Arabian Journal for Science and Engineering (2020) 45:3111–3121
https://doi.org/10.1007/s13369-019-04293-9

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

An Efficient Language-Independent Acoustic Emotion Classification System

Rajwinder Singh[1] · Harshita Puri[1] · Naveen Aggarwal[2] · Varun Gupta[1]

Received: 31 August 2019 / Accepted: 5 December 2019 / Published online: 20 December 2019

Abstract
Emotion recognition from human speech is essential to understand the convoluted human nature. For any machine to accurately decipher the intended message in the speech, it must understand the emotion of spoken words. Emotions control the modulations in the speech, and these modulations may even change the context. Through this paper, we aim to propose a system which can efficiently detect the emotions from speech. The domain of emotion recognition from human speech is very complex due to highly overlapping regions of emotions, and it sometimes becomes very difficult to distinguish between two emotions just based on voice. Such ambiguity in the label assignment is responsible for low classification accuracy in existing systems. In the proposed system, we have worked on finding both the suitable feature set as well as the classifier. The proposed system achieved 29.74% increase in classification accuracy in comparison with the baseline human accuracy on the primary dataset, i.e. 'CREMA-D'. Further, we have validated on other standard datasets such as 'EmoDB', 'RAVDESS', and 'SAVEE'. 'EmoDB' is a German language dataset, while the other two are English language datasets, which is in line with the language-independent nature of our system. When compared to the current state of the art in this domain on these datasets, the proposed system gives better accuracies for most of the cases, and for some cases, it gives comparable accuracies to baseline models or existing published work.

Keywords Emotions · Machine learning · Deep learning · Sentiment analysis · Affective computing · SVM · RNN

1 Introduction

Understanding human emotions has been a quest taken by many researchers in different fields. From philosophers to psychologists, the ultimate aim is to understand the depths of the human mind. With the advancement of computers, the desire to create agents capable of not just understanding but also acting like humans has risen. Thus, in the first step towards achieving this goal, the task of classifying human emotions was embarked. Humans express their emotions/internal mental states via various modalities, speech being one of them. Speech as a method of communication produces its desired effect not only because of its linguistic content but also due to the paralinguistic content. Even though there are many studies related to speech processing, but very little work has been done to understand emotions from the paralinguistic contents. The study of this paralinguistic part is imperative to understand the intricacies of the convoluted human speech, as the same words spoken in different emotions completely change the context of the speech, hence making the emotion analysis from speech crucial. Recognizing human emotions from paralinguistic content will also help us to make the systems language independent. Human speech contains various types of information like the message, language of the speaker, tone, emotions, pitch, etc., which all play a significant role in understanding the speech. Thus, restricting the focus only on words may sometimes lead to misinterpretation of the intended message. As phonetic structure strongly influences the accuracy of emotion recognition, it is imperative to take into consideration other aspects too [1]. There are various challenges to detect emotions from the paralinguistic content. Extraction of emotions from speech is always tricky due to overlapping regions of

Rajwinder Singh, Harshita Puri have contributed equally to this work.

✉ Rajwinder Singh
rajwindersingh12310@gmail.com

[1] Chandigarh College of Engineering and Technology (Degree Wing), Chandigarh, India

[2] UIET, Panjab University, Chandigarh, India

emotions. Further, the interpretation of emotions is highly subjective and may vary from person to person. Interpretation also changes due to the change in pitch or amplitude of the sound. For example, 'Disgust' emotion may sound as 'Anger' if we listen to it at high amplitudes. In the absence of context and language understanding, sometimes human also find it challenging to recognize the emotions. The study of emotions in speech is involved, and models designed for recognizing it have low accuracy because of the ambiguity in label assignment [2].

Moreover, emotions change with the change in intensity of speech. Thus, it becomes necessary to develop the classifiers which consider this change. In this work, we have chosen the 'CREMA-D' [3] for developing this classifier as they have recorded the dataset at three intensity levels, i.e. low, medium, and high. Using such dataset has made our classifier more robust to variations of intensity in human speech. Till date, most of the work on this dataset has been done for the task of facial emotion recognition [4] or to study the effect of data augmentation and the increase in depth of the network on the accuracy, by mainly studying the arousal, valence, and dominance [5]. Also, some work has been done related to crowdsourcing and label assignment validation using this dataset [6]. Arora et al. presented their work on preserving the speaker identity while detecting emotion from acted speech corpus [7]. Besides all this work, none of the researchers have made the paralinguistic classification the primary focus of their study using this dataset. The literature states that during crowdsourcing of 'CREMA-D' humans were able to identify audio files 40.9% times only, video only 58.2%, and audio–video only 63.69% times [3]. Moreover, humans were designated more capable than machines to detect emotions in the sequence—'Neutral', 'Happy', 'Anger', 'Disgust', 'Fear', and 'Sad' [8].

Keeping this in mind, we aimed to propose a system which acts as a determinant for classification of human speech into the above six basic emotions with an accuracy comparable or even more than that of human classification ability. For this purpose, we have worked on the audio part of 'CREMA-D' dataset. The research in the field of emotion detection from audio data is in the nascent stage, and some underlying issues needed to be resolved, such as (1) Selection of best feature set for the task; (2) Analysis of various classifiers and their parameter tuning on the validation set; (3) How well can the classifier distinguish between the male and female voice? This is essential because different genders have different pitch and amplitude levels while conveying the emotions. In this paper, we have proposed a system and presented the results of a set of experiments performed, which address all such questions. The proposed system provides an appreciable increase in the emotion classification accuracy with respect to human detection capability on 'CREMA-D'. Furthermore, to reach a reliable conclusion about the performance of a classifier in distinguishing different emotional classes, one has to validate it over more than one dataset. Thus, we have validated the results on some other standard datasets as well, such as 'EmoDB' [9], 'RAVDESS' [10], and 'SAVEE' [11]. These datasets have baseline classification accuracies of 80%, 60%, and 61%, respectively. The proposed classifier can also perform gender-based classification of acoustic emotion data.

The rest of the paper is organized as follows. Section 2 deliberates the related work, and Sect. 3 discusses the datasets used. Section 4 presents the proposed system and its two phases. Section 5 discusses the results obtained. Finally, Sect. 6 concludes the paper and provides the future scope of the work.

2 Related Work

Research in the field of emotion analysis from the human speech is growing day by day. With many new ideas emerging each day, the research community has been able to progress in the direction to bridge the gap between machines and human understanding. Researchers have proposed various theories, methodologies, and models to detect emotions in the acoustic domain efficiently. Neiberg et al. [12] worked on emotion recognition in spontaneous speech. According to them, it is more difficult to detect emotions from live or spontaneous sound than using the pre-recorded dataset. For emotion recognition in spontaneous real-time speech, they have proposed an approach in which they have used three classifiers and combined their results. They have used MFCC, Low MFCC in the range of 20–300 Hz and the pitch to accomplish the task. Their results show that two MFCCs have similar results, whereas MFCC low outperforms the pitch. They used the Gaussian mixture model (GMM) on frame level as the classifier. Blouin et al. [13] presented an approach by using LDA classifier. However, in their proposed work, the noise part is separated from the audio part by using the ASR system before the classification phase, and the sampling rate was set to 8 kHz for signal. Further to improve the uniformity of recording condition, these are coded on 16 bits and filtered in telephony band (300–3400 Hz) with 3 coefficients Butterworth filters.

Research in the domain of speech has progressed at a steady pace. However, in recent years, a new domain of automatic recognition and synthesis of multi-style speech has blossomed. Cummings et al. [14] have worked on glottal waveforms of eleven speaking styles. The domain of acoustic study is not only confined to the speech part but also spans to the other nonverbal parts of the sound; Sauter et al. [15] worked in this domain and examined the nonverbal emotional vocalization in cross-culture settings for recognizing emotions such as scream or laugh. They have presented that, regardless of the origin of the stimuli, different cultural groups of listeners reliably identify the emotions. These set of emotions are basic emotions. However, several positive emotions are recognized within—but not across the cultural groups, which may suggest that affiliative social signals are shared primarily within the same culture group members. This paper neither determined any relation between acoustics and emotions nor specified any particular technique to identify it. Instead, it provided a sociological aspect of emotions and voices across two different cultures.

Furthermore, there are various approaches proposed in the literature to identify the emotions in the audio files. The process of emotion recognition falls in three categories, the first one is frame-based processing [16] (usually with a majority voting for the final classification), second is using sequential processing [17] (considering the temporal dependencies of the acoustic signal) and the third one is using a combination of both [18]. When coming to the classification techniques for images, one is presented with a plethora of state-of-the-art classifiers as Singh et al. [19] have shown the efficiency of VGG models in the detection of wheat rust from infected crop images. To say the same for the acoustic domain would not be right, as research in this domain is still in its adolescent age. To develop the classifiers in this domain with accuracies comparable to that of humans is a challenging task. One of such classifiers, deep neural networks have always found their usability in various domains to solve a variety of problems which included but is not limited to language modelling [20], sentiment analysis [21], speech recognition [22] and neural machine translation [23]. Various researchers have hypothesized and claimed their approach to be the best for detecting emotions in audio files. Lakomkin et al. [24] used deep reinforcement learning for continuous emotion detection in audio files. They tried to recognize emotions such as anger to identify the undesired and unsafe conditions in human–robot interactions. They have worked on two factors, i.e. accuracy and latency of the classification. The agent that they have designed performs two tasks, the wait and the terminate. During the wait phase, the agent listens to the sound signal, and the listening phase starts after the terminate. They claim their approach to be better than others because the others wait for the complete utterance, but they do not wait for the utterance to complete for the classification. Thus, they claim their agent is well suited for the emergencies. They have used 15 MFCC features with their first- and second-order derivatives extracted using OpenSMILE toolkit [25] using windows of 25 ms width and 10 ms stride. For classification purpose, they have used a single-layer recurrent neural network with gated recurrent unit proposed by Bahdanau et al. [26]. There is also some of the latest work published which involves working with utterance levels.

Wang et al. [27] have used deep neural networks for learning utterance-level representations for recognizing emotion and age/gender in speech. They have also studied the silent frame and discussed assigning utterance level to these silent frames. They have proposed an utterance-level deep neural network for the classification, which gives 3.8% weighted accuracy and 2.94% unweighted accuracy. In another paper, Bothe et al. [28] have proposed an utterance-based bi-directional recurrent neural network for conversation analysis. Using the preceding utterance in the context, they were able to achieve 77% accuracy on the SwDA corpus. Rather than using the long-recorded sessions of conversations as used by most of the researchers, we have focused on dialogue clips. We have trained the classifiers on all six basic emotions, which is in contrast to [26], as they have focused only on two emotions with the samples annotated as anger or neutral. The set of these six emotions from 200 emotions achieve good accuracy when detecting emotions from human speech [8]. We have proposed a system which gives good accuracy in the classification of these six emotions in audio files.

3 Dataset Description

Vocalizations of same emotions substantially differ among humans. For instance, a sentence said in anger emotion by one having a soft voice, might not be as indicative, as that said by a hoarse-voiced person. Even more, by varying the intensity of a sentence spoken by a person in the same emotion, we can get variants of that emotion as well. For incorporating these considerations in our classifier, we needed a dataset which recorded a given sentence not only in different emotions but also at different intensities for a particular emotion. The 'CREMA-D' (Crowd-Sourced Emotional Multimodal Actors Dataset) fulfils this requirement. The data set records facial and vocal emotional expressions of twelve sentences spoken in primary emotional states ('Happy', 'Sad', 'Anger', 'Fear', 'Disgust', and 'Neutral') with varying intensities. A total of 7442 clips were recorded by 91 actors with diverse ethnic backgrounds, which then was rated by multiple raters in three modalities: audio, visual, and audio-visual. The categorical emotion labels and real intensity values for the perceived emotion were collected using crowdsourcing from 2443 raters. The human recognition of intended emotion for the audio-only, visual-only, and audio-visual data were 40.9%, 58.2%, and 63.6%, respectively. Since we have only worked with the emotion recognition from audio mode, only the audio part of the dataset was considered. The audio files belonging to six emotions, namely 'Happy', 'Sad', 'Anger', 'Fear', 'Disgust', and 'Neutral' are in.wav format. Each emotion except 'Neutral' has 1271 audio files, while 'Neutral' has 1087 files (since variable intensities could not be rated for it). Table 1 shows the

Table 1 Gender-wise and intensity-wise distribution of audio files in 'CREMA-D'

Emotions	Gender wise		Intensity wise			
	Male	Female	High	Medium	Low	Unknown
Anger	671	600	91	91	91	998
Disgust	671	600	91	91	91	998
Fear	671	600	91	91	91	998
Happy	671	600	91	91	91	998
Neutral	575	512	0	0	0	998
Sad	671	600	91	91	91	998

intensity-wise as well as the gender-wise distribution of the audio files in this dataset.

Since the dataset is clean and noise-free, the extra cleaning step is not required. Further, we have validated the proposed classifier on three commonly used datasets, namely 'EMODB' (Berlin Database of Emotional Speech) [9], 'RAVDESS' (The Ryerson Audio-Visual Database of Emotional Speech and Song) [10], and 'SAVEE' (Surrey Audio-Visual Expressed Emotion database) [11]. The 'RAVDESS' dataset, like 'CREMA-D', is also a multimodal dataset. Similar to 'CREMA-D' dataset, we have focused only on audio data, but here, audio data is both in the form of the plain dialogues as well as songs. Since our model aims for emotion recognition in speech, we have taken only dialogue audio files from it. A total number of 24 professional actors (12 male and 12 female) recorded the data in a neutral North American accent. The 'Calm', 'Happy', 'Sad', 'Angry', 'Fearful', 'Surprise', and 'Disgust' expressions were extracted, each being at two levels of emotional intensity, with an additional neutral expression. We have only considered 6 emotions from this list, leaving behind the audio files belonging to the 'Calm' and 'Surprise' category. Each recording was rated 10 times on emotional validity, intensity, and genuineness. The 247 number of raters evaluated the dataset with 60% accuracy for emotion recognition in audio-only mode. The second dataset chosen for validation was 'EMODB' (Database of German Emotional Speech). In this dataset, 10 actors (5 male and 5 female) recorded 10 German utterances in seven emotions which are 'Neutral', 'Anger', 'Fear', 'Joy', 'Sadness', 'Disgust', and 'Boredom'. Out of these, the first six were chosen for the validation purpose. Only utterances, which were recognized with an accuracy of greater than 80%, were chosen [29]. The third dataset chosen for validation was Surrey Audio-Visual Expressed Emotion ('SAVEE') dataset, which is also a multimodal dataset, from which we worked on only the audio files. The four male actors recorded a total of 480 utterances in British English. Seven emotions were recorded out of which we used the six we required. The ten subjects evaluated each audio recording, and recognition accuracy was 61%. Table 2 shows the distribution of audio files chosen from all the datasets under the six emotion categories.

Table 2 Emotion-wise distribution of audio files in four datasets

Emotions	CREMA-D	RAVDESS	EmoDB	SAVEE
Anger	1271	192	121	60
Disgust	1271	192	46	60
Fear	1271	192	69	60
Happy	1271	144	66	60
Neutral	1084	144	75	120
Sad	1271	192	59	60

4 System Description

The proposed system has different phases to detect emotions from audio files. Figure 1 shows the workflow of the proposed system. The system is divided into two phases: the first one is the feature extraction and the second one is the classification phase. The audio files of the dataset are available in.wav format, which is represented as the amplitude versus time plot in Fig. 1. The audio file is passed to the feature extractor where acoustic features are extracted from the file. Based on the extracted features, the classifier is trained to classify the audio file into six emotion categories, i.e. 'Happy', 'Sad', 'Neutral', 'Anger', 'Disgust', and 'Fear'.

4.1 Feature Extraction

Emotion recognition from speech can be performed by selecting a suitable feature set, which acts as a determinant for the different emotion categories. There are different types of information present in speech and different types of features cover different types of information stored in a very overlapped manner. The phase one of the proposed system consisted of extracting the required features from audio files. According to a broad classification, features used in speech studies are divided into the following categories: excitation source, vocal tract system, and prosodic features [30]. The excitation source features are the ones derived from the excitation source signal, and the vocal tract characteristics are suppressed to obtain such features. Which is achieved by first using filter coefficients to extract the vocal tract

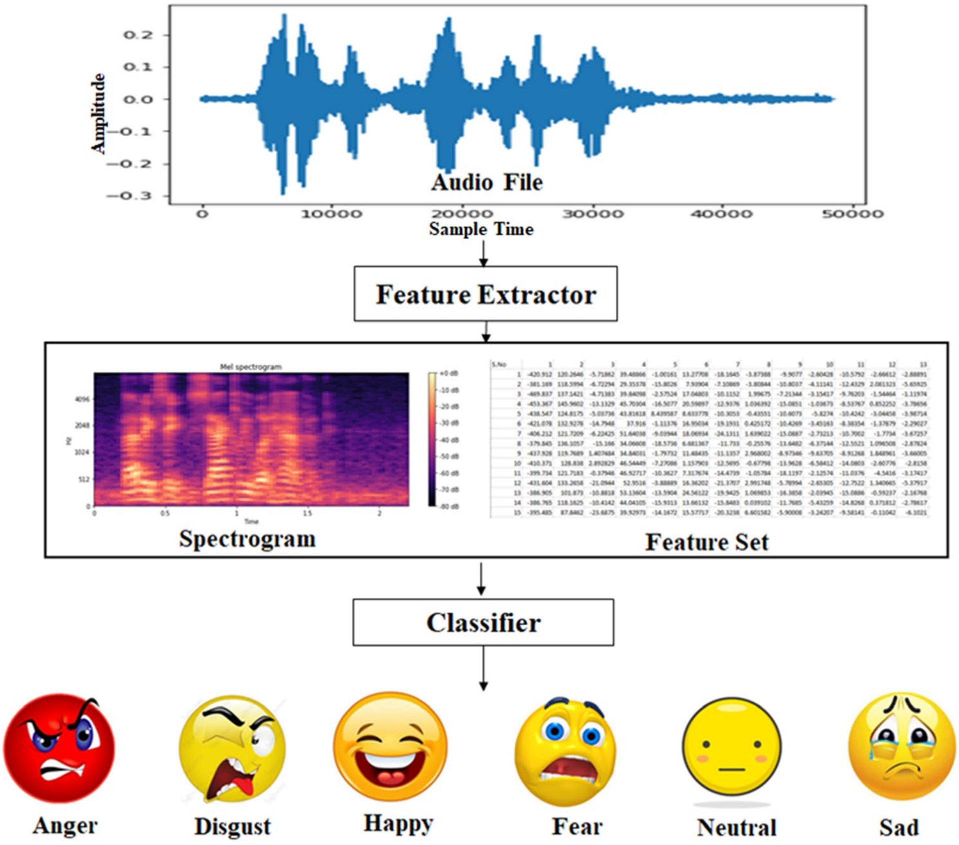

Fig. 1 System overview

characteristics and later on separating them using inverse filter formulation. Most of the time in the literature, two types of signals: LP residual signals and the Glottal Volume Velocity (GVV) signals, are explored for extraction of the excitation source features. The reason behind this is that both of these signals correlate with the excitation source information [31]. Due to the unpredictable nature of the speech signal, the LP component is mostly perceived as an error signal [32]. Also, there are high-order relations contained in the LP residual signal, and there does not exist a well-defined procedure to extract these high-order relations [33]. The second class of features known as the vocal tract features are well reflected in the frequency domain analysis of audio files. Most of the time, a segment of speech, which is in between 20 and 30 ms, is used for extracting these features. The frequency domain is directly related to the Fourier transform. The short time spectrum can be obtained from the speech frame using the Fourier transform. Taking Fourier transform on a log magnitude spectrum gives the cepstrum [34]. MFCCs (Mel Frequency Cepstral Coefficients) is the most widely used vocal tract features [35]. It has been seen that the lower-order MFCC features convey the phonetic (speech) information, while the higher-order features contain non-speech information. Therefore, in our study, we have considered the 40 MFCC coefficients, which are both the high-order as well as the low-order MFCCs, thereby removing the possibility of leaving behind any information relevant to the task of emotion recognition. Furthermore, there are various other components available in speech, which are the duration, intonation, and intensity patterns. The presence of such features makes human speech natural. There are acoustic correlates which are used as these types of features. Pitch, energy, duration, and their derivatives are used as prosodic features [36, 37]. In the proposed system, three feature sets are tested. The first one consists only the spectral features like 40 MFCC coefficients along with the mean and standard deviation values of these features. The second feature set consists only of the prosodic features which consist of pitch, zero-cross ratings (ZCR), and energy (root-mean-square energy) along with their statistical features like mean, median, standard deviation, values of upper quartile and lower quartile for pitch; mean and standard deviation of energy; mean and standard deviation of ZCR. The third one consists of a combination of both along with their statistical values. The size of feature set is *n***f*, where ‘*n*’ is the number of rows representing the audio clips and ‘*f*’ is the number of features per audio clip. The respective value of ‘*f*’ for spectral, prosodic and combined is 80, 9, and 89 feature values per row, where each row represents one audio clip. The Algorithm 1 presents the procedure to extract spectral features, and the Algorithm 2 presents the procedure to extract prosodic features. Since the speech signal is not stationary [38], framing allows the speech signal to be seen as stationary for a short-time point of view. Thus, by varying frame sizes, we get different quanta of samples under focus at a particular time. By performing experiments on these variants, we have determined the best frame size to extract the feature set for classification. Since the average size of an audio clip in this dataset is of 2 s, choosing a frame size above 16,384 samples per frame would have resulted in a frame larger than the clip, which would not be useful. On the same lines, a frame size lower than 1024 samples per frame would not be able to catch even a full phoneme production. Hence it would again be unyielding. Therefore, the features were extracted in five frame sizes: 1024, 2048, 4096, 8192, and 16,384 samples per frame with a 50% overlap, using librosa signal processing library [39].

Algorithm 1 Spectral Feature Extraction

```
procedure SpecFeatureExt()
    counter ← 0
    frame_length, hop_length, no_of_features ← Get USER INPUT
    while counter < 6 do
        Compute frame wise mfcc_values using Librosa.
        for j ←1 To no_frames do
            $mel\_vals = \sum_{i=1}^{n} mfcc\_values_i / no\_frames$
        end for
        counter ← counter +1
    end while
end procedure
```

Algorithm 2 Prosodic Feature Extraction

```
procedure ProsFeatureExt()
    counter ← 0
    frame_length, hop_length, no_of_features ← Get USER INPUT
    while counter < 6 do
        Compute frame wise zcr, rmse, pitch values using Librosa.
        Compute the statistical values.
        counter ← counter +1
    end while
end procedure
```

4.2 Classification

In this phase, the classifier classifies the audio file into one of the six emotion categories based on the values of its features. For the classification task, we have analysed two types of classifiers: support vector machines (SVMs) and the recurrent neural networks (RNNs). Keras library with tensorflow backend is used to code both the SVM and RNN models. For the SVM based classifier, 'c' parameter is set to 50 and 'gamma (γ)' to 0.01 along with the kernel as RBF. We call this classifier as 'SV-Classifier'. We performed various experiments to select appropriate values of 'c' and 'γ'. 'c' values were chosen from a range of values: 10, 50, 100, 1000 and 'γ' from: 0, 0.1, 0.01, 10, 100. After applying the grid search on these values, the results came out to be best for the values of 'c' as 50 and 'γ' as 0.01. Hence the same has been chosen for the experiments. Additionally, other kernels like Polynomial, Linear, and Sigmoid were also tested for this classifier, but RBF outperforms the other. RNNs perform better when dealing with the inputs of variable lengths and in case, classifier needs to be trained on the sequential data with long term context [40]. The audio files are of variable lengths, and the audio data are sequential in form prone to minor changes and variations. Thus, RNN was chosen as the preferred option over CNN. We have validated various variants of RNN classifier to perform the classification task by varying the number of layers and dropout rate. The six-layer RNN-based classifier with a 10% dropout rate is used, which is addressed as 'R-Classifier' in the later text. The results obtained from both the classifiers are presented in the subsequent section.

4.3 Experimental Set-up

The computational set-up used for training of the proposed classifier is CPU-based system having Intel Core i5 processor (8th generation), with 8 GB memory. Python language is used to implement the models and to perform other pre-processing tasks. The time efficiency (average time) of the classifiers was approximately 15.75 s for 'SV-Classifier' and 1594 s for 'R-Classifier' as shown in Table 3. The time efficiency is dependent on number of tuples 'n' and features 'f'; hence, the time complexity is $O(n^2 f)$. The RNN classifier is layer dependent as well.

5 Results and Discussion

This section presents and discusses the results obtained from the proposed work. The classification results obtained by applying the SVM classifier and RNN classifier on 'CREMA-D' and the validation results on the other three datasets are discussed and compared. Further, this section presents which classifier is better suited for the given particular task and what are the limitations of the

Table 3 Timing accuracy of 'SV-classifier' and 'R-classifier'

Datasets (f)	SV classifier time (s)	R-classifier time (s)
Prosodic (9)	2.18	284.94
Spectral (80)	17.83	2199.77
Combined (89)	27.24	2299.17

other. Initially, we have discussed the selection of feature sets. The ratio of training and testing data is kept as 80:20. So the testing is done on the subset of the same dataset on which the classifier is trained. Five cross-validation approaches are employed, where the results presented in this section are average of the five batches, in which every batch has random 80:20 ratio split of the data available.

5.1 Feature Set Selection

Each of three feature sets were extracted at 5 different frame sizes resulting in total 15 feature sets. Table 4 shows the validation results when these feature sets, extracted from ‘CREMA-D’, were tested with the ‘SV-Classifier’. It can be seen from the values in the table that combined feature set gives the best results when extracted with 2048 frame size from the audio files, i.e. 58.22%. The results were also tested with other kernels such as Polynomial, Linear, and Sigmoid. In each case, the classifier with RBF kernel gives the best accuracy than with other kernels. Thus, this feature set is used for the final proposed system. The overall accuracy of the ‘SV-Classifier’ for multi-classification of the audio files in one of the six emotions comes out to be 58.22%.

5.2 Binary Emotion Classification Using ‘SV-Classifier’

The ‘SV-Classifier’ with above chosen feature set was trained for binary emotion classification. For each emotion, data were grouped into two classes: class belonging to a particular emotion, and the other class representing all other combined classes. Table 5 presents the accuracy results on training and validation datasets for binary classification of each emotion.

It can be seen from Table 5 that ‘SV-Classifier’ classifies the audio files into emotions: ‘Anger’, ‘Sad’, ‘Neutral’, ‘Fear’, ‘Happy’, and ‘Disgust’ with accuracy of 90.13%, 85.43%, 84.49%, 84.35%, 81.80%, and 79.45%, respectively. The audio files belonging to ‘Anger’ emotion were the most accurately classified and that belonging to ‘Disgust’ emotions were the least accurately classified by the ‘SV-Classifier’. The accuracy for classifying disgust emotion by the classifier is the least, which may be due to the fact that disgust emotion conveyed at a higher intensity may sometimes sound as anger.

Table 4 Feature set comparison for ‘SV-Classifier’ on ‘CREMA-D’

Frame size	Prosodic (%)	Spectral (%)	Combined (%)
1024	37.68	55.94	56.54
2048	37.20	57.75	58.22
4096	38.48	56.54	57.28
8192	33.85	54.20	54.13
16,384	34.99	52.59	52.85

Table 5 Results of binary classification using ‘SV-Classifier’

Emotion class	Accuracy	
	Training (%)	Validation (%)
Anger	91.13	90.13
Disgust	79.89	79.45
Fear	83.55	84.35
Happy	82.55	81.80
Neutral	85.62	84.49
Sad	87.18	85.43

5.3 Validation of ‘SV-Classifier’ on Other Datasets

The ‘SV-Classifier’ was ran on various datasets, i.e. ‘EmoDB’, ‘RAVDESS’, ‘SAVEE’ with the same parameters. Figure 2 shows the overall accuracies achieved on each dataset. According to this figure proposed, ‘SV-Classifier’ gives the accuracy of 86.36% on ‘EmoDB’, 64.15% on ‘RAVDESS’, and 77.38% on ‘SAVEE’ dataset, respectively.

Further, Fig. 3 shows emotion-wise accuracy of binary classification by the system for different datasets.

In [41], authors achieved 94.8% recognition rate for ‘Anger’, 52.9% for ‘Happy’, and 98.6% for ‘Neutral’ emotion with an overall accuracy of 85.1% on ‘EmoDB’ dataset though results obtained in the paper are for classification of audio files in 3 classes only. This accuracy has been achieved using the one-dimensional convolutional neural networks, whereas our system gives an overall accuracy of 86.36%, which is 1.48% better than this approach. Also, the proposed system gives better results in case of both ‘Anger’ and ‘Happy’ emotions, whereas it gives comparable results in the case of ‘Neutral’ emotion. These results are even better than obtained in [42], which has the

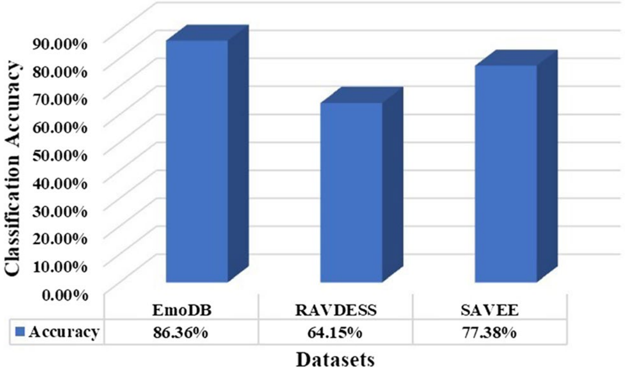

Fig. 2 Overall accuracy of classification with ‘SV-Classifier’

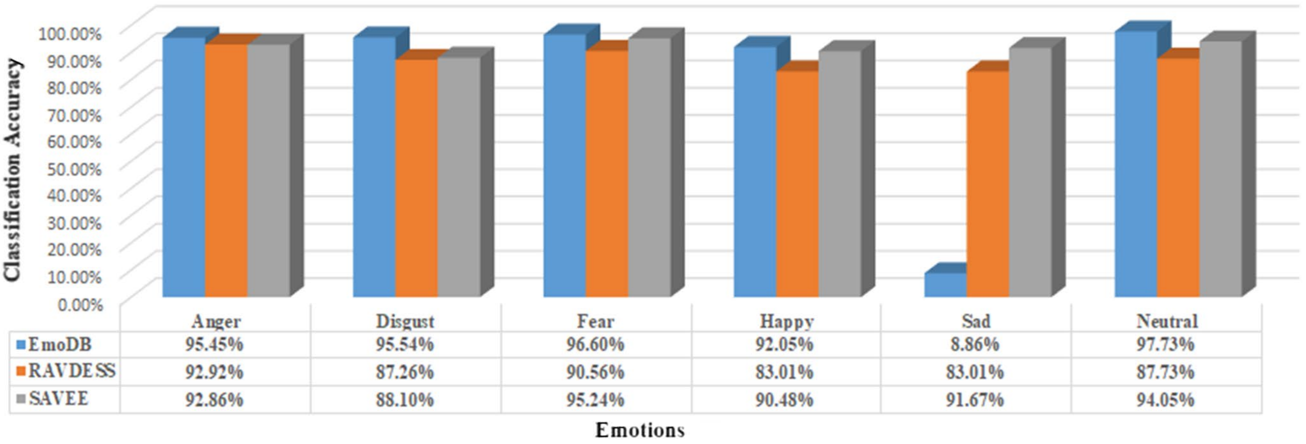

Fig. 3 Emotion-wise binary classification accuracy for different datasets

Table 6 Classification accuracies obtained by varying no. of layers and dropout rate

No. of layers	Drop-out = 10 (%)	Drop-out = 20 (%)	Drop-out = 30 (%)	Drop-out = 40 (%)	Drop-out = 50 (%)
3	45.80	46.67	44.05	44.72	46.54
6	47.41	41.97	36.13	43.04	44.12
9	35.70	31.80	30.67	31.54	36.82
12	17.24	16.84	16.55	17.07	18.98

best-reported accuracy of 83% on the same dataset. Also, for ‘SAVEE’ dataset, authors in [43] reported the best accuracy of 77.4%, whereas our system gives an equivalent accuracy of 77.38% with better classification rate for individual emotions. For the third dataset, i.e. ‘RAVDESS’ the best-reported accuracy by authors in [44] is 66.41%, and ours has a comparable accuracy of 64.15%.

5.4 Classification Using ‘R-Classifier’

For the initial experiment using RNN, the dropout rate was set to 30%, and different variants of RNN were tested by varying the number of hidden layers in the model. We started with 3 layers, and the layers were increased in the multiple of 3 up to maximum 12 layers. Dropout rate is also varied starting from 10%, with an increment of 10, up to 50%. The results are given in Table 6.

It is clear from this table that the combination of 6-layer RNN with 10% dropout gives the best result for the overall classification. Figure 4 shows the emotion-wise classification accuracy achieved by ‘R-Classifier’.

It can be seen from Fig. 4 that even with ‘R-Classifier’ the files belonging to ‘Anger’ emotion were most accurately classified. The decreasing order of classification accuracy obtained for various emotions is for ‘Anger’, ‘Neutral’, ‘Sad’, ‘Happy’, ‘Fear’, and ‘Disgust’. With this classifier also, the files belonging to ‘Disgust’ emotion were least accurately classified.

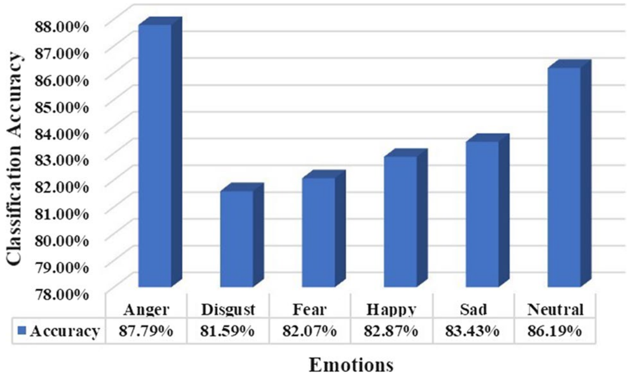

Fig. 4 Emotion-wise classification accuracy of ‘R-Classifier’

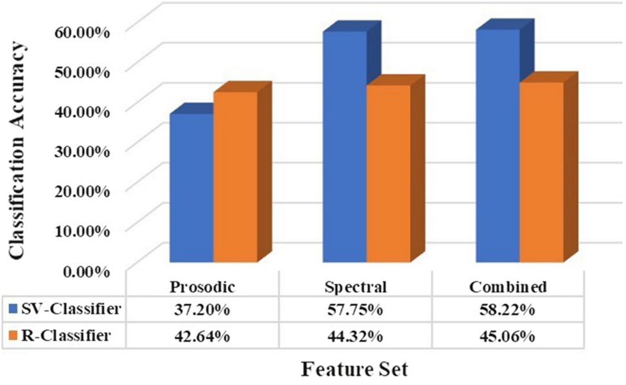

Fig. 5 Comparison of ‘SV-Classifier’ and ‘R-Classifier’

5.5 Comparison of ‘SV-Classifier’ with ‘R-Classifier’

The comparison between the ‘SV-Classifier’ and the ‘R-Classifier’ is shown in Fig. 5. It can be clearly seen that for both spectral and combined feature set the ‘SV-Classifier’ outperforms the ‘R-Classifier’. However, with the prosodic feature set, ‘R-Classifier’ performs better than the ‘SV-Classifier’. As discussed earlier, that combined feature set suits better for the given task. Thus, it can be inferred from Fig. 5 that ‘SV-Classifier’ suits better for the given classification task. The reason for ‘SV-Classifier’ being better than ‘R-Classifier’ for the task is that it requires less amount of data to train and size of the dataset available is not enough to train very deep neural networks such as RNN. The difference between the results of two classifiers is also statistically significant. To prove this, we have performed the McNemar Test [45, 46]. This test is widely used to compare supervised learning algorithms. According to this test, *p* value of the null hypothesis, *H0,* is calculated, and it is compared to the alpha value, which is mostly chosen as 0.05; if the *p* value comes greater than alpha then the null hypothesis, *H0* holds; otherwise, it is rejected. For our comparison, the null hypothesis *H0* is considered as follows:

H0 The classifiers have a similar proportion of errors and any difference in accuracy is by chance only.

The various values considered for performing the test and results obtained after applying the test are given in Table 7.

After the statistical calculations, the *p* value came out to be 1.7575235576553599e−13. This *p* value is less than alpha which shows that we can safely reject the null hypothesis and that the given models don’t have a similar proportion of results on the validation dataset. From the test, we can conclude that both the classifiers, ‘SV-Classifier’ and ‘R-Classifier’, provide variable results on the dataset and the differences in the results obtained using them are statistically significant and not by chance. Figure 6 shows the comparison of the emotion-wise binary classification accuracies of the ‘SV-Classifier’ and the ‘R-Classifier’.

It can be inferred from Fig. 6 that ‘SV-Classifier’ gives better binary classification accuracies in case of ‘Anger’, ‘Fear’, and ‘Sad’ emotions. However, for the rest of emotions ‘Disgust’, ‘Happy’, and ‘Neutral’, the ‘R-Classifier’ gives better classification results than the ‘SV-Classifier’. Although it is difficult to choose between the two classifiers for binary classification, ‘SV-Classifier’ has the overall better classification accuracy results.

Table 7 Parameter values for McNemar test

S. no.	Parameter	Value
1.	*N* (Number of pairs)	1489
2.	*K* (Number of categories)	2 (correct and incorrect)
3.	Degrees of freedom, $k(k-1)/2$	1
4.	Alpha	0.05
5.	*p* value	1.7575235576553599e−13

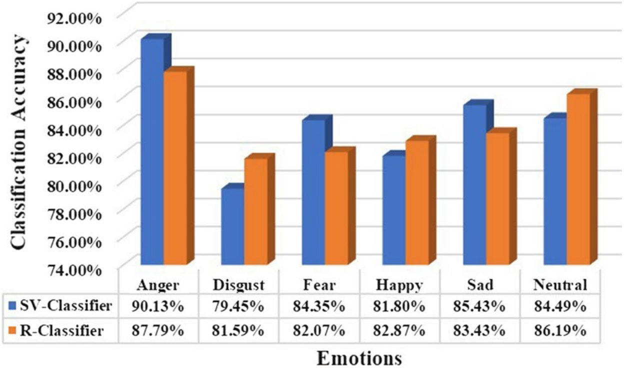

Fig. 6 Comparison of ‘SV-Classifier’ and ‘R-Classifier’ for binary classification

5.6 Classification of Audio Files According to Gender: An Added Advantage

The proposed system is also capable of classifying the audio files in two gender categories, i.e. ‘Male’ and ‘Female’. Identifying the gender from audio files is equally important because of various reasons like the amplitude or frequency range of male voice is very different from that of a female voice. Emotions conveyed through speech also vary widely according to gender. The ‘SV-Classifier’ can classify the audio files in these two categories with an accuracy of 96.24%, and the ‘R-Classifier’ performed the same task with an accuracy of 87.11%. Figure 7 shows the emotion-wise gender classification results. It can easily be observed from Fig. 7 that ‘SV-Classifier’ again outperforms the ‘R-Classifier’ for classification of audio files in two gender categories. Also it can be observed that this gender-wise classification ultimately leads to improved classification accuracies than the normal classification as

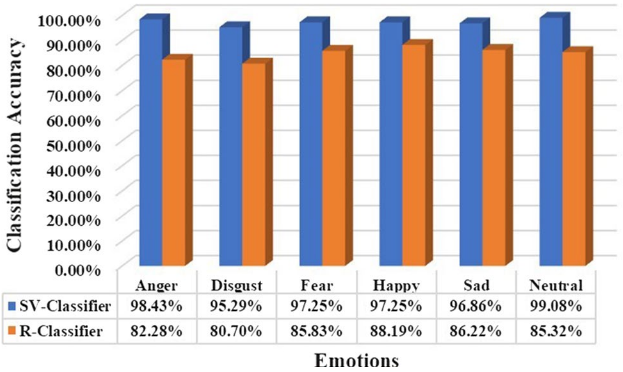

Fig. 7 Emotion-wise gender classification of audio file

done in the previous sections. The classification accuracy obtained is the maximum for the 'Neutral' emotion and is the least for the 'Disgust' emotion in case of the 'SV-Classifier', whereas for the 'R-Classifier' the maximum accuracy achieved is for the 'Happy' emotion and the least is for the 'Disgust' emotion.

6 Conclusion and Future Scope

In this paper, we have proposed a system which can efficiently classify the audio files in one of the six emotion categories. The proposed system gives an impressive increase of 29.74% in the classification accuracy in comparison with the baseline accuracy of human classification on the 'CREMA-D' dataset. The system gives a classification accuracy of 90.13%, 79.45%, 84.35%, 81.80%, 85.43%, and 84.49% for the 'Anger', 'Disgust', 'Fear', 'Happy', 'Sad', and 'Neutral' emotions, respectively. The proposed work is organized in two phases: the first one deals with finding the suitable feature set and the second one works on finding the suitable classifier. During the first phase, various feature sets based on the type of features used and frame size used for extraction of these were tested. It has been found that the combination of prosodic and spectral features along with their statistical values, extracted using a frame size of 2048 samples per frame, gives the best classification results. During the second phase, we explored and compared the support vector machine and recurrent neural network-based classifiers for the given task. In the binary classification of the emotions, there has been a significant increase in the accuracy results obtained for each emotion class in comparison with the current state-of-the-art results reported in the literature. The results obtained have further been validated on some other standard datasets such as 'RAVDESS', 'EmoDB', and 'SAVEE'. Choice of both the English as well as the German language datasets validates the language-independent aspect of this system. The proposed system gives the classification accuracy of 86.36% on 'EmoDB', 64.15% on 'RAVDESS', and 77.38% on 'SAVEE'. The proposed system has achieved an increase of 7.95% on 'EmoDB', an increase of 6.91% on 'RAVDESS', and increase of 26.85% on 'SAVEE' datasets, respectively, in the classification accuracy in comparison with the best-reported results in the literature. In most of the cases, the proposed system outperforms the previous ones, and, in some instances, it gives comparable results in comparison with the existing ones in emotion classification. The overall 'SV-Classifier' gives better classification results than the 'R-Classifier'. The same trend has been witnessed during the classification of audio files in gender categories. In this case, SV-Classifier' gives 96.24% classification accuracy and 'R-Classifier' achieves 87.11% classification accuracy. The primary reason can be due to the lack of availability of a large amount of data to train very deep neural networks such as RNN. The classification accuracy of 'R-Classifier' can be improved further if more data is made available in the future. The experiments have been performed purely on the datasets recorded in the ideal conditions, and results may vary when tested on the real-time data. Thus, there is the future scope of work to make the systems more resilient to cope with the real environmental or surrounding factors.

References

1. Vlasenko, B.; Schuller, B.; Wendemuth, A.; Rigoll, G.: On the influence of phonetic content variation for acoustic emotion recognition. In: International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, pp. 217–220. Springer, Berlin, Heidelberg (2008)
2. Scherer, K.R.: Vocal communication of emotion: a review of research paradigms. Speech Commun. **40**(1–2), 227–256 (2003)
3. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R.: 'CREMA-D': crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. **5**(4), 377–390 (2014)
4. Barsoum, E.; Zhang, C.; Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283. ACM (2016)
5. Abdelwahab, M.; Busso, C.: Study of dense network approaches for speech emotion recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5084–5088. IEEE (2018)
6. Burmania, A.; Busso, C.: A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors. In: INTERSPEECH, pp. 152–156 (2017)
7. Arora, P.; Chaspari, T.: Exploring siamese neural network architectures for preserving speaker identity in speech emotion classification. In: Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human–Machine Interaction, pp. 15–18. ACM (2018)
8. Oudeyer, P.Y.: Novel useful features and algorithms for the recognition of emotions in human speech. In: Speech Prosody 2002, International Conference (2002)
9. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
10. Livingstone, S.R.; Russo, F.A.: 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 **13**(5), e0196391 (2018)
11. Jackson, P.; Haq, S.: Surrey Audio-Visual Expressed Emotion (SAVEE) Database. University of Surrey, Guildford (2014)
12. Neiberg, D.; Elenius, K.; Karlsson, I.; Laskowski, K.: Emotion recognition in spontaneous speech. In: Proceedings of Fonetik, pp. 101–104 (2006)
13. Blouin, C.; Maffiolo, V.: A study on the automatic detection and characterization of emotion in a voice service context. In: Ninth European Conference on Speech Communication and Technology (2005)

14. Cummings, K.E.; Clements, M.A.: Analysis of the glottal excitation of emotionally styled and stressed speech. J. Acoust. Soc. Am. **98**(1), 88–98 (1995)
15. Sauter, D.A.; Eisner, F.; Ekman, P.; Scott, S.K.: Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc. Natl. Acad. Sci. **107**(6), 2408–2412 (2010)
16. Fayek, H.M.; Lech, M.; Cavedon, L.: Evaluating deep learning architectures for speech emotion recognition. Neural Netw. **92**, 60–68 (2017)
17. Huang, C.-W.; Narayanan, S.S.: Attention Assisted discovery of sub-utterance structure in speech emotion recognition. In: Proceedings of Interspeech, pp. 1387–1391 (2016)
18. Lee, J.; Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH, pp. 1537–1540 (2015)
19. Singh, R.; Rana, R.; Singh, S.K.: Performance evaluation of VGG models in detection of wheat rust. Asian J. Comput. Sci. Technol. **7**(3), 76–81 (2018)
20. Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y.: Exploring the Limits of Language Modeling (2016). arXiv:1602.02410[cs]
21. Radford, A.; Jozefowicz, R.; Sutskever, I.: Learning to Generate Reviews and Discovering Sentiment (2017). arXiv:1704.01444 [cs]
22. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; et al.: Deep Speech: Scaling Up End-to-End Speech Recognition (204). CoRR, arXiv:1412.5567
23. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; et al.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016). arXiv:1609.08144[cs]
24. Lakomkin, E.; Zamani, M.A.; Weber, C.; Magg, S.; Wermter, S.: Emorl: continuous acoustic emotion classification using deep reinforcement learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–6. IEEE (2018)
25. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B.: Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, ser. MM'13, pp. 835–838. ACM, New York
26. Bahdanau, D.; Cho, K.; Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
27. Wang, Z.Q.; Tashev, I.: Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5150–5154. IEEE (2017)
28. Bothe, C.; Magg, S.; Weber, C.; Wermter, S.: Conversational Analysis using Utterance-Level Attention-Based Bidirectional Recurrent Neural Networks (2018). arXiv preprint arXiv:1805.06242.
29. Erdem, E.S.; Sert, M.: Efficient recognition of human emotional states from audio signals. In: 2014 IEEE International Symposium on Multimedia, pp. 139–142. IEEE (2014)
30. Fourier Analysis And Synthesis. Hyperphysics.Phy-Astr.Gsu.Edu. http://hyperphysics.phy-astr.gsu.edu/hbase/Audio/fourier.html#c1. (2018). Accessed 21 Nov 2018
31. Kodukula, S.R.M.: Significance of excitation source information for speech analysis. Doctoral dissertation, Ph.D. thesis, Dept. of Computer Science, IIT, Madras (2009)
32. Yegnanarayana, B.; Murthy, P.S.; Avendaño, C.; Hermansky, H.: Enhancement of reverberant speech using LP residual. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 1, pp. 405–408. IEEE (1998)
33. Yegnanarayana, B.; Prasanna, S.M., Rao, K.S.: Speech enhancement using excitation source information. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I–541. IEEE (2002)
34. Ravindran, G.; Shenbagadevi, S.; Selvam, V.S.: Cepstral and linear prediction techniques for improving intelligibility and audibility of impaired speech. J. Biomed. Sci. Eng. **3**(01), 85 (2010)
35. Ververidis, D.; Kotropoulos, C.: Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006)
36. Bänziger, T.; Scherer, K.R.: The role of intonation in emotional expressions. Speech Commun. **46**(3–4), 252–267 (2005)
37. Cowie, R.; Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Commun. **40**(1–2), 5–32 (2003)
38. Jannat, R.; Tynes, I.; Lime, L.L.; Adorno, J.; Canavan, S.: Ubiquitous emotion recognition using audio and video data. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 956–959. ACM (2018)
39. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O.: Librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, pp. 18–25 (2015)
40. Graves, A.: Supervised sequence labelling with recurrent neural networks. Ph.D. thesis, Technische Universitat Munchen (2008)
41. Gao, M.; Dong, J.; Zhou, D.; Zhang, Q.; Yang, D.: End-to-end speech emotion recognition based on one-dimensional convolutional neural network. In: Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, pp. 78–82. ACM (2019)
42. Anjum, M.: Emotion recognition from speech for an interactive robot agent. In: 2019 IEEE/SICE International Symposium on System Integration (SII), pp. 363–368. IEEE (2019)
43. Avots, E.; Sapiński, T.; Bachmann, M.; et al.: Audiovisual emotion recognition in wild. Mach. Vis. Appl. **30**, 975 (2019). https://doi.org/10.1007/s00138-018-0960-9
44. Jannat, R.; Tynes, I.; Lime, L.L.; Adorno, J.; Canavan, S.: Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 956–959. ACM (2018)
45. Fagerland, M.W.; Lydersen, S.; Laake, P.: Statistical Analysis of Contingency Tables. Taylor & Francis/CRC, Boca Raton (2017)
46. Chow, S.C.; Shao, J.; Wang, H.; Lokhnygina, Y.: Sample size calculations in clinical research, 3rd edn. Taylor & Francis/CRC, Boca Raton (2018)