Journal of Engineering Science and Technology
Vol. 15, No. 5 (2020) 2854 - 2870

AUTOMATIC VOICE ACTIVITY DETECTION USING FUZZY-NEURO CLASSIFIER

SUHAILA N. MOHAMMED[1, 2, *], ALIA K. HASSAN[1]

[1]Department of Computer Science, University of Technology, Baghdad, Iraq
[2]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq
*Corresponding Author: suhailan.mo@sc.uobaghdad.edu.iq

Abstract

Voice Activity Detection (VAD) is considered as an important pre-processing step in speech processing systems such as speech enhancement, speech recognition, gender and age identification. VAD helps in reducing the time required to process speech data and to improve final system accuracy by focusing the work on the voiced part of the speech. An automatic technique for VAD using Fuzzy-Neuro technique (FN-AVAD) is presented in this paper. The aim of this work is to alleviate the problem of choosing the best threshold value in traditional VAD methods and achieves automaticity by combining fuzzy clustering and machine learning techniques. Four features are extracted from each speech segment, which are short term energy, zero-crossing rate, autocorrelation, and log energy. A modified version of fuzzy C-Means is then used to cluster speech segments into three clusters; two clusters for voice and one for unvoiced. After that, three feed forward neural networks are trained to adjust their weights, in which each network represents one cluster. To make the final decision regarding the class type of a given speech segment, the membership degrees of this segment in all clusters along with neural networks' decisions are given to a defuzzification step which finally gives the class type of that segment. The proposed FN-AVAD is tested on the public multimodal emotion database, Surrey Audio-Visual Expressed Emotion (SAVEE), and the error rate was 2.08%. The achieved results are comparable to the results achieved by the current published works in the literature.

Keywords: Defuzzification, Fuzzy clustering, Neural network, Speech signal, Voice activity detection.

1. Introduction

Speech signal is one of the most important signals that contain information. It includes words, emotions, intention, accent, dialect, age, speaker identity, gender, health status of the speaker, etc. [1, 2]. However, building applications to extract information conveyed within the speech signal need efficient feature extraction techniques. These features need to be extracted from the voiced part as it contains the important speech or speaker specific attributes [3]. Distinguishing of voiced speech from silence and background noise is called Voice Activity Detection (VAD). VAD is mainly used to reduce the amount of time required to process the data by removing undesired parts like silence, pause and any non-speech activity from speech signal. It is also useful for improving system computation capabilities by reducing the dimensionality of the extracted features [4, 5].

Statistical and pattern recognition approaches have been applied for deciding the class of the given segment of a speech signal [6]. Most of these methods are based on extracting a feature from speech segments to make the decision by comparing the value of that feature with a pre-defined threshold. This feature can be Mel-Frequency Cepstral Coefficient (MFCC), Zero Crossing Rate (ZCR), Cepstral domain peak, Short Term Energy (STE), or harmonic to noise ratio, etc. [7]. However, these features have their own disadvantages regarding determining the threshold in an ad hoc basis. For example, STE uses the fact that energy in a voiced segment is greater in silence/unvoiced segment. But it is not known in advance, what is the optimal threshold value that should be for precise classification, because it varies from case to case [6].

In addition, each feature has its own limitations thus, the methods proposed in the literature attempted to use a combination of features for voiced/unvoiced decision. Radmard et al. [7] have presented an approach for VAD based on the analysis of zero crossing rate, Cepstral peak, and Auto Correlation Function (ACF) peak of the speech signal segments by clustering the extracted features using K-Means algorithm with K=3 (one cluster for unvoiced and two clusters for voiced). 821 frames of speech taken from Texas Instruments/Massachusetts Institute of Technology (TIMIT) were used for testing purpose. The total error rate for voiced segments was 4.8% while the error rate for unvoiced segments was 1.1% [7].

Algabri et al. [8] used STE and ZCR with fuzzy logic to control voiced/unvoiced classification of speech segments. An error ratio equal to 2.5 % was achieved using the Arabic digits of the KSU database [8]. Roy et al. [9] decomposed the speech signal into low-frequency and high-frequency components using wavelet transform and then compute the entropy-based thresholds for these components. Two types of thresholds have been used: low-frequency thresholds (for voiced speech segments) and the high-frequency thresholds (for the unvoiced speech segments). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database was used for testing purpose and the overall accuracy was 99.3% for start-frame detection and 97.5% for end-frame detection [9]. Although the effort to find a solution for VAD problem started a long time ago and different methods have been proposed, the search continues because the accurate solution is still not found.

An automatic voiced activity detection using Fuzzy-Neuro technique (FN-AVAD) is presented in this paper. A combination of features is used to decide whether the given speech segment is voiced or unvoiced in an automatic fashion. Since the extracted features' values are not precise; it will be helpful to use Fuzzy-

Neuro technique by combining fuzzy C-Means clustering and neural network to more accurate results. Fuzzy concept allows feature approximation while neural network permits learning and behaviour reservation. The remainder of this paper is categorized as follows: In the next section, theoretical background for some of the methods used is presented. The proposed methodology is described in detail in Section 3. Section 4 illustrates the achieved results when applying the proposed method on the test database. Finally, conclusions and recommendations for future works are discussed in Section 5.

2. Theoretical Background

This section demonstrates the underlying theoretical concepts used by the proposed FN-AVAD method.

2.1. Speech framing and windowing

Generally, the sound production system of human speech needs about 12-30 milliseconds gap between two words when producing speech sound because the system needs that time to prepare itself to produce the next sound. So, it is required to break the signal into smaller short-time segments (frames). The size should be chosen within that range. In order to keep the continuity of frames, it's necessary to retain some overlaps between the continue frames. This process is called framing process. The lengths of the frame, as well as the overlap ratio between successive signal frames are related to the problem at hand. After framing the speech signal, windowing process must be performed. Windowing is the process of multiplying a waveform of speech signal segment by a time window of a given shape, to stress pre-defined characteristics of the signal and smooth the discontinuities at the beginning and end of the sampled signal. Some of the windowing functions used in speech analysis are rectangular window, Hanning window, Bartlett window, Blackman window, Kaiser window, Hamming window, etc. [10].

2.2. Speech characteristics

To separate voiced from unvoiced speech, discriminative features must be extracted from each speech frame, and these features must obey with the following speech characteristics [11, 12]:

- Lowest energy presents in the unvoiced speech waveform when compared to voiced speech.
- Less number of zero crossings associated with the voiced waveform when compared to unvoiced speech.
- No correlation among successive samples in the speech waveform for unvoiced speech.
- The absence of any signal characteristics in unvoiced speech.

2.3. Speech normalization

Data normalization is mainly used to allow the comparison of corresponding data values for different datasets in a way that eliminates the side effects of certain gross-influences [13]. Speech attribute normalization is required due to the following reasons:

(i) Different persons may speak with different loudness levels.

(ii) The speech may be recorded with different distances from the microphone which in effect reduces the level of the recorded sound.

(iii) The extracted features may have different scales that will affect the learning process of the classifier.

2.4. Fuzzy C-Means clustering algorithm

Fuzzy clustering, which is frequently used in pattern recognition, uses the concepts of the field of fuzzy logic and fuzzy set theory. Fuzzy set theory allows an element to belong to a set with a degree of membership between 0 and 1. Fuzzy clustering consists of a collection of C clusters, C_1, C_2…, C_C, and a membership matrix $\mu = \mu_{ij} \in [0,1]$, for $i = 1\ldots K$ and j = 1 ... C, where each element μ_{ij} represents the degree of membership of object i in cluster C_j. It is based on minimization of the following objective function [14]:

$$J_m = \sum_{i=1}^{K} \sum_{j=1}^{C} \mu_{ij}^m \left\| X_i - C_j \right\|^2 \quad , 1 \le m \le \infty \tag{1}$$

where K is number of data elements, C is number of clusters, m is any real number greater than 1 which represents the intensity of fuzzification, X_i is the i^{th} of d-dimensional measured data item, C_j is the d-dimension centre of the cluster, μ_{ij} is the degree of membership of X_i in the cluster j, and $\| \ \|$ is any similarity measure between data item and cluster centre. Fuzzy partitioning is carried out through an iterative process. In each iteration, the membership μ_{ij} and the cluster centres C_j are updated as follows [14]:

$$\mu_{ij} = \frac{1}{\sum_{z=1}^{C} \left(\frac{\left\| X_i - C_j \right\|}{\left\| X_i - C_z \right\|} \right)^{\frac{2}{m-1}}} \tag{2}$$

$$C_j = \frac{\sum_{i=1}^{K} \mu_{ij}^m x_i}{\sum_{i=1}^{K} \mu_{ij}^m} \tag{3}$$

2.5. Feed forward neural networks

Machine learning algorithms perform specific tasks by generalization from a set of data samples. An example of machine learning algorithms is Artificial Neural Network (ANN). ANN is interconnection of neurons in which each neuron can be defined as follows [15]:

$$O(X) = g\left(\sum_{i=0}^{K} W_i X_i\right) + b \tag{4}$$

where, X is a neuron with K input (X_0,…, X_K), O(X) is the output of X, and W_i are weights that are given for each data input(W_0,…, W_K). g is an activation function that indicates how powerful the output should be from the neuron, based on the sum of the input, and b is a bias [16]. In feed forward neural network, the input is fed forward through the network to update the weights between network neurons. The training phase in this network topology is called "back propagation", where the error is backward propagated to adjust the weights [17].

3. The Proposed FN-AVAD Method

The proposed FN-AVAD method consists of two important stages: training stage and testing stage. In training stage, the ANNs are trained with the extracted features

for the three clusters that are resulted from fuzzy clustering. Training stage involves five steps: (1) framing and windowing, (2) feature extraction, (3) feature normalization, (4) fuzzy clustering and (5) ANNs training. The purpose of testing stage is to classify the short time segments of the speech into either voiced or unvoiced classes using the pre-trained networks. Testing stage consists of six steps: (1) framing and windowing, (2) feature extraction, (3) feature normalization, (4) fuzzy clustering, (5) ANNs testing, and (6) defuzzification. Figure 1 demonstrates the general view of the proposed FN-AVAD method.

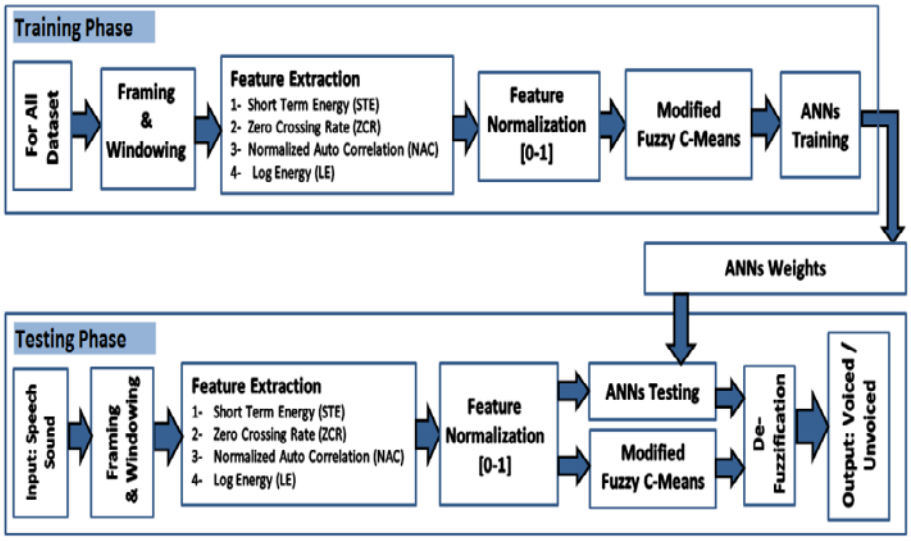


Fig. 1. General view of the proposed FN-AVAD method.

3.1. Training stage

3.1.1. Framing and windowing step

Hamming window is used for windowing process because it is simple and the frequency ripples at the ends of the window can be avoided. The mathematical representation of hamming window can be defined as in Eq. (5) [1].

$$w[n] = \begin{cases} 0.54 - 0.46\,cos\left(\frac{2\pi n}{L}\right) & 0 \le n \le L \\ 0 & otherwise \end{cases} \quad (5)$$

where L is the number of samples in each frame, n is the number of the frame under processing and $w[n]$ is the result of applying windowing process. Window length used in this work is 12 milliseconds with 50% overlapping ratio because these values gave the best framing results by trial and error.

3.1.2. Feature extraction step

In feature extraction step, the discriminated features are extracted from each speech frame. To avoid the problems of using a single feature, four features have been used in this work which is STE, ZCR, ACF and Log Energy (LE). These features are extracted because they obey speech characteristics mentioned in Section 2.2.

(a) ***Short Term Energy (STE)***

Short-term energy represents the amount of energy in the speech signal at a specific instance of time. The STE of a voiced signal is always much greater than that of unvoiced signal. STE can be defined as in Eq. (6) [11].

$$STE(n) = \frac{1}{L}\sum_{i=0}^{L-1} X_i^2 \tag{6}$$

where, n is the number of the frame under processing, L is the length of speech frame and X_i is the i^{th} sample in frame (n).

(b) ***Zero Crossing Rate (ZCR)***

Zero-crossing rate is a measure of the number of times amplitude of a speech signal crosses through a value of zero within a given interval of time. In case of unvoiced speech, the signal crosses zero more number of times than voiced speech. For silent regions, ZCR is near zero. ZCR can be defined as in Eq. (7) [4].

$$ZCR(n) = \frac{1}{2(L-1)}\sum_{i=0}^{L-1}|sgn\,(X_{i+1}) - sgn\,(X_i)| \tag{7}$$

$$\text{and, } sgn(X_i) = \begin{cases} 1, & if\ X_i \geq 0 \\ 0, & otherwise \end{cases} \tag{8}$$

where, n is the number of the frame under processing, L is the length of speech frame and X_i is the i^{th} sample in frame (n).

(c) ***Auto Correlation Function (ACF)***

The autocorrelation coefficient gives the correlation between the adjacent samples of the signal which usually varies between -1 and +1. The value of ACF for voiced signal is close to unity (1 or -1) because of the frequency concentration in the low frequencies while it is close to zero for unvoiced signal. ACF is defined as shown in Eq. (9) [6].

$$ACF(n) = \frac{\sum_{i=1}^{L} X_i\, X_{i-1}}{\sqrt{(\sum_{i=1}^{L} {X_i}^2)(\sum_{i=0}^{L-1} {X_i}^2)}} \tag{9}$$

where, n is the number of the frame under processing, L is the length of speech frame and X_i is the i^{th} sample in frame (n).

(d) ***Log Energy (LE)***

Log energy (the decibel dB) is a measure of speech loudness with respect to humans' ears. Log-energy of frame (n) can be defined as in Eq. (10) [18].

$$LE(n) = 10 * Log_{10(}\varepsilon + \frac{1}{L}\sum_{i=0}^{L-1} X_i^2) \tag{10}$$

where, n is the number of the frame under processing, L is the length of speech frame, X_i is the i^{th} sample in frame (n) and ε is any small number usually 10^{-6}.

3.1.3. Feature normalization step

Feature normalization step is required to compensate the differences among various speech cases. Features are normalized to the range of [0-1] using Eq. (11) [13].

$$V_i' = \left(\frac{V_i - Mino}{Maxo - Mino}\right) * (Maxn - Minn) + Minn \quad (11)$$

where V_i and V_i' are the *old* and *new* data values, respectively, $Mino$ and $Maxo$ are the old range confines and Minn and Maxn are the new range confines.

3.1.4. Modified fuzzy C-Means

A modified version of Fuzzy C-Means is proposed in this work. The objective of the modifications is increasing the coherence of the resulted clusters through intensify the effect of very similar objects and reduce the effect of very far objects. These modifications can be outlined with the following points:

- Since the behaviors of the extracted features are known in advance, the initial value of clusters' centers can be guessed. Clusters' centers are set to a given range through searching the feature space for a sample that holds feature with value in that range. This will lead to a guided fuzzy C-Means.
- The number of used clusters is three. One cluster for unvoiced speech with the initial center within the range [0.1-0.3]. Two clusters for voiced speech with the initial centers equal to [0.5-0.7] and [0.8-1], respectively. STE feature is used for finding the best frame sample to be set as an initial center for each cluster.
- To ensure the coherent of the resulted clusters, the membership function of fuzzy C-Means is changed. The distance of the object (X_i) from the cluster center (C_j) is computed and compared with two values (a_1 and a_2). The proposed membership function can be defined as in Eq. (12):

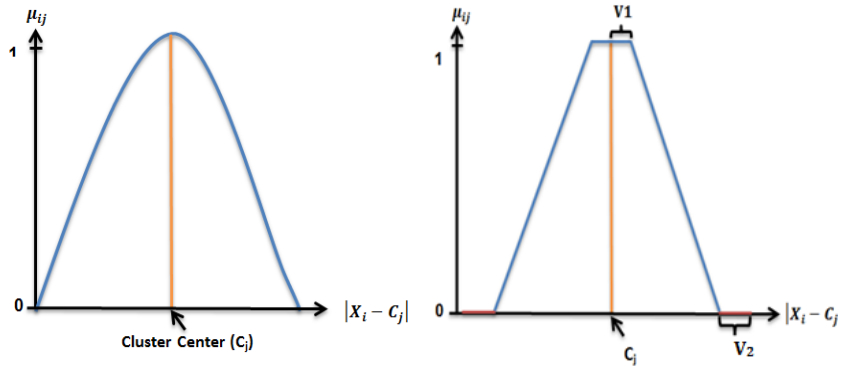
$$\mu_{ij} = \begin{cases} 1 & if\ \left(\left|X_i - C_j\right|\right) \le a_1 \\ 0 & if\ \left(\left|X_i - C_j\right|\right) \ge a_2 \\ \dfrac{1}{\sum_{z=1}^{C}\left(\dfrac{\left\|X_i - C_j\right\|}{\left\|X_i - C_z\right\|}\right)^{\frac{2}{m-1}}} & otherwise \end{cases} \quad (12)$$

where: $a_1 = Min_{1\le i\le K}\{\left|X_i - C_j\right|\} + V_1$, $a_2 = Max_{1\le i\le K}\{\left|X_i - C_j\right|\} * V_2$ and $V_1, V_2 \in$ [0-1].

In other words, if the object is very close to the cluster center (its' distance less than or equal to a_1) then its membership degree will be set to 1, to emphasis its contribution in cluster center calculation. Otherwise, if the object is so far from the cluster (its distance greater than or equal to a_2) then it will be thrown out the cluster by setting its membership degree to be 0, to ensure it will not contribute in cluster center computation. Figure 2 demonstrates the idea of modified membership function proposed in this paper.

The output of fuzzy C-Means may contain samples with a very small membership degree for some clusters. If we considered each sample showed a membership degree greater than zero as a member of the cluster (even it has a very small membership degree), we will face a problem when training ANNs because the ANNs will show similar behavior and this will lead to incorrect classification of the given speech frame. To solve this problem, a parameter called "*Fuzzification Ratio*" with value ranged from 0 to 1 is added. The membership degree of each frame within each cluster is compared with this parameter. If the membership

degree of the given frame is greater than *Fuzzification Ratio* value, then it will be member of that cluster otherwise it will be discarded.



(a) Original function. **(b) Modified function.**

Fig. 2. Fuzzy C-Means membership function.

3.1.5. ANNs training

Since the extracted features are imprecise to make voiced/unvoiced decision, it will be helpful to use fuzzy and machine learning concepts to solve VAD problem as shown in Fig. 3.

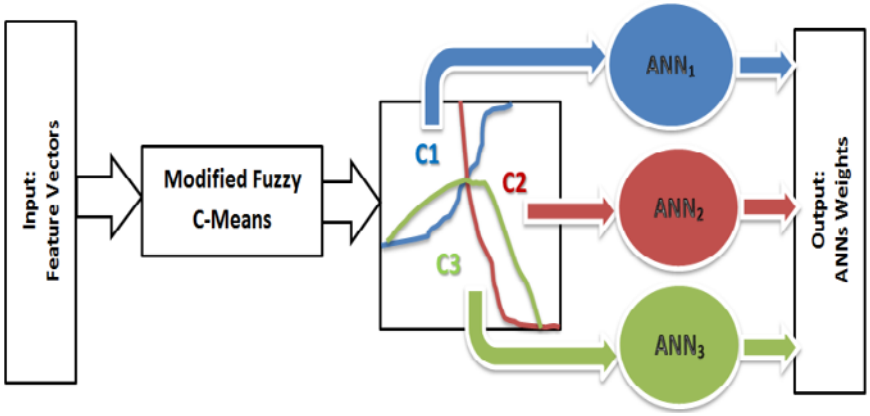


Fig. 3. Fuzzy-neuro technique used for VAD problem solution.

Training step focuses on analyzing and interpreting the patterns and structures of each cluster to enable learning and decision making. Three feed forward neural networks with the following configuration are used (as shown in Fig. 4):

Number of nodes in input layer = number of features= 4

Number of nodes in output layer= 2 (for the binary representation of {1, 2, 3})

Number of nodes in hidden layer=2

Each network is trained with the frame samples of its' corresponding cluster. The training process aims at remembering the behavior of cluster members by

adjusting network weights' values. The resulted weights of each network are then saved to a separate file, so it can be used in the testing stage.

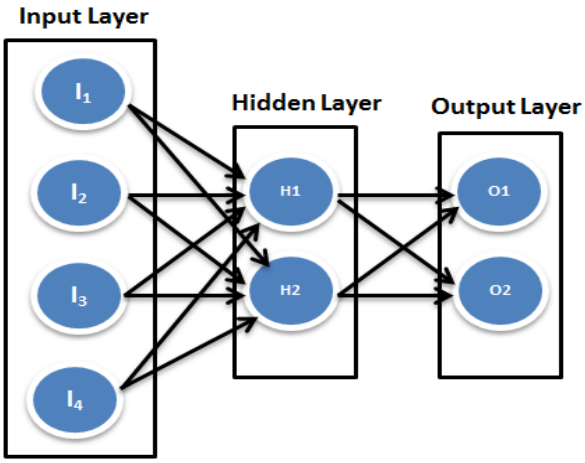


Fig. 4. ANN structure used in training step.

3.2. Testing stage

In testing stage, the short time segments of a given speech signal is classified to either voiced or unvoiced speech. This can be done in six steps as shown in Fig. 1. Framing and windowing, feature extraction, feature normalization and fuzzy clustering steps are similar to their partners' steps in the training stage. In the 5th step (ANNs testing step), each network is tested with the features extracted from the frame to be classified. The network returns either 0 (in case of the frame is not similar to the cluster member) or returns the cluster number (in case of similarity between the checked frame and its member). Finally, in defuzzification step, the membership degree of the given frame is compared with Fuzzification Ratio, if the former is greater than the later then the corresponding ANN is tested to check whether the frame is similar to these clusters' members. ANN's decisions along with the frame membership degrees are inputted to center of gravity defuzzification method. The output for frame (X) can be computed as in Eq. (13):

$$y = \frac{\sum_{i=1}^{3} \mu_{C_i}(X).\ ANN_i(X)}{\sum_{i=1}^{3} \frac{ANN_i(X)}{i} \mu_{C_i}(X)} \tag{13}$$

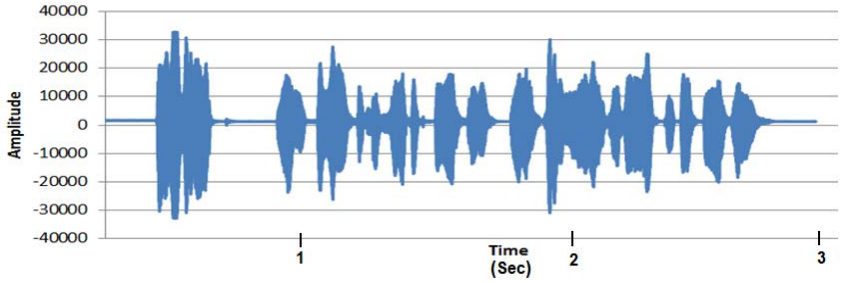
where $\mu_{C_i}(X)$ is the membership degree of X to class C_i, and $\text{ANN}_i(X)$ is the output of neural network ANN_i for input X. The summation in Eq. (13) will repeat for three times because the proposed method consists of three clusters and three ANNs.

4. Experimental Results and Discussions

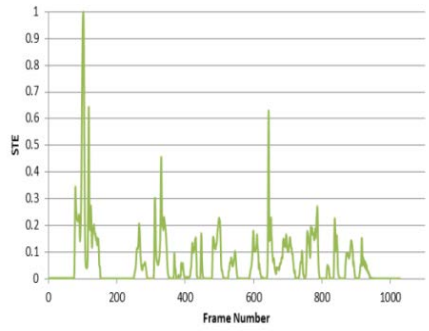
The proposed Fuzzy-Neuro VAD methodology is trained and tested on Surrey Audio-Visual Expressed Emotion (SAVEE) database. SAVEE includes recordings of seven different emotions for four male actors, a totally of 480 samples. The sentences are in British English utterances that chosen from the standard corpus of TIMIT. The data were recorded in a visual media lab with high quality audio-visual equipment. To

check the quality of the recorded videos, they were evaluated by ten subjects under audio, visual and audio-visual conditions [19].

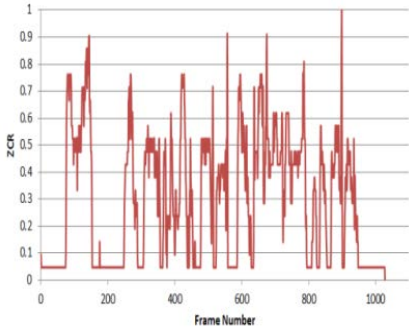
Figure 5 shows the results of feature extraction and normalization steps for the four different features using an example speech signal from the SAVEE database.



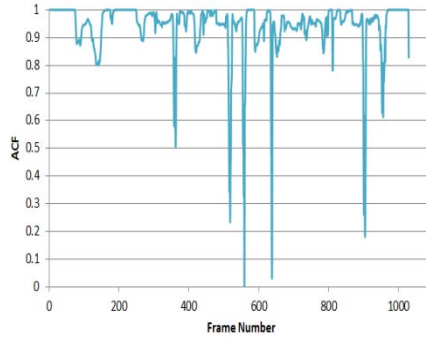
(a) Original speech signal.



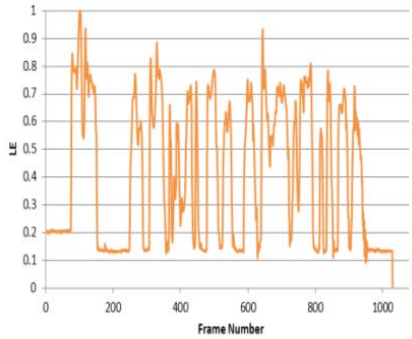
(b) STE feature.



(c) ZCR feature.



(d) ACF feature.



(e) LE feature.

Fig. 5. Feature extraction and normalization results.

Different experiments are conducted to find the best values for m, number of iterations, V_1, V_2 and Fuzzification parameters. The parameters settings that gave the best results were: *m*=1.5, *number of iterations*= 100, V_1= 0.06, V_2= 0.9, *Fuzzification Ratio*=0.5. Table 1 shows the memberships' degrees of frames numbered from 80 to

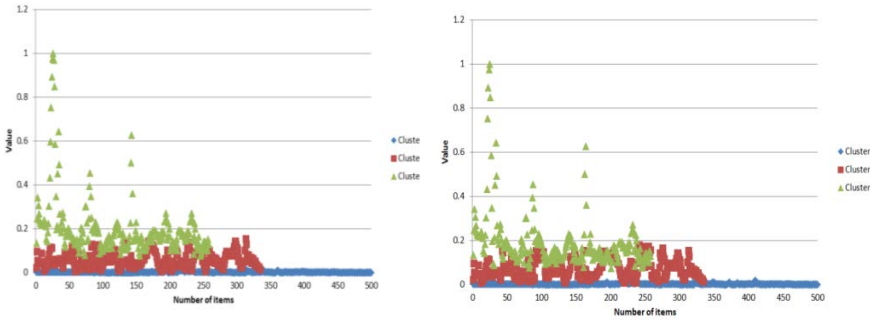
100 for an example speech signal that achieved after clustering the signal using traditional fuzzy C-Means and modified fuzzy C-Means algorithm, respectively. As shown in the table, the membership degrees are set to 1 for the frames which are very close to the cluster center, while the membership degrees are set to 0 for the frames which are very far from the cluster.

Table 1. Traditional and modified fuzzy C-Means membership degree for the frames numbered from 80 to 100 for an example speech signal.

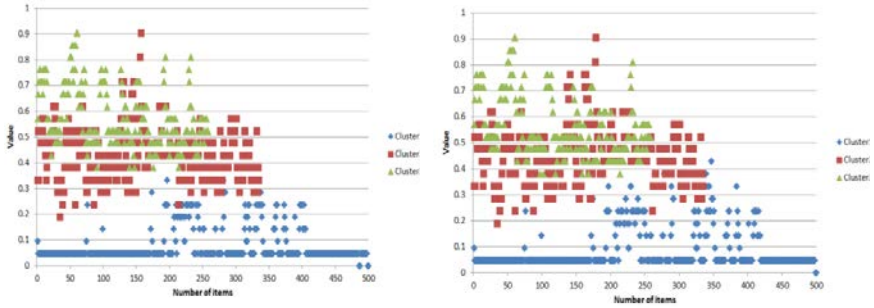
	Traditional Fuzzy C-Means				**Modified Fuzzy C-Means**		
#	Cluster1	Cluster2	Cluster3	**#**	Cluster1	Cluster2	Cluster3
1	0.00292	0.03705	0.96003	**1**	0.00542	0.09523	0.89935
2	0.00182	0.02626	0.97192	**2**	0.00379	0.07725	0.91897
3	0.00067	0.01213	0.98720	**3**	0.00181	0.04877	0.94942
4	0.00049	0.00938	0.99013	**4**	0.00144	0.04200	0.95655
5	0.00045	0.00880	0.99075	**5**	0.00137	0.04047	0.95816
6	0.00013	0.00324	0.99663	**6**	0.00061	0.02321	1
7	0.00075	0.01324	0.98601	**7**	0.00196	0.05134	0.94670
8	0.00036	0.00736	0.99228	**8**	0.00117	0.03654	1
9	0.00091	0.01542	0.98367	**9**	0.00226	0.05616	0.94158
10	0.00136	0.02106	0.97758	**10**	0.00304	0.06764	0.92932
11	0.00103	0.01702	0.98195	**11**	0.00248	0.05955	0.93797
12	0.00031	0.00660	0.99309	**12**	0.00107	0.03434	1
13	0.00001	0.00084	0.99915	**13**	0	0.00052	1
14	0.00005	0.00367	0.99628	**14**	0.00001	0.00100	1
15	0	0.00026	0.99974	**15**	0	0.00064	1
16	0.00023	0.01134	0.98843	**16**	0.00019	0.01703	1
17	0.00174	0.05085	0.94741	**17**	0.00153	0.07329	1
18	0.00687	0.11606	0.87707	**18**	0.00619	0.15289	0.84092
19	0.02776	0.34821	0.62403	**19**	0.02585	0.37518	0.59897
20	0.02810	0.22444	0.74746	**20**	0.02595	0.26451	0.70954

Emphasizing the effect of very close frames and reducing the effect of very far frames lead to more fine-tuned clusters' centers. Figure 6 demonstrates frames distribution of a sample speech signal using traditional and modified Fuzzy C-Means, for the four extracted features. The left image in the figure represents the traditional fuzzy C-Means while the right image represents the modified fuzzy C-Means. The figure clearly showed that modified fuzzy C-Means can get rid of far frames and preserve the coherence of the resulted clusters especially for ACF and LE features.

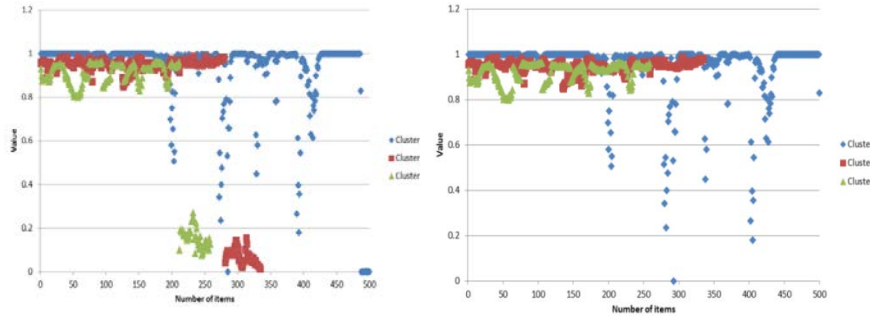
Table 2 illustrates the results of testing 20 frames (from 139-158) for an example speech signal. The table also shows values of the extracted features, membership degrees, neural networks' results and the final decision taken by defuzzification step. These frames are selected because they involve numerous frame types (i.e., voiced and unvoiced).



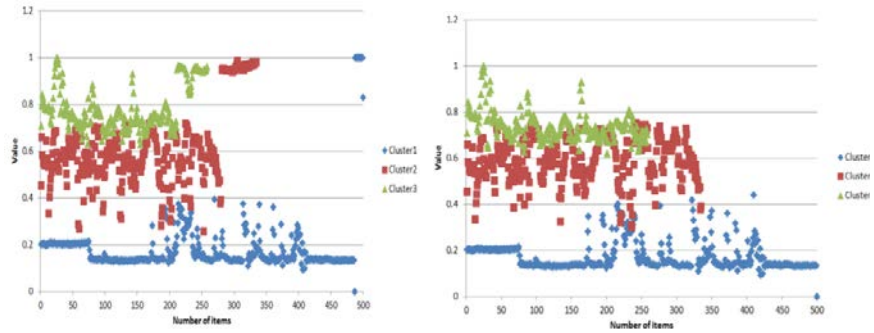
(a) STE feature.



(b) ZCR feature.



(c) ACF feature.



(d) LE Feature.

Fig. 6. The distribution of frames' samples in the resulted clusters.

Table 2. Results of 20 frames (from 139 to 158) chosen from a sample speech signal.

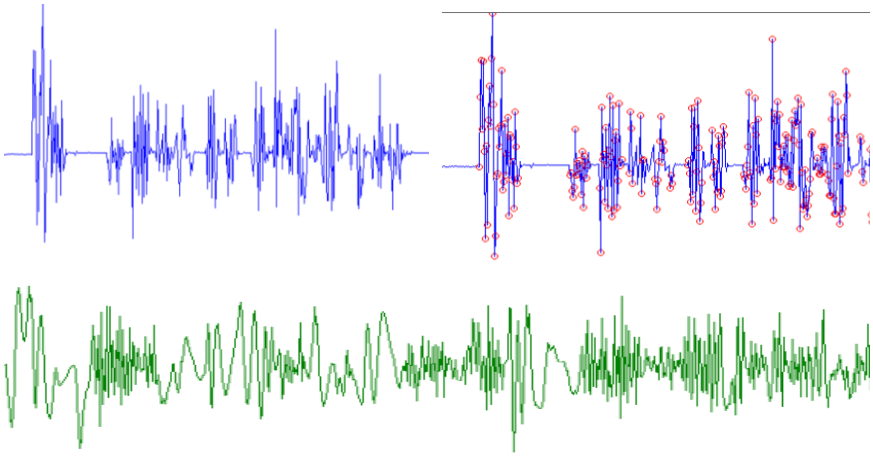
Frame No.	ZCR	STE	ACF	LE	μ_{C_1}	μ_{C_2}	μ_{C_3}	ANN_1	ANN_2	ANN_3	Decision
1	0.14691	0.71428	0.80807	0.72232	0.00176	0.0655	0.93273	0	0	3	Voiced
2	0.15734	0.85714	0.80596	0.73222	0.00466	0.09917	0.89616	0	0	3	Voiced
3	0.15284	0.76190	0.80060	0.72804	0.00249	0.07221	0.92529	0	0	3	Voiced
4	0.13474	0.76190	0.80293	0.70985	0.00377	0.12540	0.87081	0	0	3	Voiced
5	0.12812	0.85714	0.81873	0.70260	0.00704	0.18613	0.80682	0	0	3	Voiced
6	0.14115	0.90476	0.81213	0.71655	0.00745	0.15448	0.83805	0	0	3	Voiced
7	0.15182	0.90476	0.81004	0.72707	0.00628	0.12215	0.87156	0	0	3	Voiced
8	0.15062	0.66666	0.82482	0.72592	0.00073	0.03481	1	0	0	3	Voiced
9	0.12500	0.61904	0.84372	0.69905	0.00078	0.07135	1	0	0	3	Voiced
10	0.09030	0.66666	0.86046	0.65228	0.00309	0.40085	0.59605	0	0	3	Voiced
11	0.05042	0.57142	0.88885	0.56894	0.00262	1	0.20909	0	2	0	Voiced
12	0.04826	0.47619	0.92707	0.5627	0.00031	1	0.00782	0	2	0	Voiced
13	0.03962	0.47619	0.93825	0.53473	0.00088	1	0.01092	0	2	0	Voiced
14	0.00915	0.42857	0.96030	0.33368	0.20599	0.74389	0.05011	0	2	0	Voiced
15	0.00246	0.23809	0.98328	0.17812	1	0.00354	0.00074	1	0	0	Unvoiced
16	0.00165	0.04761	0.98753	0.13943	1	0.00015	0.00004	1	0	0	Unvoiced
17	0.00179	0.04761	0.99192	0.14703	1	0.00014	0.00004	1	0	0	Unvoiced
18	0.00183	0.04761	0.99342	0.14902	1	0.00014	0.00004	1	0	0	Unvoiced
19	0.00162	0.04761	0.99421	0.13827	1	0.00020	0.00005	1	0	0	Unvoiced
20	0.00169	0.04761	0.99518	0.14188	1	0.00018	0.00005	1	0	0	Unvoiced

Figure 7 shows the results of the proposed methodology when tested on three different speech signals with different speech loudness. The first image (with blue drawing) represents the original speech signal while the second image (with red circles) represents the frames which are selected as voiced parts. The third image (with green drawing) shows the final voiced part of the speech in which the frames with voiced class are gathered together in a new signal. The figure clearly demonstrates the effectiveness of the proposed FN-AVAD method.

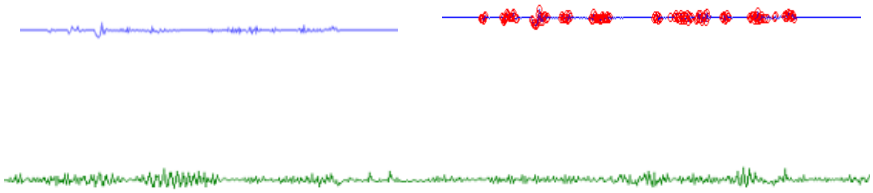
The proposed FN-AVAD method is further evaluated using error rate measure as defined in Eq. (14) [7].

$$Error\ Rate = \frac{TH-TC}{TH} \times 100 \tag{14}$$

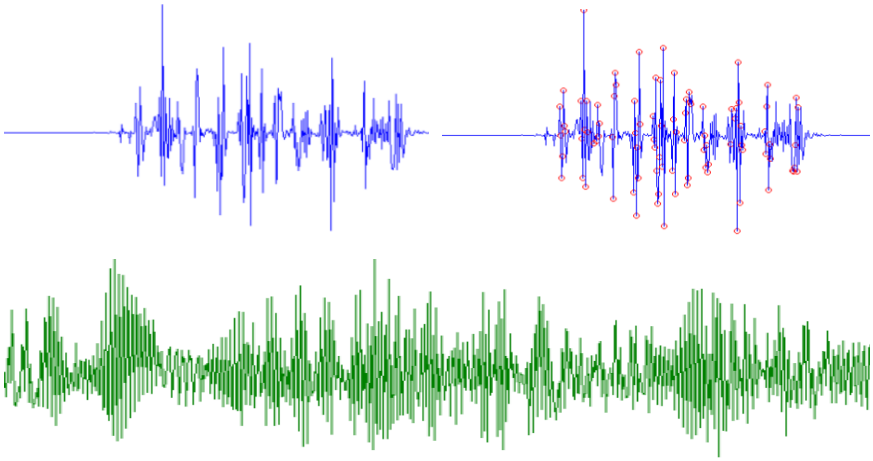
where *TH* is the number of voiced frames detected by the human and *TC* is the number of voiced frames detected using the proposed method. The final error rate achieved by the proposed VAD method was 2.08%. Table 3 shows a comparison made between the proposed method and the current research works on the VAD problem along with method used and the achieved results. The table demonstrates the effectiveness of the proposed VAD method.



(a) Speech signal with high loudness.



(b) Speech signal with low loudness.



(c) Speech signal with medium loudness.

Fig. 7. The final results of the proposed FN-AVAD method.

Table 3. Comparison between the proposed method and other works published within VAD problem.

Authors	Method	Result
Radmard et al. [7]	Analysis of the extracted features using K-Means algorithm	The total error rate for voiced segments was 4.8% while the error rate for unvoiced segments was 1.1%.
Algabri et al. [8]	STE and ZCR was used with fuzzy logic	Error rate equal to 2.5 %
Roy et al. [9]	Wavelet convolution based speech endpoint detection	Error rate for start-frame was 0.7% and end-frame was 2.5%
The proposed FN-AVAD	Fuzzy-Neuro technique	Error rate equal to 2.08%

5. Conclusions and Future Work

An automatic Fuzzy-Neuro VAD (FN-AVAD) methodology has been presented in this paper to omit the need for proper threshold selection. A modified version of fuzzy C-Means helps in achieving features' approximation by clustering speech frames into three clusters with membership degrees. On the other hand, neural networks work as a memory to remember the behaviour of each cluster by updating the network weights. The experimental results showed the effectiveness of the proposed FN-AVAD. As future work, features from spectral-domain can be extracted instead of time-domain features. In addition, to separate noise parts from silent parts, four clusters can be used with different initial ranges in which the noise has range laid between silent and voiced ranges. Recurrent deep learning techniques can also be utilized to take into account the time factor between successive frames.

Nomenclatures

$ANN_i(X)$	Output of neural network ANN_i for input X
b	Neural network bias
C	Number of clusters
C_j	Centre of the cluster j
g	Activation function
K	Number of data elements
L	Length of speech frame
m	Intensity of fuzzification
$Maxn$	Maximum value in new range confines
$Maxo$	Maximum value in old range confines
$Minn$	Minimum value in new range confines
$Mino$	Minimum value in old range confines
n	Number of the frame under processing
$O(X)$	Neural network output for input X
TC	Number of voiced frames detected using the proposed method
TH	Number of voiced frames detected by the human

$\mu_{C_i}(X)$	Membership degree of X to class C_i
μ_{ij}	Degree of membership of object i in cluster C_j
V_i	Old data value
V_i'	New data value
V_1	Maximum distance for emphasising the contribution of data element in cluster centre calculation
V_2	Minimum distance for outlier removal
W_i	Weights that are given for each data input (W_0,…, W_K)
X	Data input with K elements (X_0,…, X_K)
X_i	Data element i in the input X

Greek Symbols

ε	Any small number usually 10^{-6}

Abbreviations

ACF	Auto Correlation Function
ANN	Artificial Neural Network
FN-AVAD	Automatic Voiced Activity Detection using Fuzzy-Neuro
LE	Log Energy
MFCC	Mel-Frequency Cepstral Coefficient
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SAVEE	Surrey Audio-Visual Expressed Emotion
STE	Short Term Energy
TIMIT	Texas Instruments/Massachusetts Institute of Technology
VAD	Voice Activity Detection
ZCR	Zero Crossing Rate

References

1. Padmaja, J.N.; and Rao, R.R. (2016). A comparative study of silence and non silence regions of speech signal using prosody features for emotion recognition. *Indian Journal of Computer Science and Engineering*, 7(4), 153-161.
2. Mohammed, S.N; Jabir, A.J.;and Abbas. Z.A. (2019). Spin-Image Descriptors for Text-Independent Speaker Recognition. *In: Saeed, F.; Mohammed, F.; and Gazem, N. (eds) Emerging Trends in Intelligent Computing and Informatics. IRICT 2019. Advances in Intelligent Systems and Computing*, 1073, Springer, Cham, 216-226.
3. Jasmine, J.M.; Sandhya, S.; Ravichandran, K.; and Balasubramaniam, D. (2016). Silence removal from audio signal using framing and windowing method and analyze various parameters. *International Journal of Innovative Research in Computer and Communication Engineering*, 4(4), 6606- 6613.
4. Ong, W.Q.; Tan, A.W.C.; Vengadasalam , V.V.; Tan, C.H.; and Ooi, T.H. (2017). Real-Time Robust Voice Activity Detection Using the Upper Envelope Weighted Entropy Measure and the Dual-Rate Adaptive Nonlinear Filter. *Entropy*, 19(11), 1-21..
5. Yuxin, Z.; and Yan, D. (2014). A voice activity detection algorithm based on spectral entropy analysis of sub-frequency band. *BioTechnology: An Indian Journal*, 10(20), 12342-12348.

6. Bachu, R.G.; Kopparthi, S.; Adapa, B.; and Barkana, B.D. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. *Proceedings of the American Society for Engineering Education*. Pittsburgh, Pennsylvania, 1-7.
7. Radmard, M.; Hadavi, M.; and Nayebi, M.M. (2011). A new method of voiced/unvoiced classification based on clustering. *Journal of Signal and Information Processing*, 2, 336-347.
8. Algabri, M.; Alsulaiman, M.; Muhammad, G.; Zakariah, M.; Bencherif, M.; and Ali, Z. (2015). Voice and unvoiced classification using fuzzy logic. *Proceedings of the 2015 World Congress in Computer Science, Computer Engineering, and Applied Computing*. Las Vegas, USA, 416-420.
9. Roy, T.; Marwala, T.; and Chakraverty, S. (2019). Precise detection of speech endpoints dynamically: A wavelet convolution based approach. *Communications in Nonlinear Science and Numerical Simulation*, 67, 162-175.
10. Ibrahim, Y.A.; Odiketa, J.C.; and Ibiyemi, T.S. (2017). Preprocessing technique in automatic speech recognition for human computer interaction: An overview. *Annals. Computer Science Series Journal*, XV, 186-191.
11. Sharm, A.P. (2016). Implementation of ZCR and STE techniques for the detection of the voiced and unvoiced signals in continuous punjabi speech. *International Journal of Emerging Trends in Science and Technology*, 3(6), 4132-4135.
12. Nandhini, S.; and Shenbagavalli, A. (2014). Voiced/unvoiced detection using short term processing. *Proceedings of the International Conference on Innovations in Information, Embedded and Communication Systems*. Coimbatore, India, 39-43.
13. Witten, I.H.; Frank, E.; and Hall, M.A. (2011). Data *mining: Practical machine learning tools and techniques* (3rd ed.). USA: Elsevier.
14. Malhotra, V.K.; Kaur, H.; and Alam, M.A. (2014). An analysis of fuzzy clustering methods. *International Journal of Computer Applications*, 94(19), 9-12.
15. Haripriya, L.; and Jabbar, M.A. (2018). A survey on neural networks and its applications. *International Journal of Engineering Research in Computer Science and Engineering*, 5(4), 64-67.
16. Somwanshi, P.D.; and Chaware, S.M. (2014). A review on: advanced artificial neural networks (ANN) approach for IDS by layered method. *International Journal of Computer Science and Information Technologies*, 5(4), 5129-5131.
17. Sharma, B.; and Venugopalan, K. (2014). Comparison of neural network training functions for hematoma classification in brain CT images. *IOSR Journal of Computer Engineering*, 16(1), 31-35.
18. Dobie, R.A.; and Hemel, S.V. (2004). *Hearing loss: Determining eligibility for social security benefits*. Washington: National Academies Press.
19. Surrey Audio-Visual Expressed Emotion (SAVEE) Database Home Page. Retrieved March 1, 2019, from http://kahlan.eps.surrey.ac.uk/savee/.