Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

Localization of Speaker using Fusion Techniques and Neural Network Algorithms

Rasha H. Ali[1], Mohammed Najm Abdullah[2], Buthainah F. Abed[3], Sawsan H. Jaddoa[4]*

[1]University of Baghdad- College of Education for Women- Computer department, Baghdad, IRAQ.

[2]University of Technology, College of Engineering, Department of computer Engineering, Baghdad, IRAQ.
[3]University of Information Technology and Communications, Baghdad, IRAQ.
[4]University of Baghdad- College of Education for Women- Computer department, Baghdad, IRAQ.

*Corresponding Author: Sawsan H. Jaddoa

DOI: https://doi.org/10.31185/wjps.399
Received 01 May 2024; Accepted 12 Jun 2024; Available online 30 Jun 2024

ABSTACT: Sound source localization especially speech and speaker is sole of the most significant techniques recently because used in various applications like smart environments, industry, robots, and audio conferences. So, the usage of these techniques needs more accuracy. In this paper, a speaker localization proposed it depends on the speech signals in closed spaces by employing fusion techniques and neural networks (NN) algorithms to get more accuracy. The proposed work included finding the classification of the speaker signals, which included three phases: the preprocessing phase, the phase of the feature extraction and classification phase. Data Fusion technique used to generate the dataset of speakers. In feature extraction phase features fusion technique was used for constructing a feature vector by using Generalized Cross Correlation (GCC) for time delay estimation, Root_MUSIC, and Minimum Variance Distortion Less (MVDR) for a direction of arrival for the signal source. In the classification stage two NN algorithms used, Restricted Boltzmann Machine (RBM), which implemented using Tensor flow library and Long Short-Term Memory (LSTM), which implemented using Keras library. The experiments results shows that the accuracy of the two methods was 99.84%, 99.15% for RBM, and LSTM respectively.

Keywords: Speaker localization, Data fusion, feature fusion, RBM, LSTM.

1. INTRODUCTION

Speaker localization is crucial in various audio signal processing applications, including identifying the direction of sound or speaker, accurately capturing speech for distance talking speech recognition, and enabling robots to locate and communicate with humans in a natural manner. These are just a few examples of the applications of speech source localization [1]. When someone perceives a sound coming from the right side, the sound waves first reach the right ear, while the left ear experiences an aural "shadow" caused by the presence of the head (known as head shadow) [2]. The performance of sound localization systems can be influenced by the microphone number. The difficulty of these systems also depends on the capability of a single microphone to process data from a single sensor. Consequently, the signal captured by a single microphone may contain both noise and additional sounds along with the recorded sound. The sound localization techniques used to separate the signal from the noise or any other signals or depend on two microphones in which the systems in this type have the capability to process data from two sensors, which get two sets of the data. The geometrical method can be utilized to deals with this kind using the difference of interaural time (ITD) technique to get the multiple pairs of measured data directly [3] or rely on the microphone arrays, These microphones can be arranged in a linear, planar, or spatial configuration.. Which cover different types of microphones, and using the techniques of the

sound localization to constrain the sound from which microphone. The utilization of arrays of microphones enhances the stability and accuracy of a system. These arrays are a highly desired area of research in multichannel signal processing. they utilized in various domains, such as virtual reality environments, interactions between humans and computers, remote conferences, and Artificial auditory perception by a robot [4].

The aim of this work is the classification of the audio speech signal that resourced from the speaker in the indoor spaces with constrained dataset using the fusion techniques, which includes the data fusion and feature fusion, and NN algorithms used in the classification.

The remaining portion of this paper organized in the following manner. In section 2, the related work for the speaker localization covered. In section, 3 datasets types provided. Section 4 describes the strategies of the speaker localization. The algorithms employed by a NN discussed in section 5. The stages of the proposed approach outlined in section 6. The simulation specifics provided in section 6. The paper's inference shown in Section 7.

2. RELATED WORK

The research studies that are associated with the proposed technique included below:
In [5] the multiple Direction Of Arrival (DOA) estimations using acoustic intensity features based on convolutional recurrent neural network(CRNN) has been suggested, the work applied on the Ambisonics recordings by extracting the features from the audio signal which contains the active and reactive intensity vectors across all frequency bins in the STFT domain then used these features as input to CRNN, the classification rate was 60.9% for simulated spatial room impulse responses (SRIR). In [6], the localization of indoor acoustic sources based on the Convolutional Neural Network (CNN) by using microphone arrays have been presented. The SRP-PHAT strategies used for localization then used it for training the CNN, the results were 41.6%. In [7], the Localizing a speaker in an environment of multi-speaker has been implemented. in this work the features of each speech signal have been extracted using mask technique and using spectrum then these features are inputs for CNN is which is used for estimating the DOA the results were 58.6%. In [8], The application of deep neural networks (DNN) for localizing speakers in numerous rooms has been proposed. The Generalized Cross Correlation-PHAse Transform (GCC-PHAT) used for features extraction then the value of GCC-PHAT used as input to DNN. The result was 78% when recording in the live room. In [9] the Perception and prediction of speaker appeal have been studied using Prosodic features, Mel Frequency Cepstrum Coefficient (MFCC), Support Vector Machine (SVM) while the results were 76.38% for prosodic features and 71.12% for MFCC. Table 1 describes the researches, which connected with speaker localization with different techniques.

Table 1. Related Work Comparison

Paper Title	Dataset	Techniques	Results
Multiple DOA estimation based on CRNN.	FOA recordings	STFT, CRNN	60.9%
Localization of indoor acoustic source based on CNN by using microphones arrays.	IDIAP AV16.3 Corpus, Albayzin Phonetic Corpus	SRP-PHAT, CNN	41.6%
Localizing a target speaker in a multi-speaker environment	Librispeech	Magnitude spectrum, STFT	58.6%
Localizing Speakers in Multiple Rooms by Using DNN	DIRHA dataset	GCC-PHAT, DNN	78% when recording in the live room, and 75% when recording in the kitchen
Predication and perception of the appeal for the speaker	Irish Political Speech, which has multi, records for a single speaker.	Prosodic features, MFCC, SVM.	The accuracy is 76.38% for prosodic features, 71.12% for MFCC.

3. DATASET

The type of the dataset depends on various factors, including the language used in the recording, microphone directivity, and the distance of source microphone, the recording protocol and the population of participating subjects. The datasets listed below are used for speaker identification, verification, localization, and recognition."

3.1 English Language Speech Database for Speaker Recognition (ELSDSR)

The computer room utilized a chamber with dimensions of (8.82 in width, 11.8 in length, and 3.05 in height) to record the speech in it. The one microphone in the middle of the room, in (70*120*70) table in the front of a speaker with deflection boards. It is a relatively environment used for the speech recognition system, identification, localization, and verification. The ages of the speakers are between (24 – 63). The number of speakers is 22 (12) male and (10) female. Fig. 1 shows the setup of the room in 3D.

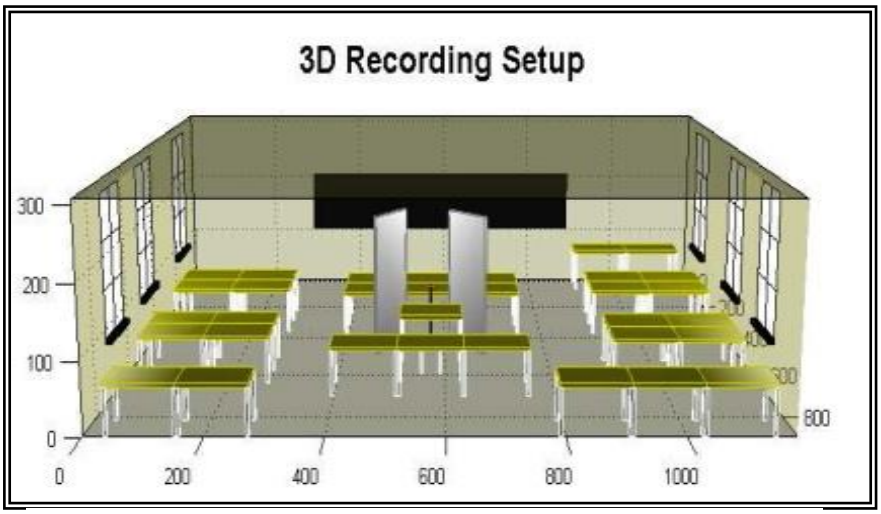

Figure.1 Setup of a room in 3D

3.2 TIMIT

It is a corpus American English speaker. It used for automatic speech and speaker recognition systems. (630) messages were the total number of voice messages (438) from male speakers and 192 from female speakers. Each speaker reads a selection of sentences in different dialects and sexes [11].

4. SPEAKER LOCALIZATION TECHNIQUES

There are different techniques that are using for speaker localization, some of these techniques depend on the estimation the shifted time between the signals, and others depend on the estimation of direction for the signal, the following are the techniques which used in this work:

4.1 Generalized Cross Correlation (GCC)

The GCC proposed by Knapp and Carter in 1976. It is the most communal among time delay of arrival (TDOA), the TDOA depend on proportional delays between two microphones [12]. GCC computes the time tag between two microphone signals and received at an array of the sensor. The function of the cross-correlation between these signals is computed by calculating the inverse of the Fast Fourier Transform for the cross-spectrum ($C_{r_1 r_2}$).The Eq. (1) shows the CCG[13]."

$$R_{r1r2}(\tau) = \sum_{k=0}^{N_f - 1} W(k) C_{r_1 r_2}(k) \exp(j2\pi \frac{k}{N_f} \tau) \quad (1)$$

Where $R_{r1r2}(\tau)$ is the cross correlation, the time lag is represented by τ , the pre filter is denoted by $W(k)$ and k is the discrete frequency, where N_f is the number of elements for the frequency vector, and $j=\sqrt{-1}$ The method used for the pre-filter is the phase transform (PHAT), which allows for the elimination of the magnitude component of the cross-spectrum, thereby preserving just the phase information. The term used for this technique is GCC-PHAT, which refers to the utilization of the PHAT pre-filtering procedure [13].

4.2 Minimum Variance Distortion Less Response (MVDR)

The MVDR is one of the algorithms for adaptive beamforming. Capon (Capon, 1969) introduces it; it is also referred to as the Capon beam former [14]. MVDR beam former works in various scenarios: diffuse noise, diffuse-plus-white noise, and spatially white noise. It also employed to calculate the beamforming coefficients by reducing the discrepancies between residual noise and interference. This achieved by imposing a set of linear restrictions to ensure that the desired signals remain undistorted [15]. MVDR algorithm relies on the steering vectors. MVDR to precisely determine the location of the intended signal amidst the interference [16] computes the weight vector. The signals received from the various array elements converted to the frequency domain using the discrete Fourier transform (DFT). The equation bellow represents the scalar output signal along with its direction.

$$y(K)=w^H.x(K) \tag{2}$$

where K is a time index.

X(k) is acomplex vector of array observation equal to [x1 (K),… xM (K)]

M is the number of array sensors

W is comlex vector of beam former weight equal to [W1 ,…, Wm]T

The observation vector can be computed using the Eq.(3).

$$X(k)=s(k)+i(k)=n(k)= s(k)a+i(k)+n(k) \tag{3}$$

Where s(k): the desired signal i(k): the interference component, n(k): the noise component, a: the signal steering vector, and s(k): the signal waveform.

The weight vector is determined by calculating the maximum value of the signal-to-interference-plus-signal ratio (SINR), which is calculated using Equation (4)[17].

$$SINR=\frac{\sigma_s^2\left|w^H a_s\right|^2}{w^H R_{i+n} w} \tag{4}$$

Where R_{i+n} is matrix of the cross spectral density, $R_{i+n}=E\{\left(i(k)+n(k)\right)\left(i(k)+n(k)\right)^H\}$

σ_s^2 is the signal power where the beam former output. The power given at the produce of a beam former. The weight vector computed by maintaining the distortion less response toward the desired signal and the output interference pulse the power of a noise. The weight vector can be computed maximizing the SINR which equivalent to the following equation [17, 18]

$$W=\arg\ \min\ w^H R_{i+n} w \tag{5}$$

(Where $w^H a=1$. the weight of the MVDR is computed using the approach of the Lagrange multiplier as shown in Eq.(6). $W(\theta)=\frac{R_{i+n}^{-1}\ a(\theta)}{a^H(\theta)R_{i+n}^{-1}\ a(\theta)}$ (6)

4.3 Root -MUSIC

The Root-Music algorithm utilized for the estimation of (DOA). The algorithm is an enhancement of the MUltiple SIgnal Classification (MUSIC) algorithm, which includes the assessment of the spatial pseudo spectrum [19]. Root-MUSIC provides more precise asymptotic estimations compared to spectral MUSIC. Root-MUSIC remains unaffected by radial estimation. The signals separated by computing the root of the polynomial derived from the noise subspace, without the need for the steering step or locating the largest peak [20]. The received signal is defined as the signal collected by all sources at a given moment as shown in Eq. (7).”

$$x(t) = A\,s(t) + n(t) \tag{7}$$

Where $x(t)$ is the received signal at time index (t), $s(t)$: the signal vector generated by a source, $n(t)$: the additive noise of white Gaussian with variance σ_n^2 .Where A : the vector of the steering which can be expanded as the following equations:

$$A = [a(\theta_1)\ a(\theta_2)\ a(\theta_3)...a(\theta_k)] \tag{8}$$

$$a(\theta_k) = \left[1 \exp(-j2\pi d\ \sin\theta_k / \lambda ... 1 \exp(-j(M-1)2\pi d\ \sin\theta_k / \lambda\right]^T \tag{9}$$

Where A= M xK matrix of steering vectors $a(\theta_k)$, $a(\theta_k)$ = Mx1 array steering vector for angle of arrival θ_k . A min step of the estimation of DOA is determining the array correlation matrix R_{xx} which calculated by the following equation:

$$R_{xx} = E\left[x.x^H\right] = AR_{ss}A^H + \sigma_n^2 I \tag{10}$$

where R_{xx} : array correlation matrix (MxM), R_{ss} : the source correlation matrix (K xK), I : the identity matrix (M xM), E[.] : the expectation operator. The MUSIC algorithm can be computed by the following equation:[21].

$$P_{MUSIC}(\theta) = \frac{1}{\left|a^H(\theta) E_N E_N{}^H a(\theta)\right|} \tag{11}$$

By the previous equation, the Root MUSIC written as following equation

$$P_{ROOT-MUSIC}(\theta) = \frac{1}{\left|a^H(\theta) C a(\theta)\right|} \tag{12}$$

where $C = E_N E_N{}^H$, a polynomial and its root are calculated 2(M-1)roots are obtained . The following equation is made use for i=1 to K, where K is only the roots which are closed to the unit circle that are applied for the estimation of the DOA.

$$\theta_i = arc\ \sin\left(-\frac{\lambda}{2\pi d}\ \arg(z_i)\right), \quad i = 1, 2, ..., K \tag{13}$$

5. NEURAL NETWORK ALGORITHMS

There are distinct kinds of algorithms that used in Artificial Neural Networks (ANN). These algorithms depend on the architecture of that algorithm such as Restricted Boltzmann Machines (RBM) and Long Short-Term Memory (LSTM).

5.1 Restricted Boltzmann Machine (RBM)

RBM invented by Geoffrey in 1986. This algorithm is useful for reducing dimensionality, regression, feature learning, classification. It is generative, and unsupervised deep learning which is depending on the probabilistic methods [22].

It contains a single layer undirected NN which consists of two binary layers, the visible layer $(v)=\{v_i\}_{i=1}^{D}$, and hidden layers $(h)=\{h_j\}_{j=1}^{J}$, the parameters are $\theta=\{w,bh,bv\}$.Where w is the weights, bh is the bias of a hidden layer, bv is the bias of a visible layer. In the Boltzmann network, there are no interactions among nodes inside a single layer of (RBM).As shown in Fig. (2), the RBM structure has 3 nodes in the layer of the visible and 4 nodes in the layer of the hidden, each node represents the neuron[23]. The objective of RBM is to find the distribution of the joint probability that maximizes the function of the log-likelihood. The model of the RBM is building using energy function as shown in Eq.(14) [24]

$$E\ (v,h;\theta)=-\sum_{i=1}^{D}\sum_{j=1}^{J}v_i w_{ij} h_j-\sum_{j=1}^{j}bh_j h_j-\sum_{i=1}^{D}bv_i v_i \quad (14)$$

Where $E\ (v,h;\theta)$: the energy, v : the visible layer, h : the hidden layer, θ : the parameters contains(w,bh,bv the weight, the bias of the hidden layer, and the bias of the visible layer respectively), D: the number of the visible layers, J : the number of the hidden layers.

$$pr(v;\theta)=\sum_{h}pr(v,h;\theta)=\sum_{h}\left(\frac{1}{Z(\theta)}\exp(-E\ (v,h;\theta))\right)=\frac{1}{Z(\theta)}\sum_{h}\exp(-E(v,h;\theta)) \quad (15)$$

Where

$pr(v;\theta)$ is the likelihood, w is the weights for hidden layer and visible layer, $bv=\{bv_i\}_{i=1}^{D}$ is the bias of the visible layer, $bh=\{bh_i\}_{i=1}^{J}$ is the bias of the hidden layer, $Z(\theta)$ is the partition function, it means the summing total potential pairs of visible vector and hidden vector, which is represented in Eq.(16).

$$Z(\theta)=\sum_{v}\sum_{h}\exp(-E\ (v,h;\theta)) \quad (16)$$

Where a conditional probability of the hidden layer and "visible layer" can be computed by using the equations (17) and (18) respectively."

$$pc(h_j=1\,|\,v)=\sigma(bh_j+\sum_{i}w_{ij}v) \quad (17)$$

$$pc(v_i=1\,|\,h)=\sigma(bv_i+\sum_{j}w_{ij}h_j) \quad (18)$$

Where $pc(h_j=1\,|\,v)$: the conditional probability for a hidden layer, h : a hidden layer, v : a visible layer, bh : the bias of the hidden layer, w :a weight of the network, $p(v_i=1\,|\,h)$: the conditional probability for the visible layer. σ is the function of a logistic sigmoid which be computed using eq.(19) [24].

$$\sigma(h)=\frac{1}{1+\exp(-h)} \quad (19)$$

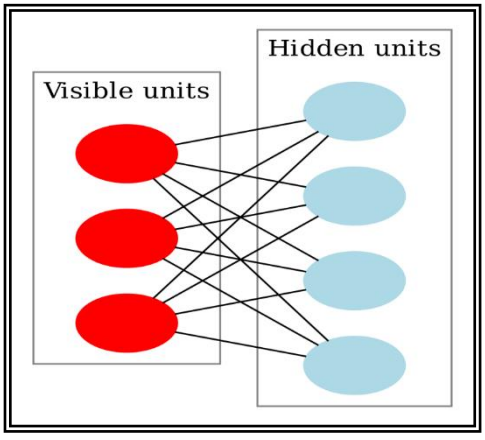

Figure.2 RBM

5.2 Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) for sequence learning tasks and has completed major success on speech recognition systems, machine translation, computer vision, and natural language processing. LSTM does not

have long-term time dependency problems by controlling information flow using the input gate, forget, and output gate. [25].The LSTM is feedbacking the information to every neuron. The feedback used to extend some type of memory that maps concatenation, it uses units called memory blocks, which hold memory cells and participate multiplicative gate units within the block. Fig. (3) displays a structure of the LSTM [26]. Three gates in an LSTM are:

- Input gate (I_t): defines which information is adding to a memory (cell state (S_t)).
- Forget gate (f_t): defines which information is removing from a memory (cell state (S_t)).
- Output gate (O_t): locates which information is using as output from a memory

Each of these gates is offered with the input vector (x_t) at every time step (t) and the output vector (h_{t-1}) of the memory cell at the previous time step ($t-1$). At every time step, the memory cell is updating. In the forward process, the initial step of LSTM is the decision about which information negligence from the previous cell state (S_{t-1}). The decision is taken it by the sigmoid layer that is called forget gate layer, while the outputs of the sigmoid are between (0 and 1) according to the output of the following equation is using for this deciding[27, 28].

$$f_t = sigmoid(W_f x_t + f_{t-1}, I_t] + b_f) \quad (20)$$

Where f_t the values of vector for the activation for the forgot gates, I_t is the input vector. The second step, the layer of LSTM is to identify which information stored in cell state S_t. This step is processed in two parts: firstly, which candidate values $\tilde{S}_t$ that are added to the cell state, this is done using eq.(21). Secondly, the values of the activation function for the input cell are computed using eq. (22).

$$\tilde{S}_t = \tanh(W_{\tilde{s}.x} x_t + W_{\tilde{s}.h} h_{t-1} + b_{\tilde{s}}) \quad (21)$$

$$I_t = sigmoid(W_{i.x} x_t + W_{i.h} h_{t-1} + b_i) \quad (22)$$

The third step, the states of the new cell (S_t) are computed based on the findings of the previous steps with $\circ$ denoting the Hadamard element wise product as shown in Eq. (23)

$$S_t = f_t \circ S_{t-1} + I_t \circ \tilde{S}_t \quad (23)$$

The last step, the memory cell output (h_t) is derived using the following equations:

$$S_t = sigmoid(W_{o.x} x_t + W_{o.h} h_{t-1} + b_o) \quad (24)$$

$$h_t = o_t \circ \tanh(S_t) \quad (25)$$

The training in the LSTM is similar to the traditional NN with type feed forward. The values of the bias and weight adjusted in manner that minimized the loss of the objective function.

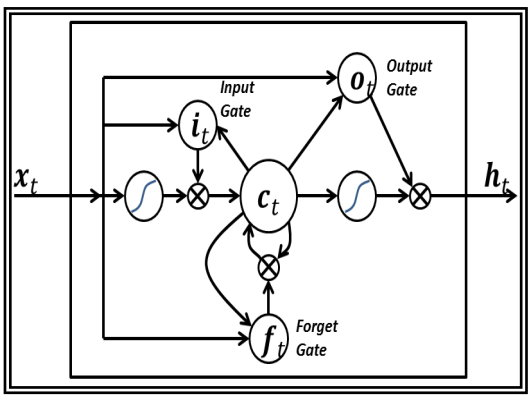

Figure.3 Structure of the LSTM

6. THE PROPOSED METHOD

The techniques of fusion are commonly used in pattern classification. The fusion can be categorized into three types:

6.1 Data fusion

it is combining the data from different sources, which produce the efficient representation of a data. There are some criteria that depend on to classify data fusion such as: - first, according to the input and output the types of data and the nature of the data. Second, according to the relations between the input data source. Third, according to the type of architecture for the data (centralized, decentralized or distributed) [29].

6.2 Feature level fusion

it is extracting the different features form the source of the data, these extracted features have been merged together to get feature fusion (FF) [30].

6.3 Decision level fusion

it is combining the results that get from different classifiers to get the final decision fusion (decision fusion (DF)), also known as score level fusion [30].

7. THE PROPOSED METHOD

The proposed method is finding the classification rate of the source direction for the speech signal in sealed spaces and when used microphone arrays which used for meeting, conference, the conversation of a telephone. Data fusion technique has been applied which used for collecting data from different resources by using 2 datasets to make the system more applicable, each data set recorded in different environments, different sampling rates, various numbers of speakers. The stages of the suggested approach contain three phases: pre-processing, feature extraction, and the classific ation phase.

7.1 The Pre-processing Stage

Because the speech signal recorded in different environments so, this stage is very important to prepare the signal to the feature extraction stage. This stage contains the following operations: reading the sound file with WAV format, each sample represented with 8 bits, normalizing the signal using pre-emphasizing. It utilized to amplify the magnitude of high-frequency range and reduce the magnitude of lower range. The pre-emphasis calculation shown in Eq. (26)[31].

$$H(z) = 1 - az^{-1} \quad (26)$$

Where a is constant, and its value (0.97).

Initially, the speech signal has been blocked into frames, each frame has a number of samples equivalent to 25(ms) and the number of samples in the overlap value equivalent to 10 (ms). The zero-padding utilized in the last frame so it has the same number of the samples. The windowing technique is used for ensuring the continuity of the signal after segmentation and decrease the discontinuities at the starting and end of each frame using a Hamming window which is defined in Eq. (27)[32]:

$$W(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \; 0 \le n \le N\text{-}1 \quad (27)$$

Where n is the Hamming window length, $N$is the samples number in the frame? Fig. (4) Illustrates the steps during the preprocessing stage.

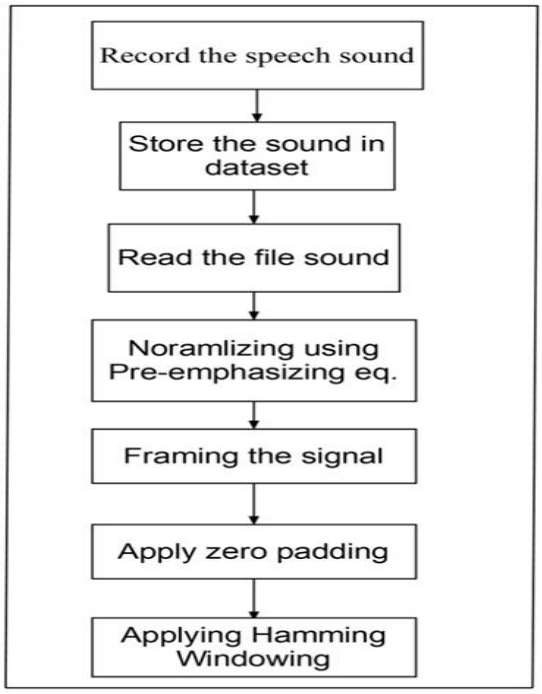

Figure.4 the Preprocessing Stage

7.2 Feature Extraction Stage

In this stage the feature vector constructed, which contains extracting the features for localization by using feature fusion technique, which achieved by concatenation three features in both techniques of the localization are the TDE, and DOA for the speech signal as shown in the following equation:

$$F_{loc} = \underbrace{\{f_{GCC}\}}_{TDE} + \underbrace{\{f_{MVDR} + f_{Root-MUSIC}\}}_{DOA} \quad (28)$$

Where F_{loc} is the features vector of speaker localization, f_{GCC} is the GCC feature for one speaker signal, f_{MVDR} is the MVDR feature for one speaker signal, $f_{Root-MUSIC}$ is the root-MUSIC feature for one speaker. Therefore, the localization features will be $F_{loc} = \{f_{GCC}, f_{MVDR}, f_{Root-MUSIC}\}$ .So after the normalizing of the speech signal.

7.3 The Classification Stage

Feature vectors, which resulted from previous stage, classified using two ANN algorithms RBM by use Tensor flow and LSTM using Keras library. The Tensorflow library is Google's engine, which has a singular method of solving problems. This singular way allows for solving ML problems very efficiently[33]. Keras is an open-source frontend library for NN, that it doings as a backbone for NN, as it has extremely good abilities for forming activation functions, the objective of Keras is to allow high-level manipulation of NN [34]. The data set contains (830) speech signal which were divided into a training set (70%) of the dataset and a testing set (30%) of the dataset. The accuracy equation has been utilized to measure the accuracy of the proposed method, as indicated in Equation (29)[35].

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (29)$$

8. THE SIMULATION AND RESULTS

The suggested method was simulated under the following hypothesis (the room dimension is [4,6], type of the room is a shoebox, the microphone is of a circular design, and used (8) microphones, the distance between microphones is (0.08),

the beam former delay is (0.050). For example, one of the speech signals from the ELSDR dataset has the following information (sampling frequency 16000, the total sample size is (238400), the number of samples is (400), the overlap is (160), number of frames is 1487. Fig. (5) Shows the simulation of the preprocessing stages on the speech signal.

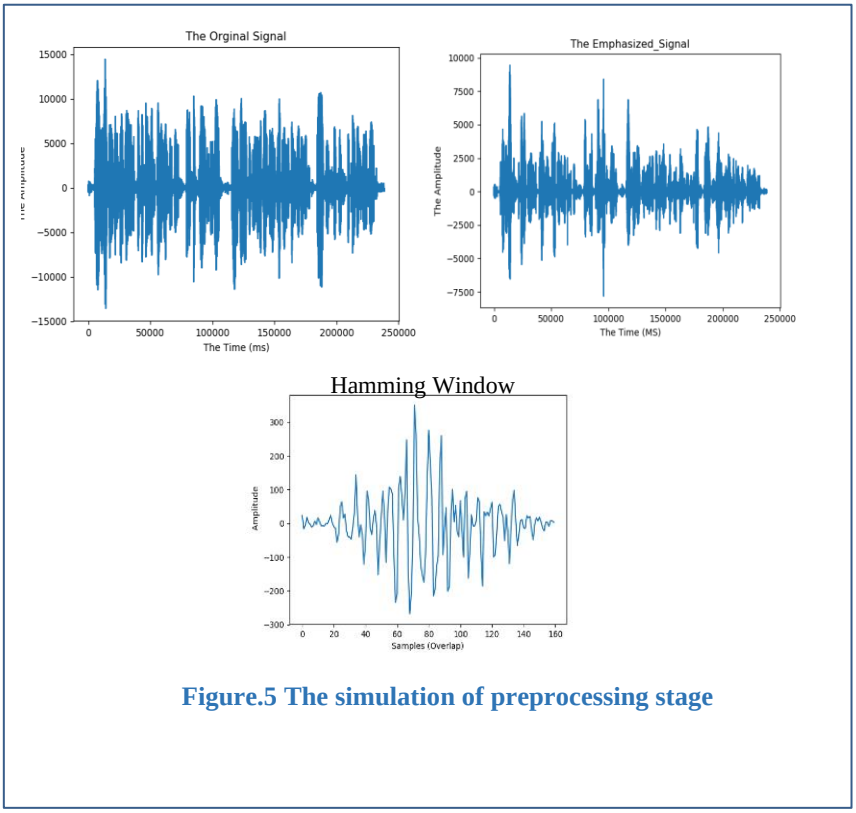

Figure.5 The simulation of preprocessing stage

In the feature extraction part, the noise signal has been generated randomly which equivalent to 2000 samples mixed with the original speech signal. The desired signal has been extracted and the features are extracted for constructing the feature vector for classification, the values of the features are (MVDR value is (19.018), Root MUSIC is (41.84), the mean of GCC vector is (131.713)). Fig. (6) Shows the noised signal and desired signal using MVDR.

The classification was achieved using the RBM neural network implemented using the Tensorflow library. The activation functions, which used in this method, are the sigmoid activation function and activation function for type rectified linear unit (Relu), the value of epoch is (10). While, the classification was achieved using LSTM neural network which was carried out using Keras library with the sequential model. The parameters of the LSTM are (an activation function is sigmoid, The LSTM model consists of (100) neurons, and the optimizer used is Adam. The work implemented using the Python (version 3.5.3) as the script of programming language. The result showed that the RBM algorithm gets a good result, which is (99.84%) while the LSTM algorithm gets (99.14%). By comparison, between the findings of a suggested method and the findings of previous works that mentioned in section 2. There is an improvement in the accuracy of a system compared with previous work. Therefore, the usage of the feature fusion technique and LSTM and RBM had contributed to an increase in the classification rate of the system compared with the methods of the classification in the previous works such as CNN, CRNN.

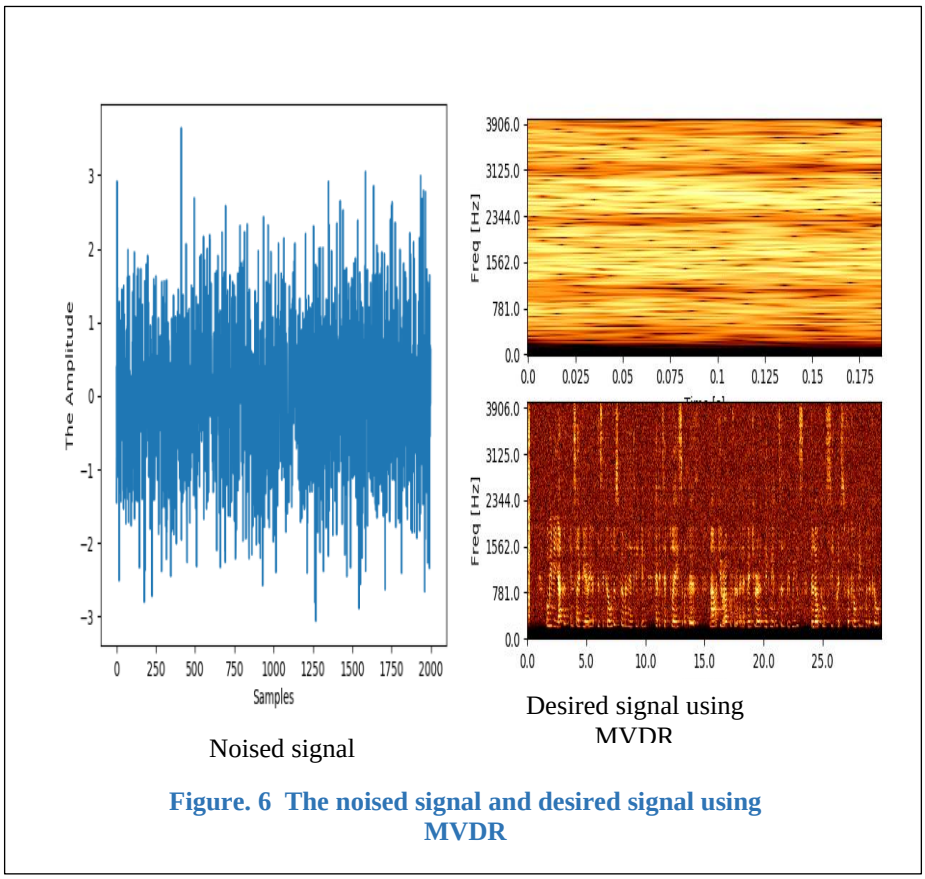

Figure. 6 The noised signal and desired signal using MVDR

9. CONCLUSION

This paper proposed a method for increasing the classification rate of the speaker localization by using fusion techniques and NN algorithms. The fusion techniques used for the localization. Firstly, the technique of data fusion which aggregating data from various resources to make the system more applicable. Secondly, feature fusion technique has been used in the feature extraction stage for constructing the feature vector of the speech signal by extracting the feature in both techniques of localization are the TDE technique using GCC and DOA technique using Root_MUSIC and MVDR. In the classification phase, two methods were been used are RBM based on TensorFlow and LSTM based on Keras for classifying the features vectors of the speech signal. The results presented improvements in the accuracy of a system when using fusion techniques. In addition, RBM had been more accurate compared with another method (LSTM) because it relies on two types of learning are unsupervised learning and supervised learning. In addition, the Tensor flow provided the speed in training of the data, while the Keras library provided more capabilities and simplicity for the achievement of the LSTM.for future work the researcher suggests proposing a system of localization and identification for multi speakers such as meetings, broadcasting news, or conferences. Also, Proposing a system for speaker localization and identification for video files in real time.

10. REFERENCES

[1] T. Kundu, "Acoustic source localization," Ultrasonics, vol. 54, pp. 25-38, 2014.
[2] J. Van Opstal, The auditory system and human sound-localization behavior: Academic Press, 2016.
[3] N. Dey and A. S. Ashour, Direction of arrival estimation and localization of multi-speech sources: Springer, 2018.
[4] C. Lenz, "Localization of Sound Sources, Studies on Mechatronics," PhD Thesis, Autonomous Systems Lab, Swiss Federal Institute of Technology Press, 2009.
[5] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," IEEE Journal of Selected Topics in Signal Processing, vol. 13, pp. 22-33, 2019.

[6] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," Sensors, vol. 18, p. 3418, 2018.
[7] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," 2018.
[8] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," Computer Speech & Language, vol. 49, pp. 83-106, 2018.
[9] A. Cullen, A. Hines, and N. Harte, "Perception and prediction of speaker appeal–A single speaker study," Computer Speech & Language, vol. 52, pp. 23-40, 2018.
[10] L. Feng, "Speaker recognition," Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2004.
[11] T. van Niedek, T. Heskes, and D. van Leeuwen, "Phonetic Classification in TensorFlow," Bachelor's Thesis, Radboud University, 2016.
[12] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-Sliding Generalized Cross-Correlation: A Sub-band Time Delay Estimation Approach," arXiv preprint arXiv:1910.08838, 2019.
[13] T. Padois, "Acoustic source localization based on the generalized cross-correlation and the generalized mean with few microphones," The Journal of the Acoustical Society of America, vol. 143, pp. EL393-EL398, 2018.
[14] C. C. Lai, S. E. Nordholm, and Y. H. Leung, A Study Into the Design of Steerable Microphone Arrays: Springer, 2017.
[15] H. Chen and L. Cao, "Multiple sound source localization using gammatone auditory filtering and direct sound componence detection," in IOP Conference Series: Earth and Environmental Science, 2017.
[16] Q. Huang, R. Hu, and Y. Fang, "Real-valued MVDR beamforming using spherical arrays with frequency invariant characteristic," Digital Signal Processing, vol. 48, pp. 239-245, 2016.
[17] Y. Xiao, J. Yin, H. Qi, H. Yin, and G. Hua, "MVDR algorithm based on estimated diagonal loading for beamforming," Mathematical Problems in Engineering, vol. 2017, 2017.
[18] S. A. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," Signal Processing, vol. 93, pp. 3264-3277, 2013.
[19] F.-G. Yan, S. Liu, J. Wang, and M. Jin, "Two-Step Root-MUSIC for Direction of Arrival Estimation without EVD/SVD Computation," International Journal of Antennas and Propagation, vol. 2018, 2018.
[20] A. Patwari and G. Reddy, "1D direction of arrival estimation using root-MUSIC and ESPRIT for dense uniform linear arrays," in Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference, pp. 667-672, 2017.
[21] Y. C. Liwei HUANG, Huiqin CHEN, "Research of DOA Estimation Based on Modified MUSIC Algorithms," Advances in Engineering Research, vol. 118, 2017.
[22] L. Bao, X. Sun, Y. Chen, G. Man, and H. Shao, "Restricted boltzmann machine-assisted estimation of distribution algorithm for complex problems," Complexity, vol. 2018, 2018.
[23] H. Hu, L. Gao, and Q. Ma, "Deep restricted boltzmann networks," arXiv preprint arXiv:1611.07917, 2016.
[24] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," Journal of Machine Learning Research, vol. 13, pp. 643-669, 2012.
[25] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in Prognostics and Health Management (ICPHM), 2017 IEEE International Conference, pp. 88-95, 2017.
[26] D. B. Silva, P. P. Cruz, and A. M. Gutierrez, "Are the long–short term memory and convolution neural networks really based on biological systems?," ICT Express, vol. 4, pp. 100-106, 2018.
[27] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research, vol. 270, pp. 654-669, 2018.
[28] J. Medina-Quero, S. Zhang, C. Nugent, and M. Espinilla, "Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition," Expert Systems with Applications, vol. 114, pp. 441-453, 2018.
[29] J. C. Dagher, A. Ashok, D. Tremblay, and K. S. Kubala, "Image data fusion systems and methods," Google Patents, OmniVision Technologies Inc, 2014.
[30] H. A. Khan, "Feature Fusion and Classifier Ensemble Technique for Robust Face Recognition," Signal Processing: An International Journal (SPIJ), vol. 11, p. 1, 2017.
[31] M. Kumar and A. G. Singh, "Performance Analysis of LPC and MFCC Techniques in Automatic Speech Recognition," 2015.
[32] P. L. Chithra and R. Aparna, "Performance Analysis of Windowing Techniques in Automatic Speech Signal Segmentation," Indian Journal of Science and Technology, vol. Vol 8, November 2015.
[33] N. McClure, TensorFlow machine learning cookbook: Packt Publishing Ltd, book, 2017.

[34] A. Nandy and M. Biswas, "Reinforcement learning with keras, tensorflow, and chainerrl," in Reinforcement Learning, ed: Springer, 2018, pp. 129-153.
[35] S. Asskali, "Polyp Detection: Effect of Early and Late Feature Fusion,"Master thesis, 2017.