IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 8, 2020, accepted April 8, 2020, date of publication May 6, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991811

Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier

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

[1]Center for Cyber-Physical Systems (C2PS), Department of 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), Ismail Shahin (ismail@sharjah.ac.ae), Youssef Iraqi (youssef.iraqi@ku.ac.ae), and Naoufel Werghi (naoufel.werghi@ku.ac.ae)

ABSTRACT This research aims to design and implement an artificial emotional intelligence system that is capable of identifying the unknown emotion of the speaker. To that end, we propose a novel framework for emotion recognition in the presence of noise and interference. Our approach accounts for energy, time and spectral parameters to examine the emotion of the speaker. However, rather than using Gammatone filterbank and short-time Fourier transform (STFT), commonly adopted in the literature, we propose employing a novel wavelet packet transform (WPT) based cochlear filterbank. Our system, coupling this representation with random forest classifier, shows superior performance over other existing algorithms when appraised on three distinct speech corpora in two different languages, and considering also stressful and noisy talking conditions.

INDEX TERMS Emotion recognition, noise reduction, cochlear filterbank, feature extraction.

I. INTRODUCTION

Artificial Emotional Intelligence or emotional AI is a hot research area in this decade. People use many of non-linguistic signs such as outward facial appearances, gestures, non-verbal communications using body language and tone of voice to express their emotions. Emotional AI aims to detect emotions just the way humans do from multiple channels. Emotional AI is often used for smart security in the banking sector, intelligent call centers and customer support, medical and forensic applications, stress and anxiety management and organizing voice mail messages based on emotions [1]. Emotional AI systems still face several challenges including: low accuracy rates of the employed classifiers, higher computational complexity of the hybrid classifier models, and scarcity in the availability of natural datasets. The objective of this work is to design and implement a novel algorithm with higher accuracy and reduced computational complexity for emotion recognition in real-world applications. The proposed algorithm uses time, frequency, and power spectral vectors as features and random forest as a classifier. The performance and the computational efficiency of the algorithm is analyzed using English and Arabic emotional speech corpora. Furthermore, Speech Under Simulated and Actual Stress (SUSAS) dataset [2] is used to evaluate the performance of the algorithm in stressful talking conditions. Speech processing modules are susceptible to noise and interference in natural environments. It might affect system performance in real-world applications. The proposed algorithm is designed in such a way to recognize the emotion of the speaker, even in the presence of noise and interference, by incorporating a noise reduction pre-processing module for noise suppression, and a pitch-based feature segregation filter. This paper is structured as follows. Literature review is given in Section II, system description is explained in Section III, experimental results are described in Section IV, and finally the conclusion is given in Section V.

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

II. LITERATURE REVIEW

Emotion recognition from speech in conjunction with noise reduction has a great impact on Natural Language Processing (NLP) for the successful implementation of an effective human-machine interaction system. Alonso *et al.* [1] proposed Support Vector Machine (SVM) classifier-based

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

emotion recognition system to categorize five emotions on the Berlin dataset [3] and obtained an accuracy of 94.9%. The emotions classified are angry, happy, neutral, sad and disgust. Luengo *et al.* [4] also introduced an SVM-based emotion recognition system and obtained an average recognition rate of 78.3% for six emotions. These emotions are angry, happy, neutral, sad, fearful and disgust [4]. Wang *et al.* introduced an emotion recognition technique using Fourier parameters instead of cepstral parameters and obtained a recognition rate of 88.8% for six distinct emotions using SVM classifier [5]. Shahin and Ba-Hutair [6] applied third-order circular suprasegmental hidden Markov models (CSPHMM3s) classifier on Mel frequency cepstral coefficents (MFCC) feature vectors to recognize talking conditions. CSPHMM3s based emotion recognition system using SUSAS dataset demonstrates an accuracy of 76.3% containing six different talking conditions.

Shukla *et al.* [7] introduced an emotion recognition system using HMM classifier along with 13-dimensional feature vectors achieving an accuracy of 93.9% in angry, lombard (speech produced in noise), neutral, and sad talking conditions in the SUSAS dataset [8]. Shahin *et al.* introduced, implemented and evaluated three distinct classifiers: HMM, Second order circular hidden markov model (CHMM2s) and Suprasegmental hidden markov model (SPHMMs) to recognize six diverse stressful conditions in the SUSAS dataset. Their work shows that the SPHMMs recognition rate is higher than that using the other two classifiers [8].

Vlassis and Likas [9] proposed a Gaussian mixture model (GMM) based emotion recognition method utilizing the global features extracted from the speech signal. GMM model obtained a recognition rate of 75% and 89.12% for speaker independent and speaker dependent emotion recognition, respectively [9]. With the advent of deep learning, deep neural network (DNN) has been employed in many latest research work [10]–[12]. Stuhlsatz *et al.* [10] introduced a ''Generalized Discriminant Analysis (GerDA)'' based on DNN for emotion recognition. Their results, averaged over nine databases, show significant improvement over the SVM based emotion recognition techniques in the literature. Han *et al.* [11] proposed emotion recognition framework using DNN and extreme learning machine and offer 20% increase in the recognition rate over classifiers such as HMMs and SVMs. Zheng *et al.* [12] introduced a structural overview of an emotion recognition system based on deep convolution neural networks (DCNNs). The MFCC based models gave an accuracy of 71.6%.

Hybrid classifier schemes have been recommended for an efficient speech emotion recognition [13]–[16]. Li *et al.* [13] proposed deep neural network - hidden Markov models (DNN-HMMs) for speech emotion recognition. Huang *et al.* [14] introduced a combined classifier that is made up of deep belief network (DBN) and SVM. Emotion recognition accuracy based on their combined classifier and using four different emotions is 86.5%. Tashev *et al.* [15] inspected combining GMM-based low-level feature extractor with a neural network. Their proposed architecture was evaluated on a Mandarin database with four emotions only: angry, happy, neutral and sad. Their results, based on GMM-DNN, gave weighted and unweighted emotion recognition accuracy of 48.0% and 41.5%, respectively [15]. Shahin *et al.* [16] introduced an emotion recognition model using Hybrid GMM-DNN classifier. Their system was evaluated on Arabic Emirati-Accented and SUSAS datasets and they obtained an average recognition rate of 83.97% and 86.67%, respectively. Kim and Park [17] proposed a multistage data selection method for speech emotion recognition from previous voice data accumulated on personal devices. Multistage data selection is conducted using log likelihood distance based measure and a universal background model [17] and obtained an average recognition rate of 83.9%. Literature shows that many of the researches obtained higher accuracy by using hybrid classifier models. However the computation complexity associated with the hybrid classification model need to be accounted for real time applications.

With respect to the previous works, we propose in this paper a combined use of Wavelet Packet Transform-based cochlear filter bank for noise suppression, multi-dimensional feature vectors, pitch based dominant feature extraction, and random forest classifier.

The major contributions of this work clearly appear in:

- The novel approach which focuses on recognizing emotions independent of text and speaker, integrated with noise reduction, multi-dimensional acoustic vectors, and random forest classifier.
- A novel perspective for Time-Frequency decomposition using the proposed WPT cochlear model.
- Designed a system which offers robust performance in both English and Arabic languages and noisy talking conditions made it suitable for real time applications.

To the best of our knowledge, this work is the first effort to recognize emotions using Arabic Emirati-emphasized speech dataset (ESD), English dataset called Speech under simulated and actual stress (SUSAS), and Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset. Furthermore, we have conducted an experimental evaluation comparing our framework with different techniques in the recent literature. The series of experiments conducted in our work are listed below:

- The system performance has been weighed using SUSAS database in both normal and noisy talking conditions.
- Random forest classifier has been compared with recent work using the RAVDEES English dataset.
- Random forest classifier has been compared with the recent hybrid GMM-DNN emotion recognition model using Arabic Emirati-accented dataset.
- The computational complexity of the proposed model has been evaluated.

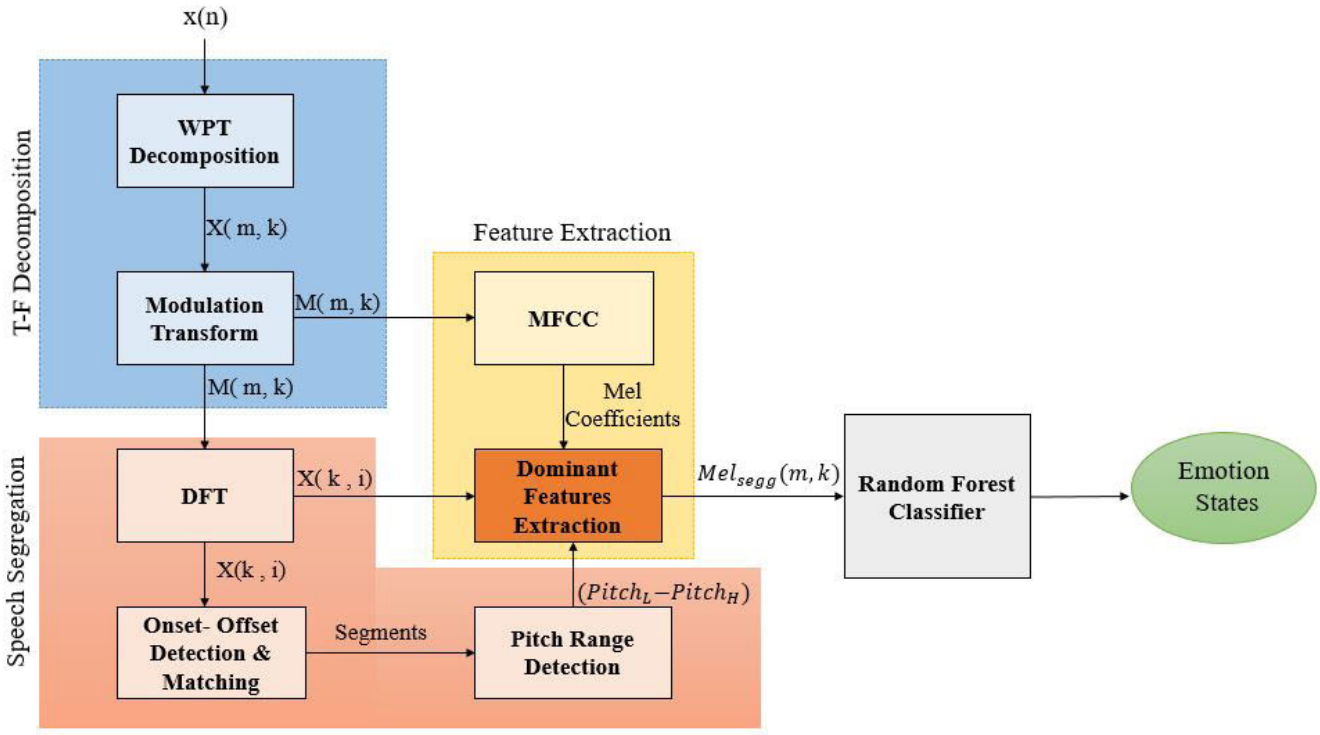

FIGURE 1. Emotion recognition basic block schematic.

III. SYSTEM DESCRIPTION

Fig.1 shows block diagram of the proposed emotion recognition system. The system consists of Time-frequency (T-F) decomposition, speech segregation, feature extraction and classifier modules. The T-F decomposition system is skilled with Wavelet packet transform (WPT) based cochlear filtering using Daubechies 4 (db4) wavelets and modulation transform techniques. MFCC computed from the output of the cochlear filter bank is used as the feature vector. Speech separation and feature selection are achieved by means of onset and offset parameters and frequency mask. Random forest is used as the classifier.

A. T-F DECOMPOSITION

The proposed framework utilizes a novel cochlear filter bank and modulation transform to obtain the T-F decomposition of the applied input signal. Several authors [18] have shown that the energy of the speech signal, computed from the output of an auditory filter bank can be used to recognize emotions in the speech. Most of the research in the field of Computational Auditory Scene Analysis (CASA) are either based on Gammatone filterbank or on Short Time Fourier Transform (STFT) filter bank. But from the state of the art, it is clear that the computational complexity associated with the Gammatone filterbank is very high [19]

STFT-based static extraction always shows a dilemma in time-frequency representation. In this work, we have used a novel WPT cochlear filter bank model for acoustic energy extraction.

1) WPT COCHLEAR FILTER BANK

Table 1 shows the critical bands of human auditory system. The proposed WPT based cochlear front end is designed to decompose the input signal into 18 channels to mimic the human cochlear model (see Fig.2). In this work, we have used Duabechies 4 wavelets to model the acoustic filter bank as it offers better performance than other wavelets [21]. Input speech signal is decomposed into frames of length 20ms. Each frame is allowed to pass through a low pass filter and high pass filter in each level of decomposition. Every level down-samples the signal by two before passing it to the next level of decomposition.

2) MODULATION TRANSFORM

The input speech signal is passed through the proposed cochlear model and the complex result obtained is the time-frequency representation of the input mixture $x(n)$, which records magnitude and phase of each point in time and frequency. This can be expressed as:

$$X(m,k) = \underset{k}{WPT}(x(n)) \tag{1}$$

The signal $X(k)$ extracted in a time slot m by the WPT decomposition process is represented as X_K and consists of modulating signal M_K and carrier signal C_K [22]. This can be expressed as

$$X_K = M_K C_K \tag{2}$$

The message signal M_K can be extracted from $X(m,k)$ by means of envelop detection since it is an amplitude

TABLE 1. Critical bands of human ear [20].

Band No.	Center Frequency	Band Width(Hz)	Band No.	Center Frequency	Band Width(Hz)
1	50	0-100	13	1850	1720-2000
2	150	100-200	14	2150	2000-2320
3	250	200-300	15	2500	2320-2700
4	350	300-400	16	2900	2700-3150
5	450	400-510	17	3400	3150-3700
6	570	510-630	18	4000	3700-4400
7	700	630-770	19	4800	4400-5300
8	840	770-920	20	5800	5300-6400
9	1000	920-1080	21	7000	6400-7700
10	1170	1080-1270	22	8500	7700-9500
11	1370	1270-1480	23	10500	9500-12000
12	1600	1480-1720	24	13500	12000-15500

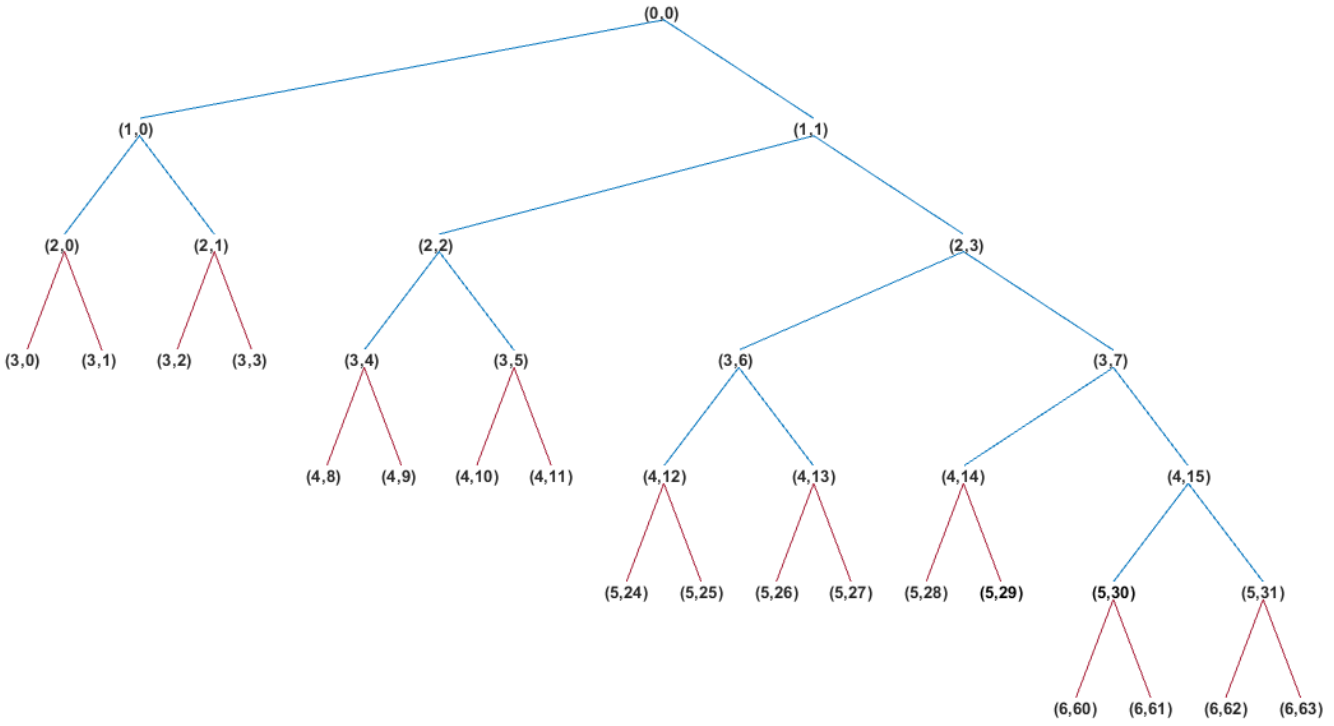

FIGURE 2. 18 channel human cochlear model.

modulated signal:

$$M_K = ED[\underset{k}{WPT}[x(n)]] \tag{3}$$

where ED denotes envelop detection. The envelope detector used here is an incoherent detector which is based on Hilbert envelope [23] as it can create a modulation spectrum with large area covered in the modulation frequency domain. It also acts as a magnitude operator for complex-valued subbands given as:

$$M_K \cong |X_K| \tag{4}$$

Operator in Equation (4) represents isomorphic or similar algebraic structures. Which means that modulating signal M_K is approximately equal to the magnitude of the complex valued subband X_K. Then, the WPT based discrete modulation transform of the signal $x(n)$ can be defined as

$$X(k,i) = DFT[ED[\underset{k}{WPT}[x(n)]] \tag{5}$$

$$= \Sigma_{m=0}^{P-1} M(m,k) e^{(\frac{-j2\pi mi}{P})}, \quad i = 0, 1, .., P-1. \tag{6}$$

DFT represents the discrete Fourier transform of length P. k and i are the acoustic frequency, and modulation frequency,

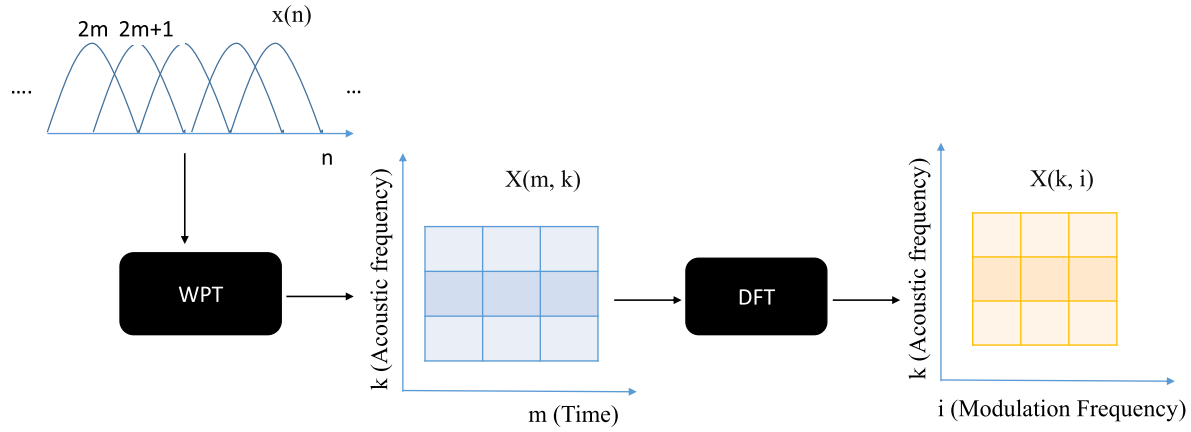

FIGURE 3. Modeling of modulation transform framework.

respectively. The entire modulation transform framework is depicted in Fig. 3.

B. SPEECH SEGREGATION

The obtained modulation transform $X(k, i)$ is smoothed by means of a low pass filter to maintain the major intensity variations while reducing the frequency fluctuations. Then, the first derivative of the smoothed signal is computed to identify the peak and valley points of the modulation spectrum. These peak and valley points are referred to as onset and offset points respectively. Onset and offset points within the human pitch range are selected and grouped. Pitch frequency range of the higher energy segment is computed to segregate the dominant signal [22].

C. FEATURE EXTRACTION

1) MFCC

Acoustic filters are designed based on the source filter model [18], [24]. The friction produced by the source excreted on to the walls of the vocal cord produces quasi periodic sound waves. It may undergo changes in spectral characteristics when it passes through the mouth, throat and nasal cavities. The entire process is basically controlled by motor areas in the brain with the feedback loop from the auditory and somatosensory system [25]–[27]. The control on these muscles altered by the emotions leads to the changes in its acoustic features. Three groups of acoustic features are analyzed for the evaluation of emotion characteristics of the speech signals. They are filter bank features, spectral features, and temporal features. Several authors [18] have mentioned that the energy of the speech signal, calculated using an auditory filter bank, can be used to recognize emotions in speech. Fig.4 shows the MFCC feature extraction from the output of the WPT cochlear filter bank. The sub-band energies are calculated using Mel filter bank, which replicates the human auditory system. Inverse Fourier transform of the log scale of the sub band energies represents the Mel frequency cepstral coefficients and represented as $\underset{k}{Mel}(m)$. The proposed algorithm uses 28 Mel filters and 16 MFCCs [28].

2) DOMINANT FEATURES EXTRACTION

Dominant features extraction is employed to segregate the original speech feature vectors from their interference part by means of a dynamic frequency mask. The mask extracts the dominant features based on the power spectral density of the input signal and the pitch frequencies of the dominant and interference parts present in the input signals. Dynamic frequency mask for each filter channel is designed as follows: We set the means of the modulation spectral energy in the dominant pitch range and the interference pitch range as E_t and E_i, respectively [22].

$$E_t = \frac{\Sigma(|X(k,i)|^2)}{(Pitch_H - Pitch_L)_{dominant}} \tag{7}$$

$$E_i = \frac{\Sigma(|X(k,i)|^2)}{(Pitch_H - Pitch_L)_{Interference}} \tag{8}$$

The filter component is designed as,

$$F = \frac{E_t}{E_t + E_i} \tag{9}$$

A linear feature selection filter can be designed with magnitude F and phase response ϕ(i)=i. The filter response of single channel in the time domain is designed as:

$$f(m) = \Sigma_{i=0}^{I-1} F e^{(j\phi(i))} e^{\frac{(j2\pi mi)}{I}} \tag{10}$$

The designed filter of linear response $f(m)$ is used to segregate the dominant set of features for classification. Segregated dominant features can be estimated by the convolution over the variable m of the designed filter response $f(m)$ of each filter channel with the set of available features $\underset{k}{Mel}(m)$ in the corresponding channel.

$$Mel_{segg}(m,k) = [\underset{k}{Mel}(m) * f(m)] \tag{11}$$

$*$ represents the convolution operator.

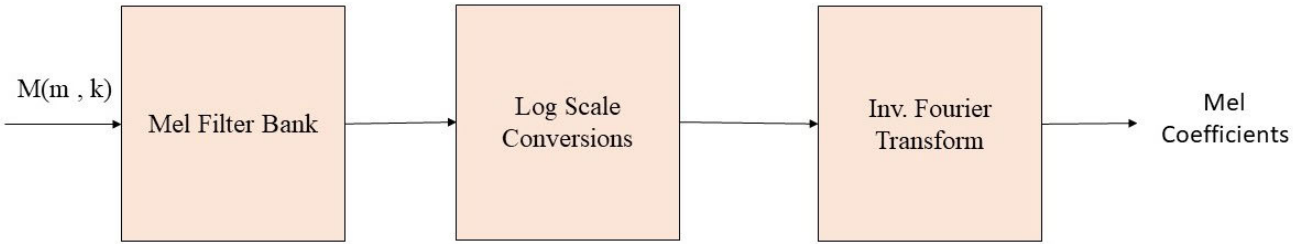

FIGURE 4. MFCC feature extraction from modulating signal $M(m, k)$.

$Mel_{segg}(m, k)$ is the set of extracted dominant feature used for classification.

D. CLASSIFICATION

The major part of machine learning strategies aims at designing a proper classification framework that defines the class an observation belongs to. The ability to precisely classify emotions is extremely valuable for various emotional intelligent applications. In such situations, appreciable performance is expected from ensemble classification algorithms. Ensemble algorithms combine more than one algorithm of same or different type for categorizing objects. Random forest is an ensemble algorithm that works on the following principle, 'a number of weak estimators, combined together, form a strong estimator'.

The random forest classifier designed here is a set of 100 decision trees. Let N and M represent the total number of emotions and the total number of samples, respectively. A set of bootstrap samples $n < N$ is selected for each decision tree. Then decision trees are built by assigning $m < M$ variables at each node until exhausting all variables. In each node of the tree, we fit the classification model with the m variables and find the cutoff. After training, predictions from the unknown samples can be determined by taking the majority vote of all predicted class from all the individual trees [29].

IV. EXPERIMENTS AND DISCUSSION

In this work, we focus on evaluating a text-independent and speaker-independent emotion recognition framework using English and Arabic languages in both stressful and noisy emotional talking conditions. We have conducted series of empirical evaluations among different emotion recognition techniques in the recent literature. Emotion recognition rate is the main criterion used in this work for evaluation, and it is defined as the ratio of the total number of correct identifications to the total number of trials.

In addition, the following experiments are conducted to study the performance and feasibility of the algorithm in real-time applications:

- Evaluation based on performance metrics such as Accuracy, Precision, Recall and F1 score.
- Empirical analysis with the recent literature.
- Statistical significance tests of the results.
- Assessment of the system performance in noisy talking conditions.
- Analysis on the computational complexity.

The algorithm is evaluated using three different datasets: The Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset, speech under simulated and actual stress (SUSAS) dataset, and Arabic Emirati emphasized speech dataset (ESD). The details of the datasets are as follows.

A. RAVDESS DATASET

The RAVDESS is a validated multimodal database of emotional speech and song [30]. The database encompasses 24 professional players which includes 12 female and 12 male speakers, uttering low lexically-matched statements in North American accent. Speech embraces angry, happy, neutral, sad, fearful, disgust, calm, and surprise expressions. Two lexically matched statements were spoken by every speaker in 60 trials constituting 1440 files. The recognition rate of the algorithm using RAVDESS dataset is evaluated using 10-fold cross validation to analyse the classification efficiency of the algorithm using English language. Fig.5 shows the emotion recognition accuracy of the algorithm using RAVDESS dataset. The results show that the highest recognition rate is obtained for neutral talking condition and it is almost stable for the other emotional conditions. In this figure, the average emotion recognition rate is 86.38%.

B. SUSAS DATASET

SUSAS dataset consists of five domains which have an array of stresses and emotions [2]. The database has two sections; one is simulated speech under stress which is termed as simulated domain. The second one is actual speech under stress which is termed as actual domain. A group of 32 speakers including 19 male and 13 female speakers in the age group 22 to 76 years were asked to pronounce more than 16,000 words [2]. All speech tokens were sampled by 16 bits analog to digital (A/D) converter at a sampling frequency of 8 kHz. The samples of signals were pre-emphasized and then segmented into frames of 20 ms each with 31.25% overlap between consecutive frames. The emphasized speech signals were implemented every 5 ms to 30 ms Hamming

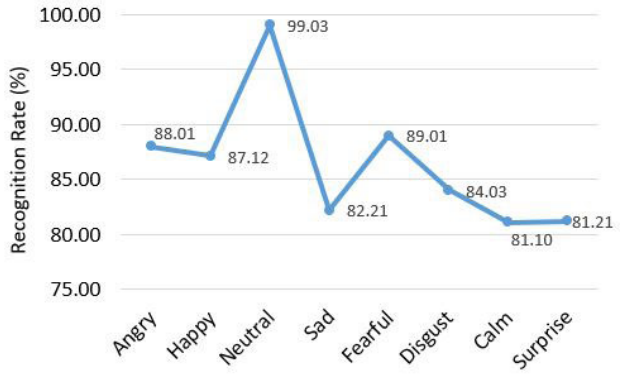

FIGURE 5. Emotion recognition rate evaluation for the proposed algorithm using RAVDESS dataset.

FIGURE 6. Emotion recognition rate evaluation for the proposed algorithm using SUSAS dataset.

window. In this work, 10-fold cross validation is used to evaluate the efficiency of the algorithm using SUSAS dataset. This section evaluates the classification efficiency of the algorithm in the stressful talking conditions. The classified talking conditions are angry, neutral, slow, loud, soft and fast. Fig 6 depicts the stressful recognition accuracy of the algorithm using SUSAS dataset. The results show a stable recognition rate for slow, loud, soft and fast talking conditions. Recognition rate obtained for angry talking condition is comparatively less than other talking conditions. The obtained average emotion recognition rate of this figure is 88.68%.

C. ESD

Emirati-Emphasized Arabic speech dataset (ESD) is an acted emotional dataset consisting of speech signals from 25 male and 25 female native Emirati speakers with ages spanning from 14 to 55-year old [16]. Each speaker uttered 8 common Emirati sentences that are heavily utilized in the United Arab Emirates society. Every speaker expressed the eight sentences in each of angry, happy, neutral, sad, fearful and disgust emotions 9 times with a span of 2 to 5 seconds. The recognition rate of the algorithm using ESD dataset is evaluated using 10-fold cross validation to analyse the classification efficiency of the algorithm using Arabic language in

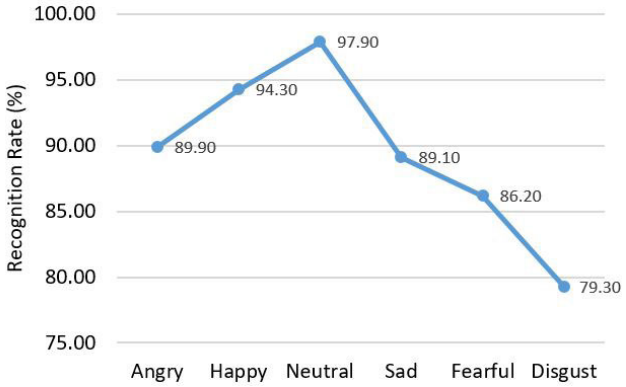

FIGURE 7. Emotion recognition rate evaluation for the proposed algorithm using ESD dataset.

Emirati accent. Fig.7 shows the emotion recognition accuracy of the proposed algorithm using ESD dataset. The results indicate that the system performance is almost stable across all the classes except disgust talking condition. The average emotion recognition rate is 89.45%.

D. ANALYSIS OF THE SYSTEM PERFORMANCE WITH THE RECENT LITERATURE

In congruence with the previous research efforts, we show the effectiveness of the proposed technique for emotion recognition in English, Arabic languages and Stressful talking conditions by using RAVDESS, SUSAS, and ESD datasets. We have used 10-fold cross-validation for the performance evaluation. Some recent research methods and their results are reported in Table 2, 3 and 4, along with our 5-fold and 10-fold cross-validation experimental results. Table 2 depicts the published state of the art emotion recognition rate on the data from RAVDESS dataset, together with their counterparts in our framework variants, using the same evaluation set-up. The results indicate that our proposed system achieves an enhancement, in the average emotion recognition rate, of 14.16%, 12.16%, 25.8% and 2.74% over the results reported by Biqiao *et al.* (St Hier) [31], Biqiao *et al.* (Mt Hier) [31], Huang *et al.* [32], and Shahin *et al.* [16], respectively.

Table 3 reports emotion recognition rate of the various classifier systems in the literature along with the proposed method using SUSAS dataset. The results show that the average recognition rate of the proposed method is 15.87%, 19.97%, 20.27%, and 2.00% superior to its counterparts in the techniques proposed by Campbel *et al.* [33], Hong and Kwong [34], Kinnunen *et al.* [35], and Shahin *et al.* [16], respectively.

Table 4 depicts the performance of recent classifier systems in recent literature using ESD. The results show an increase of the recognition rate of 4.30% and 5.64% obtained with our system variants over the hybrid GMM-DNN classifier [16].

A statistical significance test has been incorporated in this work to demonstrate whether the obtained improvement in emotion recognition rate over the GMM-DNN [16] based technique is actual or arising from statistical variations.

TABLE 2. Performance analysis using RAVDESS dataset. The best rates obtained for 4,6 and 8 classes are in bold.

Method	Feature extraction	Classification	Emotion class	Validation	Considered emotions*	Average recognition rate (%)
Z. Biqiao [31]	LLD (St Hier)	SVM	6	5 Fold	A,H,N,S,C,F	79.67
Z. Biqiao [31]	LLD (Mt Hier)	SVM	6	5 Fold	A,H,N,S,C,F	81.67
A. Huang [32]	MFCC, STFT	CNN	4	10-fold	A,H,N,S	72.20
I. Shahin [16]	MFCC	GMM-DNN	8	1:2	A,H,N,S,C,F,D,P	83.63
Proposed	MFCC	Gradient boosting	8	10-fold	A,H,N,S,C,F,D,P	85.25
Proposed	MFCC	Random forest	6	5 Fold	A,H,N,S,C,F	**93.83**
Proposed	MFCC	Random forest	4	10-fold	A,H,N,S	**98.00**
Proposed	MFCC	Random forest	8	10-fold	A,H,N,S,C,F,D,P	**86.38**
*A-Angry, H-Happy, N-Neutral, S-Sad, C-Calm, F-Fearful, D-Disgust, P-Surprise						

TABLE 3. Performance analysis using SUSAS dataset. The first and second best rate are in bold and blue (underlined), respectively.

Method	Feature extraction	Classification	Emotion class	Validation	Considered emotions*	Average recognition rate (%)
W.M Campbell [33]	MFCC	SVM	6	Not mentioned	N,A,S,L,F,O	72.80
Q.Y. Hong [34]	MFCC	GA	6	Not mentioned	N,A,S,L,F,O	68.70
T.Kinnunen [35]	MFCC	VQ	6	Not mentioned	N,A,S,L,F,O	68.40
I. Shahin [16]	MFCC	GMM-DNN	6	10-fold	N,A,S,L,F,O	86.67
Proposed	MFCC	Gradient Boosting	6	10-fold	N,A,S,L,F,O	87.97
Proposed	MFCC	Random Forest	6	10-fold	N,A,S,L,F,O	**88.67**
*A-Angry, N-Neutral, S-Slow, L-Loud, F-Fast, O-Soft						

TABLE 4. Performance Analysis using ESD dataset. The first and second best rate are in bold and blue, respectively.

Method	Feature extraction	Classification	Emotion class	Validation	Considered emotions*	Average recognition rate (%)
I. Shahin *et al.* [16]	MFCC	GMM-DNN hybrid classification	6	1:2	N,A,S,D,H,F	83.96
Proposed	MFCC	Gradient Boosting	6	10-fold	N,A,S,D,H,F	88.26
Proposed	MFCC	Random Forest	6	10-fold	N,A,S,D,H,F	**89.60**
*A-Angry, N-Neutral, S-Sad, D-Disgust, H-Happy, F-Fearful						

The statistical significance test has been done using the Student's t-distribution test defined below:

$$t_{1,2} = \frac{\bar{X_1} - \bar{X_2}}{SD_{pooled}} \tag{12}$$

$$SD_{pooled} = \sqrt{\frac{(SD_1)^2 + (SD_2)^2}{2}} \tag{13}$$

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

TABLE 5. Calculated t values between proposed method and each of GMM-DNN, SVM and MLP utilizing the RAVDESS, SUSAS and ESD datasets.

t Value	RAVDESS	SUSAS	ESD
t(Proposed, GMM-DNN)	1.66	1.63	1.69
t(Proposed, SVM)	1.70	1.65	1.76
t(Proposed, MLP)	1.64	1.66	1.91

TABLE 6. Computational Complexity associated with the GMM-DNN hybrid model vs proposed model.

Model	Average training time (sec)	Average testing time (sec)
Proposed	85,823.00	2.87
Shahin *et al.* [16]	95,921.00	4.12

Table 5 reports the computed t values between the proposed system and the other classifiers in the recent literature. Considering the critical value $t_{critical} = 1.645$ at 0.05 significant level [16], the results show that the system performance is significantly higher than the SVM, MLP and Hybrid GMM-DNN models except at the two instances corresponding to MLP with RAVDESS dataset, and the GMM-DNN with the SUSAS dataset.

E. COMPUTATIONAL COMPLEXITY ANALYSIS

In this section, we evaluate and analyze the computational complexity of the proposed emotion recognition model with the recent GMM-DNN hybrid classifier model. The computation time required in the training and evaluation phases are tabulated in Table 6. An Intel(R) core(TM) i7-3770 with a CPU @ 3.40 GHz, 4 Cores and 8 logical processors is used for the evaluation. Table 6 shows that the computational complexity of the random forest classifier is about half its counterpart in the hybrid classification model.

F. ANALYSIS OF THE SYSTEM PERFORMANCE IN NOISY TALKING CONDITIONS

Normally speech signals undergo various distortions such as noise, surface reflections, and reverberations. A human listener can identify the dominant signal from the noisy signal while most of the machine learning applications fail to address this factor. Real time NLP applications require acoustic machines and algorithms which are able to segregate the original speech signal from the other noises to ensure good performance even in noisy talking conditions. This experiment evaluates the system performance of the proposed method in such scenarios.

The original speech signal is mixed with noise signals in a ratio 2:1 and 3:1 to be used for evaluation. The various noise signals used are: other male voice, other female voice, siren noise, telephone ring, white noise, and vehicle noises. Fig. 8 shows the recognition rate of the system at different noisy conditions using RAVDESS dataset. The proposed system shows an average recognition rate of 81.83% even in the presence of noise. Fig. 9 shows the recognition rate of the system at different dominance levels. Original male and female voices are mixed with the noise signals in a ratio 2:1 and 3:1 and obtained an average recognition rate of 78.63% and 81.5% respectively at each level of dominance.

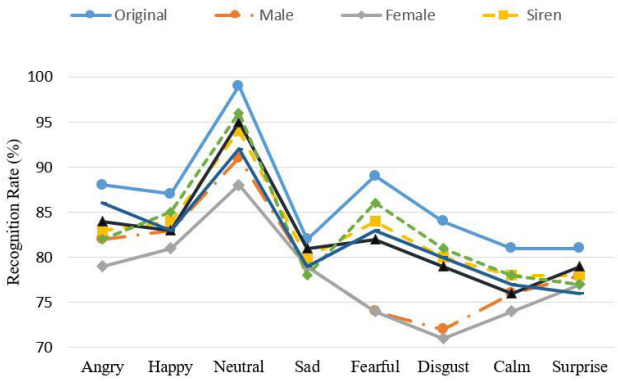

FIGURE 8. System performance on noisy talking conditions using RAVDESS dataset.

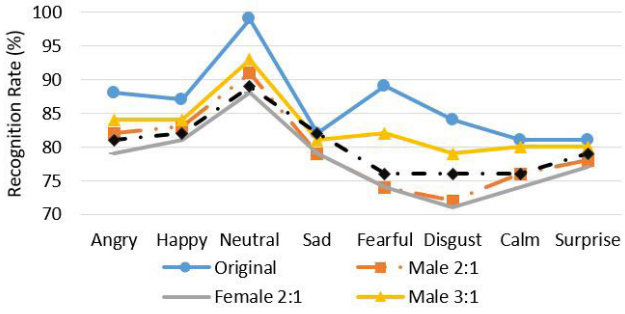

FIGURE 9. System performance on noisy talking conditions using RAVDESS dataset at different dominance level.

G. MODEL EVALUATION USING DIFFERENT PERFORMANCE METRICS

To have a finer insight on the performance of our framework, we computed a variety of metrics including the Accuracy, Precision, Recall and F1 score [36] for the different classes and across the three datasets. Accuracy is the measure of the effectiveness of a classifier in terms of detection in agreement with the actual classifications:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{14}$$

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 considers false detections. It is given by the number of correct detections over

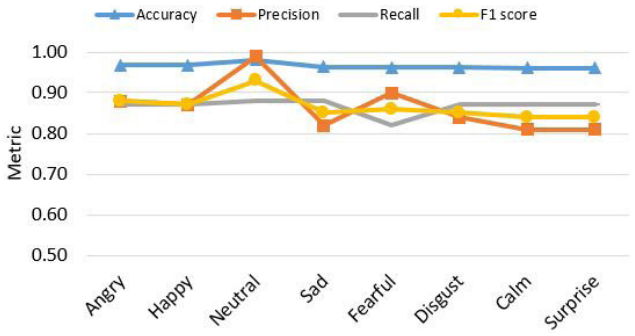

FIGURE 10. Performance evaluation parameters using RAVDESS dataset.

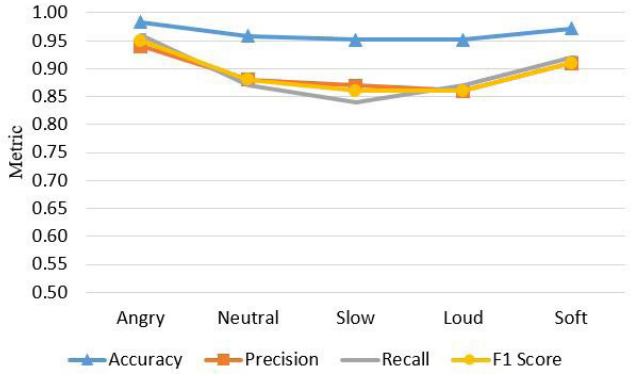

FIGURE 11. Performance evaluation parameters using SUSAS dataset.

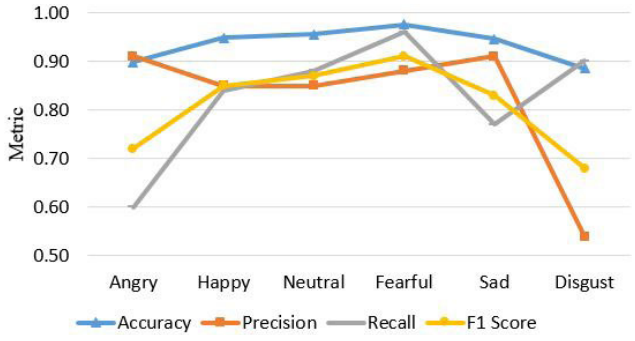

FIGURE 12. Performance evaluation parameters using ESD dataset.

all detections:

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

Recall gives the effectiveness of identifying labels per class:

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

F1 score is the harmonic mean of precision and recall:

$$F1Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{17}$$

Figure 10, 11 and 12, show the Accuracy, Precision, Recall and F1 scores, computed emotion-wise, and obtained with RAVDESS, SUSAS, and ESD datasets, respectively.

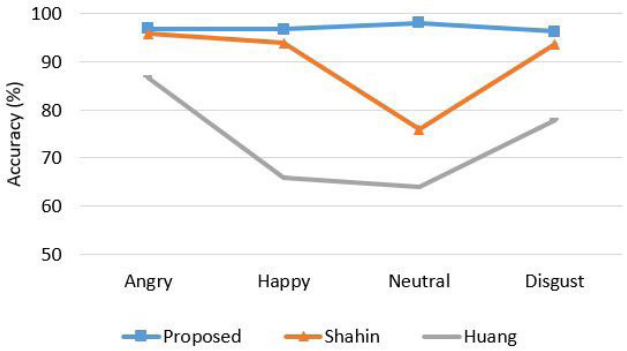

FIGURE 13. Performance analysis based on accuracy using RAVDESS dataset.

We observe that all the metrics score above 80% in the RAVDESS and SUSAS datasets. For the ESD, we notice a modest performance around 60% at the angry in the Recall and at the disgust in the Precision and the F1 score. Fig.13 shows the accuracy of our method for the four classes (angry, happy, neutral, disgust) that were reported in Huang *et al.* [32] work, together also with their counterparts in the method of Shahin *et al.* [16]. We can see that our framework scores better across all the four classes, whereas Huang *et al.* [32] exhibits low score around 65% in the happy and the neutral. Shahin's method performs well except in neutral class. We believe that the higher performance obtained by our framework, in particular for the happy and neutral, is due to the proposed WPT cochlear filterbank and speech segregation module, employed before the feature extraction, and which helps reducing the confusion between the monaural signals such as neutral and happy talking conditions. Fig.14 and 15 report the emotion-wise accuracy of the proposed system obtained using SUSAS and ESD datasets, respectively, plotted together with its counterpart in Shahin's [16] method. It is clear from Fig.14 that the proposed system performance is superior to the hybrid GMM-DNN classification model [16] in terms at each of the stressful talking conditions. In Fig. 15, we can see that our system competes well with GMM-DNN hybrid model, with a slightly lower performance at the happy, neutral and sad talking conditions. To have a finger insight on the on the lower performance obtained in these classes, we conducted a t-distribution test comparing the performance of our method and GMM-DNN classifier [16]. The results reported in Table 7 show t-values less than the threshold 1.65 for the angry, happy, neutral and sad emotions indicating no significant statistical variance between the proposed model and the GMM-DNN classifier for these talking conditions. However, the proposed system offers significant improvement in fearful and disgust talking conditions.

H. EVALUATION OF THE NOISE REDUCTION STRATEGIES

This experiment evaluates the noise reduction performance and its impact on recognizing emotions (noisy talking conditions) among various cochlear filterbank models. In this

TABLE 7. Calculated t values between proposed method and GMM-DNN model.

Emotion	Angry	Happy	Neutral	Sad	Fearful	Disgust
t(Proposed, GMM-DNN)	1.64	1.41	1.63	1.60	1.88	1.99

TABLE 8. Computed recognition rate between proposed cochlear model and each of recognition system without filterbank, Gamma-tone filterbank and STFT based filterbank.

Dataset	Number of emotion classes	Without cochlear filterbank	Gammatone	STFT	Proposed WPT
RAVDESS	8	66.34	83.41	74.31	**86.38**
SUSAS	6	65.9	82.01	72.33	**88.67**
ESD	6	69.03	84.44	75.62	**89.60**

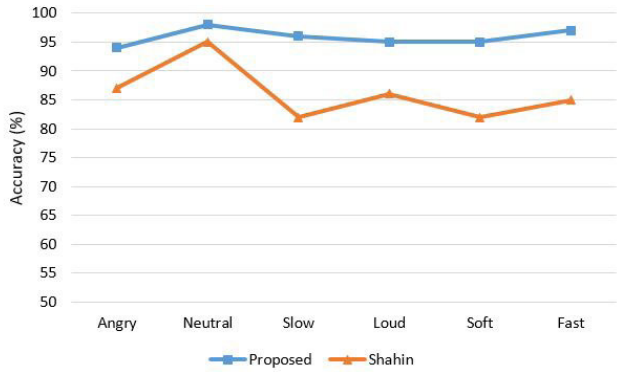

FIGURE 14. Performance analysis based on accuracy using SUSAS dataset.

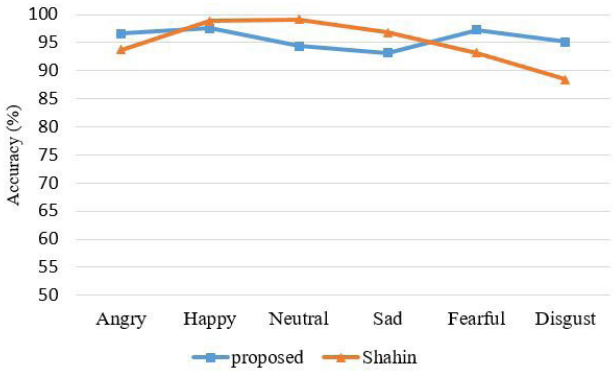

FIGURE 15. Performance analysis based on accuracy using ESD dataset.

work, we used WPT filterbank instead of conventional Gamma-tone and STFT-based filterbank models. Table 8 shows the average recognition rate obtained between the proposed model and each of emotion recognition without cochlear filterbank, emotion recognition using Gammatone filterbank and STFT based cochlear model. The original speech signal mixed with noise signals in a ratio 2:1 is used for evaluation. The various noise signals used are: other male voice, other female voice, siren, telephone ring, white noise, and vehicle noises.

TABLE 9. Computational complexity associated with the Gammatone filterbank model vs proposed model.

Model	Average training time (sec)	Average testing time (sec)
Proposed	85,823.00	2.87
Gammatone filterbank	1,15,348.00	7.32

We can see that our model is superior across all the four models. Gammatone filterbank competes well in terms of recognition rate but has a high computational complexity. Table 9 shows the computational complexity associated with each of WPT and Gammatone filter bank models. Results show that the computational complexity of the WPT-based model is about half its counterpart in the Gammatone filterbank model.

V. CONCLUSION

In this paper, we proposed a novel algorithm for emotion recognition from speech signals using WPT cochlear filter bank and random forest classifier. The proposed method is trained and evaluated in each of English and Arabic languages separately. The algorithm is evaluated in each of the stressful and noisy emotional talking conditions. Our results show that the performance of the proposed technique is superior in terms of higher accuracy and computational complexity to the other techniques in the literature. The proposed system offers better performance in noisy and stressful talking conditions with reduced computational complexity. The overall superiority of our system across the experiments conducted

with three different speech corpora in different languages gives credit to its generalization capacity. The main limitation faced during this work is that the proposed framework has been evaluated using acted and non-spontaneous speech signals. However, this is due to the lack of availability of natural emotional speech datasets. In future work, we plan to investigate multi-modal paradigms aggregating speech and facial expressions.

REFERENCES

[1] 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.

[2] 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.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 123–131.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 312–317.

[14] C. Huang, W. Gong, W. Fu, and D. Feng, "A research of speech emotion recognition based on deep belief network and SVM," *Math. Problems Eng.*, vol. 2014, pp. 1–7, Aug. 2014.

[15] 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.

[16] 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.

[17] J.-B. Kim and J.-S. Park, "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 126–134, Jun. 2016.

[18] W. J. Hardcastle, J. Laver, and F. E. Gibbon, *The Handbook Phonetic Science*, vol. 116. Hoboken, NJ, USA: Wiley, 2012.

[19] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[20] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, p. 248, 1961.

[21] A. Subasi and E. Ercelebi, "Classification of EEG signals using neural network and logistic regression," *Comput. Methods Programs Biomed.*, vol. 78, no. 2, pp. 87–99, May 2005.

[22] 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.

[23] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[24] K. Johnson and K. Johnson, "Acoustic and auditory phonetics," *Phonetica*, vol. 61, no. 1, pp. 56–58, 2004.

[25] F. Pulvermuller, M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk, and Y. Shtyrov, "Motor cortex maps articulatory features of speech sounds," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 20, pp. 7865–7870, May 2006.

[26] F. H. Guenther, "Cortical interactions underlying the production of speech sounds," *J. Commun. Disorders*, vol. 39, no. 5, pp. 350–365, Sep. 2006.

[27] R. D. Kent, R. Netsell, and J. H. Abbs, "Acoustic characteristics of dysarthria associated with cerebellar disease," *J. Speech, Lang., Hearing Res.*, vol. 22, no. 3, pp. 627–648, Sep. 1979.

[28] N. Sato and Y. Obuchi, "Emotion recognition using Mel-frequency cepstral coefficients," *Inf. Media Technol.*, vol. 2, no. 3, pp. 835–848, 2007.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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 B.Tech. and M.Tech. degrees in electronics and communication engineering from Mahatma Gandhi University, India. She was a Research Associate with the University of Sharjah. She was a Lecturer at Mahatma Gandhi University, India. She is currently a Research Engineer with the ECE Department, Khalifa University. Her research interests include artificial intelligence, deep learning, natural language processing, computer vision, and human–machine communication in the digital world. She was ranked at Second Place with honors on the M.Tech. degree in applied electronics.

ISMAIL SHAHIN (Member, IEEE) 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 at the Department of Electrical and Computer Engineering, University of Sharjah, United Arab Emirates. He has more than 60 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.

YOUSSEF IRAQI (Senior Member, IEEE) is currently an Associate Professor with the ECE Department, Khalifa University, United Arab Emirates. Before that, he was the Chair of the Computer Science Department, Dhofar University, Oman, for four years. From 2004 to 2005, he was a Research Assistant Professor with the David R. Cheriton School of Computer Sciencem, University of Waterloo, Canada. He has published more than 110 research articles in international journals and refereed conference proceedings. His research interests include adaptive resource management in multimedia wireless networks, trust and reputation management, cloud computing, and stylometry. He is on many technical program committees of international conferences and always approached for his expertise by international journals in his field. In 2008, he received the IEEE Communications Society Fred W. Ellersick Paper Award in the field of communications systems.

NAOUFEL WERGHI (Senior Member, IEEE) received the Habilitation and Ph.D. degrees in computer vision from the University of Strasbourg. He was a Lecturer with the Department of Computer Sciences, University of Glasgow. He has been a Research Fellow with the Division of Informatics, The University of Edinburgh. He has been a Visiting Professor with the University of Louisville, the University of Florence, the University of Lille, and the Korean Advanced Institute of Sciences and Technology, South Korea. He is currently an Associate Professor with the Electrical Engineering and Computer Science Department, Khalifa University of Science and Technology. His main research interests include 2D/3D image analysis and interpretation, where he has been leading several funded projects related to biometrics, medical imaging, remote sensing, and intelligent systems. He is a member of the IEEE Signal Processing Society and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is an Associate Editor of the *EURASIP Journal on Image and Video Processing*.