Abstract

Speech is considered as one of the most effective mediums to show emotions and express attitude using language. Recently, extracting the content of emotions from a signal of speech and recognise the emotions from the speech is investigated by many researchers. In this paper, we extract two sets of acoustic parameters,i.e., Mel-Frequency Cepstral Coefficients (MFCCS) and Perceptual Wavelet Packets (PWP)set. Then we take advantage of k-means to cluster both feature spaces. Subsequently, these are modelled using classification schemes, including, K-Nearest-Neighbor (KNN) and Support Vector Machines (SVM). The basic emotions are categorized as sad, angry, neutral and happy. Experimental results show acceptable robustness of the proposed system.

Keywords: Speech emotion recognition, feature extraction, PWP, MFCCs**,** pattern recognition, SVM, KNN

1 Introduction

The signals of speech are the most important human communication natural media, that they include paralinguistic information of the speakers. The speech signals transport explicit linguistic contents, as well. Recently, investigators have done their best to providing and developing methods for automatically recognising human emotions from speech signals, which is named speech emotion recognition [1]. Today, speech emotion recognition has played a pivotal role in topics such as pattern recognition, artificial intelligence and signal processing [2], [3]. On the other hand, feature extraction has been employed to bridge the efficient gap between subjective emotions and speech signals. Researchers have used a variety of hand-designed features for speech emotion recognition [4]–[5].

Speech emotion recognition has been always a popular and considerable subject. This is a challenging problem as diverse emotions can be transmitted in various forms of speech. As well as this, specifying what all features to extract from speech to analyse inherent emotions of it is a different issue in itself. The current methods which are widely used to overcome this problem are support vector machines (SVMs), K-Nearest-Neighbour (KNN), hidden Markov models (HMMs) or neural networks. Although SVM can provide an acceptable prediction with lesser effort, HMMS and neural networks are difficult to train and build. They need high time and computational power. As a result, an approach to improve the performance of SVM on the problem [6]. At this time, ensemble learning would be beneficial. Training multiple estimators and aggregating their outcomes are included in Ensemble learning [7]. Boosting and bagging (bootstrap aggregating) are considered as the outstanding methods of building ensembles that both of them sometimes include similar learners. Boosting is an iterative mechanism, while bagging is a parallel.
procedure.

The rest of this paper is prepared as follows:
Section 1 provides details about current related work. Section 2 describes the system overview, including Feature Extraction algorithms and Classification Schemas. In Section 3, we explain the parameterization of the classification methodology as well as experimental results. Finally, Section 4 concludes this work.

2. SYSTEM OVERVIEW

This section describes the feature set used (MCCFs and PWP) in this work along with the Classification schemes (HMM and SVM). The overall structure of proposed method is shown in Figure 1.

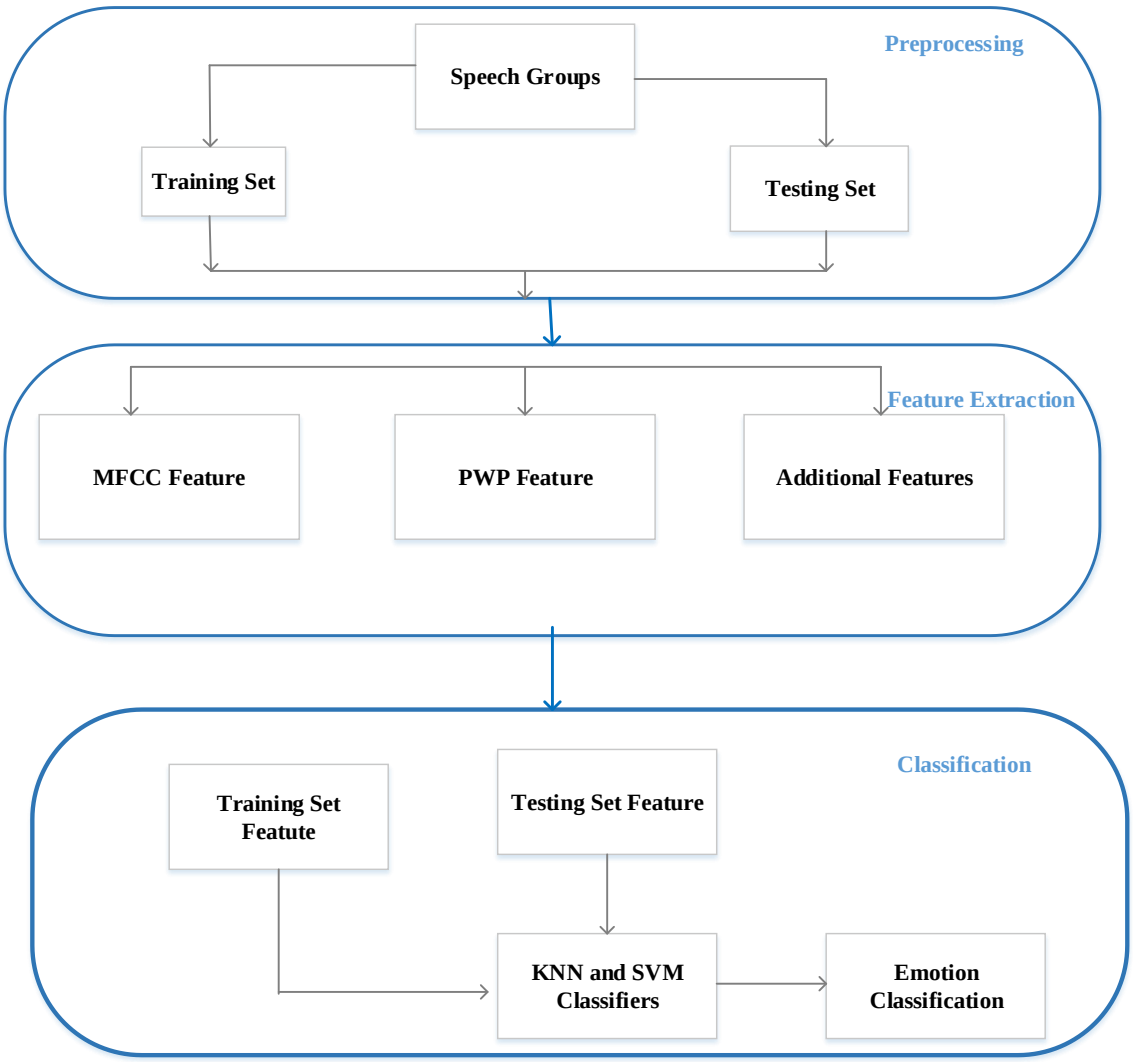

Fig. 1 Schematic diagram of the proposed system

2.1. Feature Extraction Analysis

- acoustic features [8], [9],
- language features (lexical information [10], [11])
- context information (gender, subject, culture influences [12], [13])
- hybrid features [10], [14]

Loudness, pitch and duration are widely used as prosody features [15], because they indicate the intonation patterns and stress of spoken language. Voice quality features, as the characteristic auditory colouring of an individual voice, have been shown to be discriminative in expressing positive or negative emotions [16].

The first three formants (F1, F2, F3) are commonly used voice quality features, harmonics-to-noise-ratio, spectral energy distribution, amplitude irregularity (shimmer) and pitch irregularity (jitter). Integrating voice quality and prosody features indicate better performance than utilizing prosody features alone [17], [18]. Recently, voice source parameters [20] and glottal features [19] have been employed as more progressive voice quality features for speech emotion recognition. Spectral features are the final typical acoustic features that are calculated from the short-term power spectrum of sounds, such as Log Frequency Power Coefficients (LFPC), Linear Prediction Cepstral Coefficients (LPCC) and Melfrequency Cepstral Coefficients (MFCC). Among these acoustic

features, MFCC is the most favourable spectral feature, because it can model the human auditory perception system.
We have employed two sets of feature which have indicated promising performance in generalized sound recognition tasks (Perceptual Wavelet Packets set and Mel-Frequency Cepstral Coefficients). These are described in the following paragraphs.

2.1.1 Mel-Frequency Cepstral Coefficients (MFCC)

The sound short-term power spectrum is represented by Mel-frequency Cepstrum (MFC), relying on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that concertedly create an MFC. There is a difference between the cepstrum and the Mel-frequency cepstrum. This difference is that the frequency bands are spaced on the Mel scale in the same manner, which approaches the human auditory system's response more closely than the linearly-spaced frequency bands employes in the normal cepstrum [21].

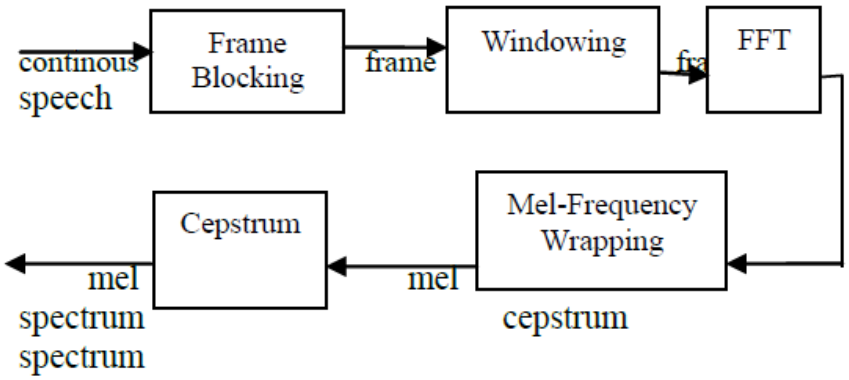

Fig 2: MFCC Extraction Process

The frame blocking step is the first step in which the ongoing speech signal is blocked into frames of N samples, with nearby frames being separated by M (M < N). The first N samples are included in the first frame. The second frame includes M next samples after the prior frame and overlaps it by N - M samples. In the third step, the individual frame is windowed in order to minimize the signal discontinuities at the first and end of each frame. Hamming window is the most widely used that can be shown as follows:

$$w(n) = 0.54 - 0.46\ cos(2\pi n / N\text{-}1), \qquad 0 <= n <= N\text{-}1 \qquad (1)$$

Fast Fourier Transform (FFT) is the last step. This is used to converts every frame of N samples from the time into the frequency domain, respectively. This step output is the spectrum. We should wrap frequently the magnitude spectrum to transform the spectrum into the Mel-frequency scale. The Mel-frequency warping is done using a Mel-filter bank, which has a set of band pass filters with spacing on the Mel-scale and constant bandwidths. This bank includes one filter for every asked Mel-frequency component, where every filter contains a triangular filter (range from zero to the Nyquist frequency) bandpass frequency response [22].

2.1.2 Perceptual Wavelet Packets (PWP):

PWP is used to capture the audio signals perceptual properties because it contains the initial filtering stage [23]. The acoustic signals are described by the specific feature set in a multi-band

manner, while wavelet packets analyse each spectral area. The diversity of each wavelet coefficient within a critical band, which is significant for the generalized sound recognition methods are captured by the parameters of PWP audio. Processing time series is the main merits of the wavelet transform, which contains nonstationary power at many various frequencies. Fourier transform has a substantial property which is the usage of sinusoids with an unlimited duration. Although wavelets like to be asymmetric and irregular, they are predictable and smooth. wavelets include a dynamic windowing method which can treat with various precision high and low-frequency data. Choosing the principal wavelet is the first step of analysing the wavelet packet. By doing so, the signal is broken up into the scaled and shifted versions of it.

2.2. Classification Schemas

Many different machine learning methods have been used to organise an acceptable classifier in order to distinguish the categories of emotion. One of the vital steps in emotional speaker recognition procedure is selecting the suitable classifier Early emotion classifiers include Artificial Neural Network (ANN) [25] and K-Nearest-Neighbor (KNN) [24]. After that statistical pattern recognition methods, such as and SVM [28], Hidden Markov Models (HMM) [27] and Gaussian Mixture Model (GMM) [26] are commonly compatible with speech emotion recognition. In [29] and [30] a few cutting edge classifiers based on sparse representation have been provided. It is obvious that each classifier has its own merits and demerits. To combine the advantages of various classifiers, ensembles of multiple classifiers have been studied for speech emotion recognition [31], [32].
Among existing models, we have employed K-Nearest-Neighbour (KNN) and Support vector machine (SVM) models.

2.2.1 K-Nearest-Neighbour (KNN)

K nearest neighbours (KNN) is an easy and simple method that stores all available cases. Then, it categorizes new cases based on a similarity measure (such as distance functions). How closely out-of-sample features resemble our training set determines how we classify a given data point. KNN has been utilized in pattern recognition and statistical prediction already in the early of 1970's as a non-parametric algorithm. An example of the k-NN classification can be seen in Figure 3.
The inside circle shows the test sample which should be classified either to the first class of green squares or to the second class of yellow triangles. If k = 3 (outside circle) it is assigned to the second class as there are two triangles and only one square inside the inner circle. If, for instance, k = 5 it is assigned to the first class (3 squares vs. 2 triangles outside the outer circle).

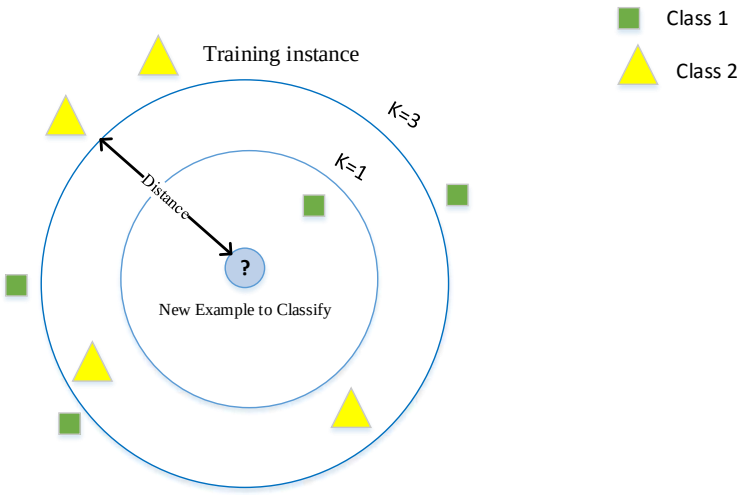

Fig. 3 KNN description

2.2.2 Support vector machine (SVM):

The main objective of SVMs is to find the hyperplanes maximizing the margin between support vectors which are extracted out of the classes of interest feature sets. SVM classifies the classes using an obvious gap with the goal being to widen it as much as possible [33]. Authors in [34] have employed SVM in order to indicate the significant effect of the emotional state upon text-independent speaker recognition. Actually, SVM provides a discriminative classifier which obtains effective results in a lot of speaker recognition tasks. SVMs are popular and widely used as they distinguish between categories (impostor and speaker) by creating a hyperplane. in the basic form of SVM is a binary classifier. The SVM efficiency depends on the kernel function and multiclass SVM method as well. To evaluate our proposed emotional speaker recognition system, we considered SVM (linear kernel).

3. EXPERIMENTAL SET-UP

In this section, we provide complete details related to the parameterization of the provided framework for categorizing the emotions of humans. The results and how these compare with classification systems widely used in the generalized sound classification literature (KNN and SVM).

Motivation:

We have aimed at satisfying the circumstance of containing both discriminative and generative pattern recognition schemes. In addition, during early experimentations, we investigated that PWP and MFCCs can capture distinct properties of the available audio signals structure. This is because these features obtained various recognition rates characterized by different misclassifications, as a result, we decided to take advantage of them concurrently.

3.1 Dataset

We employed Berlin Database of Emotional Speech which is also referred to Emo-dB [35]. Berlin dataset is created in the German language. The samples of emotion have been recorded by five males and five females who were in the age group of 21–35 years by ten different texts. Actually, recordings were carried out at the University of Berlin in an anechoic chamber. The database

formulation was a part of the DFG funded project with recordings materialized in 1997. The sentences have been recorded in anger, boredom, disgust, fear, sadness, happiness
and neutral. There are 535 emotional samples. Each sample has a length of around 2–3 s. The list of seven categories can be seen in Table 1. We consider sad, angry, neutral and happy categories.

Table 1: Description of Berlin corpus

No	Emotions	Number of Samples
1	Fear/ Anxiety	69
2	Anger	127
3	Boredom	81
4	Disgust	46
5	Happy	71
6	Neutral	79
7	Sad	62

3.2. Experimental Results

This section, suggested method is simulated using Python in order to assess its performance. We consider 0.2 percent of the data as the test set.

We used the main Python library called *pyAudioAnalysis* which is a Python library covering a variety range of audio analysis tasks. Using *pyAudioAnalysis* you can do the following tasks:

1. Extract audio features and representations (e.g. MFCC, Spectrogram, Chromagram)
2. Classify unknown sounds
3. Train, parameter tune and evaluate classifiers of audio segments
4. Detect audio events and exclude silence periods from long recordings
5. Perform supervised segmentation (joint segmentation - classification)
6. Perform unsupervised segmentation
7. Extract audio thumbnails
8. Train and use audio regression models (example application: emotion recognition)
9. Apply dimensionality reduction to visualize audio data and content similarities

We have also taken advantage of *sklearn* (for clustering by *Kmeans*, calculating *accuracy* and *Confusion Matrix*), *pywt* (for PWP feature extraction) Python libraries.

Firstly, we extract PWP feature and assigned label to the voices (Table 2).

Table 2: List of emotions and their labels

Emotion	**Label**
Angry	0
Happy	1
Neutral	2
Sad	3

The length of audio files (.wav) are not equal in Berlin dataset. As a result, to increase the accuracy we should provide a method to equation of all lengths. We calculated the length of each frame and found the minimum amount of them. Then we used the minimum frame length as the base length and equalled all frames length to that amount. After extracting PWP and MFCC features we employed K-means as the clustering method. Subsequently, these are modelled using classification schemes, including KNN and SVM. Table 3 shows the accuracy obtained by each classification method.

Table 3: The result of classification

Feature Set	Type of Classifier	Accuracy	F1	RECALL
MFCCs	SVM	18	0.32	0.56
	KNN	35	0.55	1
PWP	SVM	35	0.52	1
	KNN	36	0.52	1
pyAudioAnalysis 12 existing features + PWP + MFCCs	SVM	82	91	76
	KNN	75	80	79

In this section, we extract the PWP and MFCC features and separately we used K-means clustering algorithm (Table 4).

Table 4: The result of clustering using K-means

Feature	Accuracy
PWP	0.35
MFCCs	0.33

4. Conclusions

In this paper, we have taken advantage of Berlin dataset and the basic emotions have been categorized as happy, sad, angry and neutral. we extract two sets of acoustic parameters,i.e.,MFCCs PWP set. Then we used K-means to cluster both feature spaces. Subsequently, these are modelled using classification schemes, including, K-Nearest-Neighbour (KNN) and Support Vector Machines (SVM). The results show the accuracy of 82% and 75% using KNN and SVM classifiers on the Berlin dataset on four emotions categorizes (sad, angry, neutral and happy) when we extract PWP and MFCC features.

References:

[1] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576-1590, June 2018. doi: 10.1109/TMM.2017.2766843

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

[3] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommun. Syst.*, vol. 52, no. 3, pp. 1467–1478, 2013.

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

[5] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[6] Bhavan, Anjali & Chauhan, Pankaj & Hitkul, & Shah, Rajiv Ratn. (2019). Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems. 184. 104886. 10.1016/j.knosys.2019.104886.

[7] Zhang, Cha, and Yunqian Ma, eds. Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012

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

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

[10] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4749–4753.

[11] B. Schuller, "Recognizing affect from linguistic information in 3D continuous space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 192–205, Oct.–Dec. 2011.

[12] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 502–509, Oct. 2010.

[13] M. A. Quiros-Ramirez and T. Onisawa, "Considering cross-cultural context in the automatic recognition of emotions," *Int. J. Mach. Learn. Cybern.*, vol. 6, no. 1, pp. 119–127, 2015.

[14] H. Cao, A. Savran, R. Verma, and A. Nenkova, "Acoustic and lexical representations for affect prediction in spontaneous conversations," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 203–217, 2015.

[15] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," in *Proc. 6th Int. Conf. Spoken Language Process.*, Beijing, China, 2000, pp. 222–225.

[16] R. Tato, R. Santos, R.Kompe, and J.M. Pardo, “Emotional space improves emotion recognition,” in *Proc. Interspeech*, 2002, pp. 2029–2032.
[17] M. Lugger and B.Yang, “The relevance of voice quality features in speaker independent emotion recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, vol. 4, pp. 17–20.
[18] S. Zhang, “Emotion recognition in Chinese natural speech by combining prosody and voice quality features,” in *Proc. Adv. Neural Netw.*, 2008, pp. 457–464.
[19] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falc˜ao, “Spoken emotion recognition through optimum-path forest classification using glottal features,” *Comput. Speech Lang.*, vol. 24, no. 3, pp. 445–460, 2010.
[20] J. Sundberg, S. Patel, E. Bj¨orkner, and K. R. Scherer, “Interdependencies among voice source parameters in emotional speech,” *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 162–174, Jul.–Sep. 2011.
[21] Swain, M., Routray, A. & Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: a review. *Int J Speech Technol* **21,** 93–120 (2018) doi:10.1007/s10772-018-9491-z.

[22] " An Automatic Speaker Recognition System ", http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition.

[23] Ntalampiras, Stavros & Potamitis, Ilyas & Fakotakis, Nikos. (2010). Sound classification based on temporal feature integration. 1 - 4. 10.1109/ISCCSP.2010.5463315.
[24] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” in *Proc. 4th Int. Conf. Spoken Lang.*, 1996, vol. 3, pp. 1970–1973.
[25] J. Nicholson, K. Takahashi, and R. Nakatsu, “Emotion recognition in speech using neural networks,” *Neural Comput. Appl.*, vol. 9, no. 4, pp. 290–296, 2000.
[26] D. Ververidis and C. Kotropoulos, “Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm,” in *Proc. IEEE Int. Conf. Multimedia Expo.*, Amsterdam, The Netherlands, 2005, pp. 1500–1503.
[27] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” *Speech Commun.*, vol. 41, no. 4, pp. 603– 623, 2003.
[28] B. Schuller,G.Rigoll, andM. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp. 577–580.
[29] X. Zhao, S. Zhang, and B. Lei, “Robust emotion recognition in noisy speech via sparse representation,” *Neural Comput. Appl.*, vol. 24, no. 7/8, pp. 1539–1553, 2014.
[30] X. Zhao and S. Zhang, “Spoken emotion recognition via localityconstrained kernel sparse representation,” *Neural Comput. Appl.*, vol. 26, no. 3, pp. 735–744, 2015.
[31] D. Morrison, R.Wang, and L. C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres,” *Speech Commun.*, vol. 49, no. 2, pp. 98–112, 2007.
[32] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, 2011.

[33] Mansour, A., Chenchah, F. & Lachiri, Z. Emotional speaker recognition in real life conditions using multiple descriptors and i-vector speaker modeling technique. *Multimed Tools Appl* **78,** 6441–6458 (2019) doi:10.1007/s11042-018-6256-2.

[34] Rusu C, Ghiurcau MV, Astola J (2011) Speaker recognition in an emotional environment. Proceedings of SPAMEC
[35] Lalitha, S., Tripathi, S. & Gupta, D. Enhanced speech emotion detection using deep neural networks. *Int J Speech Technol* **22,** 497–510 (2019) doi:10.1007/s10772-018-09572-8.