FUZZY OPTIMIZATION AND MODELLING 3(2) (2022) 1-14

E-ISNN: 2676-7007

Contents lists available at FOMJ

Fuzzy Optimization and Modelling

Journal homepage: http://fomj.qaemiau.ac.ir/

Paper Type: Research Paper

Voiced-Unvoiced-Silence Detection of Speech Signal Using Combined Spectro-Temporal Features

Nafiseh Esfandian*

Department of Electrical Engineering, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran

ARTICLE INFO

Article history:
Received 14 June 2022
Revised 31 July 2022
Accepted 4 September 2022
Available online 28 October 2022

Keywords:
Weighted Gaussian Mixture Model
Clustering
Speech Segmentation
Spectro-temporal Features

ABSTRACT

This paper presents a new method for classification of voiced, unvoiced and silence segments of speech signal. In the proposed method, combination of spectro-temporal features is used for speech segmentation. Combined features are extracted using clustering in spectro-temporal domain. Multi-dimensional output of auditory model is clustered using weighted Gaussian mixture model. In this method, after extracting the main clusters for each frame, combined spectro-temporal features such as cluster's energy, energy difference of clusters and minimum value of normalized cross-correlation between clusters are used for detection of voiced, unvoiced and silence regions of speech. In the proposed algorithm, speech segmentation is performed by comparing each class of features with the appropriate threshold value. Combined spectro-temporal features are used for speech segmentation in noisy conditions. The results demonstrate performance of the proposed algorithm comparing to the other features for speech segmentation.

1. Introduction

Speech classification into voiced, unvoiced and silence (V/U/S) segments are used in many applications of speech processing [15]. Voiced, unvoiced and silence parts of speech signals are segmented by extracting various features [3, 11]. Most common features used for V/U/S detection are zero crossing rate and energy [4, 20]. Alimuradov [2] proposed the new method based on analysis of speech signal segments using the Teager energy operator and analysis of zero-crossing rate and short-term energy of the energy characteristic function. Jalil et al [13] proposed speech segmentation using short time energy, short time magnitude, zero crossing and autocorrelation function. In the speech classification algorithm was proposed by Zaw and War [33], spectral entropy and short time features such as zero crossing rate, short time energy, and linear prediction error were used. In this method, the threshold values were obtained empirically. In the recent study [7], local adaptive thresholding technique was used in for the speech segmentation. The some studies used Mel-frequency cepstral coefficients (MFCC) [29] and autocorrelation function peak [26].

* Correspondig author
E-mail address: na_esfandian@yahoo.com (Nafiseh esfandian)

DOI: 10.30495/fomj.2022.1961029.1071

In another studies, the combination of spectral and temporal features have been used for speech segmentation techniques [21]. Pitch and MFCC were used along with zero crossing and energy values for speech classification by Das et al [6]. Qi and Hunt [23] proposed speech segmentation using a multilayer feed-forward network. In this method, the feature vector was a combination of MFCC features and two time domain features, the zero-crossing rate and a nonlinear function of root-mean-square energy. In the study by Alimi and Awodele [1], time-domain features (zero-crossing, standard deviation, normalized envelope, and root-mean-square energy.) and frequency-domain features (Mel-frequency cepstral coefficients) were combined to segment the speech signal. Mondal and Barman [19] proposed automatic segmentation of V/U/S parts of speech based on waveform and frequency domain attributes. In this work, Spectral clustering using Gaussian similarity function was used to separate voiced and unvoiced parts and the energy based threshold of unvoiced sections was used to detect silence frames. Radmard et al. [24] used multi-features such as cepstral peak, zero crossing rate, and autocorrelation function (ACF) peak of short-time segments of the speech signal. These studies used conventional temporal and spectral features to improve V/U/S segmentation accuracy and clustering- based method or neural network were used to combine these features. How to combine the features and set the threshold value is the main challenges of these methods. The main goal of this paper is to extract spectro-temporal features and to compare with conventional features for speech classification. Unlike conventional features, proposed features include both time and frequency attributes of speech signal and can improve accuracy of V/U/S detection.

In the proposed technique, auditory model is used to extract spectro-temporal features. The auditory model is one of the successful models for acoustic representation of speech signal [18]. This model has been inspired according to neurological, biological findings in the human auditory system [16]. Auditory model has been simulated according to early and central stages of the auditory system of brain. In this paper, combination of clustering- based features in spectro-temporal domain is used for V/U/S separation [8]. The dimension of the spectro-temporal feature space is very large. Therefore cortical output was clustered using weighted Gaussian mixture model (WGMM). Energy of speech signal is a measure to detect voiced, unvoiced and silence region. Voiced region has higher energy in comparison to unvoiced and silence parts of speech signal. The contributions of this paper are presented as follows:

- In the proposed method, the first cluster's energy is used as main feature to classify speech signal. Also, energy difference of clusters and minimum value of normalized cross-correlation between clusters are used for V/U/S detection.
- The extracted features are combined to recognize each frame class (V/U/S) using empirical threshold. In the proposed method, different threshold values are used for each class of features. The proposed features are used for segmentation of the various sentences from the TIMIT database in noisy conditions.

The organization of this paper is as follows: The auditory model is briefly described in Section 2. Clustering of spectro-temporal features space using WGMM and extraction of combined features for speech segmentation are discussed in Section 3. The simulations' results and comparisons are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Spectro-temporal Features Extraction Using Auditory Model

In this section, auditory model is described. This model is used in many applications of speech processing in recent years. This computational model is simulated based on human auditory system. Auditory model consists of two main parts. First part of model simulates internal ear [32].

In the early stage of auditory model, one dimensional acoustic signal is converted into two-dimensional internal neural representation [34]. This stage is simulated by a bank of band-pass filters that are distributed along a logarithmic frequency axis [14]. This axis is called tonotopic axis. The filter bank includes 128 filters with the impulse response $h_{cochlea}(t;f)$ which are uniformly distributed along the tonotopic axis. The outputs of cochlear filter $y_{cochlea}(t,f)$ are converted into auditory nerve patterns $y_{an}(t,f)$ by an inner hair cell stage (IHC).

IHC stage includes of a high-pass filter, an instantaneous nonlinear compression $g_{hc}(.)$ and a time domain low-pass filter $\mu_{hc}(t)$ .In the last section, lateral inhibitory network (LIN) activity is simulated. In this part of the model, the frequency selectivity of the cochlear filters is increased. The output of LIN is calculated using a first order derivative during the tonotopic axis and $y_{LIN}(t,f)$ is obtained by using a half wave rectifier to remove the negative outputs and the final output is computed by integrating $y_{LIN}(t,f)$ along a short time window. The early stage of auditory model is shown in Figure 1. The output of the early stage was calculated by using an infinite impulse response (IIR) filter bank including 128 frequency channels between 180 and 7246 Hz at the resolution of 24 channels per octave. The time constant of 8ms was applied for the leaky time integration and filter-bank outputs were sampled every 8 ms to estimate the auditory spectrogram.

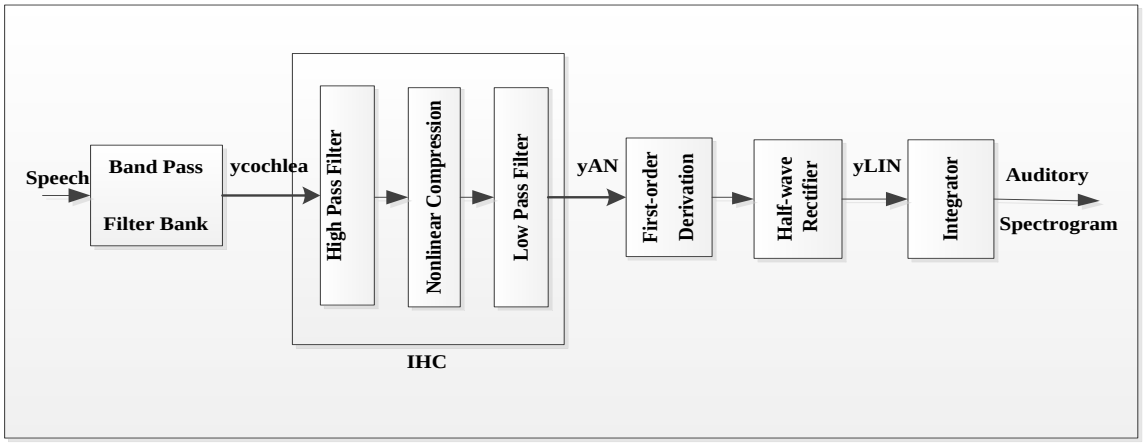

Figure 1. The early stage of auditory model

This 2D image is called auditory spectrogram. The next stage (the primary auditory cortex) is modelled by spectro-temporal filter bank that each filer tuned to a range of different rates and scales. Cortical filter bank performs two dimensional wavelet transform of auditory spectrogram. At the cortical stage, spectro- temporal modulation contents of auditory spectrogram are calculated. The cortical output is computed by convolving the spectro-temporal receptive field of filters with the auditory spectrogram. Applying two dimensional spectro-temporal receptive field (STRF) filters on the spectrogram is shown in Figure 2. There are two types of STRF which are called upward (negative rates) and downward (positive rate). The cortical output includes the four dimensions of scale, rate, frequency and time. There are many researches that primary auditory cortex determines multidimensional representation that can be used to discriminate voiced and unvoiced phonemes [17]. Temporal parameter of the filters (rate) was arranged from 2 to 32 Hz and spectral parameter of the filters (scale) was considered from 0.25 to 8 cycle/octave to estimate the spectro-temporal features.

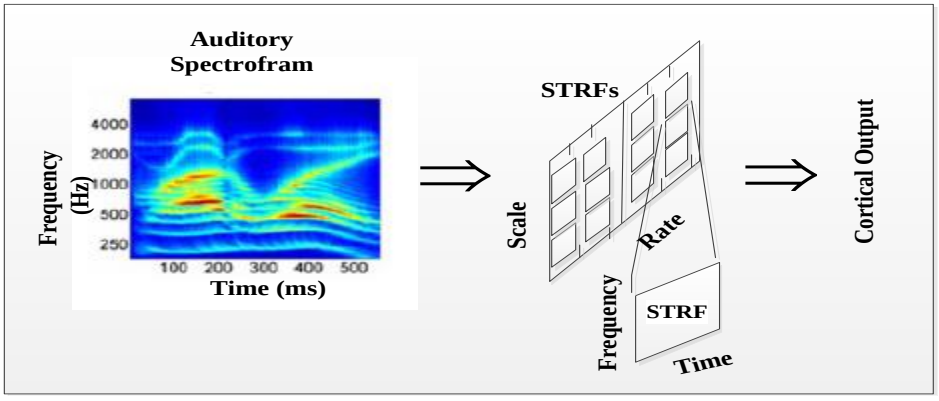

Figure 2. Cortical Stage of auditory model

3. Clustering of Cortical Output Using Weighted Gaussian Mixture Model

In this section, spectro- temporal features extraction using clustering method is described. One of the limitations of auditory model is high-dimensional features space. Therefore, feature selection is extremely important in the spectro-temporal domain. For this purpose, the output of auditory model was clustered using weighted Gaussian mixture model in the proposed algorithm. Gaussian mixture model (GMM) is one of the most popular clustering methods. In this method, a feature space is clustered using the mixture of multiple Gaussian distributions and each cluster is modeled as a Gaussian distribution [9].

The block diagram of proposed method is shown in Figure 3. In the first step of proposed model, auditory spectrogram of speech signal is calculated and then, cortical representation of speech is obtained. In the next step, weighted Gaussian mixture model (WGMM) is used for clustering of spectro-temporal features space to select valuable discriminative features. In WGMM clustering method, the spatial information of the points is considered as primary features vectors $v_i = (r_i, s_i, f_i)$. The amplitude of points is considered as weight vector, $w_i = |A_i|$, in the clustering algorithm. In the primary features vectors, r is the rate, s denotes the scale and f denotes the frequency of each point in spectro-temporal domain. In this method, three clusters are used for clustering of features space. Finally, secondary features are extracted and theses combined features are used for voiced, unvoiced and silence detection.

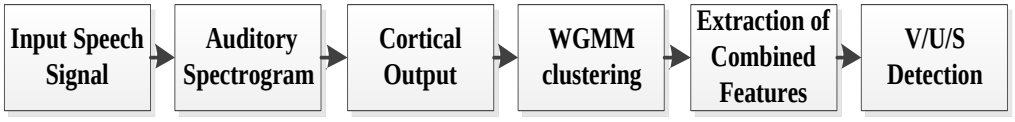

Figure 3. Block diagram of proposed method

3. 1. Extraction of combined features for speech segmentation

The main goal of this study is to combine clustering- based features in spectro-temporal domain. Three types of features were used for speech segmentation that described in this section. In the proposed algorithm, first cluster contains the highest energy in spectro-temporal space and the lowest amount of energy belongs to the third cluster. In this paper, the first cluster's energy is used as main attribute for V/U/S separation. Because, the energy of voiced part is higher than unvoiced part of the speech signal and the energy of silence region is lower than the other parts of speech. The average energy of each cluster is defined as [27, 31]:

$$E_C(i) = \frac{1}{N}\sum_{j=1}^{N}\left|A_j\right|^2 \quad (1)$$

where N is the number of samples belonging to *i*th cluster, $\left|A_j\right|$ denotes the amplitude of *j*th sample belonging

to *i*th cluster. Two threshold values are used to compare energy of the first cluster in each frame.

$$T_{E1}=\frac{1}{M}\sum_{i=1}^{M}E_{C1}(i) \quad (2)$$

$$T_{E2}=\alpha T_{E1} \qquad 0<\alpha<1 \quad (3)$$

where M is the number of speech frames and E_{C1} represents the first cluster's energy in each frame. The optimum value of α is empirically determined using a grid search strategy to optimize the accuracy rate of V/U/S detection. The best results was obtained with α =0.5. If the first cluster's energy in each frame was greater than T_{E1}, that frame was classified as voiced. If the first cluster's energy was between two values T_{E1} and T_{E2}, the frame included unvoiced and if it was lower then T_{E2}, that frame was classified as silence. The first cluster's energy of the sentences spoken by female and male speakers from TIMIT database is shown in Figure 4 (a) and (b). As it could be observed, the energy of voiced region is greater than unvoiced region and the energy of silence region is less than the other parts. The determination of the threshold was based on trial and error in this study, automatic estimation of optimum threshold was remained as an open problem for future studies.

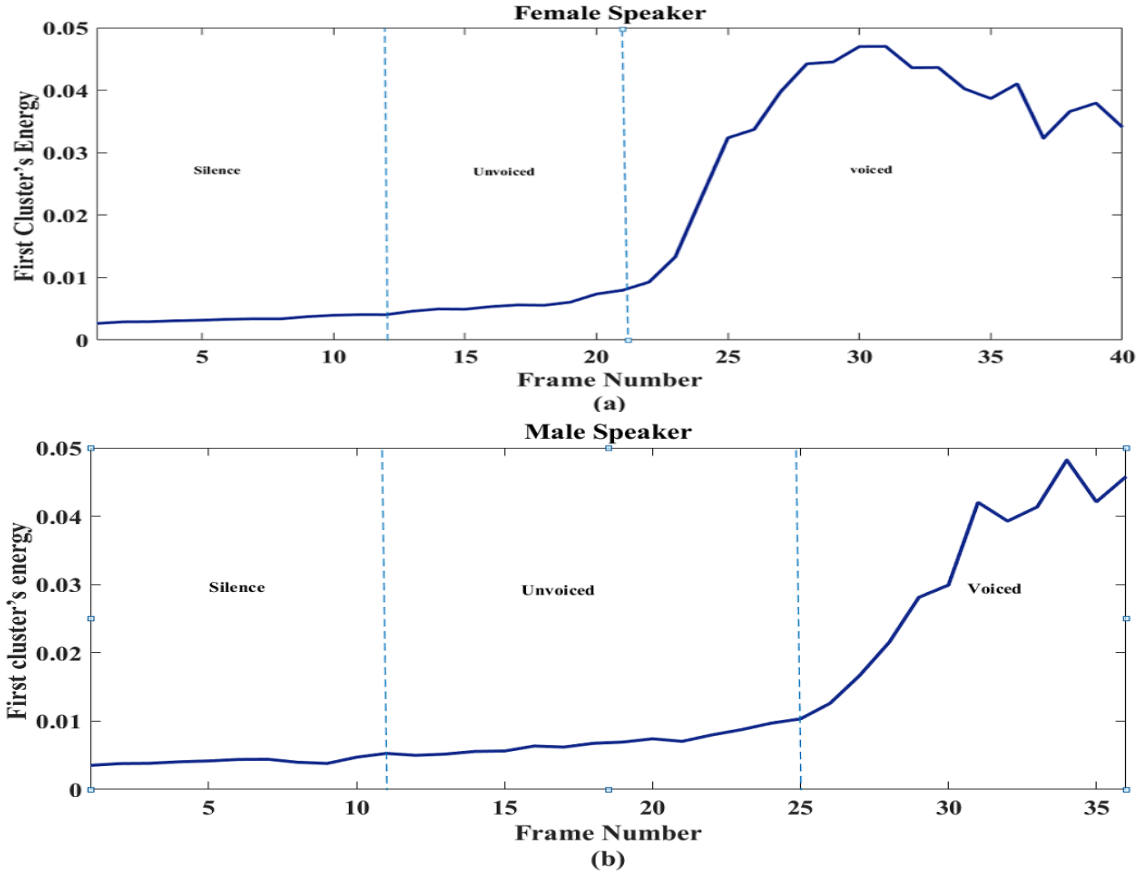

Figure 4. The first cluster's energy of speech signal

3. 2. The energy difference between the first and third clusters

In the proposed algorithm, in addition to the first cluster's energy (E_{C1}), the other features such as the energy difference between the first and third clusters and minimum value of normalizes cross-correlation (NCC)

function between the clusters over the consecutive frames were considered for V/U/S detection. The behavior of the speech signal in the spectral-temporal domain was well described by using these features. The first cluster has the highest energy and the third cluster has the lowest energy in the spectro-temporal domain. In addition, the energy of voiced segment of speech signal is greater than unvoiced and silence segments. Therefore, the energy difference between the first and third clusters in voiced parts of speech is greater than the other parts. Unvoiced segment of speech signal has the non-periodic and noise-like nature and the energy difference between the first and third clusters in unvoiced and silence regions is less than voiced segments. The energy difference between the first and third clusters (D_{13}) is calculated as:

$$D_{13} = E_{C1} - E_{C3} \tag{4}$$

In the above equation, E_{C1} and E_{C3} are the energy of first and third clusters respectively. The energy difference between the first and third clusters of speech signal spoken by female and male speakers is shown in Figure 5 (a) and (b). It can be observed, the energy difference between the first and third clusters (D_{13}) in the voiced region is greater than the other regions. Therefore, in the proposed method, this measure is used for separation of voiced regions from unvoiced and silence regions. The threshold value on the energy difference is defined as:

$$T_D = \frac{1}{M}\sum_{i=1}^{M} D_{13}(i) \tag{5}$$

where, M is the number of frames. If the energy difference between the first and third clusters was higher than threshold value (T_D), this frame was classified as voiced; otherwise, the frame contains unvoiced or silence segments of speech signal.

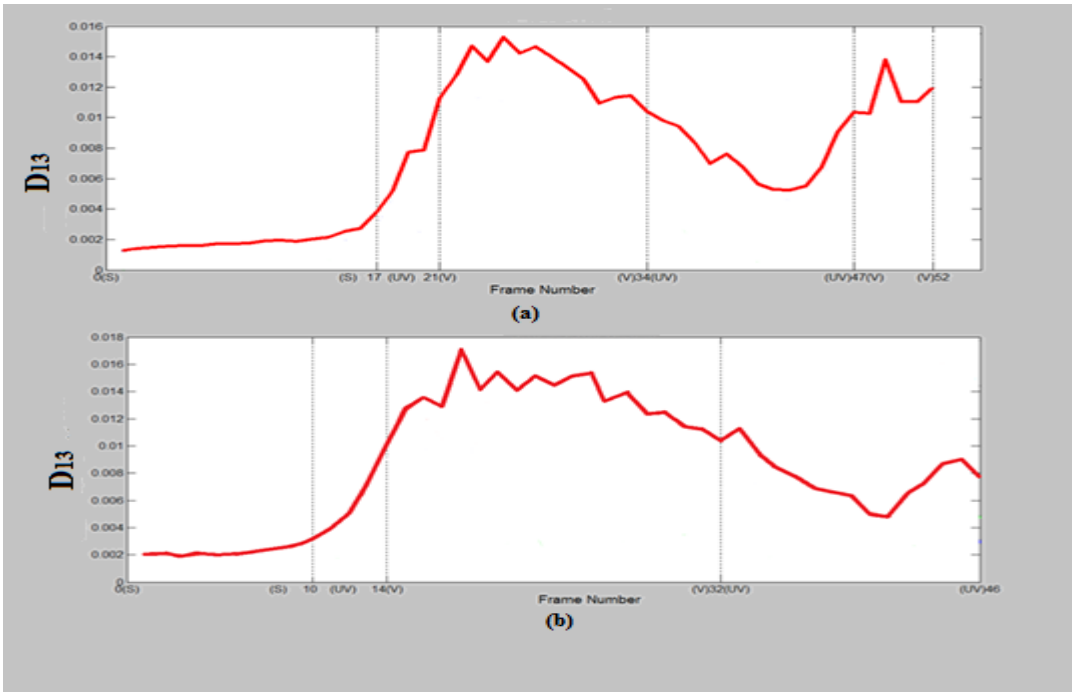

Figure 5. The energy difference between the first and third clusters

3. 3. Normalized cross-correlation between the clusters over the consecutive frames

The energy difference between the first and third clusters (D_{13}) is used as a measure to recognize voiced segments of speech signal. In the proposed algorithm, the other feature was used to distinguish between

unvoiced and silence. For this purpose, normalized cross-correlation (NCC) was used as a measure to evaluate degree of similarity between the clusters over the consecutive frames. The spatial information of points (the mean and variance vectors of cluster centers) in the feature space was used to calculate normalized cross-correlation. The mean and variance vectors of cluster centers were sorted using energy measure. Normalized cross-correlation of the clusters is defined as [15]:

$$NCC(i,j)=\frac{\sum_{n=1}^{K}C_i(n)C_j(n)}{\sum_{n=1}^{K}\sqrt{|C_i)n)|^2}\sum_{n=1}^{K}\sqrt{|C_j)n)|^2}} \tag{6}$$

where $NCC(i,j)$ is normalized cross correlation between cluster C_i in the current frame and C_j in the next frame. $C_i=(\mu_i,\sigma_i)$ includes the mean vector, $\mu_i=(\mu_{ri},\mu_{si},\mu_{fi},\mu_{|A_i|})$, and variance vector of ith cluster, $\sigma_i=(\sigma_{ri},\sigma_{si},\sigma_{fi},\sigma_{|A_i|})$. The secondary features vectors contain 8 attributes ($K=8$). NCC is a 3×3 matrix defined as:

$$NCC_{n,n+1}=\begin{pmatrix} C_1^nC_1^{n+1} & C_1^nC_2^{n+1} & C_1^nC_3^{n+1} \\ C_2^nC_1^{n+1} & C_2^nC_2^{n+1} & C_2^nC_3^{n+1} \\ C_3^nC_1^{n+1} & C_3^nC_2^{n+1} & C_3^nC_3^{n+1} \end{pmatrix} \tag{7}$$

where $NCC_{n,n+1}$ is normalized cross correlation between 3 clusters of nth frame and 3 clusters of n+1th frame. Due to the noise-like nature of unvoiced region of speech, the minimum similarity between the clusters of these segments is greater than the other segments of speech. Therefore, the minimum value of NCC matrix was considered as a measure to recognize unvoiced region from the other parts of speech. The minimum value of NCC matrix of speech signal spoken by female and male speakers is shown in Figure 6 (a) and (b). This Figure depicts that $\min NCC$ in unvoiced region is greater than voiced and silence region of speech signal.

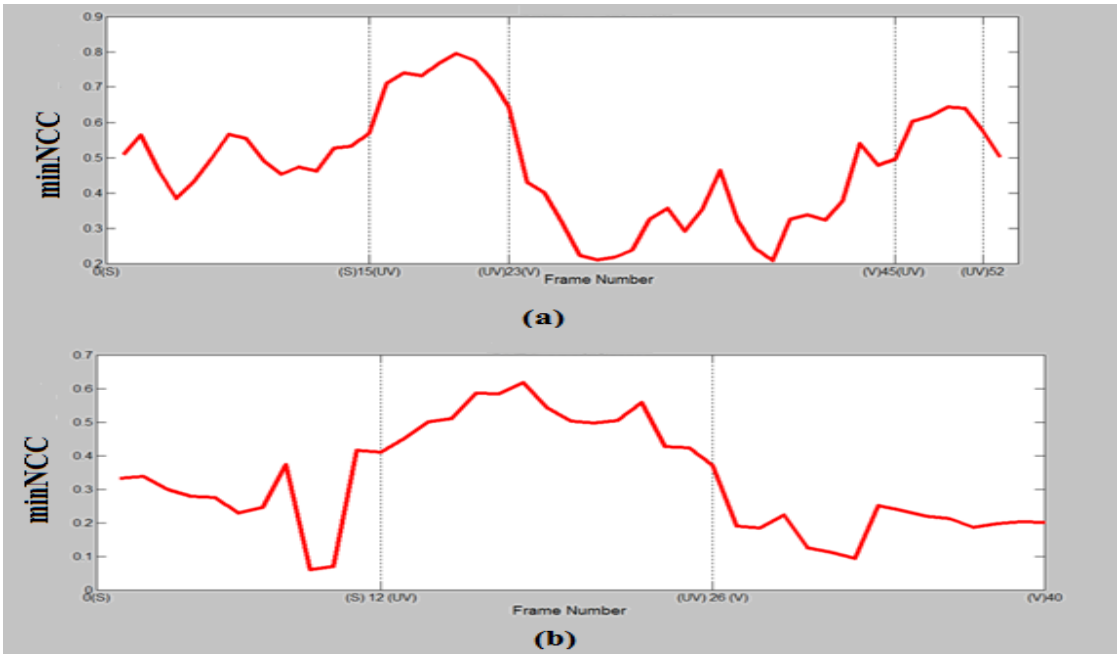

Figure 6. The minimum value of NCC matrix

The minimum value of NCC matrix was compared with threshold value to detect unvoiced region of speech. $\min NCC$ of unvoiced frames is higher than threshold value (T_N). Threshold value (T_N) is calculated from

equation (8). In this equation, M is the number of speech frames. The flow chart of speech classification using proposed features is shown in Figure (7). The proposed algorithm was implemented in MATLAB 2021b.

$$T_N = \frac{1}{M}\sum_{i=1}^{M} \min NCC(i) \tag{8}$$

4. Simulations Results

In this section, simulation results and discuss about efficiency of proposed features are described. The evaluation results were obtained using 200 sentences; female and male speakers were randomly selected from the TIMIT database of each eight dialects [10]. 3 different types of noises (White, car and babble) from the NOISEX database [30] were added to the clean speech signal taken from TIMIT dataset. The noisy speech signal was windowed using a window of 16 ms and the cortical output of auditory model was clustered using WGMM. Then the first cluster's energy (E_{C1}), the energy difference between the first and third clusters (D_{13}) and minimum value of normalizes cross-correlation ($\min NCC$) were extracted as secondary features. The class of speech signal (V/U/S) in each frame was determined using the combined features in comparison to threshold values. TIMIT database is manually annotated at the phone level. The label of each frame (V/U/S) was recognized by proposed algorithm and compared with the TIMIT manual labels for the same utterance. The accuracy of the proposed system is evaluated using accuracy measure defined as:

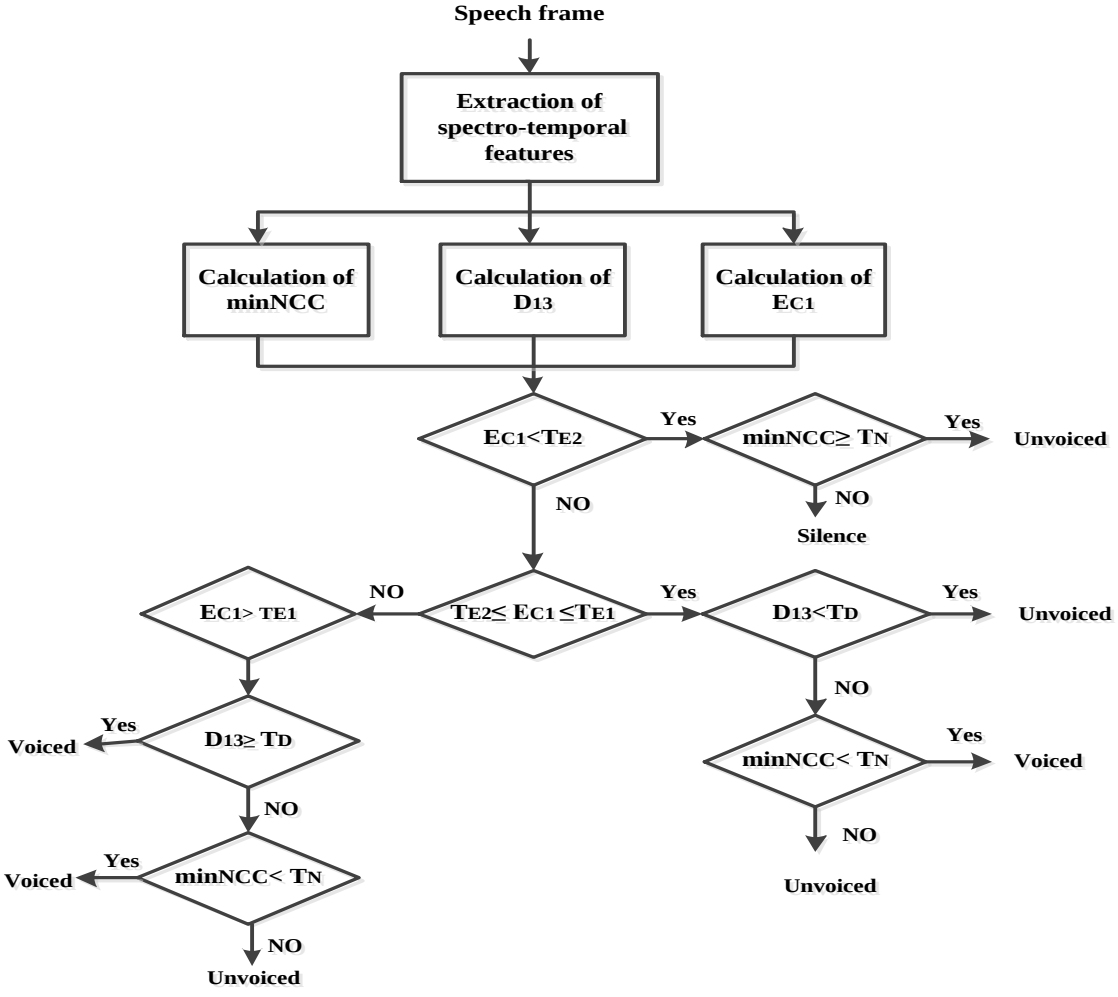

Figure 7. The flow chart of speech classification using proposed features

$$P_{S\,2S} = \frac{N_{S\,2S}}{N_S} \times 100 \quad (9)$$

$$P_{V\,2V} = \frac{N_{V\,2V}}{N_V} \times 100 \quad (10)$$

$$P_{U\,2U} = \frac{N_{U\,2U}}{N_U} \times 100 \quad (11)$$

$$P_{T\,2T} = \frac{N_{T\,2T}}{N_T} \times 100 \quad (12)$$

where, $P_{S\,2S}$, $P_{V\,2V}$ and $P_{U\,2U}$ respectively are the accuracy rate of silence, voiced and unvoiced classification and $P_{T\,2T}$ is the total accuracy rate of speech classification. In these equation, $N_{S\,2S}$, $N_{V\,2V}$ and $N_{U\,2U}$ respectively are the number of silence, voiced and unvoiced correctly recognized frames. $N_{T\,2T}$ is the total number of correctly classified frames. Also, N_S , N_V and N_U respectively are the number of silence, voiced and unvoiced frames and N_T is the total number of speech frames. Table 1 presents the mean energy of the clusters in each part of speech signal (V/U/S). As it can be observed, the mean energy of first cluster in voiced region is higher than the other parts of speech signal. The lowest energy belongs to silence region of speech signal.

Table 1. The mean energy of the clusters in each part of speech signal

Class	First cluster	Second cluster	Third cluster
Voiced	0.4	0.3	0.1
Unvoiced	0.03	0.01	0.008
Silence	0.002	0.001	0.0005

The results of speech classification using white noise with SNR of 10dB are presented in Table 2. In this Table, the accuracy rate of speech classification using proposed features was compared to the conventional features such as short time energy (STE), fundamental frequency (F_0) and zero-crossing rate (ZCR) [13, 15]. The results present that the accuracy rate of V/U/S segmentation is improved using proposed features.

Table 2. Comparison of proposed features with conventional features

Accuracy rate	STE, F_0 and ZCR	Proposed features
$P_{V\,2V}$	94.5	98.9
$P_{U\,2U}$	91.9	97.8
$P_{S\,2S}$	93.6	98.2
$P_{T\,2T}$	93.3	98.3

In Table 3, speech segmentation results of proposed features were compared to combination of conventional features such as energy, zero-crossing rate, Mel Frequency Cepstral Coefficients (MFCC) and Pitch [5, 25]. In this Table, the accuracy rate of V/U/S detection was obtained using white noise with SNR of 5dB. As it can be observed, the better results were obtained using proposed spectro-temporal features in comparison to conventional features.

Table 3. Accuracy rate of proposed features in comparison to conventional features

Accuracy rate	Conventional features	Proposed features
$P_{V\,2V}$	91.2	96.8
$P_{U\,2U}$	89.5	94.1
$P_{S\,2S}$	90.6	95.9
$P_{T\,2T}$	90.4	95.6

V/U/S detection results for speech including white, car and babble noises at different signal-to-noise ratios (SNRs) are shown in Figures 8, 9 and 10. In these Figures, proposed algorithm was compared to speech segmentation based on the autocorrelation function (ACF), cepstrum, zero-crossing rate and energy of the signal

[12, 28]. The results demonstrate that the accuracy rate of speech segmentation are improved using the proposed features especially at low signal-to-noise ratios (SNRs).

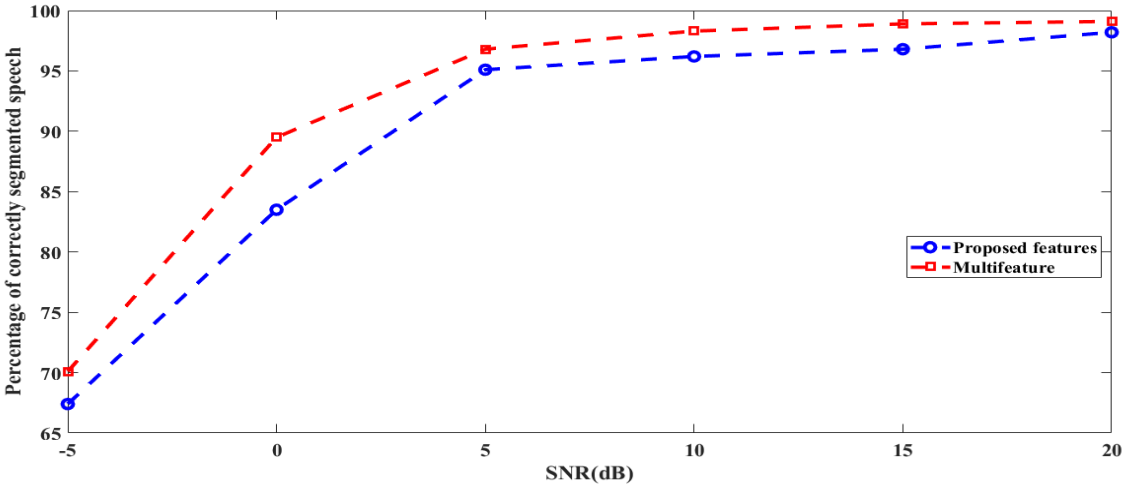

Figure 8. Accuracy rate of proposed features in comparison to multifeature for speech with white noise

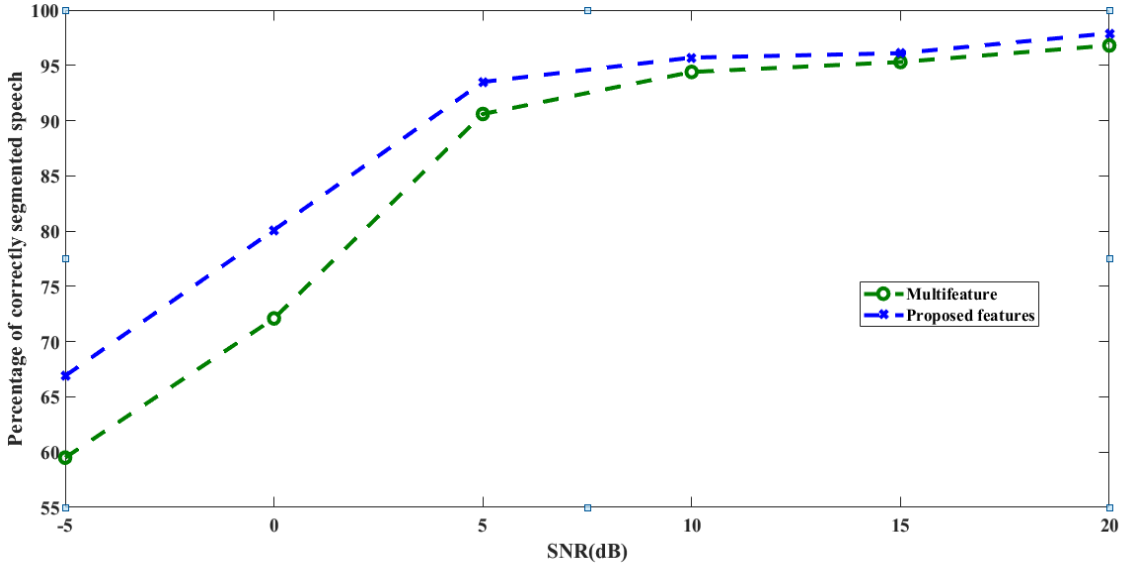

Figure 9. Accuracy rate of proposed features in comparison to multifeature for speech with car noise

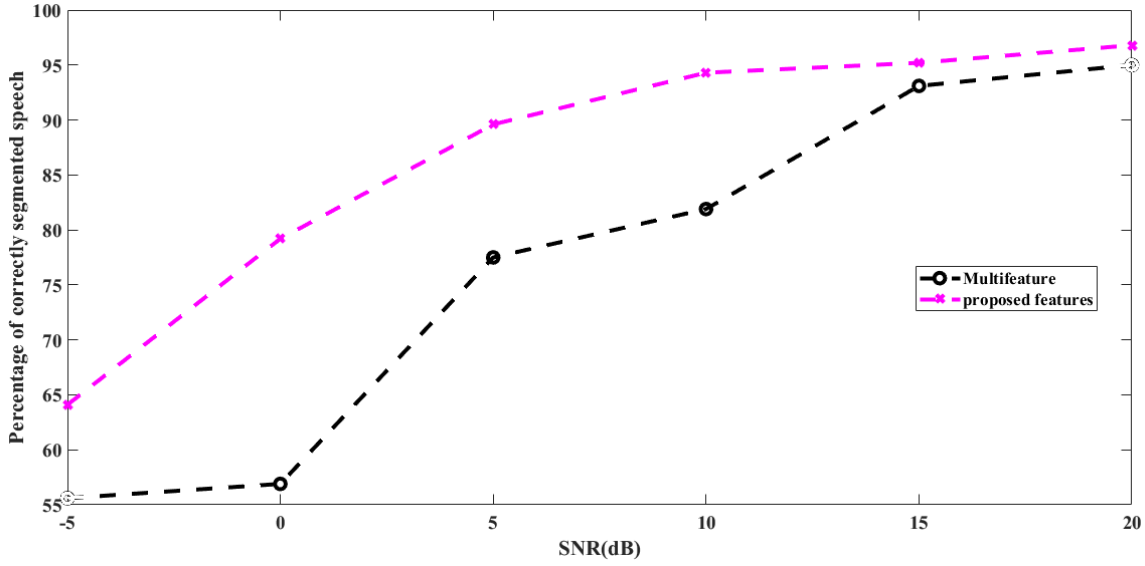

Figure 10. Accuracy rate of proposed features in comparison to multifeature for speech with babble noise

The frame error rate (FER) of voiced and unvoiced detection was shown in Figure 11. In this Figure, the proposed algorithm was compared to existing methods [19, 22]. Qi et al. [22] proposed the method for V/U/S classification using the support vector machine (SVM). In this method, Voiced and unvoiced parts of speech was classified by using four characteristic parameters, including the full-band energy, the low-band energy, the period level parameter and the zero-crossing rate. In the other method [19], time and frequency domain features were used speech segmentation. Short time energy of unvoiced segments was compared with threshold value to separate silence from speech. The spectral clustering was used to detect voiced and unvoiced parts of the speech signal. The results demonstrate frame error rate was improved using the proposed algorithm. In Figure 12, the proposed features were compared with existing method presented by Prasetio et al. [21]. In this method, the short time energy and spectral centroid were proposed as the features and the automatic multi-scale algorithm as the signal peak detection. As it could be observed, the frame error rate was decreased using the proposed features.

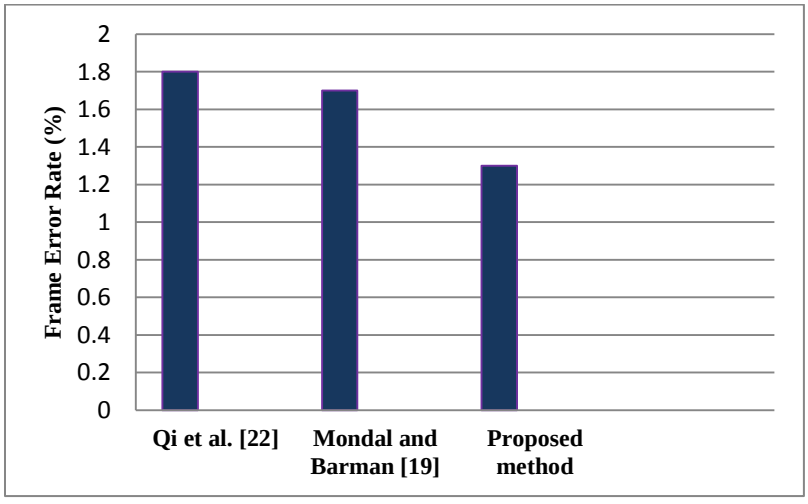

Figure 11. Frame error rate of proposed method in comparison to existing method [19, 22]

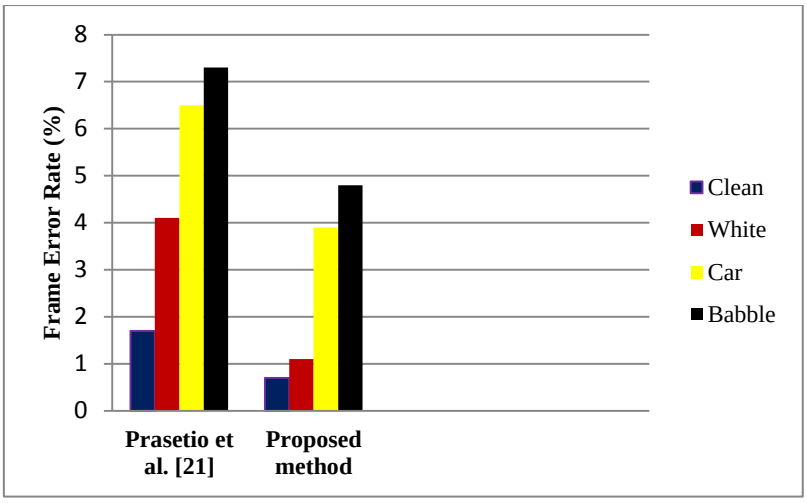

Figure 12. Frame error rate of proposed method in comparison to existing method [21]

5. Conclusion

This paper presented a new algorithm for speech segmentation based on spectro-temporal features. In the proposed algorithm, spectro-temporal features space was clustered using WGMM. The first cluster's energy, the energy difference between the first and third clusters and minimum value of normalizes cross-correlation function between the clusters were considered for V/U/S classification. The V/U/S segmentation results presented the performance of proposed algorithm in comparison to existing method especially at low SNRs. In this paper, empirical threshold value was used for each frame of speech signal. However the performance measure of V/U/S detection depends on empirical threshold value. In the future research, the threshold value can be updated in each frame of noisy speech. Also, the deep learning-based methods will be used to cluster in spectro-temporal features space.

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Alimi, S., & Awodele, O. (2022). Voice Activity Detection: Fusion of Time and Frequency Domain Features with A SVM Classifier. *Computer Engineering and Intelligent Systems*, 13(3),20-29.
2. Alimuradov, A. K. (2021). Enhancement of Speech Signal Segmentation Using Teager Energy Operator. *2021 23rd International Conference on Digital Signal Processing and its Applications (DSPA)*, 1-7.
3. Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) zone conference proceedings*, 1-7.
4. Caruntu, A., Toderean, G., & Nica, A. (2005). Automatic silence/unvoiced/voiced classification of speech using a modified Teager energy feature. *WSEAS international conference on Dynamical systems and control*, 62-65.
5. Chebbi, S., & Jebara, S. B. (2018). On the use of pitch-based features for fear emotion detection from speech. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1-6.
6. Das, B. K., Das, A., & Bhattacharjee, U. (2014). Detection of voiced, unvoiced and silence regions of assamese speech by using acoustic features. *Int J Comput Trends Technol*, *14*(2), 43.
7. Endah, S. N., Kusumaningrum, R., Adhy, S., & Ulfattah, R. A. (2021). Automatic speech recognition by using local adaptive thresholding in continuous speech segmentation. *Journal of Physics: Conference Series*, 1943(1), 1-8.

8. Esfandian, N., & Hosseinpour, K. (2021). A Clustering-Based Approach for Features Extraction in Spectro-Temporal Domain Using Artificial Neural Network. *International Journal of Engineering*, *34*(2), 452-457.
9. Esfandian, N., Razzazi, F., & Behrad, A. (2012). A clustering based feature selection method in spectro-temporal domain for speech recognition. *Engineering Applications of Artificial Intelligence*, *25*(6), 1194-1202.
10. Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. *in Technical Report NISTIR 4930, National Institute of Standards and Technology*.
11. Graf, S., Herbig, T., Buck, M., & Schmidt, G. (2015). Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, *2015*(1), 1-15.
12. Gupta, P., & Sengupta, S. (2018). Voiced/Unvoiced Decision with a Comparative Study of Two Pitch Detection Techniques. *International Research Journal of Engineering and Technology (IRJET)*,5(7),2223-2229.
13. Jalil, M., Butt, F. A., & Malik, A. (2013). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. *2013 The international conference on technological advances in electrical, electronics and computer engineering (TAEECE)*, 208-212.
14. Lu, K., Liu, W., Zan, P., David, S. V., Fritz, J. B., & Shamma, S. A. (2018). Implicit memory for complex sounds in higher auditory cortex of the ferret. *Journal of Neuroscience*, *38*(46), 9955-9966.
15. Mehrotra, T., Shukla, N., Chaudhary, T., Rajput, G. K., Altuwairiqi, M., & Asif Shah, M. (2022). Improved Frame-Wise Segmentation of Audio Signals for Smart Hearing Aid Using Particle Swarm Optimization-Based Clustering. *Mathematical Problems in Engineering*, 2022, 1-9.
16. Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences*, *111*(18), 6792-6797.
17. Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899-909.
18. Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(3), 920-930.
19. Mondal, S., & Barman, A. D. (2015). Clustering based voiced-unvoiced-silence detection in speech using temporal and spectral parameters. *2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. 390-394).
20. Nandhini, S., & Shenbagavalli, A. (2014). Voiced/unvoiced detection using short term processing. *International Journal of Computer Applications*, *975*, 39-43.
21. Prasetio, B. H., Widasari, E. R., & Tamura, H. (2021). Automatic Multiscale-based Peak Detection on Short Time Energy and Spectral Centroid Feature Extraction for Conversational Speech Segmentation. *6th International Conference on Sustainable Information Engineering and Technology 2021*, 44-49.
22. Qi, F., Bao, C., & Liu, Y. (2004). A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech. In *2004 international symposium on Chinese spoken language processing*, pp. 77-80.
23. Qi, Y., & Hunt, B. R. (1993). Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE Transactions on speech and audio processing*, 1(2), 250-255.
24. Radmard, M., Hadavi, M., & Nayebi, M. M. (2011). A new method of voiced/unvoiced classification based on clustering. *Journal of Signal and Information Processing*, *2*(04), 336-347.
25. Ruggles, D. R., Tausend, A. N., Shamma, S. A., & Oxenham, A. J. (2018). Cortical markers of auditory stream segregation revealed for streaming based on tonotopy but not pitch. *The Journal of the Acoustical Society of America*, *144*(4), 2424-2433.
26. Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, 107020.
27. Sharma, P., & Rajpoot, A. K. (2013). Automatic identification of silence, unvoiced and voiced chunks in speech. *Journal of Computer Science & Information Technology (CS & IT)*, *3*(5), 87-96.
28. Sharma, S., Sharma, A., Malhotra, R., & Rattan, P. (2021). Voice Activity Detection using windowing and updated K-Means Clustering Algorithm. *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*. 114-118.
29. Tan, Z. H., & Dehak, N. (2020). rVAD: An unsupervised segment-based robust voice activity detection method. *Computer speech & language*, *59*, 1-21.

30. Varga, A., & Steeneken, H. J. (1993). 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*, *12*(3), 247-251.
31. Verma, A., Jana, S., Ravela, R. R., Bansal, H., & Nune, K. K. (2021). Performance Comparison of Soft Computing Algorithms for Voiced and Unvoiced Speech Detection. *Advances in Systems Engineering*, 437-445.
32. Yen, F. Z., Huang, M. C., & Chi, T. S. (2015). A two-stage singing voice separation algorithm using spectro-temporal modulation features. In *Sixteenth Annual Conference of the International Speech Communication Association*, 3321-3324.
33. Zaw, T. H., & War, N. (2017). The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection. *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 1-5.
34. Zulfiqar, I., Moerel, M., & Formisano, E. (2020). Spectro-temporal processing in a two-stream computational model of auditory cortex. *Frontiers in computational neuroscience*, *13*, 1-18.

Esfandian, N. (2022). Bi-Level Portfolio Optimization Considering Fundamental Analysis in Fuzzy Uncertainty Environments. *Fuzzy Optimization and Modelling Journal*, 3(2), 1-14.

https://doi.org/ 10.30495/FOMJ.2022.1961029.1071

Received: 14 June 2022 Revised: 31 July 2022 Accepted: 4 September 2022