Exploring the mel scale features using supervised learning classifiers for emotion classification

Kalpana Rangra* and Monit Kapoor

Department of Cybernetics,
School of Computer Science,
University of Petroleum and Energy Studies,
Dehradun, 248007, India
Email: krangra@ddn.upes.ac.in
Email: mkapoor@ddn.upes.ac.in
*Corresponding author

Abstract: Human emotions are inherently ambiguous and impure but emotions are important while considering the human uttered speech. The role of human speech is intensified by the aspect of the emotion it conveys. There are several characteristics of speech that differentiates it among different utterances. Various prosodic features like pitch, timbre, loudness and vocal tone categorise speech into several emotions and other domains. The sample speech is changed when it is subjected to various emotional environments. Researches support various experimental analyses for phonetics and prosodic parameters that quantify the quality of speech. Identification of different emotional states of an actor (speaker) can also be done on the basis of mel scale. MFCC is one such variant to study the emotional aspects of the utterances by the speaker. The paper implements a model to identify several emotional states from MFCC for two datasets. The work classifies emotions for two datasets on the basis of MFCC features and gives the comparison of both. This work implements a classification model based on dataset minimisation that is done by taking the mean of features for the improvement of classification accuracy on different machine learning algorithms.

Keywords: speech recognition; emotion recognition; mel-frequency cepstral coefficient; MFCC; machine learning; supervised learning; ANN.

Reference to this paper should be made as follows: Rangra, K. and Kapoor, M. (2021) 'Exploring the mel scale features using supervised learning classifiers for emotion classification', *Int. J. Applied Pattern Recognition*, Vol. 6, No. 3, pp.232–253.

Biographical notes: Kalpana Rangra is an Assistant Professor in University of Petroleum and Energy Studies with seven years of teaching experience. His areas of expertise include data mining, machine learning, emotion recognition, and personality detection.

Monit Kapoor is a Professor and HOD at the Department of Cybernetics at University of Petroleum and Energy Studies. He is a focused individual with flair for teaching/faculty/mentoring roles and active desire for research in computer engineering with expertise in areas like DevOps, cloud computing solutions, IOT, wireless networks and artificial intelligence.

Copyright © 2021 Inderscience Enterprises Ltd.

1 Introduction

The vocal acoustics are rich with emotional cues for analysing the speaker's emotional state. Each emotion is associated with tone of the speaker. Emotions can be detected from text and sounds both. Each of them has different approaches to identify the emotional state of the speaker. Intelligent and rational decision making is also affected by emotions so they greatly define the interpersonal relations. Emotions can easily handle the communication bridge among the speaker and the listener. The emotional in utterances bring more clarity and increase efficiency of the interaction among individuals. Emotions carry considerable information of the mental state of human being and are pivot for engaging discussions among groups (Koolagudi and Rao, 2012). The information hidden in emotions paved the door for evolution of a speech recognition field commonly referred as automatic speech recognition (ASR). Researchers have proposed several models for interpreting and retrieval of emotions from images of speakers face and his expressions, voice and tone of the speaker conversations and few other have also discussed about physiological signals (Marechal et al., 2019). Emotions have great role in communication since they express the speaker's intentions to listeners in conversation. There are several spoken language interfaces available today that support ASR. Availability of such systems are base to collect the samples for speech recognition (Rao et al., 2013). The speech systems available currently can process the naturally spoken utterances recorded for analysis with high accuracy but the emotional component in speech processing makes the ASR systems more realistic and meaningful. There are several real world applications that may benefit from emotional context of the utterances. Few of them to be enlisted include entertainment, indexing audio files based on emotions, HCI-based systems, etc. (Koolagudi et al., 2012).

Some of the selected features can be trained to classify, recognise and predict emotions. There are several emotions that can be extracted from the utterances. Few of the universally enlisted among them are happiness, fear, sadness, anger, neutral and surprise. These emotions can be recognised by any intelligent system constraining to finite computational resources. Exploring the emotional domains in speech make the human computer interactions more real and efficient. Analysis of utterances for voice and speech is practically plausible and enhance human conversations. The results of emotion detection can be broadly applied to wide areas including e-learning platforms, car-board systems, medical field and so on.

The remaining sections of the paper are as follows, Section 2 is a detailed literature survey of the related works in the area, Section 3 contains the problem statement, Section 4 carried the details of system implemented and results achieved under problem solution, and finally Section 5 is the conclusions. In this paper, we have tried to establish the proof of concept for using only 12 mel-frequency cepstral coefficient (MFCC) from 39, have identified which 12 MFCC are to be used for speech and emotion analysis. The dataset used for this experimentation is EMODB. Various supervised learning approaches have been implemented for classifying emotions from two databases EMODB and Surrey audio-visual expressed emotion (SAVEE).

2 Literature review

Spectral analysis is a promising technique for detecting emotions from sample speech. Prosodic features of speech signals can also be used for analysing emotions since they contain emotional information. Researchers explored the role and context of emotion by using an 88 feature set called eGmaps (Latif et al., 2018). Speech patterns can be obtained from collaboration of various speech features obtained from speaker utterances. The importance of feature selection is pivot in differentiating several emotions of same speaker from a speech (Lee and Narayanan, 2005) and it is dependent on selecting best features from the signal. Various available human languages vary in ascent, sentences, and speaking styles (Banse and Scherer, 1996) that makes it difficult to identify the emotions from utterances. Various aspects of spoken languages alter the extracted features of the sound signal. It is possible that a sample speech may have more than one emotion which means that each emotion corresponds to different part of same speech signal. This results in challenge to define boundaries among emotions. An attempt has been made to study the multilingual emotion classification models in Hozjan and Kačič (2003).

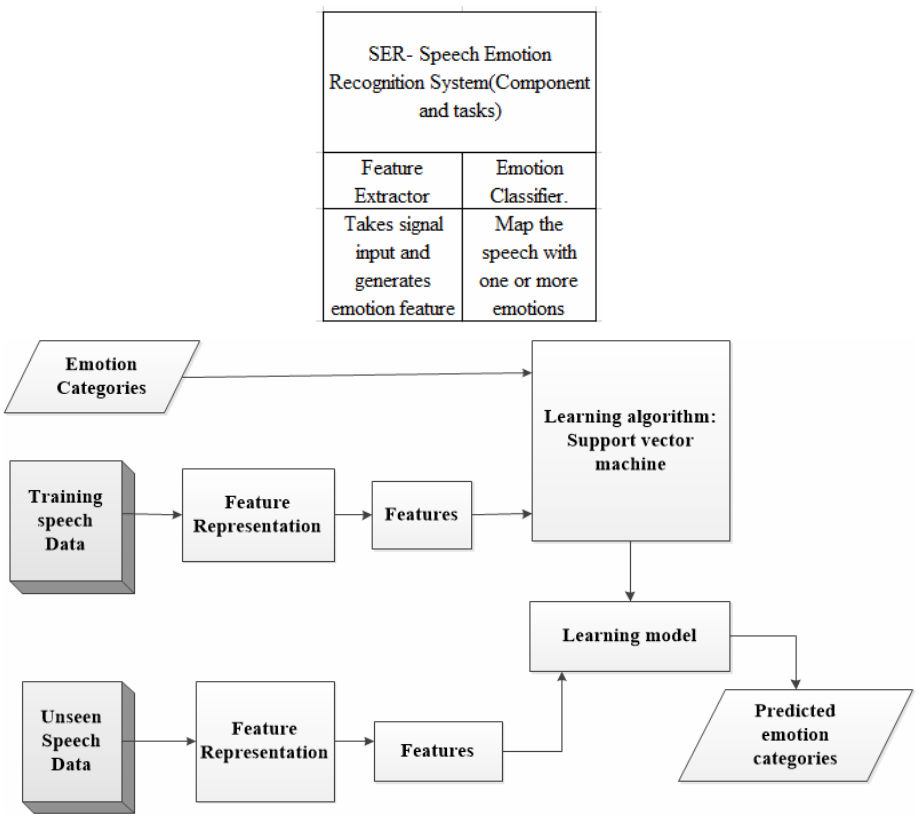
There has been a substantial advancement in technology that has affected the growth and maturity of the emotion recognition fields (Ramakrishnan, 2012; Sebe et al., 2005). Few of them can be named as call centres, remote education (Zhang et al., 2007). Existing speech recognition system can be improved by exploring cepstral features with acoustic features (Jing et al., 2018).

Authors identified classes of features extracted from emotional speech features electrography and speech signals. There is an important aspect of SER that includes characterisation of emotional content of speech (Albornoz et al., 2011). Several speech features are obtained from speech acoustic analysis and can be used to detect and predict emotions (Özseven et al., 2018). The aim of selecting speech features is determination of properties that can improve rate of classification for emotions from a feature set (Kishore and Satish, 2013). The machine learning approaches are flexible enough to adapt themselves to any model that studies emotion and show good performance perform well predicting tasks based on selected features (Yousefpour et al., 2017).

2.1 Speech, emotion and classification

There are different terms entwined with emotions such as mood, temperament, personality, sentiment and motivation (Shu et al., 2018). Emotions can be understood as the complex feeling of the mind that results in physiological and physiological changes. Human thoughts and behaviour is influenced by emotions and there are instances of changes in body when it encounters different emotional states (Oosterwijk et al., 2012). Pessoa (2011) and Winkielman et al. (2014) proves that there is considerable influence of emotional syndromes human actions and reactions. Several applications created by researchers have emotion detection as integral component by for identification of behavioural patterns (Fernández-Caballero et al., 2016; Guan et al., 2018). Speech features can be extracted from various sources to accomplish predictive analytics. The sources can be vocal tract, excitation source and prosodic extraction.

Figure 1 Framework for supervised emotion classification



Source: Cen et al. (2016)

Emotions can be broadly studied using discrete and continuous approaches. Various classes of emotions pertain to discrete emotions and continuous approach of studying emotions combine is combination of several psychological dimensions on different axis for emotion derivation (Shuman and Scherer, 2015). Speech emotion recognition generally identify emotions on the basis of categorical approach that depend on the usage common words of daily usage. Researchers have derived emotions from expressions of face, speech and various physiological signals. The facial expressions are great source of finding emotions (Ekman, 2005, 1993; Langner et al., 2010; Bourke et al., 2010; Van den Stock et al., 2007) since human face conveys emotions very aptly without uttering a word. Voice recording are potentially important for expressing speakers mental state and their intentions. Speech features can be studied as vectors for detection of emotion from a dataset (Banse and Scherer, 1996; Gulzar et al., 2014; Shrawankar and Thakare, 2013). The autonomous nervous system allows to asses emotion and thus physiological signals like ECG, RSP BP can be utilised to recognise emotions for people and can possibly contribute to cure mental illness (Shu et al., 2018).

2.2 Emotional speech databases

There is need of suitable databases to train the emotion recognition systems. Researchers suggest several existing databases aligned with the task of detecting and classifying the emotions. These databases can be categorised into three broad domains that cover acted emotions, natural emotions and felicitated emotions (Haamer et al., 2018). Out of the three mentioned domains enacted emotions are frequently supported by research since they are strong and reliable. EMODB is one of such highly used database for emotional classification. SAVEE is yet another enacted emotion database used for studying emotions. EMODB is a Berlin database for emotions while SAVEE is an English database specifying various emotions.

2.2.1 Emotion corpus

The task of training and testing for current work utilised two databases:

1 Berlin

2 SAVEE datasets.

Berlin database created in 1999 and consist of utterances spoken by various actors. EMODB has different number of spoken utterances for seven emotions (Lalitha et al., 2015). Emotions of database included anger, boredom, disgust, fear, happy, neutral and sad. The dataset contains more than 500 utterances spoken by 51 male and 60 female actors from age 21 to 35 years. The emotions labelled in EMODB are enlisted in Table 1.

Table 1 EMODB labels

Letter	*Emotion (German)*	*Emotion (English)*
W	Ärger (Wut)	Anger
L	Langeweile	Boredom
E	Ekel	Disgust
A	Angst	Fear/anxiety
F	Freude	Happiness
T	Trauer	Sadness
N	Neutral	

The increasing demand of research in speech analysis led to development of SAVEE (Jackson and Haq, 2014; Liu et al., 2018) database recordings for the helping the study of automatic emotion recognition system. The database contains recordings from four male actors in seven different emotions, 480 British English utterances in total. TIMIT corpus was used for sentence selection and contains phonetically-balanced emotions. The data were recorded in a visual media lab with efficient audio-visual equipment. The recordings were then processed and labelled. Ten subjects under audio, visual and audio-visual conditions evaluated quality of performance, of the recordings. The actors of the database utterances were four male speakers annotated as DC, JK, JE and KL. The speakers who contributed for recordings were postgraduate students and researchers at the University of Surrey. The age of speakers lie between 27 to 31 years. Seven discrete categories of emotions described as anger, disgust, fear, happiness, sadness and surprise were recorded (Ekman et al., 1987). The focused research was carried out on recognising

the discrete emotions (Zeng et al., 2009). Table 2 compares the features of both the datasets used in experimental analysis.

Table 2 Comparison of EMODB and SAVEE

Attributes	*EMODB*	*SAVEE*
No. of speakers	111	4
Age of speakers	21 to 35 years	27 to 31 years.
No. of utterances	500+	480
Language	German	British English
Emotions	Happy, angry, anxious, fearful, bored disgusted, neutral	Anger, disgust, fear, happiness, sadness, surprise, neutral

2.3 Feature extraction (Koduru et al., 2020; Kumar et al., 2011; Tiwari, 2010)

The core step for recognising speech or emotions from speech is extracting the features of speech. The process of feature extraction refers to identifying the components of the vocals from the audio signal. The audio signal is good source of linguistic information if the noise is discarded in the signal. There is another interpretation of feature extraction which says that it is characterisation and recognition of information specifically related to the actor's (speaker) mood, age and gender. The general process of feature extraction involves transformation of raw signal into feature vectors, which suppress the redundancies and emphasise on the speaker specific properties. The properties are like pitch, amplitude and frequency. The speaker dependencies such as health, voice tone, speech rate and acoustical noise variations may vary the speech signal during the testing and training sessions due to Gulzar et al. (2014), Dave (2013) and Yankayi (2016). The shape of the vocal tract filters the sounds generated by human beings which if determined efficiently can be used to derive phoneme representation of the speech sample with high accuracy.

The features to be extracted from speech can be studied under three categories named as high level features which may include phones, lexicon, accent, pronunciation; prosodic and spectra-temporal features that can be studied as pitch energy duration, rhythm, temporal features, and short-term spectral and prosodic features pertaining to spectrum glottal pulse (Ananthakrishnan and Narayanan, 2008; Kinnunen and Li, 2010; Wang et al., 2013). Short-term spectral features aid in better prediction with higher accuracies for various applications. The spectrogram analysis can be used for information extraction from the short-term spectral features. Linear predictive cepstral coefficients (LPCC), mel-frequency discrete wavelet coefficient (MFDWC) and MFCCs are most commonly used short-term spectral features for speech analysis (Gulzar et al., 2014; Dave, 2013; Lyons, 2014).

2.3.1 MFCC

MFCC are considered as the commonly used acoustic features for the task of identifying the speaker and the properties of the speech. MFCC takes into account human perception sensitivity with respect to frequencies. The combination of both is best for speech identification and differentiation. The importance of MFCC is inspired by the fact that the shape of vocal tract that includes tongue, teeth, throat, etc. filters the sound generated by

human speakers. The accuracy in determining the shape enables easy analysis of the sound that comes out through the vocal tract. The accurate in determination of the shape of sound can help in finding the phonetic information. The task of MFCC is to accurately represent envelop of the short time power spectrum of the sound when it traverse through the vocal tract (Tiwari, 2010; Lyons, 2014; Palo et al., 2018).

MFCCs were identified as a feature and are widely applicable to ASR and identification of speaker. The correlation among the actual and heard signal frequencies can be derived efficiently by incorporating mel scale. Davis and Mermelstein were pioneer in identifying MFCC as sound feature in the 1980s. MFCC ever since its discovery has been considered important feature for analysis of speech signals. There are few other features along with MFCC, like linear prediction coefficients (LPCs) and LPCCs that were coined before MFCC and remained the main features for ASR, especially with classification algorithm such as HMM (Yazici et al., 2018). In practice 8 to 12 or 13, MFCC are considered for representing the shape of spectrum and hence are used for speech analysis (Wang et al., 2002). MFCC are highly preferred choices in ASR systems (Davis and Mermelstein, 1990). Palaz et al. (2019) and Passricha and Aggarwal (2020) found that MFCC is effective for end to end acoustic modelling using CNN. MFCC is widely used feature while considering speech modelling (Vimala and Radha, 2014; Dalmiya et al., 2013). MFCC-based comparative study of speech recognition techniques was conducted by authors who found that MFCC with HMM gave recognition accuracy of 85% and with deep neural networks the score was 82.2% (NithyaKalyani and Jothilakshmi, 2019). Computation of MFCCs includes a conversion of the Fourier coefficients to mel scale (Stevens et al., 1937). Mel scale is the most popular variant used today, even if there is no theoretical reason that the mel scale is superior to the other scales (Mitrović et al., 2010).

2.4 Decision tree classifiers

There are several machine learning algorithms that can be applied for recognising speech emotions. The algorithms can be used independently or in hybrid mode for classifying emotions. Decision tree are one of the machine learning algorithms that can be used for classification task (Caruana and Niculescu-Mizil, 2006; Kotsiantis, 2007). The decision tree uses the supervised learning approach that works on labelled data. The data is split into train and test subsets for carrying out the classification task. The current work uses random forest, KNN and extreme gradient boosting (XGBoost) algorithms for classifying emotions. All the mentioned algorithms are the variations of decision tree classifiers and a brief description of each classifier is given below (Özseven et al., 2018).

2.4.1 Random forest

One of the most flexible and easily implementable learning algorithms in machine learning is random forest. The algorithm provides better solutions over basic decision trees. The random forest depends on few parameters which if tuned can provide good results. The algorithm is widely used due to its simple and flexible aspect of implementation. Random forest supports both regression and classification task while modelling a solution. It is supervised learning technique that creates random forests. The ensemble decision trees are referred to forest and mostly use bagging for training (Davis and Mermelstein, 1990; Palaz et al., 2019). The importance of regressive bagging lies in

the fact that it increases the overall results. Multiple decision trees are build and together to increase efficiency in random forest algorithm.

Random forest generation uses same hyper parameters that are used for decision tree or a bagging classifier. The class of classifier does not require combining the decision trees to bagging classification algorithm. The algorithm proceeds by searching for the best feature from available features subset. The selected feature will then be used for splitting the node. The node split diversifies and enhances the results. The relative importance of each feature is measured while prediction. SK-learn tool can be used to measure a feature importance. The tool reduces impurity at the tree nodes that use the feature, across all trees in forest. Score for each feature is automatically computed after training. Features and observations are randomly selected by random forest and averaged for building several decision trees. The decision uses rules and facts for decision making and trees from over fitting. Random forest prevents it by creating subsets and combining them to subtrees. The only limitation of random forest is slow computation which is affected by number of trees build by random forest (Luckner et al., 2017).

2.4.2 XGBOOST (Sutton, 2012)

XGBoost uses gradient boosting technique to ensemble decision trees. XGBoost is stands for 'extreme gradient boosting'. Small, medium structured and tabular data uses XGBoost for classification. XGBoost is studied as improvisation upon the base GBM framework. Optimisation and algorithmic techniques are used to improve the base framework of GBM. Regularisation is used to enhance the performance of algorithm by preventing data overfitting. The algorithm automatically learns best missing values depending on the training loss and handle variety of patterns of sparsity more efficiently. It also has built-in cross-validation method at each iteration.

XGboost is sequential tree building algorithm implemented by parallelisation. The interchangeable nature of the loops determines the base of building algorithm.

The external loop is responsible for maintaining the tree count, and features are calculated by the internal loop. Loops are interchangeable and thus enhance run time performance. All the instances are globally scanned, initialised and sorting is done using parallel threads. This switch of loops increases the algorithmic performance. The parallelisation overheads in computation are offset. The tree splitting within GBM framework for stopping the split is greedy in nature. Splitting of tree at node depends on the negative loss criterion at the point of split. XGBoost uses 'max_depth' parameter as specified instead of criterion first, and pruning of the trees is done backward. The computational performance is improved significantly by using this 'depth-first' approach (Luckner et al., 2017).

2.4.3 KNN

The early description of KNN was found in 1950. KNN is labour intensive approach for large datasets. It was used for pattern recognition initially. The learning of KNN is based on the comparison test data with train data such that both have similarities. A set of N attributes describe the tuple data. An n-dimensional space is used to store all the training tuples where each of them corresponds to a point in space. The pattern space for k training tuples that are closest to unknown tuple is identified by the K-nearest neighbour

classifier. The closest found points are referred to nearest neighbours and Euclidian distance defines the nearness of the neighbouring clones (Sutton, 2012).

The Euclidean distance between two Co tuples represented by $A_1 = \{a_{11}, a_{12}, \ldots, a_{1n}\}$, $A_2 = \{a_{21}, a_{22}, \ldots, a_{2n}\}$ is obtained using following calculation:

$$distance(A_1, A_2) = \sqrt{\sum_{i=1}^{n} (a_{1i} - a_{2i})^2} \tag{1}$$

Thus, the difference of values of attribute in A_1 and A_2 is obtained. The difference is then squared to accumulate total distance count. Attributes with large ranges can outweighs attributes within small ranges (binary attributes). Normalisation is applied to each attribute value to resolve the issue. Min max normalisation is one of the normalisation technique that transforms the value y of a numeric attribute A to V_0 in the range [0, 1] by computing

$$u' = (u - \min_A) / (\max_A - \min_A) \tag{2}$$

where values defined for attribute A are $\min_A$ and $\max_A$. Most common class is assigned to unknown tuple among its K-nearest neighbours. If the value of K equates 1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space. Real value prediction is returned by KNN for unknown value tuples. The unknown values the classifier of KNN returns the average of the real valued labels associated with K-nearest neighbours of unknown tuple (Sutton, 2012; Deng et al., 2016; Okfalisa et al., 2018).

3 Problem formulation

Literature quotes MFCC as important speech feature for analysing and classifying various aspects of speech. Some of them quote that only 13 MFCC are sufficient features to be considered for experimentation. There is no experimental validation for this statement of considering only 13 MFCC. Moreover, there is no sufficient research on the identification of these 12 MFCC from the extracted 20 base features of mel scale. MFCC also have derivatives of base features named as delta and double delta. The aim of the work is to establish experimental proof of considering only 8 to 13 MFCC from extracted 39 features of MFCC (Cen et al., 2016; Wang et al., 2002). The current work does experimental analysis on MFCC obtained from EMODB, a Berlin database that consisted of more than 500 utterances, the utterances were from 111 speakers that included both male and female speakers from various age groups.

4 Experiments and results

Following simulations and experiments were conducted. The flow graph in Figure 2 shows the steps carried out for the experiment.

The experiment was done on MFCC extracted from the EMODB and SAVEE dataset. Four subsets of MFCC features comprising 20 cepstral constants were analysed for feature importance. This was done to identify which 13 MFCC should be used for speech

emotion analysis. The result of each subset using different classifiers areas shown in Tables 4 and 5.

Figure 2 Flow graph of the emotion classification using decision tree classifiers (see online version for colours)

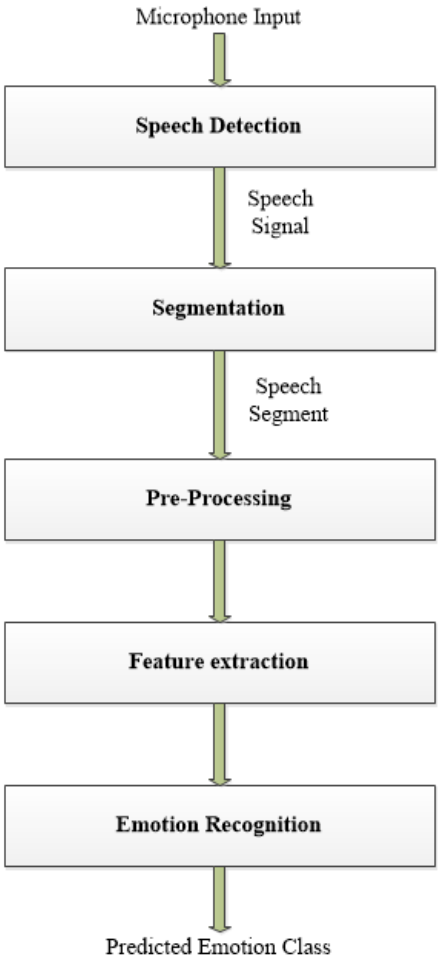


Table 3 Label encoded emotions for EMODB

Emotion EMODB	*Emotion SAVEE*	*Encoded label*
Fear/anxiety	Anger	0
Disgust	Disgust	1
Happiness	Fear	2
Boredom	Happiness	3
Neutral	Sadness	4
Sadness	Surprise	5
Anger	Neutral	6

Table 4 Results of EMODB with subsets of MFCC

MFCC	*M0–M19*	*M0–M12*	*M15–M17*	*M6–M19*
RF	52%	50%	47%	41%
KNN	56%	44%	45%	42%
XGB	40%	38%	28%	27%

Table 5 Results of SAVEE with subsets of MFCC

MFCC	*M0–M19*	*M0–M12*	*M15–M17*	*M6–M19*
RF	57%	55%	50%	50%
KNN	67%	64%	55%	50%
XGB	30%	38%	30%	30%

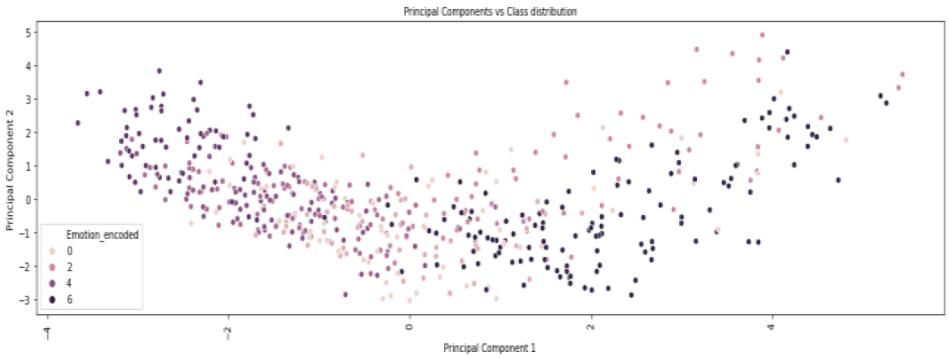
Each subset was used to classify emotions using supervised learning algorithm (variants of decision trees). It was observed that the results obtained using 20 MFCC over set of first 13 was very near to each other. There was no effective and substantial difference in the accuracy scores for classification while using 13 and 20 MFCC subsets. Increased number of features often increases the complexity of the system and so if 13 MFCC are used instead of 19 MFCC, the results will not suffer much loss. The validation of the experiment was also done by extracting important features from PCA analysis. It was seen that most of the important features corresponded to initial 13 MFCC extracted from dataset. The accuracy score of classification on EMODB using M0–M12 was 52%, 50%, 47%, 41% using subsets M0–M19, M10–M12, M15–M17 and M16–M19, respectively using random forest classifier. KNN shows 56%, 44%, 45% and 42% accuracy score using subsets M0–M19, M10–M12, M15–M17 and M16–M19, respectively. XGB showed poor performance on the original extracted dataset for classification task. SAVEE results as shown in Table 4 depicts accuracy sores of RF using subsets M0–M19, M10–M12, M15–M17 and M16–M19 are 57%, 55%, 50% and 50%, respectively. For KNN, the performance of four subsets is 67%, 54%, 55% and 50%. XGB showed poor performance on all the subsets. The results obtained in Table 4 and Table 5 clearly indicates that selecting M0–M12 would be a better choice for features from MFCC dataset. It can be seen from the results that first 13 mel coefficients can successfully be used for playing with speech over using 20 features. This selection shall only optimise the results but also reduce the complexity of the model thereby reducing the computation time.

Thereafter, the dataset was minimised and re-experimented for feature importance and classification. The accuracy of results for classification increased effectively but the set of important features still contained features M0 to M12 that initial 13 features. The reason behind this is that as the sound signal passes through the vocal tract and comes out as the utterance there is a subsequent addition of noise to the originally generated signal. Addition of noise disturbs the energy whose log is computed as the base of MFF extraction. The induction of noise imputes the signal at later levels so the original signal remains intact for usage in analysis. The features present here can be successfully used for better results as compared those extracted towards the end of the sample of each speech signal.

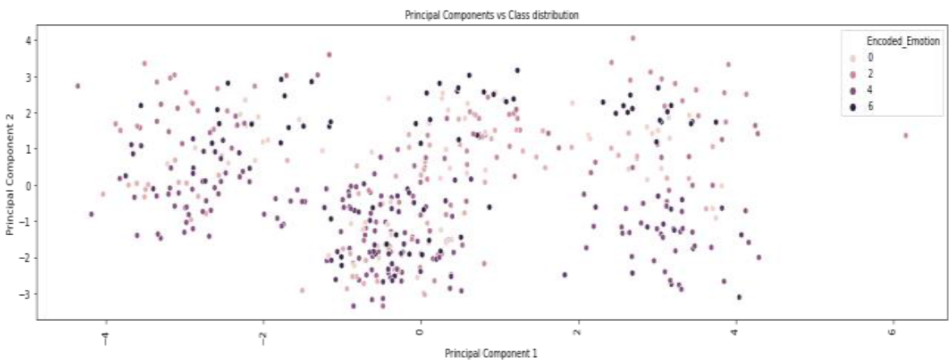
The results in Table 4 and Table 5 show that M0–M19 and M0–M13 has nearly similar results for accuracy on classifiers. The datasets in Table 3 and Table 4 used the non-manipulated MFCC extracted from the speech utterances in EMODB and SAVEE.

For further analysis, the datasets were minimised and preprocessed using min max scaling. The important features identified for the minimised data using principle component analysis are in Figure 3 and Figure 4. The *x* axis of the plot shows various classes of emotion and *y* axis plot shows the MFCC features.

Figure 3 Principle component vs. class distribution for (a) EMODB and (b) SAVEE (see online version for colours)



(a)



(b)

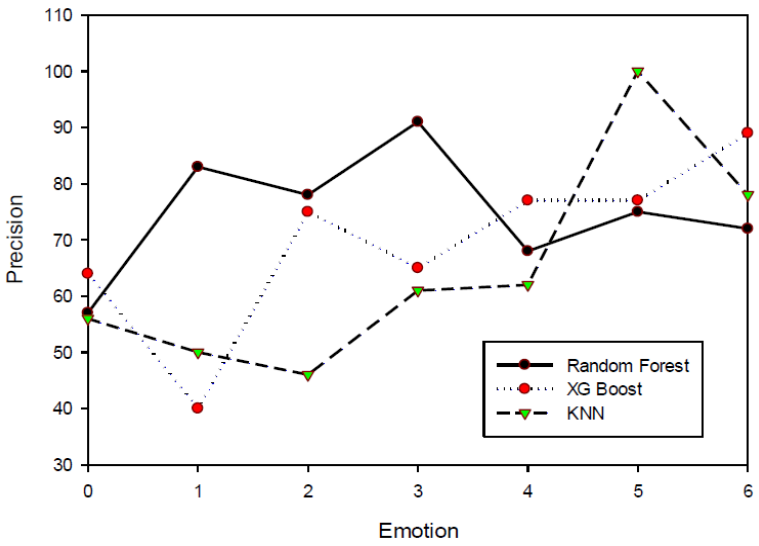
The minimum number of MFCC for each methodology was 13 that belonged to first subset of MFCC features. For EMODB M0–M13 were significantly important and for SAVEE, M0–M8 were identified as most significant features.

The results of using 13 MFCC from minimised datasets EMODB and SAVEE are shown in Figures 4(a), 4(b) and 4(c). SAVEE results can be seen in Figures 6(a), 6(b) and 6(c). The plots in Figures 4(a), 4(b), 4(c), 6(a), 6(b) and 6(c) clearly displays the precision, recall and F1-scores for emotions in both the datasets using variants of decision trees.

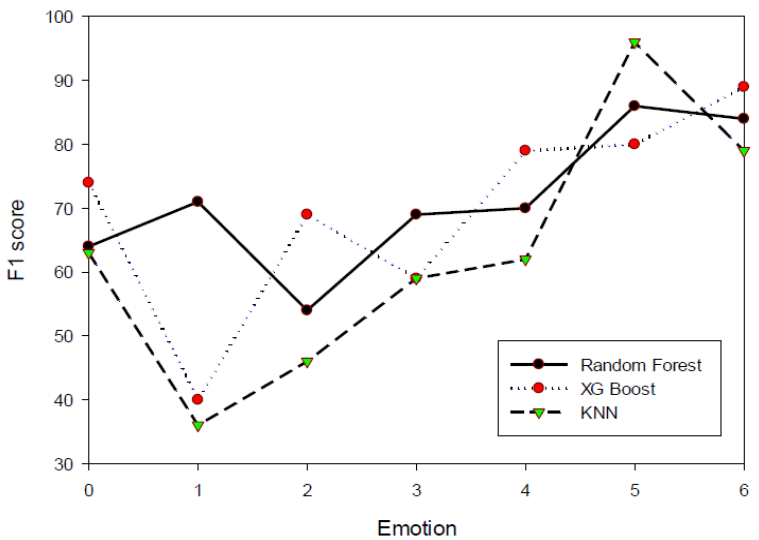
Results showed that for EMODB all the three classifiers defined fairly variable results. Boredom (91%), anger (74%) and sadness (100%) had highest precision for random forest XGB and KNN respectively in EMODB. For SAVEE, a higher precision

rate for disgust (84%) was identified using random forest. KNN and XGBoost identified sadness more precisely over remaining six emotions where the scores for sadness were 84%, 70% and 74% with KNN, random forest and XGBoost, respectively. A common conclusion was obtained from the results of both the datasets that sadness was commonly identified with highest precision using KNN. So KNN can be effective for studying the emotion sadness in emotion analysis.

Figure 4 Emotion vs. (a) precision score, (b) F1-score and (c) recall score for first 13 MFCC using KNN, random forest and XGB classifiers on EMODB (see online version for colours)

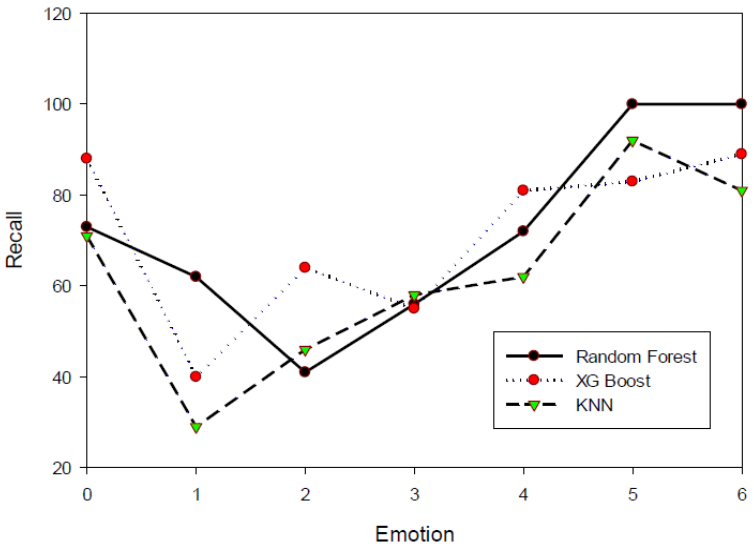


(a)



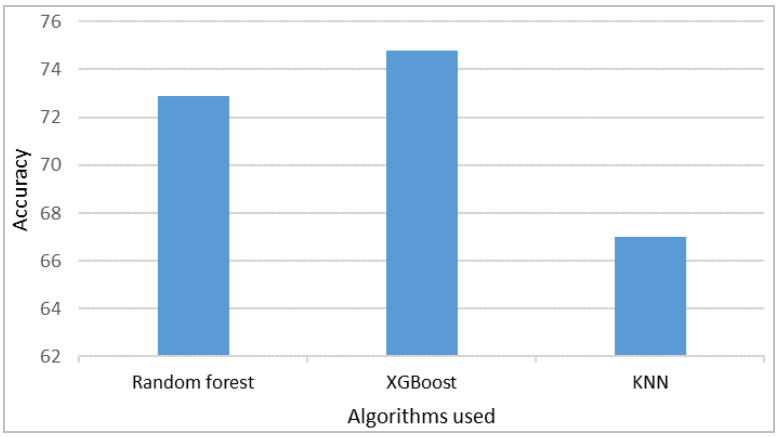
(b)

Figure 4 Emotion vs. (a) precision score, (b) F1-score and (c) recall score for first 13 MFCC using KNN, random forest and XGB classifiers on EMODB (continued) (see online version for colours)



(c)

Figure 5 Accuracy score for emotion classification using KNN, random forest and XGB classifiers on EMODB (see online version for colours)

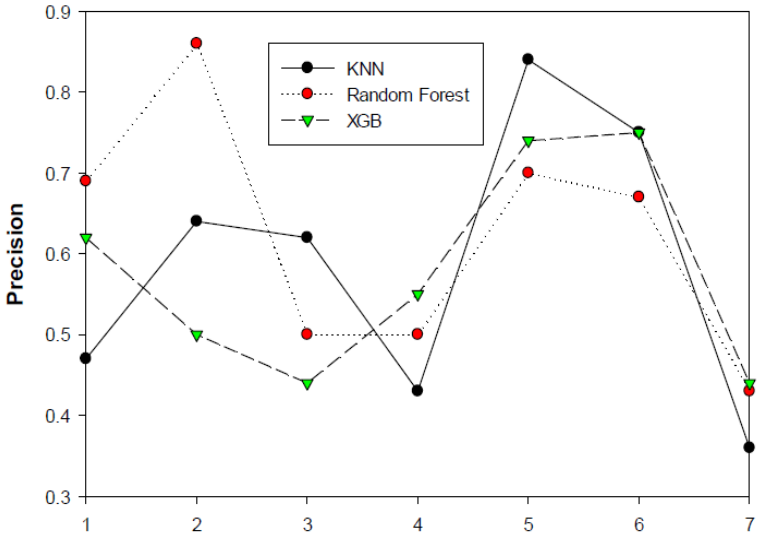


4.1 Effect of noise on emotions

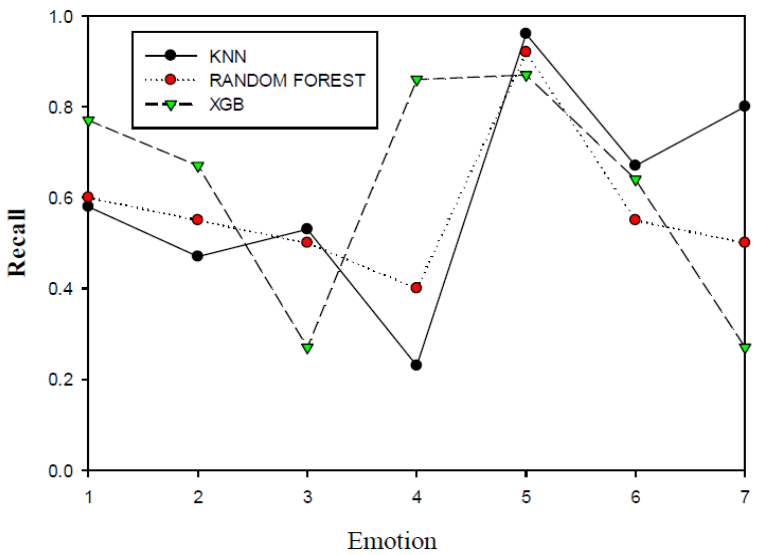
To investigate the effect of noise on several emotions the noise of +2 dB and –5 dB was introduced in the voice samples for both datasets. It was seen that the presence of noise detrimentally effected the emotion prediction as compared to clean samples. A higher slope of MFCC identified fear, happiness and anger emotions more correctly. Anger emotion was less influenced by mild noise in comparison with fear and joy. Neutral and

sadness emotion showed deterioration in accuracy for SNR –5 dB. It was observed that on introducing noise, very few utterances were classified correctly for emotions as compared to clean conditions. There were large differences in accuracies for emotion prediction. The results for the effect of noise addition are shown in Figure 9.

Figure 6 (a) Emotion vs. precision score for first 13 MFCC using KNN, random forest and XGB on SAVEE (b) Emotion vs. recall score for first 13 MFCC using KNN, random forest and XGB on SAVEE (c) Emotion vs. F1 score for first 13 MFCC KNN, random forest and XGB on SAVEE (d) Emotion vs. recall score for 13 MFCC using three decision tree classifiers on EMODB (see online version for colours)

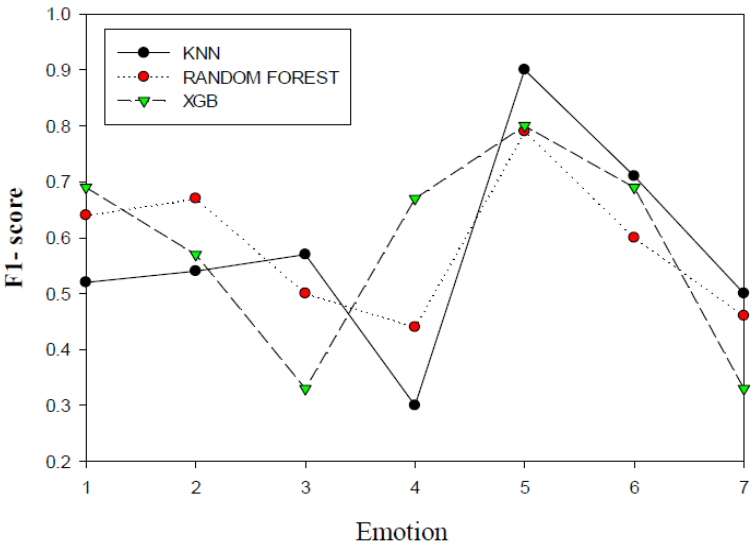


(a)

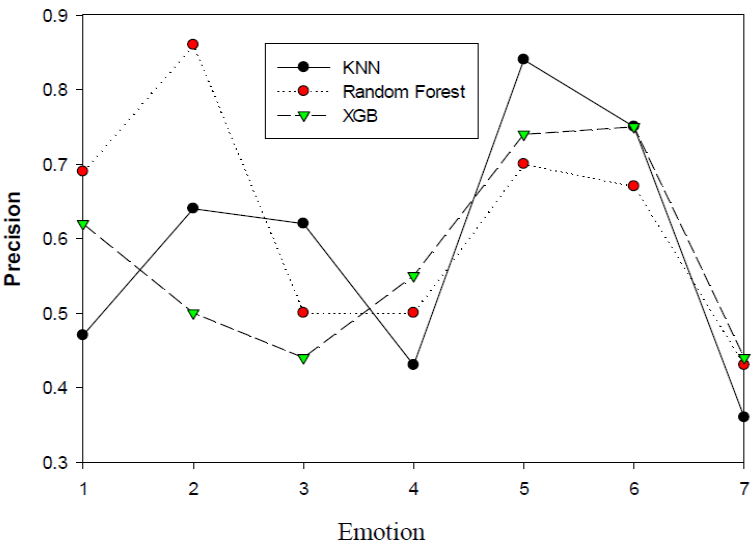


(b)

Figure 6 (a) Emotion vs. precision score for first 13 MFCC using KNN, random forest and XGB on SAVEE (b) Emotion vs. recall score for first 13 MFCC using KNN, random forest and XGB on SAVEE (c) Emotion vs. F1 score for first 13 MFCC KNN, random forest and XGB on SAVEE (d) Emotion vs. recall score for 13 MFCC using three decision tree classifiers on EMODB (continued) (see online version for colours)



(c)



(d)

Figure 7 Accuracy score obtained with first 13 MFCC using KNN, random forest and XGB (see online version for colours)

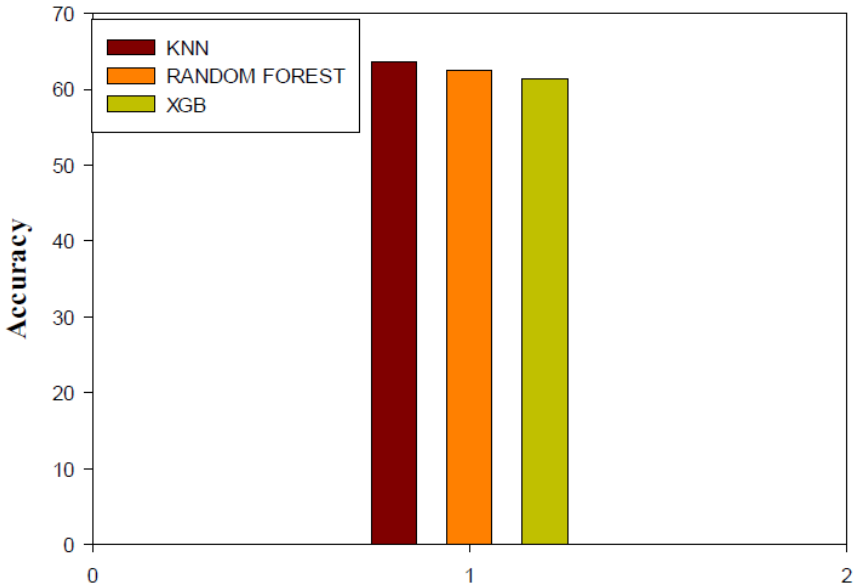
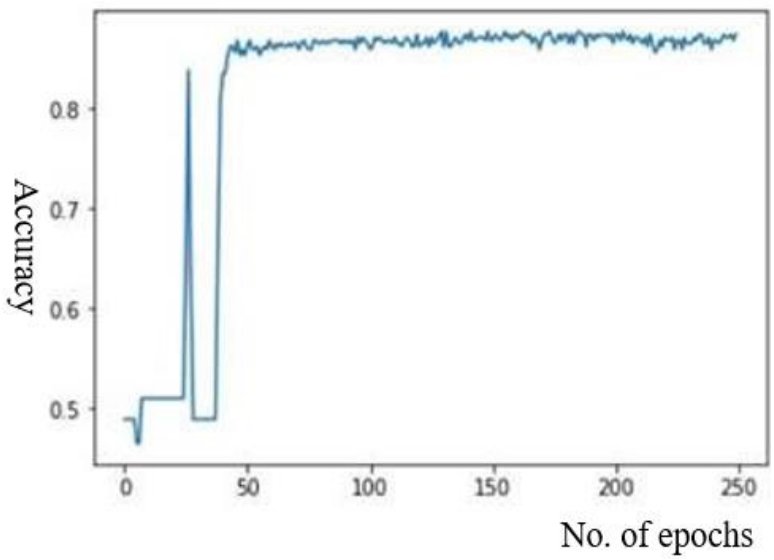


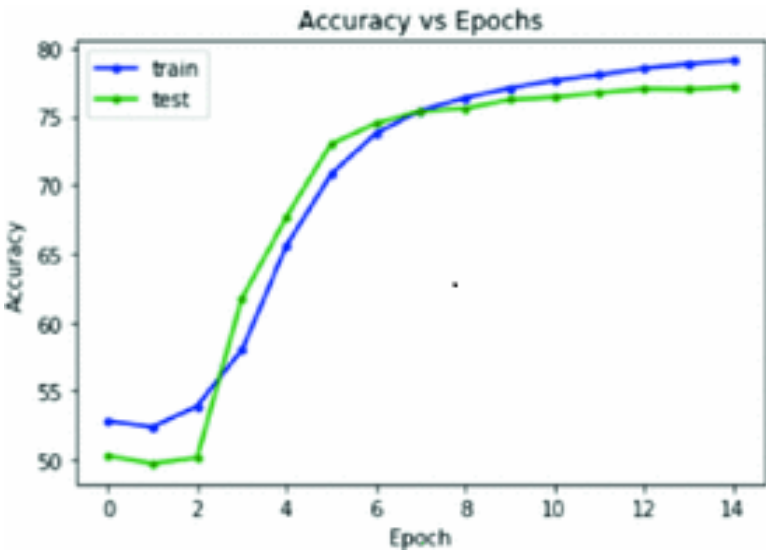
Figure 8 Results of deep learning (see online version for colours)



(a)

Note: (a) The plot shows accuracy score for deep learning using ANN on EMODB and (b) the plot shows accuracy score for deep learning using ANN on SAVEE.

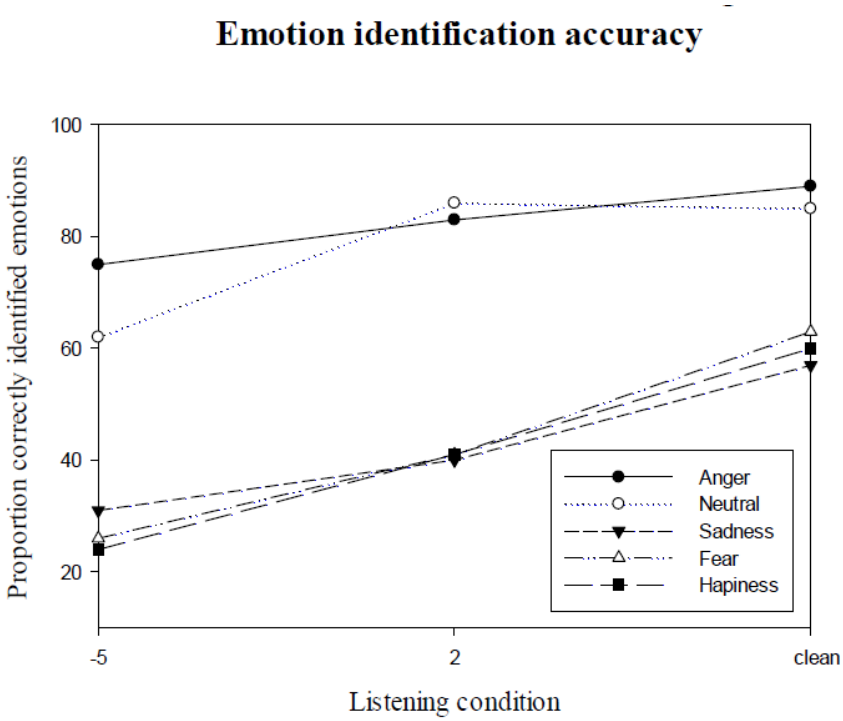
Figure 8 Results of deep learning (continued) (see online version for colours)



(b)

Note: (a) The plot shows accuracy score for deep learning using ANN on EMODB and (b) the plot shows accuracy score for deep learning using ANN on SAVEE.

Figure 9 Effect of noises of emotions



5 Conclusions

Speech features have always remained the one of the regressively studied topic in research. Speech or utterances contain vital information regarding the intention, emotion and psychology of the speaker. The paper studied the use of one such speech feature called MFCC and utilised it to classify emotions using two datasets. The work also tried to establish the importance of using first 13 MFCC when we have a set of 20 mel constants that can be extracted for speech based on the vocal physiology of human mouth. Accuracy scores for emotion classification using variants of decision tree approach have been obtained for EMODB and SAVEE for two datasets. KNN was identified as the common classification algorithm for both datasets. The score of sadness as obtained from KNN were highest for both the datasets. The results of the experiments can be utilised for predicting emotions and personality of the speaker. The results can be integrated with various applications pertaining to human psychology and medical treatments. Further along with MFCC, another acoustic feature like mean, intensity, and energy can be used to predict emotions under the effect of several noises. The speaker identification can also be added to the work as a future research perspective.

Acknowledgements

The contributions of individual authors to the creation of a scientific article (ghostwriting policy) are as follows: Kalpana Rangra carried out the overall simulation and the optimisation, and organised and executed the experiments. Monit Kapoor is responsible for responsible for data pre-processing, article preparation and revision.

References

Albornoz, E.M., Milone, D.H. and Rufiner, H.L. (2011) 'Spoken emotion recognition using hierarchical classifiers', *Computer Speech and Language*, Vol. 25, No. 3, pp.556–570 [online] https://doi.org/10.1016/j.csl.2010.10.001.

Ananthakrishnan, S. and Narayanan, S.S. (2008) 'Automatic prosodic event detection using acoustic, lexical, and syntactic evidence', *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16, No. 1, pp.216–228 [online] https://doi.org/10.1109/tasl.2007.907570.

Banse, R. and Scherer, K.R. (1996) 'Acoustic profiles in vocal emotion expression', *Journal of Personality and Social Psychology*, Vol. 70, No. 3, pp.614–636.

Bourke, C., Douglas, K. and Porter, R. (2010) 'Processing of facial emotion expression in major depression: a review', *Australian and New Zealand Journal of Psychiatry*, Vol. 44, No. 8, pp.681–696.

Caruana, R. and Niculescu-Mizil, A. (2006) 'An empirical comparison of supervised learning algorithms', in *ACM International Conference Proceeding Series*.

Cen, L., Wu, F., Yu, Z.L. and Hu, F. (2016) 'A real-time speech emotion recognition system and its application in online learning', in *Emotions, Technology, Design, and Learning*, pp.27–46.

Dalmiya, C.P., Dharun, V.S. and Rajesh, K.P. (2013) 'An efficient method for Tamil speech recognition using MFCC and DTW for mobile applications', in *2013 IEEE Conference on Information and Communication Technologies, ICT 2013*.

Dave, N. (2013) 'Feature extraction methods LPC, PLP and MFCC in speech recognition', *International Journal for Advance Research in Engineering and Technology*, Vol. 1, No. 6.

Davis, S.B. and Mermelstein, P. (1990) 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', in *Readings in Speech Recognition*, pp.65–74.

Deng, Z., Zhu, X., Cheng, D., Zong, M. and Zhang, S. (2016) 'Efficient kNN classification algorithm for big data', *Neurocomputing*, Vol. 195, pp.143–148.

Ekman, P. (1993) 'Facial expression and emotion', *American Psychologist*, Vol. 48, No. 8, pp.384–392.

Ekman, P. (2005) 'Basic emotions', in *Handbook of Cognition and Emotion*, pp.45–60.

Ekman, P. et al. (1987) 'Universals and cultural differences in the judgments of facial expressions of emotion', *Journal of Personality and Social Psychology*, Vol. 53, No. 4, pp.712–717.

Fernández-Caballero, A. et al. (2016) 'Smart environment architecture for emotion detection and regulation', *Journal of Biomedical Informatics*, Vol. 64, pp.55–73.

Guan, H., Liu, Z., Wang, L., Dang, J. and Yu, R. (2018) 'Speech emotion recognition considering local dynamic features', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Gulzar, T., Singh, A. and Sharma, S. (2014) 'Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks', *International Journal of Computer Applications*, Vol. 101, No. 12, pp.22–27.

Haamer, R.E., Rusadze, E., Lüsi, I., Ahmed, T., Escalera, S. and Anbarjafari, G. (2018) 'Review on emotion recognition databases', *Human-Robot Interaction – Theory and Application*, Vol. 3, pp.39–63.

Hozjan, V. and Kačič, Z. (2003) 'Context-independent multilingual emotion recognition from speech signals', *International Journal of Speech Technology*, Vol. 6, No. 3, pp.311–320.

Jackson, P. and Haq, S. (2014) *Surrey Audio-visual Expressed Emotion (SAVEE) Database*, University of Surrey, Guildford, UK.

Jing, S., Mao, X. and Chen, L. (2018) 'Prominence features: effective emotional features for speech emotion recognition', *Digital Signal Processing: A Review Journal*, October, Vol. 72, pp.216–231.

Kinnunen, T. and Li, H. (2010) 'An overview of text-independent speaker recognition: from features to supervectors', *Speech Communication*, Vol. 52, No. 1, pp.12–40.

Kishore, K.V.K. and Satish, P.K. (2013) 'Emotion recognition in speech using MFCC and wavelet features', in *Proceedings of the 2013 3rd IEEE International Advance Computing Conference, IACC 2013*.

Koduru, A., Valiveti, H.B. and Budati, A.K. (2020) 'Feature extraction algorithms to improve the speech emotion recognition rate', *International Journal of Speech Technology*, Vol. 23, No. 1, pp.45–55.

Koolagudi, S.G., Barthwal, A., Devliyal, S. and Rao, K.S. (2012) 'Real life emotion classification from speech using Gaussian mixture models', in *International Conference on Contemporary Computing*, pp.250–261, Springer, Berlin, Heidelberg.

Koolagudi, S.G.K. and Rao, K.S. (2012) 'Emotion recognition from speech: a review', *International Journal of Speech Technology*, Vol. 15, No. 2, pp.99–117.

Kotsiantis, S.B. (2007) 'Supervised machine learning: a review of classification techniques', *Informatica (Ljubljana)*, Vol. 160, No. 1, pp.3–24.

Kumar, K., Kim, C. and Stern, R.M. (2011) 'Delta-spectral cepstral coefficients for robust speech recognition', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*.

Lalitha, S., Geyasruti, D., Narayanan, R. and Shravani, M. (2015) 'Emotion detection using MFCC and Cepstrum features', *Procedia Computer Science*, Vol. 70, pp.29–35.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T. and van Knippenberg, A. (2010) 'Presentation and validation of the radboud faces database', *Cognition and Emotion*, Vol. 24, No. 8, pp.1377–1388.

Latif, S., Rana, R., Younis, S., Qadir, J. and Epps, J. (2018) 'Transfer learning for improving speech emotion classification accuracy', *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, January, Vol. 2018, pp.257–261.

Lee, C.M. and Narayanan, S.S. (2005) 'Toward detecting emotions in spoken dialogs', *IEEE Transactions on Speech and Audio Processing*.

Liu, Z.T., Xie, Q., Wu, M., Cao, W.H., Mei, Y. and Mao, J.W. (2018) 'Speech emotion recognition based on an improved brain emotion learning model', *Neurocomputing*, Vol. 309, pp.145–156.

Luckner, M., Topolski, B. and Mazurek, M. (2017) 'Application of XGboost algorithm in fingerprinting localisation task', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Lyons, J. (2014) 'Mel frequency cepstral coefficient', *Practical Cryptography*.

Marechal, C. et al. (2019) 'Survey on AI-based multimodal methods for emotion detection', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Mitrović, D., Zeppelzauer, M. and Breiteneder, C. (2010) 'Features for content-based audio retrieval', in *Advances in Computers*, Vol. 78, pp.71–150, Elsevier.

NithyaKalyani, A. and Jothilakshmi, S. (2019) 'Speech summarization for Tamil language', in *Intelligent Speech Signal Processing*, pp.113–138, Academic Press.

Okfalisa, Gazalba, I., Mustakim and Reza, N.G.I. (2018) 'Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification', in *Proceedings – 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*.

Oosterwijk, S., Lindquist, K.A., Anderson, E., Dautoff, R., Moriguchi, Y. and Barrett, L.F. (2012) 'States of mind: emotions, body feelings, and thoughts share distributed neural networks', *NeuroImage*, Vol. 62, No. 3, pp.2110–2128.

Özseven, A., Düğenci, T. and Durmuşoğlu, M. (2018) 'A content analysis of the research approaches in speech emotion', *International Journal of Engineering Sciences & Research Technology*, Vol. 7, No. 1, pp.1–26.

Palaz, D., Magimai-Doss, M. and Collobert, R. (2019) 'End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition', *Speech Communication*, Vol. 108, pp.15–32.

Palo, H.K., Chandra, M. and Mohanty, M.N. (2018) 'Recognition of human speech emotion using variants of mel-frequency cepstral coefficients', *Lecture Notes in Electrical Engineering*, Vol. 442, pp.491–498.

Passricha, V. and Aggarwal, R.K. (2020) 'A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 11, No. 2, pp.675–691.

Pessoa, L. (2011) 'Emotion and cognition and the amygdala: from "what is it?" to "what's to be done?"', *Neuropsychologia*, Vol. 48, No. 12, pp.3416–3429.

Ramakrishnan, S. (2012) 'Recognition of emotion from speech: a review', in *Speech Enhancement, Modeling and Recognition – Algorithms and Applications*, Vol. 7, pp.121–137.

Rao, K.S., Koolagudi, S.G. and Vempada, R.R. (2013) 'Emotion recognition from speech using global and local prosodic features', *International Journal of Speech Technology*, Vol. 16, No. 2, pp.143–160.

Sebe, N., Cohen, I. and Huang, T.S. (2005) 'Multimodal emotion recognition', in *Handbook of Pattern Recognition and Computer Vision*, 3rd ed.

Shrawankar, U. and Thakare, V.M. (2013) *Techniques for Feature Extraction in Speech Recognition System: A Comparative Study*, arXiv preprint arXiv:1305.1145.

Shu, L. et al. (2018) 'A review of emotion recognition using physiological signals', *Sensors*, Vol. 18, No. 7, p.2074, Switzerland.

Shuman, V. and Scherer, K.R. (2015) 'Emotions, psychological structure of', in *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., ScienceDirect.

Stevens, S.S., Volkmann, J. and Newman, E.B. (1937) 'A scale for the measurement of the psychological magnitude pitch', *Journal of the Acoustical Society of America*, Vol. 8, No. 3, pp.185.

Sutton, O. (2012) *Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction*, University Lectures, University of Leicester.

Tiwari, V. (2010) 'MFCC and its applications in speaker recognition', *International Journal on Emerging Technologies*, Vol. 1, No. 1, pp.19–22.

Van den Stock, J., Righart, R. and de Gelder, B. (2007) 'Body expressions influence recognition of emotions in the face and voice', *Emotion*, Vol. 7, No. 3, p.487.

Vimala, C. and Radha, V. (2014) 'Suitable feature extraction and speech recognition technique for isolated Tamil spoken words', *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 1, pp.378–383.

Wang, W.Y., Biadsy, F., Rosenberg, A. and Hirschberg, J. (2013) 'Automatic detection of speaker state: lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification', *Computer Speech and Language*, Vol. 27, pp.168–189.

Wang, X., Dong, Y., Hakkinen, J. and Viikki, O. (2002) 'Noise robust Chinese speech recognition using feature vector normalization and higher-order cepstral coefficients', *WCC 2000-ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, IEEE, Vol. 2, pp.738–741.

Winkielman, P., Niedenthal, P., Wielgosz, J., Eelen, J. and Kavanagh, L.C. (2014) 'Embodiment of cognition and emotion', in *APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition*, pp.151–175.

Yankayi, M. (2016) *Feature Extraction Mel Frequency Cepstral Coefficients (MFCC)*, pp.1–6, Juniper Publisher.

Yazici, M., Basurra, S. and Gaber, M. (2018) 'Edge machine learning: enabling smart internet of things applications', *Big Data and Cognitive Computing*, Vol. 2, No. 3, p.26.

Yousefpour, A., Ibrahim, R. and Hamed, H.N.A. (2017) 'Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis', *Expert Systems with Applications*, Vol. 75, pp.80–93.

Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S. (2009) 'A survey of affect recognition methods: audio, visual, and spontaneous expressions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 1, pp.39–58.

Zhang, Q., Wang, Y., Wang, L. and Wang, G. (2007) 'Research on speech emotion recognition in e-learning by using neural networks method', in *2007 IEEE International Conference on Control and Automation, ICCA*.