IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 22, 2021, accepted May 2, 2021, date of publication June 3, 2021, date of current version June 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3086062

An Enhanced Emotion Recognition Algorithm Using Pitch Correlogram, Deep Sparse Matrix Representation and Random Forest Classifier

SHIBANI HAMSA[1], YOUSSEF IRAQI[1], (Senior Member, IEEE), ISMAIL SHAHIN[2], (Member, IEEE), AND NAOUFEL WERGHI[1], (Senior Member, IEEE)

[1]Center for Cyber-Physical Systems (C2PS), Department of Electrical and Computer Engineering (ECE), Khalifa University of Science, Technology and Research, Abu Dhabi, United Arab Emirates

[2]Department of Electrical Engineering, University of Sharjah, Sharjah, United Arab Emirates

Corresponding authors: Shibani Hamsa (shibani.koya@ku.ac.ae), Youssef Iraqi (youssef.Iraqi@ku.ac.ae), Ismail Shahin (ismail@sharjah.ac.ae), and Naoufel Werghi (naoufel.werghi@ku.ac.ae)

ABSTRACT This work presents an approach for text-independent and speaker-independent emotion recognition from speech in real application situations such as noisy and stressful talking conditions. We have incorporated a new way for feature extraction, representation, and noise reduction, replacing the frequently used cepstral features in the literature. The proposed algorithm is modeled as the combination of pitch-correlogram-based noise reduction pre-processing module, sparse-dense decomposition-based feature representation, and random forest classifier. The work is assessed in terms of efficiency and computational complexity using English and Arabic datasets corresponding to noisy and stressful talking conditions. Our system yields significant improvement in results in comparison with other techniques based on the same classifier model. The proposed network architecture achieves significant rise in performance correspond to the recent literature on benchmark datasets.

INDEX TERMS Emotion recognition, feature extraction, noise reduction, random forest classifier.

I. INTRODUCTION

Recognition of human emotion from the speech is a hot research topic with broad area of potential applications, such as intelligent human-computer interaction, smart environments, intelligent call centers, and security in the banking sector [1]. Affective computing using audio is portrayed as perceiving the emotional context of speakers from their discourse. Emotions play an essential role in interpersonal communications and it is significantly important in taking intelligent decisions. In the literature, several modalities such as facial expressions, speech gestures, and biological signals have been considered to identify human emotions. Compared to other physiological parameters, sound signals can be collected easily and economically. This makes audio signals an accepted source for affective computing [2]. The human-computer interface has been established rapidly in the advanced world and thereby identifying the emotions of the human interface plays a fundamental role. For instance, emotional intelligent machines have been widely industrialized and were applied to a series of real-time applications [2]. Perception of the emotional condition is still one of the significant challenges faced by the human-machine interface researches. Variations in facial expressions, body gestures, lexical signs and other biological changes are accounted for perceiving emotions. Speech production is a physiological process by which the thoughts are transformed to audio signals and the changes in mental condition will be reflected in these audio signals in addition to the face of the speaker. The change in pitch and modulation of the speech will likewise be an indication of the mental and emotional status of the speaker. [3].

Emotion recognition using acoustic signals has been considered as a ''pattern recognition problem''. Most of the ''speech emotion recognition'' models use a two-stage algorithm for ''pattern recognition: feature extraction and

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

classification'' based on the selected features. Different combinations of energy, pitch, time, and spectral features are coupled to classifiers based on the applications and strategies employed in the literature [4]. In this paper, we introduce deep sparse matrix representation (DSMR) feature extraction method as an alternative to the commonly used ''Mel frequency cepstral coefficients (MFCC)'' feature extraction approaches in the literature. The proposed algorithm is modeled as the combination of pitch correlogram-based noise reduction pre-processing module, sparse-dense decomposition based features, feature selection based on sparse principal component analysis (SPCA) and random forest classifier.

The objective of this work is to introduce a novel approach for efficient emotion classification in noisy and stressful talking conditions by replacing the commonly used MFCC features with deep sparse factorization method. The performance of the algorithm in terms of efficiency and computational complexity is assessed in English and Arabic datasets corresponding to noisy and stressful talking conditions. A noise suppression module is incorporated along with the pattern recognition framework for ensuring better performances even in challenging talking conditions. The rest of the paper is organized as follows. Literature review is incorporated in II. Section III explains the model description, Section V discusses the results and analysis part, and Section VI provides the conclusion.

II. LITERATURE REVIEW

Implementing emotion recognition with improved noise immunity is critical in the development of real-time speech processing applications [5]. ''Alonso *et al.*'' [6] and ''Luengo *et al.*'' [7] introduced recognition frameworks based on ''Mel Frequency Cepstral Coefficient (MFCC)'' features cascaded to ''support vector machine (SVM) classifier and obtained an accuracy of 94.9% and 78.3% for classifying five emotions such as anger, happy, neutral, sad and boredom in the Berlin dataset [8]''. Wang *et al.* [9] used the same classifier model trained and tested using Fourier parameters instead of MFCC and attained 73.5% for classifying 6 emotions respectively.The emotions classified are neutral, sad, angry, fear, boredom and happy. Campbell *et al.* [10], Hong and Kwong [11], Kinnunen *et al.* [12] and Shahin and Ba-Hutair [13] applied MFCC features to distinct classifiers for the human affect evaluation in the stressful speaking situations using ''Speech Under Simulated and Actual Stress (SUSAS) dataset''. The emotions classified are ''slow, angry,loud, neutral,soft and fast''. ''Hong and Kwong [11], Kinnunen *et al.*'' [12] reported recognition rates of 68.70% using genetic algorithm (GA) and 68.40% using vector quantization (VQ), respectively. However, Campbell *et al.* [10] achieved a recognition rate of 72.80% using SVM classifier. ''Shahin and Ba-Hutair'' [13] used MFCC features on ''third-order circular supra-segmental hidden Markov models (CSPHMM3s)'' and obtained 76.3% recognition rate for classifying six emotions [14]. Shukla *et al.* [15] coupled 13-dimensional feature vectors to the Hidden Markov Model (HMM) classifier for classifying the four distinct emotion classes in the ''SUSAS dataset'' and obtained an accuracy of 93.9% in the ''stressful talking conditions''. The emotions observed are ''neutral, angry, sad and lombard''. Vlassis and Likas [16] proposed an algorithm for emotion recognition which utilized the global features along with the Gaussian mixture model (GMM) classifier. This work is performed on data taken separately from known distributions and reported a recognition rate of 75%. Several recent works employed Deep neural network (DNN) [17]–[19] showing significant improvement in results over the conventional classifier model utilizing cepstral coefficients. EmoDB dataset having emotions ''angry, boredom, disgust, fear, happy, neutral and sadness'' is used in [17] and ''Interactive Emotional Dyadic Motion Capture (IEMOCAP) database'' with five emotions: excitement, frustration, happiness, neutral and surprise is used in [18] and [19].The above mentioned models using MFCC feature vector demonstrated an improvement in accuracy of around 30% using deep learning techniques than the other conventional classifier models. France *et al.* [20] utilized pitch, amplitude modulation, and formant features for diagnosing the depression level in male and female patients. Their work reported an average recognition score of 55% using clinical data for classifying four depression levels such as control/dysthymic, control/Major depressed, dysthymic/major depressed and control/dysthymic/major depressed. Palo *et al.* [21] classified low and high arousal emotion classes using cepstral coefficients such as linear prediction coefficient (LPC), MFCC, and linear prediction cepstral coefficient (LPCC). They attained a recognition rate of 48.60%, 80.00% and 54.50%, respectively for classifying 4 emotional classes Boredom, angry, sad and surprise are the emotions classified using Multilayer Perception (MLP) technique. The database is prepared using 10 subjects aged 6 to 13 years. Lee and Narayanan [22] classified a call center data based on emotions using pitch, energy, duration, and formant features. Forward Selection (FS) method is employed to extract features and Linear Discriminant Classifiers (LDC) followed by k-nearest neighborhood classifiers (k-NN) are used for classification of emotions. The speech dataset consists of around 7200 utterances. Wu and Liang [23] evaluated multiple classifiers by means of pitch, intensity, formants, shimmer, and MFCC features and achieved maximum recognition rate of 78.16% using SVM classifier using their private dataset. Kuang and Li [24] used pitch, energy, formants, LPCC and MFCC features, with a classifier fusion method based on Dempster–Shafer evidence theory for the recognition of angry, sad, surprise, and disgust emotions. Their model obtained a recognition score of 89.64% using Berlin emotional database.

Liu *et al.* [25] proposed an emotion recognition technique using updated ''brain emotional learning (BEL) model revived by the sentimental handling mechanism of limbic structure in the brain''. They attained an accuracy of 64.60% for ''speaker-independent emotion recognition [25] using

FAU Aibo dataset''. Torres-Boza *et al.* [26] demonstrated an approach using Sparse hierarchical coding (SC) for sentimental evaluation. They have used VAM-Audio and AVEC 2012 challenge database for the feature set evaluation.

Zhang *et al.* [27] used low level descriptors (LLD) and attained 81.67% for classifying six distinct class in the RAVDESS dataset. ''Directed Acyclic Graph SVM (DAGSVM) is adopted as single-task multi-class emotion classifier''. Emotions such as angry, happy, fearful were easier to categorize as compared to neutral, calm and sad. Whereas, in the experiment using same dataset, the model propose by Huang and Bao [28] reported 72.20% for classifying 4 classes.

The MFCC-based emotion recognition model proposed by Shahin *et al.* [29] reached an ''average recognition rate'' of 83.97% and 86.67% on ''Emirati-emphasized Arabic speech dataset (ESD)'' and ''SUSAS dataset'', respectively. Emotions classified on ''ESD dataset'' are neutral, happy, fearful, sad, disgusted and angry Hamsa *et al.* [30] used MFCC features and random forest classifier along with ''computational auditory scene analysis (CASA)'' based speech segregation system to achieve high performance with reduced computational complexity. They obtained an average recognition rate of 86.38%, 88.67% and 89.60% using RAVDESS, SUSAS and ESD dataset, respectively [30]. Emotions classified on RAVDESS dataset are neutral, happy,angry, sad, surprise,fearful, calm and disgust. On ''SUSAS dataset'', five emotions such as angry, neutral, slow, loud and soft are classified.

In this paper, we propose an ''emotion recognition system'' using correlogram-based noise suppression module, deep sparse representation of speech segments and random forest classifier.

The significant achievements of the work are explicitly seen in:

- An enhanced proposal for emotion recognition using pitch correlogram-based noise-reduction, deep sparse matrix representation (DSMR) based features, SPCA feature selection and random forest classifier.
- An enhanced approach for noise reduction and feature extraction to replace the conventional cepstral coefficients.
- A design of a computationally less complex, real-time application system without compromise in performance.

III. MODEL DESCRIPTION

Fig. 1 shows the block schematic representation of the proposed emotion recognition system. The system consists of noise reduction pre-processing, feature extraction and classifier modules. Auto-correlation based correlogram approach is employed for noise suppression. Feature extraction and selections are by means of DSMR model and sparse principal component analysis (SPCA). A detailed description of these stages will be provided in the next subsections.

TABLE 1. Pseudo code - Pitch determination.

Steps:
1) T-F decomposition of input sample.
2) Separate X_low and X_high using HPF.
3) Compute DFT of X_low and X_high.
4) Obtain the periodicity of the signals using auto-correlation.
5) Convert to time domain using IDFT.
6) Calculate SACF of each signal.
7) Detect the lowest and highest onset positions to determine the pitch ranges of interference and dominant signals.

A. NOISE REDUCTION AND PRE-PROCESSING

Fig. 2 shows the schematic representation of speech segregation pre-processing module designed for noise reduction. The noise reduction system consists of ''time-frequency (T-F) decomposition'', pitch determination and segmentation, and re-synthesis modules.

1) T-F DECOMPOSITION

In this paper, the proposed algorithm achieves T-F decomposition by means of ''short time Fourier transform (STFT)'' based cochlear filter-bank. The input speech signal is disintegrated into small segments of duration 20ms. Fourier transform of all segments are computed separately, to obtain the T-F relationship of the audio signal. It can be denoted as:

$$\begin{aligned}\mathbf{X}(m,k) &= \underset{m,k}{S}\, TFT(\mathbf{x}(n)) \\ &= \sum \mathbf{x}(n)\mathbf{w}(n-m)exp(-j\omega n) \end{aligned} \tag{1}$$

$\mathbf{X}(m,k)$ is the narrow band signal coming out of the k th band with the time index m. $\mathbf{x}(n)$ and $\mathbf{w}(n)$ represents the input signal and window function, respectively [31].

2) PITCH DETERMINATION

In this work, an ''auto-correlation method'' is adopted for ''pitch estimation''. Fig. 3 shows the schematic representation of pitch estimation and the pseudo code is given in Table 1. The input signal is allowed to pass through separate channels by means of a high pass filter (HPF) with 12 dB per octave attenuation in the stop band. Then the ''discrete Fourier transform (DFT)'' of the low frequency signal X_{low} and X_{high} are computed separately. The estimated periodicity of each signals are combined and converted to time domain by means of ''inverse discrete time Fourier transform (IDFT)''. Generalised auto-correlation method is used for the periodicity detection. The summation of the generalized auto-correlation of both high and low frequency frames corresponds to the SACF. The Summary auto-correlation function (SACF) is set as s:

$$s = IDFT[(|DFT(\mathbf{X}_{low})|)^k + IDFT(|DFT(\mathbf{X}_{high})|)^k] \tag{2}$$

$$s = IDFT[(|DFT(\mathbf{X}_{low})|)^k + (|DFT(\mathbf{X}_{high})|)^k] \tag{3}$$

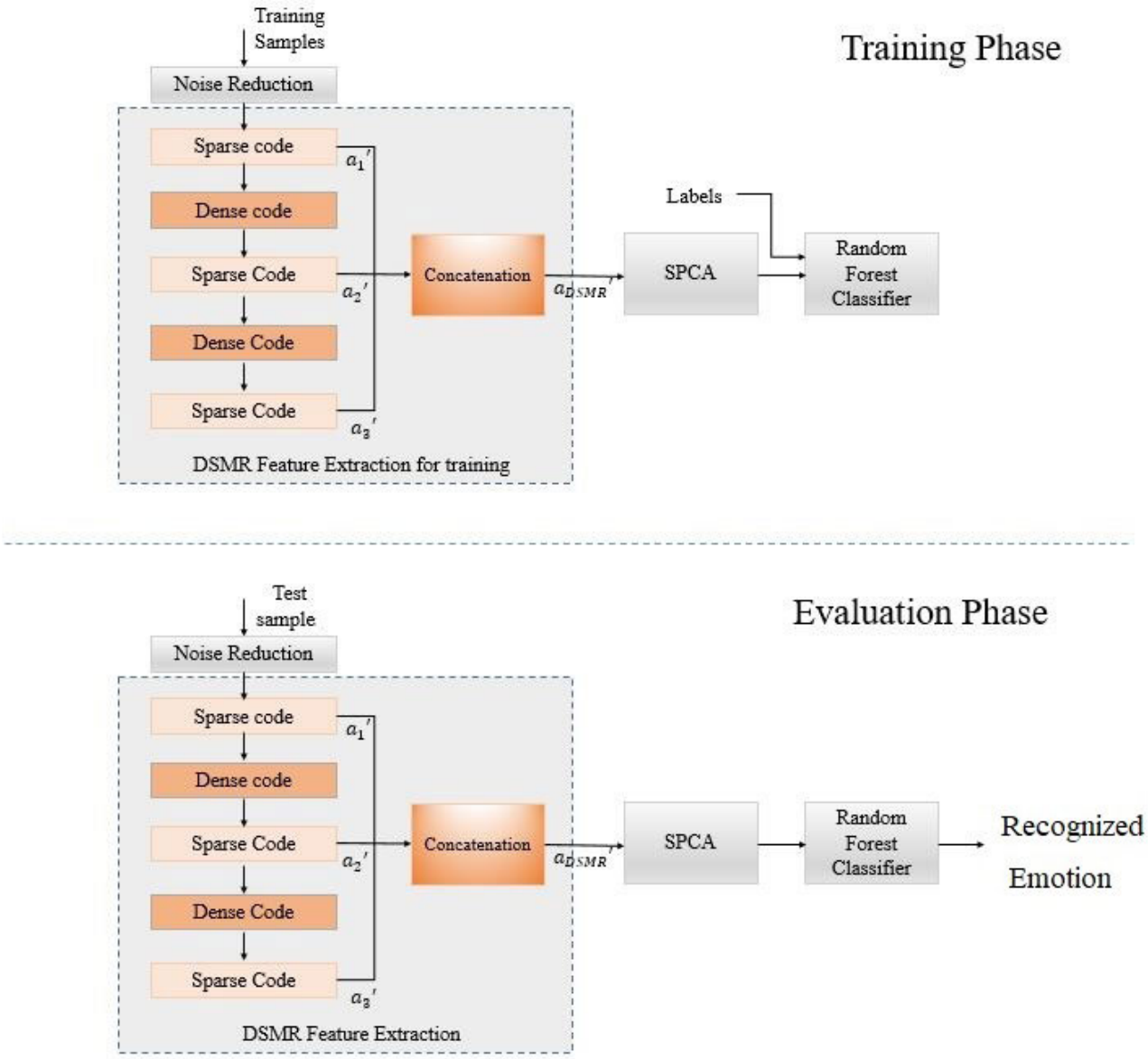


FIGURE 1. Emotion recognition block schematic with sparse representation.

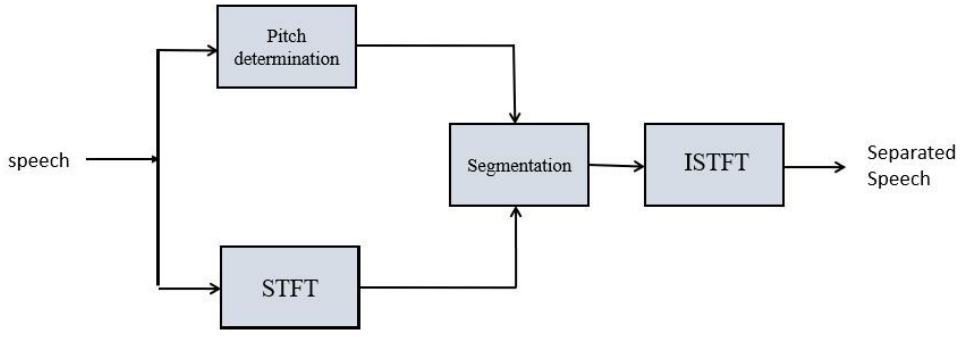


FIGURE 2. Original speech segregation and noise reduction.

where k is the frequency domain compression factor and its value is less than 2. Each fundamental frequency's period is represented by a peak in the SACF. As a result, the highest peak in the resulting SACF signal corresponds to the pitch of the dominant speech. As a result, the detection of the highest and lowest onset positions determines the pitch range of both target and interference speech. s is further intensified by passing it through a negative clipper to segregate the SACF positive values. Then, the signal is interpolated and subtracted from the SACF positive values. The resultant signal is again interpolated and clipped to its positive values to attain the enhanced summary auto-correlation function (ESACF). Time delay corresponding to the peak value of this ESACF gives the pitch of the frame, P_f, where f represents the frame number. The P_f with maximum probability of occurrence is marked as the dominant pitch frequency P_d. These dominant pitch values are used only to segregate the dominant signal from other noise and interference prior to feature extraction.

3) SEGMENTATION AND RE-SYNTHESIS

A binary mask is designed for the segmentation of dominant and noisy T-F parts present in the input signal. The transfer function of ideal binary mask (IBM) is set as the product of binary mask value of each frame with a cosine window,

$$TF_{IBM}(i,j) = tf_{IBM}(i,j)cos(win(i)) \tag{4}$$

where tf_{IBM} represents the $20ms$ duration binary mask of each frame and is defined as follows:

$$tf_{IBM}(i,j) = \begin{cases} 1 & \text{if } P = P_d \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where P and P_d are the pitch and dominant pitch values, respectively.

The original speech separation is done by the convolution of the input signal with the IBM transfer function,

$$\mathbf{X}(m,k)_{dominant} = (\mathbf{X}(m,k) * TF_{IBM})_{(m,k)} \tag{6}$$

The dominant signal after noise reduction, $x(n)_{dominant}$ can be re-synthesized by applying ''inverse short time Fourier transform (ISTFT)''.

$$\mathbf{x}(n)_{dominant} = \underset{k}{ISTFT}(\mathbf{X}(m,k)_{dominant}) \tag{7}$$

B. FEATURE EXTRACTION

In this work, we used a novel DSMR approach for modeling the feature vectors. The proposed DSMR model is employed to highlight the desirable information present in the various speech segments, which is further utilized for emotion recognition. As shown in Fig. 1, a five-layer DSMR model is designed for efficient feature extraction and representation. Dense layers are used as an intermediate layer and are incorporated between two sparse layers. Each dense layer plays an important role to reduce the ''computational complexity'' by controlling the data over-fitting. The feature representations obtained from each sparse layer contains complementary

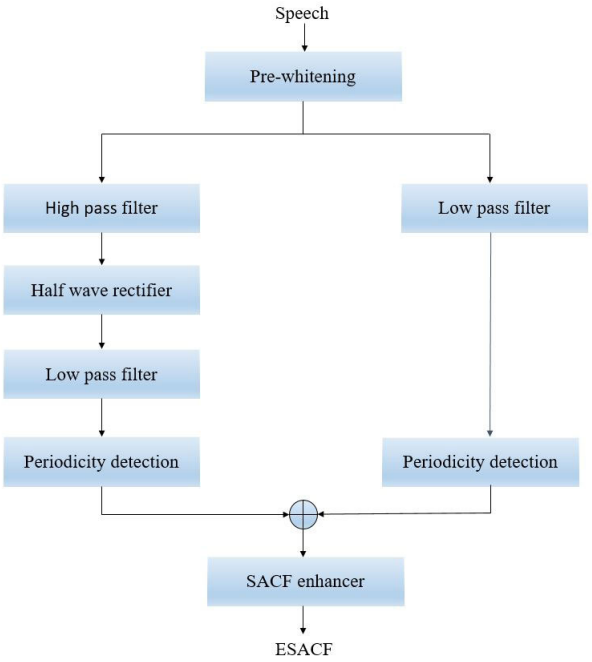


FIGURE 3. Pitch estimation using auto-correlation approach.

information and they are concatenated to obtain the complete feature vector. Afterwards, SPCA is employed to reduce the concatenated feature dimension without loss of information.

In general, the segregated speech signal is decomposed into $l+1$ factors as:

$$\begin{aligned} X_{dominant} &= D_1 D_2 D_3 \ldots D_l A_l \\ A_{l-1} &= D_l A_l \\ &\vdots \\ A_2 &= D_3 \ldots D_l A_l \\ A_1 &= D_2 \ldots D_l A_l \end{aligned} \tag{8}$$

where, $l = 1, 2, \ldots 5$, since we used a 5 layer DSMR model.

$$\begin{aligned} A_4 &= D_5 A_5 \\ A_3 &= D_4 D_5 A_5 \\ A_2 &= D_3 \ldots D_5 A_5 \\ A_1 &= D_2 \ldots D_5 A_5 \end{aligned}$$

An efficient emotion recognition algorithm requires large amount of data sufficient for deep learning. Hence we have incorporated an exemplar-based dictionary [32] in the first sparse layer to accommodate the whole training data.

The segregated input signal is represented as l factors in the first sparse layer. The representation includes $l-1$ dictionaries denoted as D, and a matrix A. Then the first dense layer (second layer) representation is obtained by using method of optimal directions (MOD) algorithm for sparse approximation [33].

$$(\hat{D}_2 \hat{A}_2) = \underset{D_2 A_2}{argmin} ||\hat{A}_1 - D_2 A_2||_2 \tag{9}$$

The representation obtained at the 1st dense layer is used to generate the 2nd sparse layer (third layer) representation using K-SVD [34] deep learning algorithm. D_2 and A_2 are the dictionary and matrix at the 2nd layer.

$$(\hat{D}_3\hat{A}_3) = \underset{D_3A_3}{argmin} \forall_i \, g(a_{3i}) \quad s.t \; ||\hat{A}_2 - D_3A_3||_2 \leq \lambda \tag{10}$$

In the proposed DSMR representation, even layers (dense layers) are using MOD algorithm and odd layers (sparse layers) except the first layer are using K-SVD algorithm. The process repeats until the sparse layer representation converges to the threshold λ. The value of λ is set to achieve the error tolerance and in this work the value used for error tolerance is 0.001. The 2nd dense layer (fourth layer) representation using MOD is obtained as:

$$(\hat{D}_4\hat{A}_4) = \underset{D_4A_4}{argmin} ||\hat{A}_3 - D_4A_4||_2 \tag{11}$$

The 3rd sparse layer (fifth layer) representation using K-SVD is obtained as:

$$((\hat{D}_5\hat{A}_5)) = \underset{D_4A_4}{argmin} \forall_i \, g(a_{5i}) \quad s.t \; ||\hat{A}_4 - D_5A_5||_2 \leq \lambda \tag{12}$$

Deep sparse layer dictionaries are stored and are defined as:

$$D_q{}' = \prod_{l=1}^{m} \hat{D}_l, \tag{13}$$

where q represents the three sparse layer among the l layers. Since we use a five-layer DSMR model in this work, the sparse dictionaries are defined as,

$$D_1{}' = \hat{D}_1$$
$$D_3{}' = \hat{D}_1\hat{D}_2\hat{D}_3$$
$$D_5{}' = \hat{D}_1\hat{D}_2\hat{D}_3\hat{D}_4\hat{D}_5$$

The representations obtained from each sparse layers are set as:

$$a'_q = \underset{a_m}{argmin} ||a_q|| \quad s.t \; ||x'_{dominant} - D'_q a_q||_2 \leq \lambda \tag{14}$$

The final DSMR representation is obtained by concatenating the representations of each sparse layer; $a_1{}'$ of length n_1, $a_2{}'$ of length n_2 and $a_3{}'$, of length n_3.

$$\underset{m}{a'_{DSMR}} = [a'_1 a'_2 a'_3] \tag{15}$$

The concatenation results in a high dimensional feature vector $\underset{m}{a'_{DSMR}}$. This may cause higher computational complexity. Hence, we used SPCA to reduce the feature dimensions. With "Sparse PCA", we expect to find a collection of corresponding "sparse principal components" that aid in visualization nearly as well as PCA while revealing some unique structure. To accomplish this, we ran the "block coordinate ascent algorithm" [35] on the data with a set of penalty parameter values ρ. A plot of the variance described by the "first sparse principal component (PC)" as a function of its cardinality was obtained. We then choose a cardinality that can explain at least 90% of the variation described by the first PCA principal variable. The covariance matrix was then deflated by eliminating the part resulting from the first sparse PC, and the process was repeated to obtain the "second sparse PC". Similarly, we solved for the third sparse PC.

C. CLASSIFICATION

We adopted the Random Forest classifier [36] as a predictive model. Random forest makes decisions by adding the predictions from its base models. Each base model is a simple decision tree. Random forests are able to capture non-linear interaction between the features and the target. The sparse features are switched to a linear model for the ease of Random forest classifier.

In this work, selected DSMR feature vectors are cascaded to RF classifier for the classification and recognition of emotion classes. Let the speech dataset S contains p number of signals with q number of feature representations. F indicates the features used for classifications and E represents the p dimensional output emotion classes.

During training, the number of decision trees are decided on the basis of random choices made by considering the factors such as performance and computational complexity of the classifier. In this work, we have used 100 decision trees with randomly selected signals and features. The set of bootstrap samples s selected from whole set S of size p contains $\frac{p}{100}$ number of signals. Then, decision trees are allocated by delegating values at each node until depleting all variables. All decision trees are involved to find the sample possibly matching the test sample in the evaluation phase and each decision tree made a single vote from its own perspective. Finally, we employ a majority voting technique over all the individual trees prediction to determine the emotion class [37].

IV. EVALUATION

A. DATASETS

Our work is appraised using four distinct speech corpora: RAVDESS, SUSAS, "Arabic Emirati emphasized speech dataset (ESD)" and speech emotion annotated data for emotion recognition systems (SAVEE) datasets. The details of the datasets and the methodology used for evaluation are as follows.

1) RAVDESS CORPUS

The "RAVDESS" is an authenticated multimodal speech corpus of emotional discourse and song [38]. The database includes 24 professional artists in a gender ratio of 1:2, articulating statements in a poor lexical match with a North American accent. Speech recognizes emotions such

as ''angry, happy, neutral, sad, fearful, disgust, calm, and surprise.'' Any speaker in 60 trials spoke two lexical complement utterances, resulting in 1,440 files. Each audio file is 3 seconds long and contains speech classified as one specific emotion.In this work, we have used ''RAVDESS dataset'' to examine the performance of the proposed framework in the English language.

2) SUSAS CORPUS

SUSAS is a public speech corpus captured in English that contains articulations of over 16,000 words from 19 male and 13 female performers ranging in age from 22 to 76 [39]. The ''SUSAS dataset'' is used in this work to evaluate the efficiency of our proposed model in stressful talking situations such as angry, neutral, slow, loud, soft, and fast.In this study, twenty different words were spoken twice by twenty different speakers in seven stressful talking environments. This were combined in a 2:1 to 3:1 ratio with the other speech signals in the same database and then included in this work. Ten separate words pronounced by the same ten speakers twice under six challenging talking situations were combined in the ratios 2:1 and 3:1 with various noise signals.

3) ESD CORPUS

Emirati-Emphasized Arabic speech dataset (ESD) is a private simulated emotional speech corpus that includes 50 speakers with 25 male and 25 female actors. Every actor repeated the 8 casual sentences among the Emirati community 9 times with a duration of 2-5 seconds in each of ''angry, happy, neutral, sad, fearful and disgust emotions''. During the training stage, the first four sentences were used, while in the testing phase, the remaining four utterances were utilized.

4) SAVEE CORPUS

The SAVEE dataset was recorded from the ''post graduate students and researchers of the University of Surrey age spanning from 27 to 31 years''. This dataset includes 7 emotion classes including neutral talking conditions resulting into corpus of 480 British English utterances. The text material consists of ''3 common, 2 emotion-specific and 10 generic sentences that are different for each emotions and phonetically-balanced''. In addition, ''3 common and 6 emotion specific sentences'' were considered in the normal class.

B. EVALUATION METHODOLOGY

For minimizing over-fitting and optimizing generalized contrast with the state of the art, the proposed study was tested using basic hold-out validation and K-fold cross validation techniques.

An experimental inquiry was divided into two sections to ensure that the experiment's mechanism could be readily observed. Various feature space-based emotion detection was performed on independent datasets during step 1, in the first part of the experiment. The aim of this section is to evaluate the training efficiency for each of the datasets used in the experiment. The datasets were divided according to the 80/20 law (80% of data are used for model training/validation and 20% for testing). Furthermore, 80 percent of the dataset was divided according to the same system, i.e., 80% was used for model preparation and 20% for model testing.Test sets contain samples of the same emotion uttered by the same speaker, with a repetitive emotion sample recording. For a repeated emotion sample recording, test sets include recordings of the same emotion spoken by the same voice. The overall network accuracy and the averaged F1 score were determined to test the classifier results. In accordance with the state-of-the-art, In congruence with the state-of-the-art, the viability of the suggested model is illustrated based on the benchmark outcomes on different languages in normal and various challenging speaking contexts.

To that end, all datasets have the same learning and evaluation strategies (as described earlier). Tests were conducted on 20% of the data from each emotion database used in network training. For network training and evaluation, no test data was used. With the approach of K-fold validation j, the available data is split into K partitions of equal size. For each partition, train a model on the remaining $k - 1$ partitions, and evaluate it on partition j. The final score is then the average of the k score obtained. In this work we have employed 10 fold and 5 fold cross trial validation approach for the performance analysis with reference to the published state-of-the-art using ''RAVDESS, SUSAS, ESD and SAVEE datasets''.

The below listed assessments were carried out to show the effectiveness of our suggested framework for text-independent and speaker-independent emotion recognition:

- Performance evaluation in terms of Accuracy, Precision, Recall and F1 score.
- Experimental comparisons with the recent literature.
- Evaluation based on statistical significance of the results.
- Evaluation of the proposed model in terms of noise susceptibility.
- Analysis on the computational complexity.
- Evaluation of system performance with and without noise reduction module.

V. EXPERIMENTS AND DISCUSSION

In this work, we have used ''RAVDESS dataset'' to examine the performance of the proposed framework in the English language. Fig. 4 interprets the classification ability of the proposed approach using ''RAVDESS dataset''. Emotion recognition rate obtained based on random forest classifier model using cepstral features [30], proposed model without noise reduction module, and proposed model with noise suppression using ''RAVDESS dataset''s are shown in Fig. 4. The results clearly indicate that, the best performance is achieved in neutral ''talking condition and obtained a recognition rate'' of about 90%. For the emotion classes, the obtained recognition rate is between 86% and 90%, respectively. The average

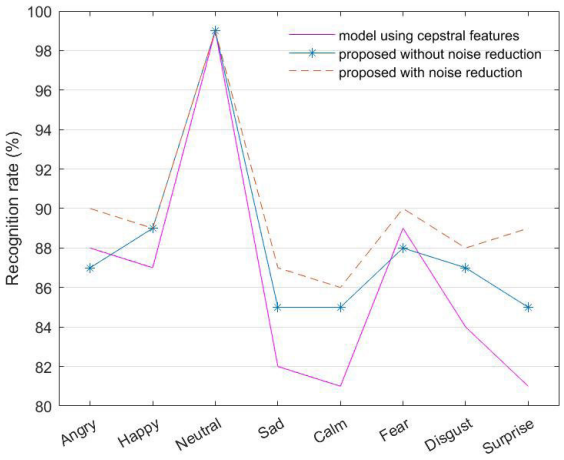


FIGURE 4. Emotion recognition rate obtained based on random forest classifier model using cepstral features [30], proposed model without noise reduction module, and proposed model with noise suppression using RAVDESS dataset.

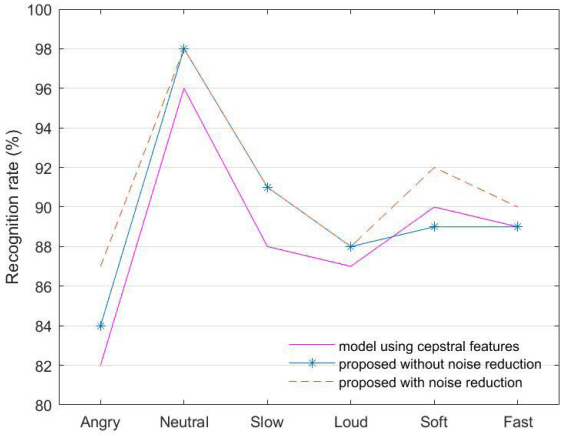


FIGURE 5. Emotion recognition rate obtained based on random forest classifier model using cepstral features [30], proposed model without noise reduction module, and proposed model with noise suppression using "SUSAS dataset".

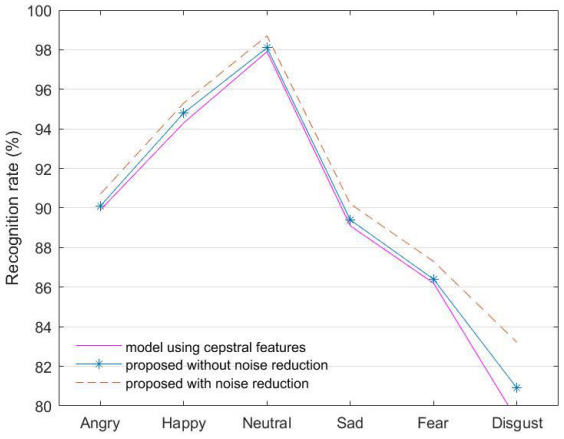


FIGURE 6. Emotion recognition rate obtained based on random forest classifier model using cepstral features [30], proposed model without noise reduction module, and proposed model with noise suppression using "ESD dataset".

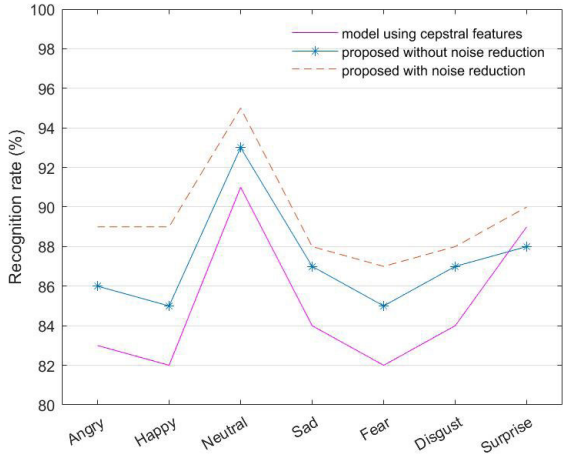


FIGURE 7. Emotion recognition rate obtained based on random forest classifier model using cepstral features [30], proposed model without noise reduction module, and proposed model with noise suppression using SAVEE dataset.

emotion recognition rate obtained for the proposed model with noise reduction module is 89.75%, which is significantly higher than the same classifier model using cepstral features. Fig. 4 states the effectiveness of the proposed model over the same classifier model using cepstral coefficients [30].

Fig. 5 depicts the stress recognition rate of the proposed algorithm using "SUSAS dataset". Fig. 5 shows the emotion recognition rate obtained based on random forest classifier model using cepstral features [30], proposed model without noise reduction module, and proposed model with noise suppression. Results show the superiority of the proposed model in terms of performance over the same classifier model using cepstral features [30]. The proposed model achieves a stable recognition rate in all the stressful "talking conditions and obtained an average recognition rate" of 91.00%.

Fig. 6 shows the performance statistics of the model using cepstral features [30], proposed model without noise reduction, and proposed model with noise reduction modules using "ESD dataset". Recognition rate obtained for the same classifier model using cepstral features [30] and the proposed model without noise reduction shows almost identical performance without any significant deviation. However, the proposed model with noise suppression slightly improves its performance and exhibits some significant improvement in terms of recognition rate. The proposed model achieves an average recognition rate of 90.90%.

Fig. 7 demonstrates the recognition score obtained based on the random forest classifier model using cepstral coefficients [30], proposed model without noise reduction module, and proposed technique with noise suppression module respectively, using SAVEE dataset. The average emotion recognition rate attained based on the proposed model with noise reduction module is 89.75%.

Fig. 5 proposed method with and without noise reduction module achieves almost identical performance in-terms of recognition rate in some talking conditions. In this experiment, the performance of the models are assessed without adding any external noise. i.e, the obtained emotion

TABLE 2. Evaluation based on statistical significance using distinct speech corpus.

t Value	"RAVDESS"	"SUSAS"	ESD	"SAVEE"
t (Proposed, Random forest [30])	1.67	1.76	1.69	1.89
t (Proposed, Hybrid GMM-DNN [29])	1.74	1.77	1.72	1.93

TABLE 3. Performance evaluation with the state-of-the-art using "RAVDESS dataset". The first and second best rates obtained for each class are in "bold and underlined respectively".

Method	Features	Classifier	Validation	Emotions*	Average recognition rate (%)
Z. Biqiao [27]	LLD (St Hier)	SVM	Cross validation (5-fold)	An,Ha,Ne,Sa,Ca,Fe	79.67
Z. Biqiao [27]	LLD (Mt Hier)	SVM	Cross validation (5-fold)	An,Ha,Ne,Sa,Ca,Fe	81.67
Y. Gao [41]	MFCC,Pitch, LSP, ZCR	SVM	Cross validation (10-fold)	An,Ha,Ne,Sa,Ca,Fe	79.28
A. Huang [28]	MFCC, STFT	CNN	Cross validation (10-fold)	An,Ha,Ne,Sa	72.20
H.Holmstrom [42]	low level descriptors	CNN	80:20 ratio	An,Ha,Ne,Sa,Ca,Fe,Di,Su	39.00
I. Shahin [29]	MFCC	GMM-DNN	1:2 ratio	An,Ha,Ne,Sa,Ca,Fe,Di,Su	83.63
S. Hamsa [30]	MFCC	RF**	Cross validation (10-fold)	An,Ha,Ne,Sa,Ca,Fe,Di,Su	<u>86.38</u>
Proposed	DSMR	RF**	Cross validation (5-fold)	An,Ha,Ne,Sa,Ca,Fe	**95.33**
Proposed	DSMR	RF**	Cross validation (10-fold)	An,Ha,Ne,Sa	**98.89**
Proposed	DSMR	RF**	Cross validation (10-fold)	An,Ha,Ne,Sa,Ca,Fe,Di,Su	**89.75**
*An-Angry, Ha-Happy, Ne-Neutral, Sa-Sad, Ca-Calm, Fe-Fearful, Di-Disgust, Su-Surprise, **RF- Random Forest					

recognition rates are influenced only by the implicit noise present in the speech samples. For example, in Fig. 5, the emotion recognition rate obtained in angry, soft and fast talking conditions obtained better results using the proposed approach with noise reduction module than the proposed technique without noise reduction module. The results indicate the presence of implicit noise and the ability of the proposed noise reduction module to segregate these noises present in the speech signals. The influence of implicit noise in "RAVDESS dataset" is evident in all emotions except noise in Fig. 4. The results given in Fig.4, Fig. 5, Fig. 6 and Fig. 7 indicate that the proposed model with noise reduction module is effective in segregating the implicit noise present in the speech samples and achieves better performance in such conditions.

A. ANALYSIS OF STATISTICAL SIGNIFICANCE

A statistical significance test has been conducted to check whether the results are actual or emerging from analytical variations. We used Student's t-distribution test for the evaluation [40].

$$t_{1,2} = \frac{\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2}{SD_{pooled}} \tag{16}$$

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

where $\bar{X_1}, \bar{X_2}$ are the means and SD_1 and SD_2 are the standard deviations of two sets of length n, respectively.

The proposed model is compared with each of hybrid GMM-DNN classifier model [29] and Random forest classifier model utilizing cepstral features [30]. Table 2 "reports the t values s between our system and other algorithms in the state-of-the-art using the four datasets." In view of the critical value $t_{critical} = 1.645$ at 0.05 significant level [29], the outcomes show that the framework execution is significantly higher than the efficient classifier models using cepstral features in the literature Shahin *et al.* [29], and Hamsa *et al.* [30].

B. EXAMINATION OF THE PROPOSED MODEL PERFORMANCE WITH THE RECENT LITERATURE

Table 3 clearly illustrates the performance analysis of the proposed model with reference to the published state-of-the-art using RAVDESS corpus based on the same evaluation setup used by us. The results convey that the proposed DSMR model attains a rise, in terms of average emotion recognition rate of, 15.66%, 13.66%, 26.69%, 6.12% and 3.37% over the outcomes indicated by "Biqiao *et al.* (St Hier) [27], Biqiao *et al.* (Mt Hier) [27], Huang and Bao [28], Shahin *et al.*" [29] and Hamsa *et al.* [30], respectively.

The proposed framework is trained and evaluated in stressful talking c onditions using "SUSAS dataset". We have implemented the same evaluation strategy for the

TABLE 4. Performance evaluation with the state-of-the-art using "SUSAS dataset". The first and second best rates obtained for each class are in "bold and underlined respectively".

Method	Features	Classifier	Validation	"Emotions*"	"Average recognition rate" (%)
"W.M Campbell" [10]	"MFCC"	SVM	Not available	Ne,An,Sl,Lo,Fa,So	72.80
"Q.Y. Hong" [11]	"MFCC"	GA	Not available	Ne,An,Sl,Lo,Fa,So	68.70
"T. Kinnunen" [12]	"MFCC"	VQ	Not available	Ne,An,Sl,Lo,Fa,So	68.40
"I. Shahin" [29]	"MFCC"	GMM-DNN	Cross validation (10-fold)	Ne,An,Sl,Lo,Fa,So	86.67
S. Hamsa [30]	DSMR	RF**	Cross validation (10-fold)	Ne,An,Sl,Lo,Fa,So	87.97
C.K Yogesh [43]	Hybrid BBO PSO features	ELM Kernal	70:30	Ne,An,Sl,Lo,Fa,So	<u>90.09</u>
Proposed	DSMR	RF**	Cross validation (10-fold)	Ne,An,Sl,Lo,Fa,So	**91.79**
Proposed	DSMR	RF**	1:2 ratio	Ne,An,Sl,Lo,Fa,So	**91.00**
*An-Angry, Ne-Neutral, Sl-Slow, Lo-Loud, Fa-Fast, So-Soft; **RF-Random Forest					

TABLE 5. Performance evaluation with the state-of-the-art using "ESD dataset". The first and second best rates obtained for each class are in "bold and underlined respectively".

Method	Features	Classifier	Validation	"Emotions"	"Average recognition rate" (%)
I. Shahin [29]	MFCC	GMM-DNN hybrid classification	1:2 ratio	Ne,An,Sa,Di,Ha,Fe	83.96
S.Hamsa [30]	MFCC	Gradient Boosting	Cross validateion (10-fold)	Ne,An,Sa,Di,Ha,Fe	88.26
S.Hamsa [30]	MFCC	RF	Cross validation (10-fold)	Ne,An,Sa,Di,Ha,Fe	<u>89.60</u>
Proposed	DSMR	RF	1:2 ratio	Ne,An,Sa,Di,Ha,Fe	**91.52**
Proposed	DSMR	RF	Cross validation (10-fold)	Ne,An,Sa,Di,Ha,Fe	**90.90**
*An-Angry, Ne-Neutral, Sa-Sad, Di-Disgust, Ha-Happy, Fe-Fearful, **RF-Random Forest					

performance analysis and the results are reported in Table 4. The techniques proposed by Campbel *et al.* [10], Hong and Kwong [11], and Kinnunen *et al.* [12] reported average recognition rates underneath 75%. However, Shahin *et al.* [29], and Hamsa *et al.* [30] accounted 86.67% and 87.97%, respectively. Our model overtook their results with a recognition rate of 91%, which is about 4.33% and 3.03% superior than Shahin *et al.* [29], and Hamsa *et al.* [30], respectively.

In view of previous research endeavour, we indicate the virtue of the DSMR model for emotion recognition with the obtained results in Emirati accented Arabic language using ''ESD dataset''. Table 5 exhibits the performance of the proposed DSMR model with reference to other models using cepstral features for classification using ''ESD dataset''. The results show an increase in recognition rate of about 6.94%, 2.64% and 1.03%, respectively over the hybrid GMM-DNN [29], gradient boosting [30] and random forest classifier [30] models using cepstral features for emotion recognition. Table 6 shows the recognition score obtained for the various techniques in the state-of-the-art using SAVEE dataset. Clearly, our approach achieved more remarkable results than the recent literatures.

Table 7 shows the performance evaluation of the proposed feature extraction method with reference to the various feature extraction approaches employed in the state-of-the-art. The results indicate that the proposed model offers better performance than the various commonly used feature extraction techniques.

C. PERFORMANCE ANALYSIS IN TERMS OF COMPUTATIONAL COMPLEXITY

This section provides the theoretical analysis on computational complexity of the DSMR-based emotion recognition algorithm, which is important from an implementation point of view in real-time applications. Computational complexities of the proposed, random forest [30], and hybrid GMM-DNN [29] classifier-models utilizing cepstral features for emotion recognition are analyzed. Table 8 reports the computation times for the training and the evaluation, whereby we can notice that the proposed model achieves

TABLE 6. Performance evaluation with the state-of-the-art using SAVEE dataset. The first and second best rates obtained for each class are in "bold and underlined respectively".

Method	Features	Classifier	Validation	"Emotions*"	"Average recognition rate" (%)
E.Avots [44]	Spectrograms	SVM	Cross validation (10-fold)	An,Ha,Ne,Sa,Fe,Di,Su	77.40
K. Paliwal [45]	88 features	SVM	Cross validation (10-fold)	An,Ha,Ne,Sa,Fe,Di,Su	43.47
K. Paliwal [45]	88 features	RF	Cross validation (10-fold)	An,Ha,Ne,Sa,Fe,Di,Su	65.28
N.Hajarolasvadi [46]	Spectrogram	3D-CNN	Cross validation (10-fold)	An,Ha,Ne,Sa,Fe,Di,Su	81.05
S.Hamsa [30]	MFCC	RF	Cross validation (10-fold)	An,Ha,Ne,Sa,Fe,Di,Su	<u>85.60</u>
C.K Yogesh [43]	Hybrid BBO PSO features	ELM Kernal	70:30	An,Ha,Ne,Sa,Fe,Di,Su	62.38
Proposed	DSMR	RF	Cross validation (10-fold)	An,Ha,Ne,Sa,Fe,Di,Su	**89.75**
*An-Angry, Ha-Happy, Ne-Neutral, Sa-Sad, Ca-Calm, Fe-Fearful, Di-Disgust, Su-Surprise; **RF-Random Forest					

TABLE 7. Performance evaluation with the state-of-the-art feature extraction techniques using "RAVDESS dataset" and Random Forest classifier. The first and second best rates obtained for each class are in "bold and underlined respectively".

"Feature extraction"	"Average emotion recognition rate (%)"
MFCC	86.38
DWT-MFCC	86.79
MFCC-PNCC	87.54
x-vector	83.12
d-vector	84.78
PNCC-GFCC	85.68
Proposed DSMR	89.75

TABLE 8. Computational complexity associated with the hybrid GMM-DNN [29] and random forest [30] classifier models using cepstral features vs. proposed model.

Method	Average training time (sec)	Average testing time (sec)
Proposed	84,128.13	2.46
Random forest [30]	85,822.74	2.87
Hybrid GMM-DNN [29]	95,921.45	4.12

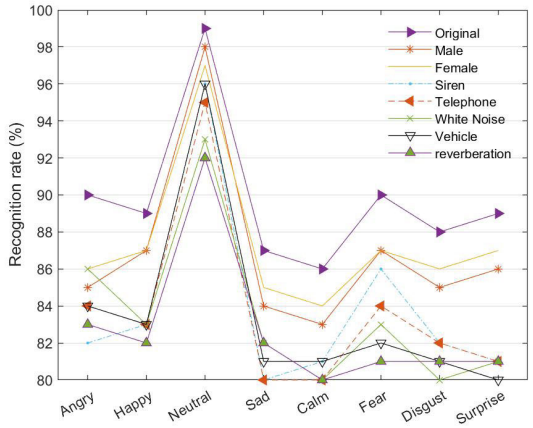
minimum computational complexity with associated recognition rate. These figures were obtained with an Intel core i7-3770, 3.40 GHz, 4 Cores machine.

D. EVALUATION OF THE SYSTEM PERFORMANCE IN NOISY AND REVERBERATED TALKING CONDITIONS

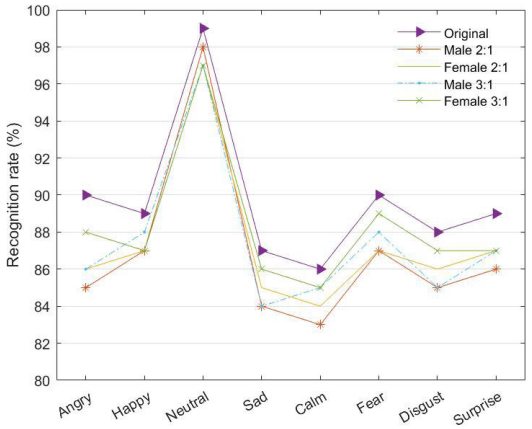
It has been known for many years that the speakers will increase their sound in the presence of background noise. Human listeners are intelligent enough to listen to the dominant voice rather than its noisy counter part. A noise reduction pre-processing module has been added in this work to develop this intelligibility in the proposed algorithm computationally. Here, we appraise the intelligibility of the proposed model to recognize the emotions in noisy talking conditions. We have used various noise samples such as ''other male voice, other female voice, siren noise, telephone ring, white noise, and vehicle noises''. The original speech signal is mixed with the noise at different dominance levels, in a ratio of 2:1 and 3:1 and are used in the evaluation phase. The reverberant emotional speech data is obtained based on convolution method. Specifically, impulse responses recording of different environments are convoluted with the clean data, in order to create the reverberant emotional speech data.

Fig. 8a indicates the obtained recognition rate at different noisy conditions using ''RAVDESS dataset''. Fig. 8b indicates the performance of the proposed model in the normal and noisy talking conditions at diverse dominance levels using ''RAVDESS dataset''. The proposed DSMR model reported an average recognition rates of 86.87% and 87.38%, respectively, for male and female voiced speech noises at 2:1 dominance levels. Whereas, 87.5% and 88.25% are obtained, respectively for male and female noises at 3:1 dominance levels. The results clearly illustrate that the performance of the system increases with the increase in dominance levels.

Fig.9a communicates the recognition rate obtained for the model at different noisy conditions using ''SUSAS dataset''. Our proposed framework attains an average recognition rate of 85.77% in noisy talking environments. Fig. 9b displays the performance of the proposed model in the normal and noisy talking conditions at distinct dominance levels using ''SUSAS dataset''. The proposed model ensures recognition rates of 86.45% and 85.32% respectively, for male and female voiced speech noises at 2:1 dominance levels. Then again the performance is improved to 87.33% and 88.32% respectively

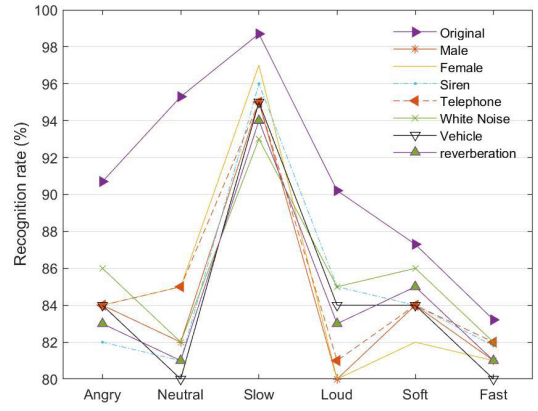


(a) Emotion recognition rate obtained in the presence of various noise.

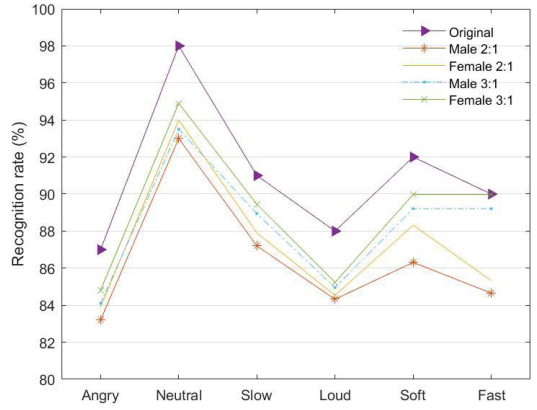


(b) Evaluation at various dominance levels.

FIGURE 8. Evaluation in terms of noise immunity using RAVDESS dataset.

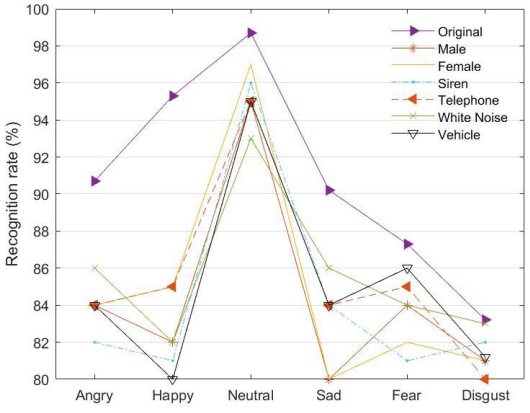


(a) Emotion recognition rate obtained in the presence of various noise.

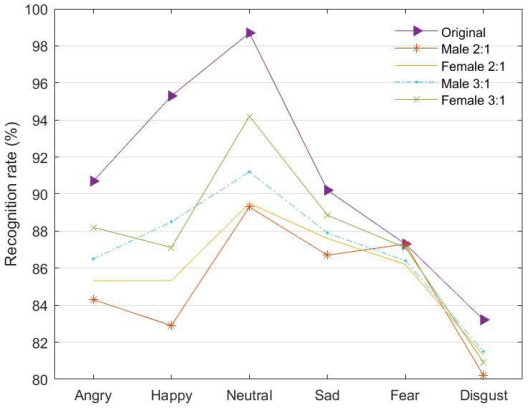


(b) Evaluation at various dominance levels.

FIGURE 9. Evaluation in terms of noise immunity using "SUSAS dataset".

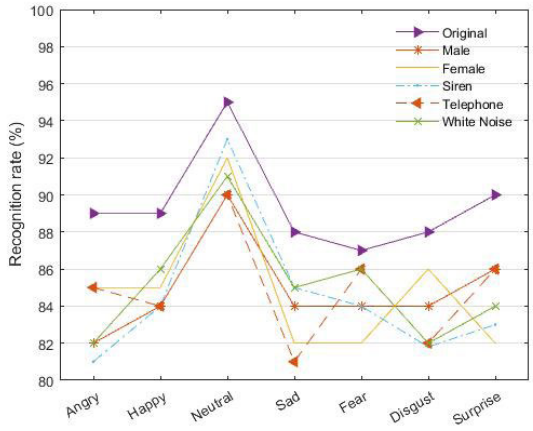


(a) Emotion recognition rate obtained in the presence of various noise.

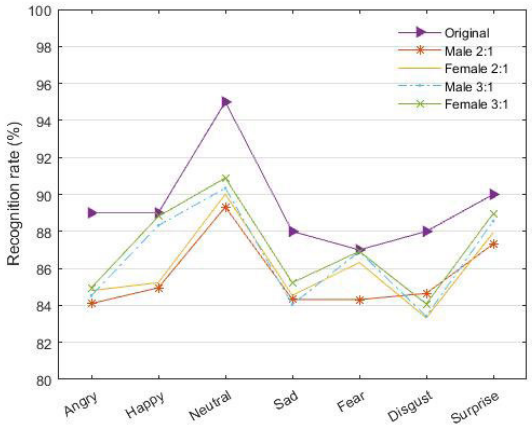


(b) Evaluation at various dominance levels.

FIGURE 10. Evaluation in terms of noise immunity using "ESD dataset".



(a) Emotion recognition rate obtained in the presence of various noise.



(b) Evaluation at various dominance levels.

FIGURE 11. Evaluation in terms of noise immunity using SAVEE dataset.

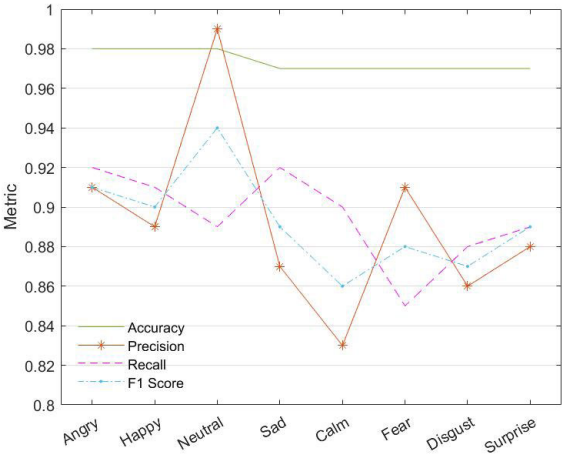


FIGURE 12. Performance evaluation parameters using RAVDESS dataset.

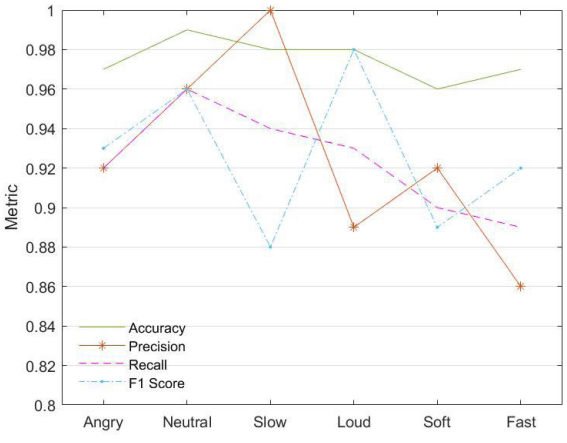


FIGURE 13. Performance evaluation parameters using "SUSAS dataset".

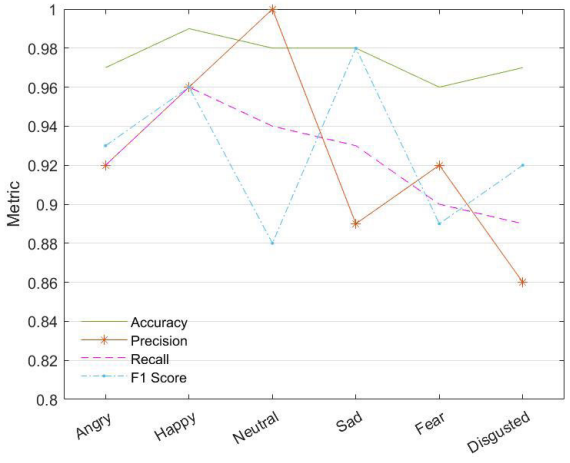


FIGURE 14. Performance evaluation parameters using "ESD dataset".

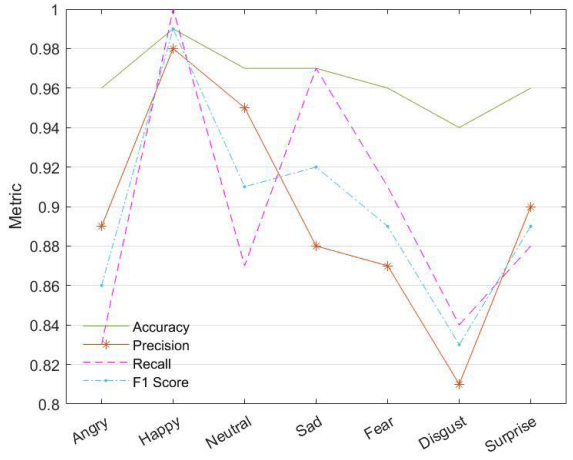
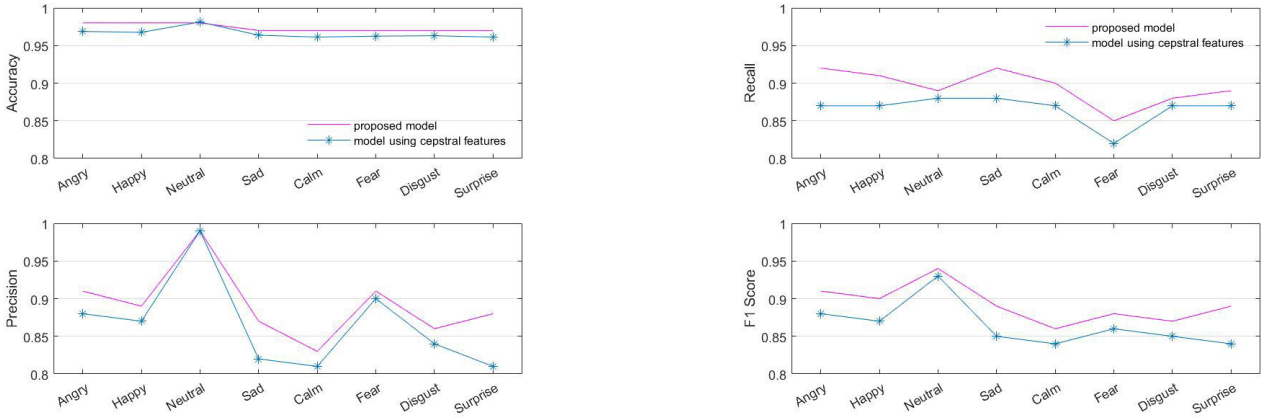


FIGURE 15. Performance evaluation parameters using SAVEE dataset.

for speech signal mixed with male and female voiced speech noises in a ratio 3:1.

Fig. 10a illustrates the performance of the proposed DSMR model at different noisy conditions using "ESD dataset" and obtained an average recognition rate of 85.80%. Fig. 10b represents the performance of the proposed model in the normal and noisy talking conditions at different dominance levels using "ESD dataset". Average recognition rates of



(a) Accuracy and Precision

(b) Recall and F1 score

FIGURE 16. Performance evaluation matrices obtained for the proposed model and random forest classifier model using cepstral features.

86.87% and 87.38% are obtained for male and female voiced speech noises at 2:1 dominance levels. Under 3:1 dominance levels, 87.5% and 88.25% are attained for male and female voiced speech noises respectively.

Fig. 11a exhibits the noise susceptibility of our model using SAVEE dataset and reported an average rate of 85.51%. Fig. 11b represents the performance of the proposed model in the normal and noisy talking conditions at different dominance levels using SAVEE dataset. The proposed model yields an average recognition rate of 85.51% in noisy talking environments. From Fig. 11a, it is apparent that the recognition rate shows an increasing trend with the increase in dominance levels.

E. PERFORMANCE EVALUATION METRICS

In this section, different performance evaluation metrics such as accuracy, precision, F1 score and recall [47] values are accounted to have a clear portrayal of the system performance.

The performance metric accuracy measures the number of observations classified correctly [47]:

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{18}$$

''where tp, tn, fp and fn represents the true positive, true negative, false positive and false negative values, respectively, obtained from the confusion matrix''. Precision is the ratio of number of all correctly predicted observation and all cases in the data [47]:

$$Precision = \frac{tp}{tp + fp} \tag{19}$$

Recall is the ratio of all correctly identified positive observations to all actual positive classes [47]:

$$Recall = \frac{tp}{tp + fn} \tag{20}$$

$$F1Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{21}$$

F1 score is the weighted average of precision and recall that takes into account the false positive and false negative [47].

Fig. 12, 13, 14 and 15 show the emotion-wise Accuracy, Precision, Recall and F1 scores, obtained with ''RAVDESS, SUSAS, ESD and SAVEE datasets'', respectively. We have noticed that all the metrics score above 80% in all the distinct speech datasets. The results indicate the highest performance obtained in neutral class, whereas the comparatively least performance can be found in disgust emotion class. Fig. 16 clearly illustrates the performance evaluation matrices obtained for the proposed model and random forest classifier model using cepstral features [30]. The given results are based on the evaluation using ''RAVDESS dataset''. We can spot a remarkable enhancement in the performance of the proposed model than the model using cepstral features.

VI. CONCLUSION

This paper presents an algorithm for emotion recognition using deep sparse metric representation and random forest classifier. An auto-correlation based noise reduction module is also incorporated to ensure better performance in noisy real-time applications. The proposed model is evaluated using English and Arabic languages in noisy and stressful talking situations to have a clear view of its performance in real applications. Our results show that the proposed DSMR-random forest model has higher emotion recognition rate, Precision, Recall and F1 score than those of other models in the recent literature. All models are evaluated using four distinct datasets including ESD private Arabic dataset, SUSAS, RAVDESS and SAVEE public English datasets. The performance of the proposed model has been improved significantly by the use of DSMR features in the normal and challenging talking conditions. The computational complexity of the model is controlled by the usage of SPCA feature selection

approach. The key obstacles encountered in this research, however, are due to the lack of connectivity to natural standard emotional datasets.

REFERENCES

[1] J. Liu, Y. Xu, S. Seneff, and V. Zue, "CityBrowser II: A multimodal restaurant guide in mandarin," in *Proc. 6th Int. Symp. Chin. Spoken Lang. Process.*, Dec. 2008, pp. 1–4.

[2] A. Rawat and P. K. Mishra, "Emotion recognition through speech using neural network," *Int. J*, vol. 5, pp. 422–428, May 2015.

[3] T. Özseven, "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition," *Appl. Acoust.*, vol. 142, pp. 70–77, Dec. 2018.

[4] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, Nov. 2011.

[5] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107141.

[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, no. 24, pp. 9554–9564, Dec. 2015.

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

[8] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiaion," in *Proc. LREC*, 2006, pp. 1123–1126.

[9] K. Wang, N. An, B. Nan 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] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, nos. 2–3, pp. 210–229, Apr. 2006.

[11] Q. Y. Hong and S. Kwong, "A genetic classification method for speaker recognition," *Eng. Appl. Artif. Intell.*, vol. 18, no. 1, pp. 13–19, Feb. 2005.

[12] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 277–288, Jan. 2006.

[13] 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, no. 1, pp. 77–90, Mar. 2015.

[14] 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, nos. 1–2, pp. 59–71, Jul. 2012.

[15] 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, Dec. 2016.

[16] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to Gaussian mixture modeling," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 29, no. 4, pp. 393–399, Jul. 1999.

[17] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, 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] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. 15th Annu. Conf. Int. speech Commun. Assoc.*, 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] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000.

[21] H. Palo, M. N. Mohanty, and M. Chandra, "Use of different features for emotion recognition using MLP network," in *Computational Vision and Robotics*. Springer, 2015, pp. 7–15.

[22] C. Min 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.

[23] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 10–21, Jan. 2011.

[24] Y. Kuang and L. Li, "Speech emotion recognition of decision fusion based on DS evidence theory," in *Proc. IEEE 4th Int. Conf. Softw. Eng. Service Sci.*, May 2013, pp. 795–798.

[25] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018.

[26] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Commun.*, vol. 99, pp. 80–89, May 2018.

[27] B. Zhang, G. Essl, and E. M. Provost, "Recognizing emotion from singing and speaking using shared models," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 139–145.

[28] A. Huang and P. Bao, "Human vocal sentiment analysis," 2019, *arXiv:1905.08632*. [Online]. Available: http://arxiv.org/abs/1905.08632

[29] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion recognition using hybrid Gaussian mixture model and deep neural network," *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[30] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, "Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier," *IEEE Access*, vol. 8, pp. 96994–97006, 2020.

[31] A. Mahmoodzadeh, H. R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Single channel speech separation with a frame-based pitch range estimation method in modulation frequency," in *Proc. 5th Int. Symp. Telecommun.*, Dec. 2010, pp. 609–613.

[32] D. Baby, T. Virtanen, T. Barker, and H. Van Hamme, "Coupled dictionary training for exemplar-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2883–2887.

[33] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5, Mar. 1999, pp. 2443–2446.

[34] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Oct. 2006.

[35] B. Liu, T. Xie, Y. Xu, M. Ghavamzadeh, Y. Chow, D. Lyu, and D. Yoon, "A block coordinate ascent algorithm for mean-variance optimization," 2018, *arXiv:1809.02292*. [Online]. Available: http://arxiv.org/abs/1809.02292

[36] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005.

[37] I. Barandiaran, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 1–22, Aug. 1998.

[38] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[39] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, 1997, pp. 1743–1746.

[40] J. A. Grissom, D. Kalogrides, and S. Loeb, "Using student test scores to measure principal performance," *Educ. Eval. Policy Anal.*, vol. 37, no. 1, pp. 3–28, Mar. 2015.

[41] Y. Gao, B. Li, N. Wang, and T. Zhu, "Speech emotion recognition using local and global features," in *Proc. Int. Conf. Brain Informat.* Springer, 2017, pp. 3–13.

[42] H. Holmström and V. Zars, "Effect of feature extraction when classifying emotions in speech—An applied study," Tech. Rep., 2018.

[43] Y. C. K., M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, and K. Polat, "Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech," *Appl. Soft Comput.*, vol. 56, pp. 217–232, Jul. 2017.

[44] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 975–985, Jul. 2019.

[45] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, Jan. 2019.

[46] N. Hajarolasvadi and H. Demirel, "3D CNN-based speech emotion recognition using K-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, May 2019.
[47] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," in *Proc. NESUG Health Care Life Sci.*, Baltimore, MD, USA, vol. 19, 2010, p. 67.

SHIBANI HAMSA received the bachelor's and master's degrees in electronics and communication engineering from Mahatma Gandhi University, India. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Khalifa University. From 2019 to 2021, she worked as a Research Engineer with the ECE Department, Khalifa University, after a stint as a Research Associate with the University of Sharjah. She worked as a Lecturer with Mahatma Gandhi University. Her areas of research interests include artificial intelligence, deep learning, and natural language processing.

YOUSSEF IRAQI (Senior Member, IEEE) is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, Khalifa University, United Arab Emirates. Before joining KU, he was the Chair of the Computer Science Department, Dhofar University, Oman, for four years. Before that, he was a Research Assistant Professor with the School of Computer Science, University of Waterloo, Canada. He has published more than 110 research articles. His research interests include adaptive resource management in wireless networks, blockchain, trust and reputation management, and stylometry. In 2008, he received the IEEE Communications Society Fred W. Ellersick Paper Award in the field of communications systems.

ISMAIL SHAHIN (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Southern Illinois University, Carbondale, USA, in 1992, 1994, and 1998, respectively. He is currently an Associate Professor with the Department of Electrical Engineering, University of Sharjah, United Arab Emirates. His research interests include speech recognition, speaker recognition under neutral, stressful, emotional talking conditions, emotion and talking condition recognition, gender recognition using voice, accent recognition, and natural language processing. He received many research grants, where his major role was the main principal investigator of all the projects. He has more than 80 journal and conference publications. He has remarkable contribution in organizing many conferences, symposiums, and workshops. Finally, he taught wide range of undergraduate and graduate electrical and computer engineering courses.

NAOUFEL WERGHI (Senior Member, IEEE) received the Habilitation and Ph.D. degrees in computer vision from the University of Strasbourg. He is currently an Associate Professor with the Electrical Engineering and Computer Science Department, Khalifa University, United Arab Emirates. His research interests include computer vision, machine learning with application to medicine, biometry, and remote intelligent systems, where he has been leading several funded projects. He received four papers awards and published more than 170 journal and conference papers. He served as the Publication Chair for the IEEE International Conference on Image Processing 2020. He is an Associate Editor for the *Journal on Image and Video Processing* (Eurasip).