IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 29, 2019, accepted February 18, 2019, date of publication February 25, 2019, date of current version March 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901352

Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network

ISMAIL SHAHIN, ALI BOU NASSIF, AND SHIBANI HAMSA
Department of Electrical and Computer Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates

Corresponding authors: Ismail Shahin (ismail@sharjah.ac.ae), Ali Bou Nassif (anassif@sharjah.ac.ae), and Shibani Hamsa (shibani.h@gmail.com)

This work was supported by the University of Sharjah through the two competitive research projects entitled ‘‘Emotion Recognition in each of Stressful and Emotional Talking Environments Using Artificial Models’’ under Grant 1602040348-P and ‘‘Capturing, Studying, and Analyzing Arabic Emirati-Accented Speech Database in Stressful and Emotional Talking Environments for Different Applications’’ under Grant 1602040349-P.

ABSTRACT This paper aims at recognizing emotions for a text-independent and speaker-independent emotion recognition system based on a novel classifier, which is a hybrid of a cascaded Gaussian mixture model and deep neural network (GMM-DNN). This hybrid classifier has been assessed for emotion recognition on ‘‘Emirati speech database (Arabic United Arab Emirates Database)’’ with six different emotions. The sequential GMM-DNN classifier has been contrasted with support vector machines (SVMs) and multilayer perceptron (MLP) classifiers, and its performance accuracy is indexed at 83.97%, while the other two perform at 80.33% and 69.78% using SVMs and MLP, respectively. These results demonstrate that the hybrid classifier significantly gives higher emotion recognition accuracy than SVMs and MLP classifiers. Our GMM-DNN model yields the results similar to those obtained by human judges in a subjective assessment context. Also, the performance of the classifier has been tested using two distinct emotional databases and in normal and noisy talking conditions. The dominant signal mask provided by the hybrid classifier offers better system performance in the presence of noisy signals.

INDEX TERMS Deep neural network, emotion recognition, Gaussian mixture model.

I. INTRODUCTION

Emotion is regarded as both a mental and physiological state. Gestures, facial expressions and speech can convey human emotions. Speech emotion is the major method of interaction among human beings, but it is not confined to linguistic statements as it contains emotional content that is critical to human interaction [1]. Hence, emotion recognition is critical. Recognizing emotions of speakers is an integral part of an intelligent human computer interaction system [2]. Emotion recognition is often utilized for intelligent security in smart banking, clever customer care, criminal investigations, robotics, smart education, ranking voice mail messages according to emotion [3], operator performance assessment [3], and distant logging on a server or accessing private library files on a server [1], [4].

Even a human intelligent system fails to offer 100% accuracy in classifying emotions in speech due to subjectivity; consequently, it is excessive to assume that a machine is capable of giving a more accurate classification. Two challenges for speech emotion recognition stem from i) scarcity of natural emotional speech datasets and ii) low accuracy rates of employed classifiers [5].

This study aims at developing a sequential GMM-DNN classifier to enhance emotion recognition accuracy (text-independent and speaker-independent) using Arabic speech dataset in Emirati accent. This database has been collected in this work to evaluate GMM-DNN. Furthermore, four experiments have been conducted to assess the GMM-DNN classifier.

This paper unfolds as follows: First, literature review is presented in Section II. Then, the description of the ‘‘Emirati Speech Database (ESD)’’ is provided in Section III. The information of feature extraction is covered in Section IV. The model description of the proposed hybrid classifier GMM-DNN is explained in Section V. Emotion recognition algorithm based on GMM-DNN is given in Section VI. The attained results along with further conducted experiments are presented in Section VII. Finally, conclusions are given in Section VIII.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

II. LITERATURE REVIEW

Alonso *et al.* [6], Luengo *et al.* [7], and Cao *et al.* [8] extracted spectral, prosody, and pitch characteristics of the Berlin Emotional Speech (BES) database and used the Support Vector Machines (SVMs) as a classifier. Alonso *et al.* [6] obtained 94.9% emotion recognition accuracy using five emotions: "anger, happiness, neutrality, boredom, and sadness". Luengo *et al.* [7] reported 78.3% emotion recognition accuracy utilizing seven emotional states: "anger, boredom, disgust, fear, happiness, neutrality, and sadness". Cao *et al.* [8] achieved 82.1% emotion recognition accuracy using seven emotions: "anger, disgust, fear, happiness, neutrality, sadness, and boredom". Wang *et al.* [9] reported an emotion recognition accuracy of 88.8% using prosody features in an SVM based classification. They used six distinct emotions: "happiness, sadness, anger, boredom, anxiety, and neutrality". A few studies used the "Speech Under Simulated and Actual Stress (SUSAS) database" with a "Hidden Markov Model (HMM)" as a classifier [10]–[12]. Shahin and Ba-Hutair [10] used "Mel Frequency Cepstral Coefficients (MFCCs)" for feature extraction and HMM as a classifier. They attained an average recognition accuracy of 76.3% using six different talking conditions: "neutrality, anger, slowness, loudness, softness, and fastness") based on "Second-Order Circular Suprasegmental Hidden Markov Models (CSPHMM2s)". Shukla *et al.* [11] used 13 dimensional features of the SUSAS database and HMM as a classifier to achieve a 93.9% speaker-dependent talking condition recognition performance in four diverse talking conditions: neutral, angry, sad, and Lombard conditions. Deb and Dandapat [12] obtained a 72.8% speaker-dependent emotion recognition accuracy using the breathiness features and MFCC of the SUSAS database and HMM as a classifier. They experimented in five different stress conditions: "anger, happiness, Lombard, neutrality, and sadness". Shahin devoted in one of his work [13] on investigating and improving "talking condition recognition in stressful and emotional environments (completely two separate environments)" based on three distinct and independent classifiers. These classifiers are: "HMMs, Second-Order Circular Hidden Markov Models (CHMM2s), and Suprasegmental Hidden Markov Models (SPHMMs)". The "stressful talking environments" used in his work are comprised of "neutral, shouted, slow, loud, soft, and fast talking conditions", while the "emotional talking environments" are made up of "neutral, angry, sad, happy, disgusted, and fear emotions". The reported results in his work demonstrate that SPHMMs lead each of HMMs and CHMM2s in enhancing talking condition recognition in stressful and emotional environments. In another work [14], Shahin improved emotion recognition performance by merging emotion recognizer and gender recognizer into one recognizer combining both HMMs and SPHMMs as classifiers. He achieved 86.8% as an average emotion recognition performance using six basic emotions: "neutrality, anger, sadness, happiness, disgust, and fear".

Emotion recognition based on Gaussian Mixture Model (GMM) has been studied in many research [15], [16]. Cheng and Duan [15] classified five emotional states: neutrality, happiness, anger, sadness, and surprise based on GMM. They combined 60 basic features to generate the feature vector. Then, the features that were extracted by Principal Component Analysis (PCA) were sent into the improved GMM to be classified and recognized. Their results demonstrated that the chosen features are efficient for emotion recognition [15]. El Ayadi *et al.* [16] proposed Gaussian Mixture Vector Autoregressive Model (GMVAM) for emotion recognition. They assessed their proposed statistical classifier on Berlin emotional speech dataset. They reported emotion recognition accuracy of 76% using six different emotions: "neutrality, anger, fear, happiness, boredom, and sadness" [16].

Deep Neural Network (DNN) has been utilized as a classifier in many studies of emotion recognition [17]–[19]. Stuhlsatz *et al.* [17] introduced and used a "Generalized Discriminant Analysis (GerDA)" based on DNN to identify unknown emotions. Their results, averaged over nine different speech databases, demonstrated greatly significant emotion recognition enhancement compared to SVMs for the two-class arousal and valence. Kun *et al.* [18] proposed recognizing speech emotions using DNN and extreme learning machine and obtained 20% accuracy improvement compared to other approaches such as HMMs and SVMs. They evaluated their approach using five distinct emotions of the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. The five emotions are neutrality, happiness, surprise, excitement, and frustration. Zheng *et al.* [19] introduced a systematic framework to apply an effectively emotion recognition system based on Deep Convolution Neural Networks (DCNNs) using labeled training audio data. They achieved, based on DCNNs with two convolution and two pooling layers, 40% emotion recognition accuracy using IEMOCAP corpus with five various emotions: neutrality, happiness, surprise, excitement, and frustration [19].

Some multiple classifier schemes have been proposed, utilized, and assessed for speech emotion recognition [20]–[22]. Li *et al.* [20] proposed Deep Neural Network Hidden Markov Models (DNN-HMMs) for speech emotion recognition. They explored their proposed classifier with each of "Restricted Boltzman Machine (RBM)" based unsupervised pre-training and discriminative pre-training. They tested their experiments on eNTERFAC'05 (using "anger, happiness, sadness, fear, surprise and disgust") and Berlin (using "anger, boredom, disgust, fear, happiness, sadness, and neutrality") databases using DNN-HMMs with RBM based unsupervised pre-training and discriminative pre-training, respectively. Their results demonstrated that when the number of hidden layers as well as hidden units are appropriately set, the DNN-HMMs could expand the labeling capability of GMM-HMMs. Hence, among all the models, the DNN-HMMs with discriminative pre-training achieve the optimum results [20]. Huang *et al.* [21] introduced a combined classifier that is made up of Deep Belief Network (DBN) and Support Vector Machine (SVM). Emotion recognition accuracy based on their combined classifier and using four different emotions (anger, surprise, happiness, and sadness) is 86.5% [21]. Tashev *et al.* [22] investigated combining a ''GMM-based low-level feature extractor'' with a neural network that serves as a high level feature extractor. The benefit of their suggested framework is that it combines the quick growing neural network-based solutions with the classic statistical methods implemented to emotion recognition. Their proposed architecture was evaluated on a Mandarin database with four emotions only: neutral, happy, sad, and angry. Their results, based on GMM-DNN, gave weighted and un-weighted emotion recognition accuracy of 48.0% and 41.5%, respectively [22].

In this research, we focus on recognizing text-independent and speaker-independent emotions using Arabic speech corpus in Emirati accent based on a proposed hybrid classifier called cascaded GMM-DNN (GMM followed by DNN). The current research is different from [23]. In [23], Shahin *et al.* proposed, implemented, and tested GMM-DNN as a classifier for text-independent speaker identification in emotional talking environments. Our main contribution in this work clearly appears in utilizing Emirati database to assess a novel classifier that it is combined and integrated from both GMM and DNN to recognize emotions. To the best of our knowledge, this work is the first effort to recognize emotions using Emirati-accented dataset based on a cascaded GMM-DNN classifier. Furthermore, we conducted empirical evaluation among different classifiers such as GMM-DNN, DNN-GMM, DNN alone and GMM alone. In addition, we conducted four experiments to evaluate the proposed GMM-DNN model for emotion recognition as follows:

1. In experiment 1, we evaluate our proposed GMM-DNN classifier on the SUSAS dataset which is a public English dataset [24].
2. In experiment 2, a ''subjective assessment'' of results based on GMM-DNN utilizing the Emirati speech database (ESD) has been carried out with ten non-professional Arabic audience members (human judges).
3. In experiment 3, the system performance has been assessed using two distinct emotional databases in both normal and noisy talking conditions.
4. In experiment 4, GMM-DNN has been contrasted using ESD (local database) with DNN-GMM, GMM alone, and DNN alone for emotion recognition.

III. EMIRATI SPEECH DATABASE

In this paper, we build an Emirati speech database (ESD) to evaluate GMM-DNN for emotion recognition. A group of local Emirati speakers (15 men and 15 women of ages ranging between 14 and 55 years) participated to construct ESD using the ''Emirati Arabic-emphasized speech database''. Eight sentences commonly used in the UAE society were spoken by every speaker 9 times in various emotions: ''neutrality, happiness, sadness, disgust, anger, and fear'' with a range of 2 – 5 seconds. These speakers were not trained to avoid falsified expressions. Table 1 shows the eight sentence; in the column on right side is the Emirati version and in the left column is the English translation. This database was collected in two different scheduled sessions: ''training session and testing session''.

TABLE 1. Emirati dataset and its English version.

No.	English Translation	Emirati Accent
1.	I'm leaving now, may God keep you safe.	فداعة الرحمن بترخص عنكم الحينه.
2.	The one whose hand is in the water is not the same as he/she whose hand is in fire.	اللي ايده في الماي مب نفس اللي ايده في الضو.
3.	Where do you want to go today?	وين تبون تسيرون اليوم؟
4.	The weather is nice, let's sit outdoors.	قوموا نيلس في الحوي , الجو غاوي برع.
5.	What's in the pot, the spoon gets out.	اللي في الجدر يطلعه الملاس.
6.	Welcome millions, and they are not enough.	مرحبا ملايين ولا يسدن.
7.	Get ready, I will pick you up tomorrow.	زهب عمرك بخطف عليك باجر.
8.	He/she who doesn't know the value of the falcon, will grill it.	اللي ما يعرف الصقر يشويه.

The recording took place in the ''United Arab Emirates at the University of Sharjah, College of Communication''. An acquisition board with a ''16-bit linear coding analog-to-digital converter'' has been used to sample the captured speech signals at a 44.6 kHz frequency. Then, the signals were down-sampled to 16 kHz, pre-emphasized and divided into frames of 25 ms a piece with 31.25% overlap between consecutive frames. Typical frame sizes in speech processing range from 20 ms to 40 ms with 50% (+/ − 20%) overlap between consecutive frames.

IV. FEATURE EXTRACTION

Mel Frequency Cepstral Coefficient (MFCC) is the most commonly used feature extraction techniques in speaker [25], [26] and emotion [27], [28] recognition. MFCC gives the logarithmic perception of onset and pitch of the human auditory system. The computation of MFCC is shown in the block diagram of Fig. 1 [29].

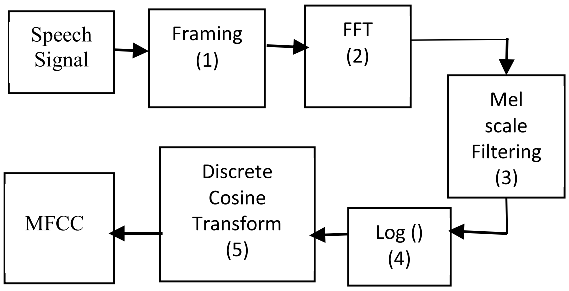

FIGURE 1. Block diagram of MFCC algorithm.

Mel frequency m is computed from the normal frequency (f) as [29],

$$m = 2595\, log(1 + \frac{f}{100}) \quad (1)$$

In this study, Mel frequency has been computed using the following five steps [29]:

1. The first step is to divide the signal into 25 ms frames. Then, the frame length for a 16 kHz input signal is S(n) = 0.02 × 516000 = 400 samples. In this method of framing, S(n) is transformed into $S_i(n)$, where *i* shows the frame number.
2. The second step is to take the "Fourier transform" of the framed signal, $S_i(n)$. $S_i(k)$ can be found as,

   $$S_{I}\,(k) = \Sigma_{n=1}^{k} S_i\,(n)\,h\,(n)\,e^{-j2\pi kn/N},\ 1 \leq \mathrm{k} \leq \mathrm{N} \quad (2)$$

   where "*h*(*n*) is the impulse response of the Hamming window and k is the DFT length". "The power spectral estimate" of the signal $S_i(n)$ is,

   $$P_i(k) = \frac{1}{N}\,[S_i(k)]^2 \quad (3)$$

3. The third step is to enumerate the "Mel spaced filter bank". This consists of a group of 25 triangular filters that can be implemented to the "periodogram power spectral estimate" which designates the energy level in each filter bank.
4. The fourth step is to compute the Log values of the 26 filter bank energies of *Step*3.
5. The last step is to compute the "Discrete Cosine Transform (DCT) of the 26 filter bank energies" of *Step 4* to acquire "Spectral Coefficients".

V. MODEL DESCRIPTION

Studies that involve recognizing emotions utilize diverse classifiers, algorithms, and models at the classification stage to spot the constancy in the classification outcomes and to select the optimum classifier for a particular distinctive attribute [30]. Morrison *et al.* [31] tested different classification methods: "SVM, Multilayer Perceptron (MLP), k-Nearest Neighbor (k-NN), Stacking C, and vote". The most often used classifiers for the recognition of stress and emotion are: k-NN, SVM, GMM, HMM, and MLP. Of all classifiers, GMM classifier is popular for "speaker identification and language recognition" since the classification in GMM is considered as computationally efficient. Furthermore, GMM yields better approximation for randomly formed densities [32]. We, therefore, decided to use a hybrid model that consisted of both GMM and DNN for the recognition of emotions.

A. GMM MODEL

The use of "Gaussian Mixture Model (GMM)" in forming emotion identification is inspired by the interpretation that the elements of Gaussian characterize some general emotion-dependent spectral shapes and the ability of Gaussian mixtures to express random densities. The Gaussian mixture emotion model is interpreted as a non-parametric and multivariate pdf model which has the capability to model random feature distributions [32].

The "Gaussian mixture density model" is represented as the weighted sum of *M* component densities as shown in Fig. 2 [32]. The following equation defines the Gaussian Mixture Density [32],

$$P(\overline{x}|\lambda) = \sum\nolimits_{i=1}^{M} P_i b_i(\overline{x}) \quad (4)$$

where $\overline{x}$ is the D-dimensional random vector, $b_i\,(\overline{x})$ represents the component densities for $i = 1, \ldots, M$, P_i represents the component probabilities, and λ is the GMM tag.

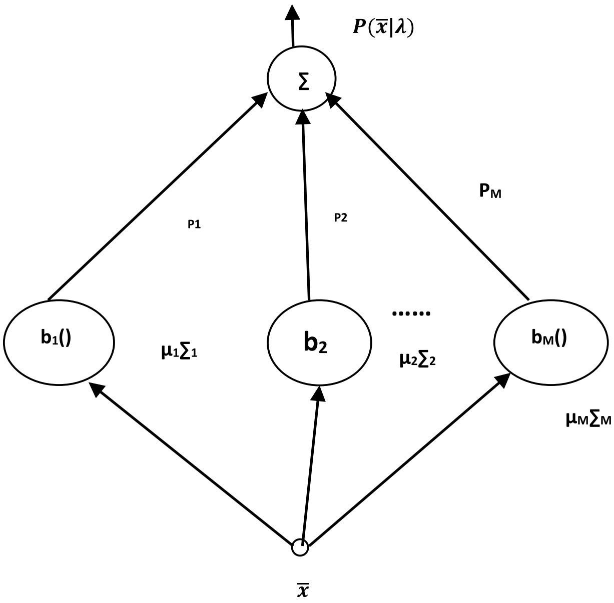

FIGURE 2. GMM model [32].

The "component density" can be expressed as [32],

$$b_i\,(\overline{x}) = \frac{1}{2\pi^{D/2}|\sum_i|^{1/2}} \exp\left\{\frac{-1}{2}(\overline{x} - \overline{\mu)}' \sum\nolimits_i^{-1} (\overline{x} - \overline{\mu}_i)\right\} \quad (5)$$

The GMM tag is represented by the "Gaussian mixture density parameters mean $\overline{\mu}_i$, covariance Σ_i , and the mixture weights p_i".

$$\lambda = \left\{P_i, \overline{\mu}_i, \Sigma_i\right\} \quad \text{where i} = 1, \ldots, M \quad (6)$$

The "feature vectors" are obtained from the test speech signals for emotion recognition, which are then partitioned into intersecting segments of *T* feature vectors. In this research, we apply the following steps to train models:

1. "GMM training" is initialized with the beginning tag λ.
2. Calculate the next tag $\overline{\lambda}$, thus, $p(X|\overline{\lambda}) \geq p(X|\lambda)$.
3. Replicate the method to get the convergence [32],

$$p(i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \quad (7)$$

The ''mixture weights'' are described as,

$$\overline{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda) \tag{8}$$

The mean is given by,

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} \tag{9}$$

The variance is defined as,

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} - \overline{\mu}_i^2 \tag{10}$$

The utterance set $S = \{1, 2, \ldots, s\}$ is symbolized by GMM tags: $\lambda_1, \lambda_2, \ldots, \lambda_s$.

In the next step, the classification task consists of choosing the class with the highest probability which is defined as,

$$\hat{S} = \underset{1\le k\le S}{\arg\max}\, P(\lambda_k|X) = \underset{1\le k\le S}{\arg\max}\, \frac{p(X|\lambda_k)P(\lambda_k)}{p(X)} \tag{11}$$

B. DEEP NEURAL NETWORK

Neural networks are a set of algorithms inspired by how the human brain recognizes patterns. Because of clustering and classification properties, neural networks include a broad span of applications in the development of machine learning. Neural networks help to group unlabeled data according to similarities among the samples. In some situations, in order to get a more precise classification, the features extracted by neural networks may be processed by other algorithms or vice versa. This shows the importance of deep neural networks to machine learning [33].

Incorporating hidden layers with a vast number of neurons in a DNN has proven to greatly enhance the modeling capabilities of the DNN and hence found many closely optimal configurations [34]. Even in the case where parameter learning was trapped into a local optimum, the subsequent DNN is still able to perform quite well since the possibility of having a poor local optimum becomes lower and lower as the number of neurons used is large. However, utilizing deep neural networks would necessitate high computational power during the training phase. Since massive computational capabilities were not easily available in the past, it was not until current years that researchers have begun seriously studying and employing deep neural networks.

There are many different deep learning algorithms, two of these popular algorithms are: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [34].

DNN is ''a feed-forward, artificial neural network'' with multiple layers of hidden units between its inputs and outputs. The vector $_{\mathrm{xj}}$ is the input to the hidden layer. The activation function of the hidden layer is a logistic function that converts the input vector to a scalar state yj which is then sent to the next step as given below [33],

$$y_j = logistics(x_j) \tag{12}$$

$$= \frac{1}{1+e^{-x_j}} \tag{13}$$

$$x_j = b_j + \sum y_j w_{ij} \tag{14}$$

where i is an index that represents the lower layers, w_{ij} is the weight between the layers i and j. Then, the classification probability is defined as,

$$P_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)} \tag{15}$$

where k represents the overall index.

VI. EMOTION RECOGNITION ALGORITHM BASED ON CASCADED GMM-DN

Fig. 3 shows the basic training and testing procedure for the cascaded GMM-DNN based emotion recognition system

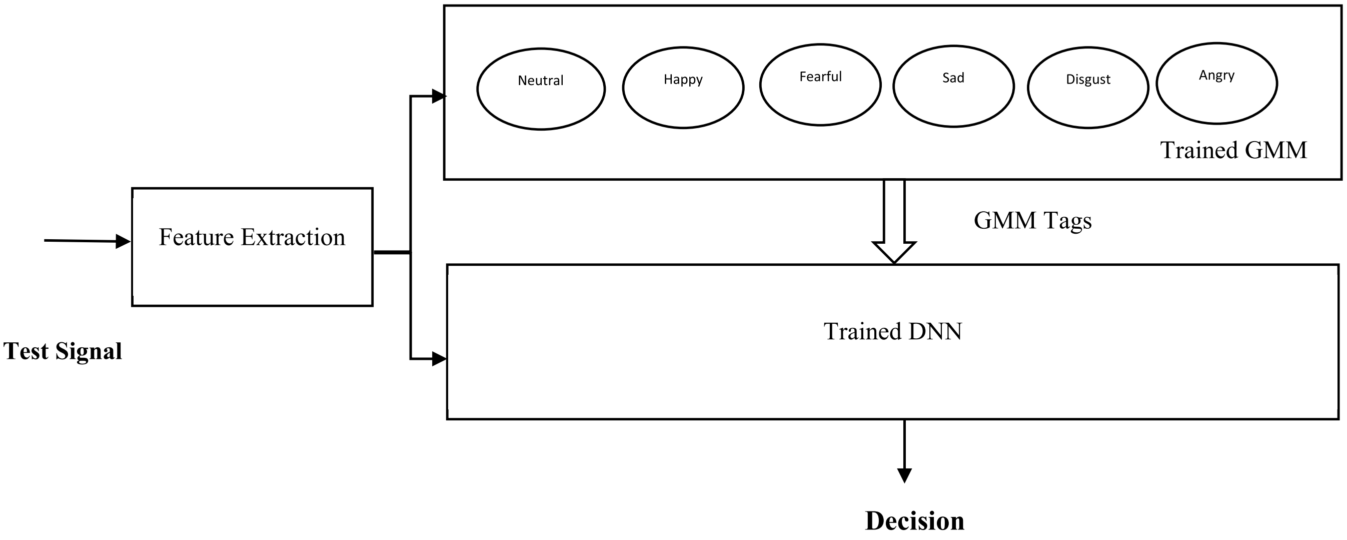

FIGURE 3. Block diagram of emotion recognition system based on GMM-DNN.

in this research. In the training phase, one reference model per emotion has been obtained utilizing twenty out of the thirty speakers talking the first 4 sentences of ESD where every sentence is spoken 9 times. Hence, every emotion is characterized by one "reference emotion model". Therefore, the total number of speech samples utilized in this phase is 4,320 (first 20 speakers × first 4 sentences × 9 repetitions Œ6 emotions). The MFCC features are obtained from each speech sample at the feature extraction stage. The classifier stage followed by the feature extraction stage is based on cascaded GMM-DNN model. The features extracted from the training dataset are used to train the cascaded GMM-DNN based classification.

In the "testing phase", the entire number of speech samples utilized is 2,160 (last (remaining) 10 speakers × last 4 sentences × 9 repetitions ×6 emotions). Therefore, our work is a text-independent and speaker-independent emotion identification problem. The MFCC features are obtained from each speech sample at the feature extraction stage of the "testing phase". The "log likelihood distance" between the training features and the "GMM tag" is competed to identify the emotional state and, thus, produces a recent group of features using the DNN classifier for the final decision.

The proposed GMM-DNN classifier is designed in a cascaded structure. GMM recognition is based on the log probability. In the training phase, the GMM classifier evaluates and stores the log probability of the voice vectors used for training. In the testing phase, the log probability of the test samples is competed with the stored data and assigned a binary 0 or a binary 1 to each emotion, which is referred to as a GMM Tag. These GMM tags are fed into the DNN whose input layer is based on the size of the GMM output. Using the ESD, six emotions are evaluated. Hence GMM outputs six GMM tags for every test speech signal and used it as the input of DNN. The DNN that has been used is a convolutional neural network with four hidden layers with 256 "rectified linear hidden units and the gradient descend method" to learn the "weights in DNN". The trained DNN yields a "probability distribution P" across the entire range of emotions. Next, the "decision block" chooses the specific model that has the maximum probability value.

VII. RESULTS AND DISCUSSION

This paper proposes a new hybrid classifier named GMM-DNN for emotion recognition using a collected Emirati speech dataset. Fig. 4 illustrates emotion recognition accuracy of the GMM-DNN classifier which performs almost ideally in each of neutral and happy emotional situations. GMM-DNN yields poor accuracy for "disgust" emotion. Based on this figure, the average emotion recognition accuracy using the novel GMM-DNN classifier is 83.97% over all the six emotions of Emirati speech database (ESD). Our attained average emotion recognition accuracy using six emotions is greater than the weighted and un-weighted emotion recognition accuracy reported by Tashev *et al.* [22] using only four emotions by 74.9% and 102.3%, respectively.

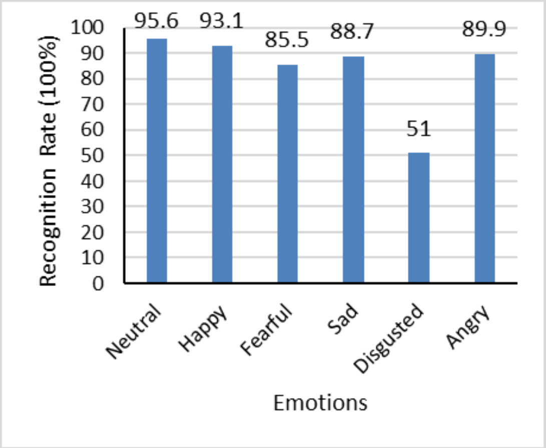

FIGURE 4. Emotion recognition accuracy assessment of the proposed GMM-DNN classifier using ESD.

It is evident that our proposed classifier is superior to their framework.

Table 2 presents a confusion matrix which represents a percentage of confusion of an unknown (to be recognized) emotion with the remaining emotions based on GMM-DNN using ESD. The rate of confusion is less for neutral, happy, sad, fearful, and angry emotional conditions. The table shows poor performance for "disgust" emotion. This is because the disgust emotion is greatly confused with fear and anger emotions (total confusion percentage is 49%). On the other hand, the disgust emotion is not confused at all with neutrality, happiness, or sadness.

TABLE 2. Confusion percentage of an unknown (to be recognized) emotion with the other emotions based on GMM-DNN using ESD (%).

Emotion	Neutral	Happy	Fearful	Sad	Disgusted	Angry
Neutral	95.6	1.9	0	0	0	0
Happy	2.1	93.1	0	0	0	1.3
Fearful	0	0	85.5	0	24.9	3.9
Sad	2.3	0	2.5	88.7	0	4.9
Disgusted	0	2.1	7.9	11.3	51	0
Angry	0	2.9	4.1	0	24.1	89.9

The proposed classifier technique is compared with MLP and SVM classifiers. The MLP used is a feed forward neural network classifier trained with static back propagation algorithm. The utilized SVM classifier classifies the data through a set of support vectors. This helps to minimize the structural complexity with a minimum average error. Table 3 and Table 4 illustrate "confusion matrices" in view of MLP and SVMs as classifiers, respectively. Analysis of Tables 2, 3 and 4 shows that the proposed classifier yields a significant improvement in emotion recognition performance compared to that using MLP and SVM. The performance analysis of GMM-DNN, MLP, and SVM is graphically represented in Fig. 5. This figure demonstrates that the

TABLE 3. Confusion percentage of an unknown (to be recognized) emotion with the other emotions based on mlp using ESD (%).

Emotion	Neutral	Happy	Fearful	Sad	Disgusted	Angry
Neutral	76.4	14	0	0	0	0
Happy	12.8	75.9	0	0	0	0
Fearful	10.8	2.9	65.1	0	0	0
Sad	0	0	11.1	84.3	32.5	11.1
Disgusted	0	3.6	13.9	15.7	45.7	17.6
Angry	0	2.6	9.9	0	21.8	71.3

TABLE 4. Confusion percentage of an unknown (to be recognized) emotion with the other emotions based on SVMs using ESD (%).

Emotion	Neutral	Happy	Fearful	Sad	Disgusted	Angry
Neutral	90.3	3.2	0	0	0	0
Happy	9.7	87.9	0	0	0	4.8
Fearful	0	0	81.2	2.7	1.7	5.3
Sad	0	0	5.1	87.7	39.9	0
Disgusted	0	3.2	6.9	9.6	49.9	4.9
Angry	0	5.7	6.8	0	8.5	85.0

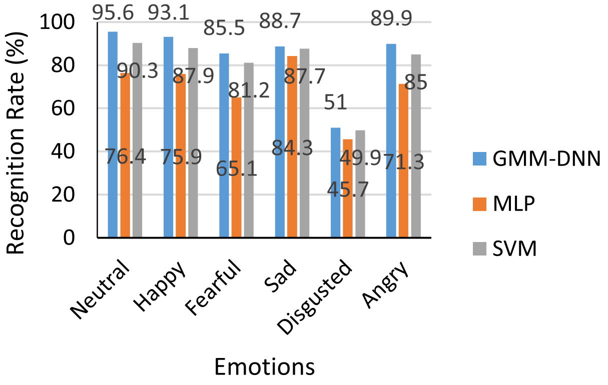

FIGURE 5. Emotion recognition accuracy assessment using ESD based on GMM-DNN, MLP, and SVM.

average emotion recognition accuracy is 83.97%, 80.33%, and 69.78% based, respectively, on GMM-DNN, SVM, and MLP. Therefore, GMM-DNN leads each of SVM and MLP for emotion recognition.

A "statistical significance test" has been implemented in this work to demonstrate whether emotion recognition performance differences (emotion recognition accuracy based on GMM-DNN and that based on each of SVM and MLP) are actual or easily arise from statistical variations. This evaluation is carried out utilizing the "Student's t Distribution test" given by [35],

$$t_{1,2} = \frac{\overline{X}_1 - \overline{X}_2}{SD_{Pooled}} \tag{16}$$

where $\bar{x}_1$ "is the mean of the first sample of size n, $\bar{x}_2$ is the mean of the second sample of the same size, and $\mathrm{SD}_{\mathrm{pooled}}$ is the pooled standard deviation of the two samples given as" [35],

$$SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}} \tag{17}$$

where "SD_1 is the standard deviation of the first sample of size n and SD_2 is the standard deviation of the second sample of equal size".

In this research, the "computed t values" between the proposed GMM-DNN classifier and each of "SVM and MLP" using the ESD are arranged in Table 5. We can notice that every "figured t value" is higher than "the tabulated critical value $_{t0.05}$ = 1.645 at 0.05 significance level" [35]. Hence, emotion recognition performance based on GMM-DNN significantly leads that based on " SVM and MLP".

TABLE 5. Calculated t values between GMM-DNN and each of SVMs and MLP utilizing ESD.

"t $_{1,2}$"	"Calculated t value"
t $_{\text{GMM-DNN, SVMs}}$	1.753
t $_{\text{GMM-DNN, MLPs}}$	1.892

The Studen's t-test is a parametric test that can be used when data is normally distributed. In our classifiers, the distribution of data is skewed and in this case it would be better to use non-parametric tests. Parametric tests compare models based on the mean; however, non-parametric tests contrast models based on the medians. To conduct a thorough comparison between the proposed models, we conducted two non-parametric tests, Kruskal Wallis [36] and Wilcoxon [37] tests, in addition to the parametric t-test. The Kruskal Wallis test compares the three models at once. On the other hand, the Wilcoxon test compares two models to see if they are statistically different.

In this work, our results show that, based on the Kruskal Wallis test, the three models are statistically different at "95% confidence level" (p-value = 0.022) [36]. The results of the Wilcoxon test [37] are presented in Table 6.

TABLE 6. Non-parametric Wilcoxon test.

Classifier	GMM-DNN	SVM	MLP
GMM-DNN	NA	0.037	0.037
SVM	0.037	NA	0.054
MLP	0.037	0.054	NA

The results demonstrate that GMM-DNN is statistically different (shaded cells) from SVM, as well as MLP based on "95% confidence level" (p-value = 0.037).

However, the SVM model is not statistically different from the MLP model at the "95% confidence level" (p-value = 0.054). Based on the t-test, Kruskal Wallis and Wilcoxon tests, it is noticed that the proposed GMM-DNN model is statistically different from the SVM and MLP models.

Four additional experiments have been independently executed to analyze the attained emotion recognition accuracy based on GMM-DNN. The four experiments are

Experiment 1: In this experiment, the "SUSAS" database has been utilized to test the three classifiers for stressful talking condition recognition. The major purpose of the "SUSAS" database was initially for speech processing in "neutral and stressful speaking environments" [28]. The stress domain of the SUSAS includes: "i) talking styles (slow, fast, soft, loud, angry, clear, question), ii) single tracking task or speech produced in noise (Lombard effect), iii) dual tracking computer response task, iv) actual subject motion-fear tasks (G-force, Lombard effect, noise, fear), and v) psychiatric analysis data (speech under depression, fear, anxiety)" [24]. "Angry and shouted talking conditions" are utilized as substitutes since, in our life, it is not easy to isolate them [38]. Thirty diverse utterances of seven speakers uttered in each of "neutral and stressful talking conditions (angry, slow, loud, soft, and fast)" are chosen to assess each of "GMM-DNN, SVM, and MLP" for stressful talking condition recognition.

Percentage of confusion matrix utilizing "SUSAS dataset based on GMM-DNN, SVM, and MLP" is shown in Table 7. Using this table, the accuracy of stressful talking condition recognition based on "GMM-DNN, SVM, and MLP" is 86.67%, 76.50%, and 77.00%, respectively. These numbers clearly show that GMM-DNN performance is greater than that of SVM and MLP for stressful talking condition recognition using the SUSAS database by 13.3% and 12.6%, respectively.

Experiment 2: A "casual and subjective assessment" of GMM-DNN using the ESD has been accomplished with ten non-professional "Arabic audience members (human judges)". A total of 120 speech samples (5 speakers × 4 sentences × 6 emotions) are utilized in this training phase. The "assessment phase" progresses independently with every listener instructed to recognize the unknown emotion. The graphical illustration of this accuracy analysis based on the "subjective evaluation" is demonstrated in Table 8. The confusion matrix illustrates that human listener accuracy is similar to GMM-DNN performance except for the disgust emotion. The analysis based on the average emotion recognition accuracy of the novel GMM-DNN classifier and human judges is given in Fig. 6. The average emotion recognition accuracy based on the subjective assessment is 89.20% which is close to that attained based on GMM-DNN (83.97%).

Experiment 3: During the "evaluation phase", test data is mixed with some noise in a ratio 2:1 and the achieved results are displayed in Table 9. Arbitrarily picked speech samples from every emotion mixed with interference signal are used as the test data and, thus, the final outputs are competed with the other classifiers. The interference used in this experiment is mixed with other male and female voiced speech signals with reduced dominance level (two speech mixtures and 3 speech mixtures), white noise, siren noise, and telephone noise.

Normal speech and distorted data are used in the testing phase. Distorted data is obtained by mixing the original speech with various noises at a dominance level of 2:1. Each speech sample is mixed with other male speech sample, other female speech, white noise, siren noise, telephone noise,

TABLE 7. Confusion percentage of an unknown (to be recognized) emotion with the other emotions based on GMM-DNN, SVM, and MLP using SUSAS database (%).

			GMM-DNN			
"Talking condition"	**"Neutral"**	**"Angry"**	**"Slow"**	**"Loud"**	**"Soft"**	**"Fast"**
Neutral	98	2	1	0	0	0
Angry	0	77	2	11	0	10
Slow	1	1	84	0	12	0
Loud	0	13	1	87	0	4
Soft	1	3	10	0	88	0
Fast	0	4	2	2	0	86
			SVM			
"Talking condition"	**"Neutral"**	**"Angry"**	**"Slow"**	**"Loud"**	**"Soft"**	**"Fast"**
Neutral	95	4	4	2	1	1
Angry	1	67	5	15	2	13
Slow	1	1	71	2	17	1
Loud	1	18	2	73	2	6
Soft	1	2	14	3	75	1
Fast	1	8	4	5	3	78
			MLP			
"Talking condition"	**"Neutral"**	**"Angry"**	**"Slow"**	**"Loud"**	**"Soft"**	**"Fast"**
Neutral	96	5	3	3	2	1
Angry	1	70	6	12	3	13
Slow	1	2	73	6	17	1
Loud	0	18	1	72	2	6
Soft	2	2	15	3	73	1
Fast	0	3	2	4	3	78

TABLE 8. Confusion percentage of an unknown (to be recognized) emotion with the other emotions based on human listener (%).

Emotions	Neutral	Happy	Fearful	Sad	Disgusted	Angry
Neutral	97.6	1.9	0	0	0	0
Happy	2.4	95.1	0	0	0	2.3
Fearful	0	0	85.5	1	14.9	4.9
Sad	0	0	2.5	90	0	4.9
Disgusted	0	1.1	7.9	9	79.1	0
Angry	0	1.9	4.1	0	6	87.9

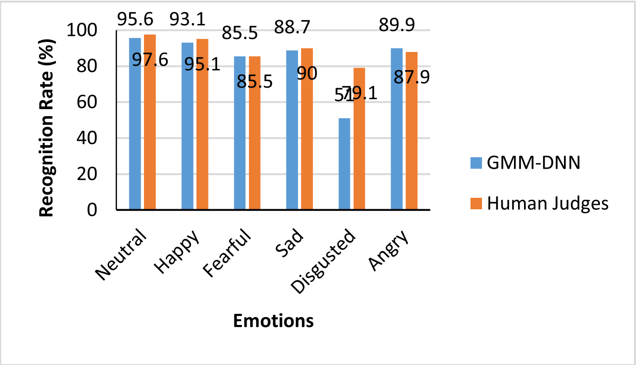

FIGURE 6. Emotion recognition accuracy assessment using the ESD based on GMM-DNN and human listene.

TABLE 9. Performance analysis of emotion identification using normal and distorted data (%).

Emotion	Emotion Identification Performance (%)					
	GMM-DNN		MLP		SVM	
	Normal	Distorted	Normal	Distorted	Normal	Distorted
Neutral	95.6	94.3	76.4	75.1	90.3	89.1
Happy	93.1	92.5	75.9	74.9	87.9	86.6
Fearful	85.5	84.1	65.1	62.3	81.2	79.5
Sad	88.7	86.6	84.3	81.3	87.7	84.9
Disgusted	51.0	49.2	45.7	42.3	49.9	45.5
Angry	89.9	88.8	71.3	68.8	85.0	82.4
Average	**84.0**	**82.6**	**69.8**	**67.5**	**80.3**	**78.0**

and the average value obtained is displayed in Table 9. The GMM-DNN performance does not show a considerable change in the recognition accuracy even in the presence of interference. This is accomplished by the use of ''GMM emotion identification tag''. This founds a mask for the dominant signal from other interference vectors. It is apparent from this table that emotion identification accuracy for normal speech signals based on GMM-DNN is greater than that based on MLP and SVM by 20.3% and 4.6%, respectively. In the case of distorted speech signals, GMM-DNN leads MLP and SVM by 22.4% and 5.9%, respectively.

Experiment 4: In this experiment, the ESD has been used to evaluate GMM-DNN, DNN-GMM, GMM alone, and DNN alone for emotion recognition. Emotion recognition accuracy using the ESD based on GMM-DNN, DNN-GMM, GMM alone, and DNN alone is shown in Figure 7. Based on this figure, the average emotion recognition performance based on GMM-DNN, DNN-GMM, GMM alone, and DNN alone is 83.97%, 83.00%, 70.46%, and 81.48%, respectively. This experiment apparently demonstrates that the two classifiers GMM-DNN and DNN-GMM yield almost the same results of emotion recognition performance. Also, it is evident from this experiment that a hybrid classifier of both GMM and DNN leads each of GMM alone and DNN alone.

Regarding the impact of combining the two classifiers (GMM and DNN) on time performance, it was found that GMM alone is the fastest model among the four models based on the average training time, followed by GMM-DNN, then DNN alone, while DNN-GMM is the slowest one. This shows that the GMM-DNN model not only gives better accuracy than the DNN model, but the computational time of the GMM-DNN model is less than the DNN model. This is because in the GMM-DNN model, DNN is fed with additional input which is the output of GMM. This makes the classification of GMM-DNN faster than DNN alone. Specifically, GMM used here is meant for coarse tuning and DNN is

TABLE 10. Ratio of computational training time using gym alone, GMM-DNN, DNN alone, and DNN-GMM.

classifier	GMM alone	GMM-DNN	DNN alone	DNN-GMM
Ratio	1	2	6	7

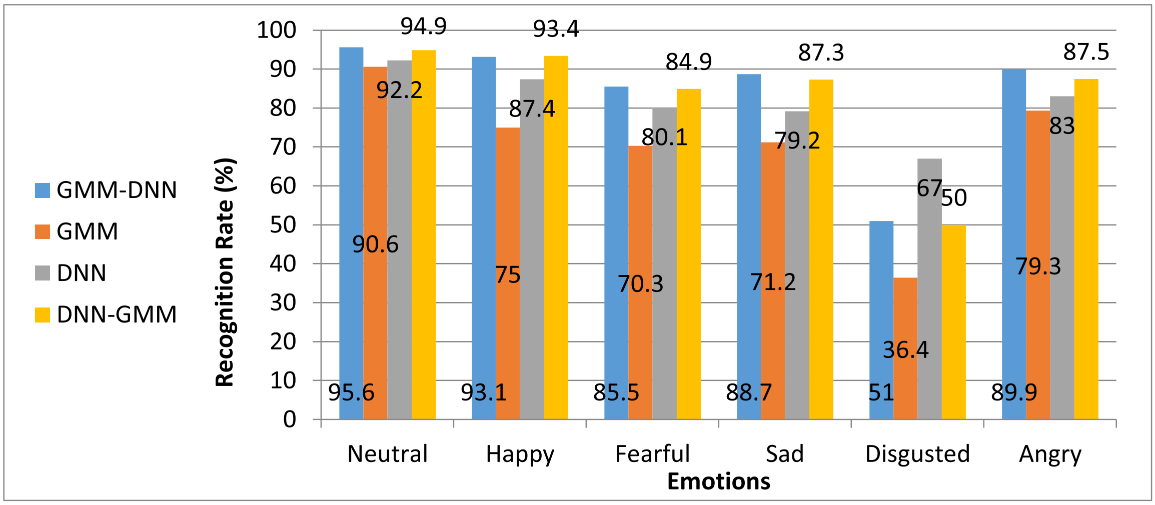

FIGURE 7. Emotion recognition accuracy assessment using the ESD based on GMM-DNN, DNN-GMM, GMM alone, and DNN alone.

for fine tuning. In DNN alone, data is processed in the DNN hidden layers, while in GMM-DNN, DNN is processed with the GMM tags only. Hence, the computational complexity is reduced. Table 10 displays the ratio of computational training time using GMM alone, GMM-DNN, DNN alone, and DNN-GMM.

VIII. CONCLUSION

Novel sequential GMM-DNN based classifier has been proposed, executed, and assessed for the purpose of improving emotion recognition performance. Two distinct and separate speech datasets (collected Emirati-accented and SUSAS) have been utilized to test the proposed classifier. This work demonstrates that the proposed model gives better results than the greatly used ''SVM and MLP'' based classifiers. Also, our results show that GMM-DNN gives results that are close to those given by DNN-GMM and each one of these two classifiers outperforms the GMM and DNN classifiers. The proposed classifier performs well even in the presence of noise and interference. Consequently, this increases the acceptance of the GMM-DNN model in human-computer intelligent interaction. The algorithm based on ''GMM tag based feature vector'' reduction aids in minimizing the complexity of DNN classifier; hence, enhancing emotion recognition accuracy.

The performance of ''GMM-DNN, SVM, and MLP'' classifiers in the recognition of the disgust emotion is not as good as in the recognition of other emotions. This is because disgust emotion is highly confused with fear and anger emotions. A more massive study is underway for the enhancement of recognition of the disgust emotion. A future work will develop a system that incorporates Computational Auditory Scene Analysis (CASA) and that is more suitable to interruptive and noisy emotional talking conditions.

The main limitation of our work is that the collected Emirati speech dataset is acted and unspontaneous due to the difficulty of capturing spontaneous emotions from human beings. This is similar to the majority captured databases which are acted. This limitation does not alter our attained results since the proposed GMM-DNN classifier has been evaluated on two different speech databases and subjected subjective assessment comparison. Our results demonstrate that the GMM-DNN model excels in both datasets.

''Conflict of Interest: The authors declare that they have no conflict of interest''.

''Informed consent: This study does not involve any experiments on animals''.

Statement of Ethics: The authors have permission from the University of Sharjah to collect speech database from UAE citizens based on the ''two competitive research'' grants entitled ''Emotion Recognition in each of Stressful and Emotional Talking Environments Using Artificial Models, No. 1602040348-P and Capturing, Studying, and Analyzing Arabic Emirati-Accented Speech Database in Stressful and Emotional Talking Environments for Different Applications, No. 1602040349-P''.

''Consent of Parent' of Minors'': This study includes very few speakers who are less than 18 years old. A consent from the minors' parents was provided before conducting the experiments''.

REFERENCES

[1] S. G. Koolagudi and K. S. Rao, ''Emotion recognition from speech: A review,'' *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[2] R. Cowie *et al.*, ''Emotion recognition in human-computer interaction,'' *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[3] V. A. Petrushin, ''Emotion recognition in speech signal: Experimental study, development, and application,'' in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, pp. 222–225.

[4] N. Fragopanagos and J. G. Taylor, ''Emotion recognition in human–computer interaction,'' *Neural Netw.*, vol. 18, no. 4, pp. 389–405, 2015.

[5] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, ''Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,'' *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2015. doi: 10.1007/s10462-012-9368-5.

[6] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, ''New approach in quantification of emotional intensity from the speech signal: Emotional temperature,'' *Expert Syst. Appl.*, vol. 42, pp. 9554–9564, Dec. 2015.

[7] I. Luengo, E. Navas, and I. Hernaez, ''Feature analysis and evaluation for automatic emotion identification in speech,'' *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[8] H. Cao, R. Verma, and A. Nenkova, ''Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,'' *Comput. Speech Lang.*, vol. 29, pp. 186–202, Jan. 2015.

[9] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, ''Speech emotion recognition using Fourier parameters,'' *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, Jan. 2015.

[10] I. Shahin and M. N. Ba-Hutair, ''Talking condition recognition in stressful and emotional talking environments based on CSPHMM2s,'' *Int. J. Speech Technol.*, vol. 18, pp. 77–90, Mar. 2015.

[11] S. Shukla, S. Dandapat, and S. R. M. Prasanna, ''A subspace projection approach for analysis of speech under stressed condition,'' *Circuits, Syst., Signal Process.*, vol. 35, no. 12, pp. 4486–4500, 2016.

[12] S. Deb and S. Dandapat, ''A novel breathiness feature for analysis and classification of speech under stress,'' in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb./Mar. 2015, pp. 1–5.

[13] I. Shahin, ''Studying and enhancing talking condition recognition in stressful and emotional talking environments based on HMMs, CHMM2s and SPHMMs,'' *J. Multimodal User Interfaces*, vol. 6, no. 1, pp. 59–71, Jun. 2012. doi: 10.1007/s12193-011-0082-4.

[14] I. M. A. Shahin, ''Gender-dependent emotion recognition based on HMMs and SPHMMs,'' *Int. J. Speech Technol.*, vol. 16, pp. 133–141, Jun. 2013.

[15] X. Cheng and Q. Duan, ''Speech emotion recognition using Gaussian mixture model,'' in *Proc. 2nd Int. Conf. Comput. Appl. Syst. Modeling*, 2012, pp. 1222–1225.

[16] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, ''Speech emotion recognition using Gaussian mixture vector autoregressive models,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Apr. 2007, pp. IV-957–IV-960.

[17] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zieıke, G. Meier, and B. Schuller, ''Deep neural networks for acoustic emotion recognition: Raising the benchmarks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.

[18] H. Kun, Y. Dong, and T. Ivan, ''Speech emotion recognition using deep neural network and extreme learning machine,'' in *Proc. INTERSPEECH*, 2014, pp. 223–227.

[19] W. Q. Zheng, J. S. Yu, and Y. X. Zou, ''An experimental study of speech emotion recognition based on deep convolutional neural networks,'' in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 827–831.

[20] L. Li *et al.*, ''Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition,'' in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2013, pp. 312–317.

[21] C. Huang, W. Gong, W. Fu, and D. Feng, ''A research of speech emotion Mathematical recognition based on deep belief network and SVM,'' *Problems Eng.*, vol. 2014, no. 1, Aug. 2014, Art. no. 749604.

[22] I. J. Tashev, Z.-Q. Wang, and K. Godin, "Speech emotion recognition based on Gaussian mixture models and deep neural networks," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2017, pp. 1–4.

[23] I. Shahin, A. B. Nassif, and S. Hamsa, "Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments," *Neural Comput. Appl.*. doi: 10.1007/s00521-018-3760-2.

[24] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Int. Conf. Speech Commun. Technol. (EUROSPEECH)*, Rhodes, Greece, vol. 4, Sep. 1997, pp. 1743–1746.

[25] I. Shahin, A. B. Nassif, and M. Bahutair, "Emirati-accented speaker identification in each of neutral and shouted talking environments," *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 265–278, Jun. 2018. doi: 10.1007/s10772-018-9502-0.

[26] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 90–100, Jan. 2010.

[27] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[28] I. Shahin, "Employing emotion cues to verify speakers in emotional talking environments," *J. Intell. Syst.*, vol. 25, no. 1, pp. 3–17, Jan. 2016. doi: 10.1515/jisys-2014-0118.

[29] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machines," *Int. J. Smart Home*, vol. 6, no. 2, pp. 101–108, Apr. 2012.

[30] H. Altun and G. Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8197–8203, May 2009.

[31] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, Feb. 2007.

[32] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[33] P. Mat jka *et al.*, "Analysis of DNN approaches to speaker identification," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5100–5104.

[34] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[35] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics*. London, U.K.: Collier-Macmillan, ch. 4, 1970.

[36] N. Breslow, "A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship," *Biometrika*, vol. 57, no. 3, pp. 579–594, 1970.

[37] E. A. Gehan, "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples," *Biometrika*, vol. 52, nos. 1–2, pp. 203–223, 1965.

[38] I. Shahin, "Employing second-order circular suprasegmental hidden Markov models to enhance speaker identification performance in shouted talking environments," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, no. 1, 2010, Art. no. 862138. doi: 10.1155/2010/862138.

ISMAIL SHAHIN received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Southern Illinois University at Carbondale, USA, in 1992, 1994, and 1998, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Sharjah, United Arab Emirates. He has more than 55 journal and conference publications. His research interests include speech recognition, speaker recognition under neutral, stressful, and emotional talking conditions, emotion and talking condition recognition, gender recognition using voice, and accent recognition. He has remarkable contribution in organizing many conferences, symposiums, and workshops.

ALI BOU NASSIF received the master's degree in computer science and the Ph.D. degree in electrical and computer engineering from Western University, Canada, in 2009 and 2012, respectively, where he is also an Adjunct Research Professor. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering and the Assistant Dean of Graduate Studies at the University of Sharjah, UAE. His research interest includes the applications of statistical and artificial intelligence models in different areas, such as software engineering, electrical engineering, e-learning, security, signal processing, and social media. He is a registered Professional Engineer in Ontario and a member of the IEEE Computer Society.

SHIBANI HAMSA received the B.Tech. and M.Tech. degrees in electronics and communication engineering from Mahatma Gandhi University, India, and ranked at Second Place on the M.Tech. degree in applied electronics. She is currently a Research Assistant of electrical and computer engineering with the University of Sharjah. Her research interests include artificial intelligence, cybersecurity, deep learning, speech processing, image processing, smart systems, and human-machine communication in the digital world.