی‌های خوشه ها و مقدار آستانه به روز رسانی شده در هر قاب تشخیص داده می شود. به دلیل بالا بـودن انـرژی خوشـه اول، از خوشه اول به عنوان بخش اصلی گفتار در محاسبات استفاده می‌گردد. کارایی روش پیشنهادی برای جدا کردن بخـش‌هـای سـکوت از گفتـار در شـرایط مختلف نویزی ارزیابی شد. میزان جا به جایی خوشه‌ها در فضای طیفی- زمانی به عنوان معیاری برای تعیین مقاوم به نویز بودن ویژگی‌ها در نظـر گرفتـه شد. با توجه به نتایج، نرخ بخش بندی نواحی سکوت از گفتار با استفاده از روش پیشنهادی در مقایسه با ویژگی‌هـای حـوزه زمـان و فرکـانس در نسـبت سیگنال به نویزهای پایین بهبود یافته است.

کلمات کلیدی: ویژگی‌های طیفی- زمانی، مدل شنیداری، مدل مخلوط گوسی، خوشه‌بندی K میانگین وزن دار، تشخیص نواحی فعال گفتار.