Analysis of the Mel scale features using classification of big data and speech signals

Volodymyr Osadchyy

dept. of computer science, UCF
Orlando, Florida, USA

Ruslan V. Skuratovskii

Interregional Academy of Personnel Management, Kyiv, Ukraine and Institute for applied system analysis,
Kiev, 03056, Ukraine

Aled Williams,

Cardiff University,
Cardiff, UK

Abstract: The role of human speech is intensified by the emotion it conveys. The parameterization of the vector obtained from the sentence divided into the containing emotional-informational part and the informational part is effectively applied. There are several characteristics and features of speech that differentiate it among utterances, i.e. various prosodic features like pitch, timbre, loudness and vocal tone which categorize speech into several emotions. They were supplemented by us with a new classification feature of speech, which consists in dividing a sentence into an emotionally loaded part of the sentence and a part that carries only informational load. Therefore, the sample speech is changed when it is subjected to various emotional environments. As the identification of the speaker's emotional states can be done based on the Mel scale, MFCC is one such variant to study the emotional aspects of a speaker's utterances. In this work, we implement a model to identify several emotional states from MFCC for two datasets, classify emotions for them on the basis of MFCC features and give the comparison of both. Overall, this work implements the classification model based on dataset minimization that is done by taking the mean of features for the improvement of the classification accuracy rate in different machine learning algorithms.

***Keywords*: Machine Learning; Speech Recognition; Emotion recognition; MFCC; supervised learning; decision trees.**

1. Introduction

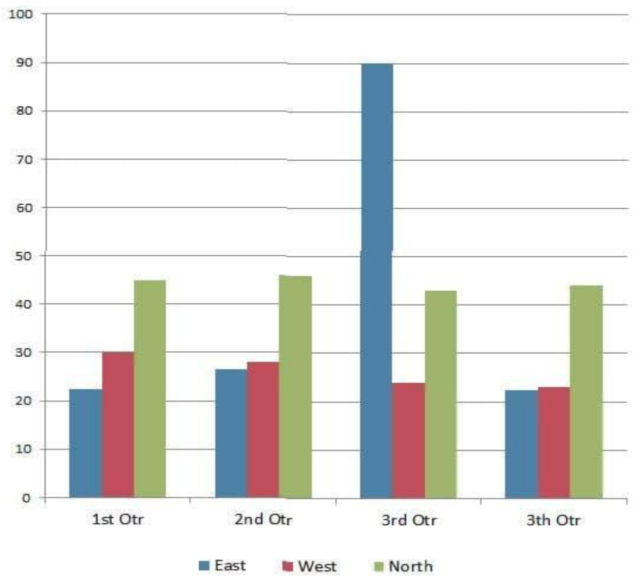
To obtain emotion score differences we use Miller function and scale. The vocal acoustics are full of emotional cues to analyze the speaker's emotional state. Since the expression of emotions occurs most often either at the beginning or at the end of a sentence, a sentence was divided by us into two parts. To the first part we refer a beginning and an end of a sentence, referred to the emotional content expressing by an author; to the second part we refer a middle of a sentence containing only the informational and narrative part. It serves to vector parameterization in KNN for speech emotions. Each emotion is associated with tone of the speaker. Emotions can be recognized both from text and sound. Each of them has different approaches to identify the emotional state of the speaker. Emotions also greatly define the interpersonal relations by affecting intelligent and rational decision-making. The emotions are a communication bridge among the speaker and the listener. An interaction between individuals is way more clear and effective when the emotions are used in utterances. They play a pivotal role in engaging of a human being in a group discussion and can tell a lot about the one's mental state [1]. The information hidden in emotions ignited the process of the evolution of the speech recognition field commonly referred as automatic speech recognition. Several models of retrieval and interpretation of emotions from images of speaker's face and the recordings of his expressions, voice and tone during a conversation has been proposed by researchers. The utilization of physiological signals in the same manner has also been discussed [2]. The significance of emotions in communication can hardly be overestimated since they express the speaker's intentions to his listeners. There are several spoken language interfaces available today that support automatic speech recognition. By collecting the samples, such systems are providing a base for the speech recognition field [3]. The currently available speech systems are able of processing naturally spoken utterances with high accuracy, but the lack of emotional component makes the ASR systems less realistic and meaningful. There are several real-world fields that may benefit from the recognition of the emotional context of an utterance, such as entertainment, emotion-based audio file indexation, HCI-based systems etc. [4].

Some of the selected features can be trained to classify, recognize 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 recognized by any intelligent system, constrained by computational resources. The implementation of the emotional sector of speech makes the human-computer interactions more real and efficient. The analysis of voice and speech for the sake of enhancing the quality of human conversations is reasonable and within the bounds of possibility. The results of emotion detection can be broadly applied in e-learning platforms, car-board systems, medical field etc.

The remaining sections have the following content: Section 2 contains literature observation on the topic, Section 3 is dedicated to the description of the problem, Section 4 carries the details of method implementation and results achieved under problem solution and finally Section 5 is the conclusion. In this paper we prove efficiency of the concept of using only 12 MFCC from 39, have identified which 12 MFCC to use for speech and emotion analysis. The dataset used for this experimentation is EMODB. Variants of supervised learning approaches have been implemented to classify emotions from two databases EMODB and SAVEE.

2. Literature Review

As well known KNN makes prognostications on time by quick calculating the similarity between an input sample and each training exemplar. Spectral analysis is a promising technique for detecting emotions from sample speech. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. Spectral analysis is a promising method of emotion detection from samples of speeches. Prosodic features of speech signals can also be used for analyzing emotions since they contain emotional information. Researchers explored the role and context of emotionы by using a set of 88 features called eGmaps [5]. Speech patterns can be obtained from combination of various speech features acquired from speaker’s utterances. Feature selection plays a pivotal role in the differentiation of different emotions of the same speaker from his speech[6] and it relies on selecting the best features from the signal. Different human languages have different accents, structures of sentences, and speaking styles [7] thus making the identification of emotions from utterances challenging. 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 a different part of the same speech signal, which complicates the setting of boundaries between emotions. An attempt has been made to study models of the multilingual emotion classification in literature [8].



The substantial advancement in technology has boosted the development of the emotion recognition fields [9] [10]. Call-centers and remote education for example [11] Existing speech recognition systems can be improved by implementing spectral analysis[12].

Authors have identified classes of features extracted from speech electrography and signals.There is an important aspect of SER that includes characterization of emotional content of a speech [13]. Several speech features are obtained from speech acoustic analysis and can be used to detect and predict emotions [14].The aim of selecting speech features is to determine properties that can improve the rate of classification emotions from a set [15]. The machine learning models are flexible enough to adapt themselves to any model that studies emotions and show good performance in predicting tasks based on selected features [16].

Authors identified classes of features extracted from emotional speech features electrography and speech signals. There is an important aspect of SER that includes characterization of emotional content of speech [13]. Several speech features are obtained from speech acoustic analysis and can be used to detect and predict emotions [14].The aim of selecting speech features is determination of properties that can improve rate of classification for emotions from a feature set [15]. 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 [16].

2.1. *Speech, Emotion and Classification*

Emotions are intertwined with mood, temperament, personality, sentiment and motivation [17]. Emotions can be understood as a complex feeling of the mind that results in physiological and physiological changes. Human thoughts and behavior is influenced by emotions and there are instances of changes in body when it encounters different emotional states [18]. Literature [19], [20] proves that there is a considerable influence of emotional syndromes on human actions and reactions. Several applications created by researchers rely on emotion detection as an integral component for identification of behavioral patterns [21], [22]. Speech features can be extracted from various sources to accomplish predictive analytics. The list of sources includes vocal tract, excitation source and prosodic extraction.

SER[23]– Speech Emotion Recognition System(Component and tasks)	
Feature Extractor	Emotion Classifier.
Takes signal input and generates emotion feature	Map the speech with one or more emotions

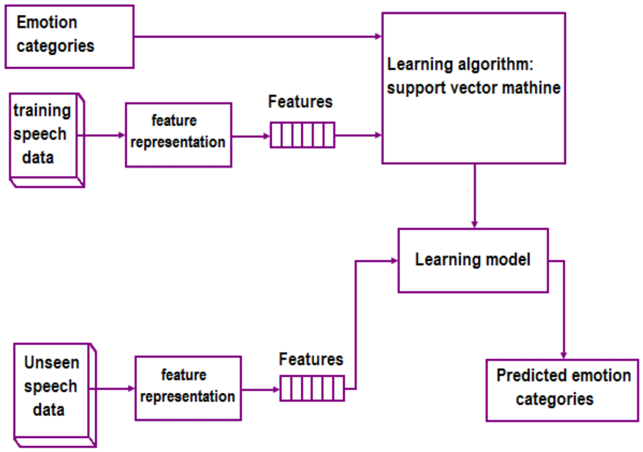


Fig.1. Framework for supervised emotion classification [23].

Emotions can be broadly studied using both discrete and continuous approaches. Different classes represent different kinds of emotions and the continuous approach of studying of emotions is a derivation based on combination of several psychological measurements on different axes [24]. Speech Emotion Recognition in the main identifies emotions on the basis of categorical approach, which depends on usage of common words. Researchers derive emotions from expressions of face, speech and various physiological signals. The analysis of facial expressions is a great way to find emotions [25–29] since human face displays emotions very aptly even without a single word uttered. Voice recordings are potentially important for expressing the speakers' mental state and their intentions. Speech features can be studied as vectors for detection of emotion from a data set. [30–32]. The autonomous nervous system allows to assess an emotion and thus utilize physiological signals like ECG, RSP BP to recognize people's emotions and possibly help cure mental illnesses [17].

We took into consideration the line of temperament. We also integrate the state in which a representative of this type of temperament may be. There are 8 axes of Miller are applied by us. Also it is optimal to take 8 axes.

Consciously controlling the volume level, for example, to emphasise a secret, even an angry person can make the voice calmer and quieter in order to show that it was a secret, a question or the essence of the issue. On the contrary, in order to highlight the characterisation of some hero of his story, the speaker can consciously increase the amplitude of the voice.

Physiological features of voice and hearing. It takes into account the average amplitude (frequency) and other characteristics of the person's voice, obtained on the basis of data from the database.

Given the above assumptions, if we wish to approach the study of temperament, we first need an understanding of emotion. Such understanding is not easy to come by a glance at a typical textbook of psychology will show that "emotion" is used to refer to a ragbag of apparently disconnected facts and is never itself clearly defined at all. Yet, within one branch of psychology, namely, animal learning theory, there has long been a reasonably clear consensus that emotions consist of states elicited by stimuli or events which have the capacity to serve as reinforcers for instrumental behavior. This, for example, is the framework within which Miller analyzed the concept of "fear" and its role in avoidance learning [69]. The term temperament has considerable overlap with dimensions of personality and with emotional, cognitive, and behavioral functions.

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 data bases can be categorized into three broad domains that cover acted emotions, natural emotions and felicitated emotions [33]. Out of the three mentioned domains enacted emotions are frequently supported by research since they are strong and reliable. EMODBis 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. Statistical Data Corpus

Three databases has been utilized for training and testing:

1. BERLIN DATABASE
2. SENTIMENT DATABASE English
3. SAVEE DATASET

Berlin Database was created in 1999 and consists of utterances spoken by various actors. EMODB has different number of spoken utterances for seven emotions[34].Emotions included in the database are anger, boredom, disgust, fear, happiness, indifference, sadness. The dataset contains more than 500 utterances spoken by 51 male and 60 female actors from the age 21 to 35 years. The emotions labelled in EMODB are listed in Table 1.

Surrey Audio-Visual Expressed Emotion (SAVEE) [35], [36]. The increasing demand of research in speech analysis led to the development of SAVEE database recordings to help study automatic emotion recognition system. The database contains recordings from 4 male actors in 7 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. 10 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, KL. The speakers who contributed for recordings were postgraduate students and researchers at the University of Surrey. The age of speakers lies between 27 to 31 years. Seven discrete categories of emotions described as anger, disgust, fear, happiness, sadness and surprise were recorded [37]. The focused research was carried out on recognizing the discrete emotions [38]. Table 2 compares the features of both the datasets used in experimental analysis.

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	

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	angry, happy, anxious, fearful, bored disgusted, neutral	anger, disgust, fear, sadness, happiness, surprise, neutral

2.3 Feature Extraction [39]–[41]

The core step for recognizing 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 characterization and recognition of information specifically related to the actor’s (speaker) mood, age, gender. The general process of feature extraction involves transformation of raw signal into feature vectors, which suppress the redundancies and emphasize on the speaker specific properties. The properties are like pitch, amplitude, 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 [31], [42], [43]. 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 [44]–[46]. 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) , Mel-Frequency Cepstral Coefficients (MFCC) are most commonly used short term spectral features for speech analysis [31], [42], [47].

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 [41], [47], [48].

Mel Frequency Cepstral Coefficients (MFCCs) were identified as a feature and is widely applicable to automatic speech recognition and identification of speaker. The correlation among the actual and heard signal frequencies can be derived efficiently by incorporating Mel scale. Davis and Murmelstein were pioneer in identifying MFCC as sound feature in the 1980's. 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 Linear Prediction Cepstral Coefficients (LPCCs) that were coined before MFCC and remained the main features for automatic speech recognition (ASR), especially with classification algorithm such as HMM [49]. In practice 8 to 12 or 13 MFCC are considered for representing the shape of spectrum and hence are used for speech analysis [50].MFCC are highly preferred choices in automatic speech recognition systems [51]. Authors found that MFCC is effective for end to end acoustic modelling using CNN [52], [53]. MFCC is widely used feature while considering speech modelling [54], [55]. MFCC based comparative study of speech recognition techniques was conducted by authors who found that MFCC with HMM gave recognition accuracy of 85 percent and with deep neural networks the score was 82.2 percent [56]. Computation of MFCCs includes a conversion of the Fourier coefficients to Mel-scale [57]. Mel-scale are the most popular variant used today, even if there is no theoretical reason that the Mel-scale is superior to the other scales [58].

2.4. Decision Tree Classifiers

There are several machine learning algorithms that can be applied for recognizing 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 [59], [60]. 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 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 [14].

2.4.1. Random Forest

One of the most flexible and easily implementable learning algorithm 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 [51, 52]. 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 [61].

2.4.2.XGBOOST

XGBoost [61] 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. Optimization and algorithmic techniques are used to improve the base framework of GBM. Regularization 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 parallelization. The interchangeable nature of the loops determine 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, initialized and sorting is done using parallel threads. This switch of loops increase the algorithmic performance. The parallelization 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 improves [61].

2.4.3. Sklearn KNeighbors classifier and 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 neighbor classifier. The closest found points are referred to nearest neighbors and euclidian distance defines the nearness of the neighboring clones [62].

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

$$d\left(A_1, A_2\right) = \sqrt{\sum_{i=1}^{n}\left(a_{1i} - a_{2i}\right)^2} \qquad (1)$$

And in case of parametrized KNN we use in particular case the following generalization of formula (1):

$$d\left(A_1, A_2(y)\right) = \sqrt{\sum_{i=1}^{n} w_i \left(a_{1i} - a_{2i}\right)^2}$$

After we choose a class y maximizing this distance. Weights depend on the neighbor's number $w(x_i)=w(i)$. For brevity, we denote these quantities by w_i . In general case we utilize some realizations of classification by formula

$$a(x)=\arg\max_{y\in Y}[x(i)=y]w_i, \quad (2)$$

where $x(i)$ are points near testing point y and Y is assumed by us class of object y .

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).

To normalize data, we will use Z-scaling based on the mean value and standard deviation; dividing the difference between the variable and the mean by the standard deviation. In practice, minimax scaler and Z-scaling have similar applicability and are often interchangeable. However, when calculating the distances between points or vectors, Z-scaling is used in most cases. And the minimax is useful for visualization, for example, to transfer the features encoding the color of the pixel into a range of [0… 255] [2].

Recall that Z-scaling based on the mean and standard deviation is dividing the difference between variable and mean by standard deviation;

$$z=\frac{x-\mu}{\sigma} \quad (3)$$

where μ is an expected value and σ is the standard deviation of the value.

Because the K-Nearest Neighbours Algorithm (hereinafter the KNN algorithm) is about the distance from a point to a class, Z-scaling is usually used for its application, as it is known that in calculating the distances between points or vectors in most cases the result of Z- scaling is much more accurate.

The main advantage of Z-scaling is that it preserves the normal distribution of a random value.

Z-normalization keeps a distribution normal if it was, and a non-normal distribution converts to a non-normal distribution too. As a result of such a transformation, we get a value with 0-th mean (mean) and 1 dispersion.

Normalization is applied to each attribute value to resolve the issue.

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 are 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 [62]–[64].

Min-max scaler keeps outliers so we have to use robust scaler of statistics that are robust to autliers. As an alternative normalizatio we propose to use Z-normalization (Z-scaler). This normalization holds a normal distribution.

3. Problem Statement

Authoritative literary sources mention MFCC as an important feature for analyzing and classifying various aspects of speech. Some of them state that only 13 MFCC features are sufficient enough to be considered for experimentation. There is currently no experimental validation for this statement. 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 [23], [50]. The current work conducts experimental analysis on MFCC obtained from EMODB, a Berlin database that consists of more than 500 utterances, which were recorded from 111 both male and female speakers from various age groups.

4. Experimental confirmation of results

To reach more effectiveness we utilize parametric KNN method. To present a number row, you have to look at the dynamics of the signal change. In large sentences, emotions are placed either at the end of the sentence or at the beginning. Therefore, when parameterizing the distance vector, it is important to set weighting factors in such a way as to distinguish the significance of the start of the sentence distance and the distance to the end of the vector coordinate.

When applying discriminant recognition techniques to the object, the feature vector is displayed in five-dimensional space. The training matrix will be defined as the data matrix:

$$W=\begin{pmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} & w_{16} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} & w_{26} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} & w_{36} \\ w_{41} & w_{42} & w_{43} & w_{44} & w_{45} & w_{46} \\ w_{51} & w_{52} & w_{53} & w_{54} & w_{55} & w_{56} \end{pmatrix}$$

We highlight five characteristics: in addition to the three mentioned in the beginning, we will highlight the beginning of the sentence and its end, as they are emotionally loaded. They carry not only an information load, but also an emotional one. When analyzing the beginning and the end of a sentence, we will take into account how much each feature has changed relative to the average characteristics of the speaker. Therefore, forming a data-vector with 5 coordinate we substitute in the *i*-th coordinate characterizing the amplitude the ratio $\frac{a_i}{M(a)}=v_i$ of the amplitude of the current phrase a_i to the average value $M(a)$ of the amplitude of the person's voice.

The columns of the matrix W contain the weighting factors [66]-[67] that most characterize this class of emotion, and the rows contain the features extracted from the phrase. For example, high amplitude is most characteristic for the emotion of aggression. The corresponding weight coefficient in the aggression column will be larger. Knowing the average values of the features of speech, we construct the matrix based on the changes in the parameters.

The following simulations and experiments were performed. The flow graph in Fig 2 shows the steps carried out for the experiment.

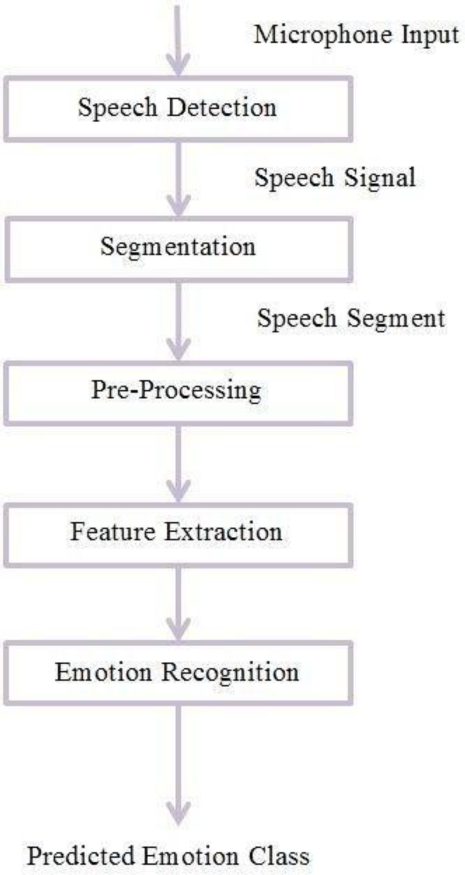


Fig 2. Flowgraph of the emotion classification using decision tree classifiers.

The experiment was done on MFCC extracted from the EMODB and SAVEE data set. Four subsets of MFCC features comprising 20 cepstral constants were analyzed 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 Table 3. 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 increase 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%, 50% respectively. For KNN the performance of four subsets is 67%, 54%, 55%, 50%.XGB showed poor performance on all the subsets. The results obtained in Table 4 and Table 5 clearly indicate that selecting M0-M12 would be a better choice for features from MFCC data set. 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 optimize the results but also reduce the complexity of the model thereby reducing the computation time.

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

There after the dataset was minimized 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 analaysis [65]. 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.

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%

The results in table 4 and Table 5 shows 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 in EMODB and SAVEE.

For further analysis the datasets were minimized and preprocessed using min max scaling. The important features identified for the minimized data using principle component analysis are in Fig 5 and Fig6. The x-axis of the plot shows various classes of emotion and y axis plot shows the MFCC features.

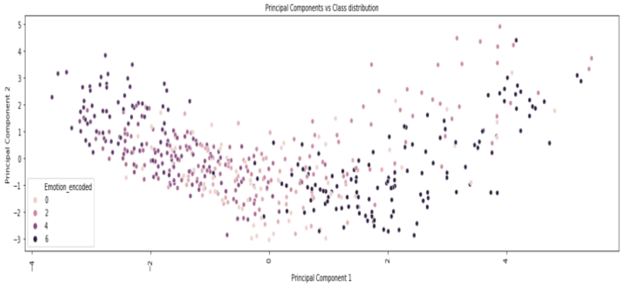


Fig.5. Principle component Vs Class Distribution for EMODB.

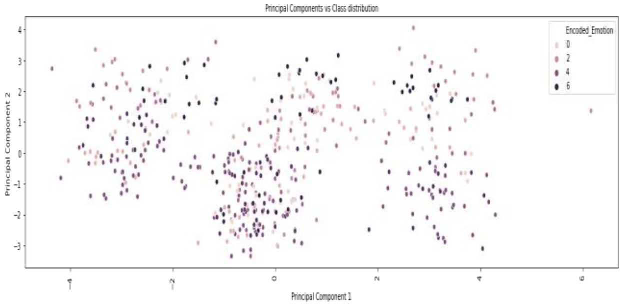


Fig.6. Principle component Vs Class Distribution for SAVEE.

The results of using 13 MFCC from minimized datasets EMODB and SAVEE are shown in Fig. 6(a), 6(b), 6(c). SAVEE results can be seen in Fig. 8(a), 8(b), 8(c). The plots in the mentioned figures 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.

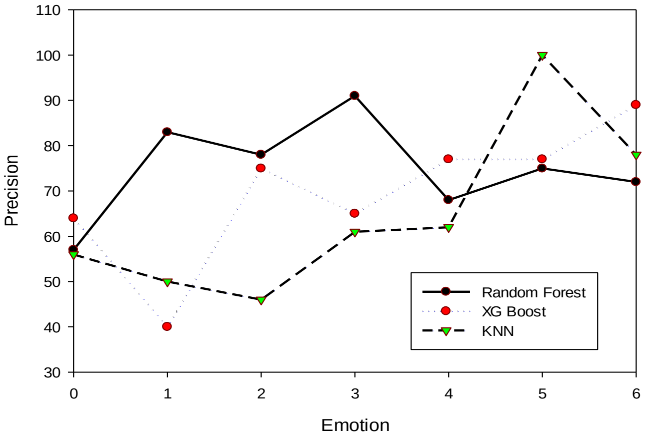


Fig.6 (a). Emotion Vs Precision score for first 13 MFCC using KNN, Random Forest and XGB classifiers on EMODB

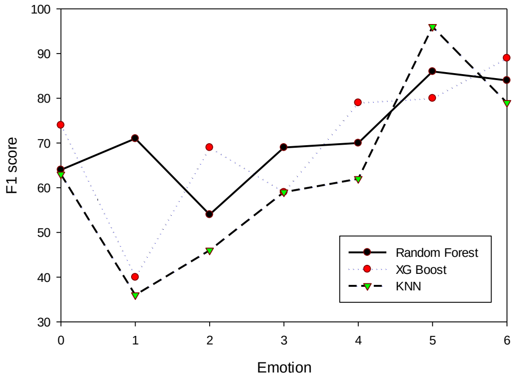


Fig.6 (b). Emotion Vs F1-score for first 13 MFCC using KNN, Random Forest and XGB classifiers on EMODB.

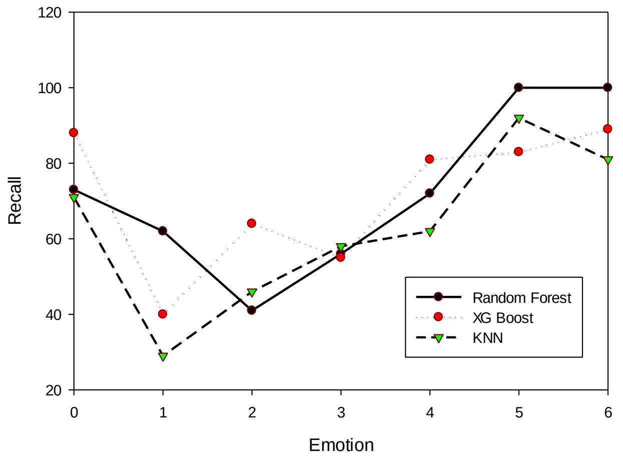


Fig.6 (c). Emotion Vs Recall score for first 13 MFCC KNN, Random Forest and XGB classifiers on EMODB.

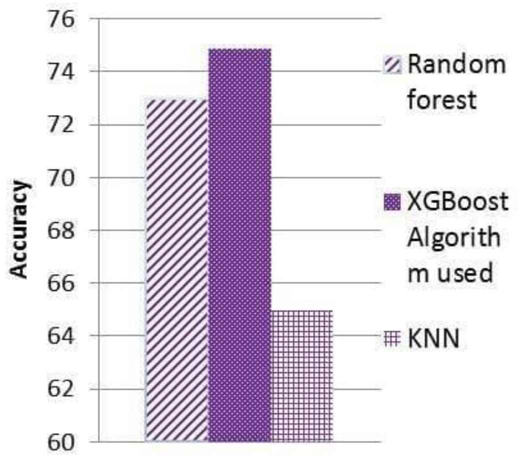


Fig.7. Accuracy score for Emotion classification using KNN, Random Forest and XGB classifiers on EMODB.

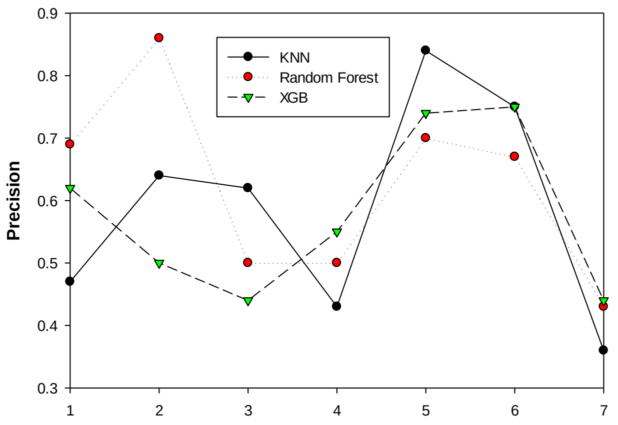
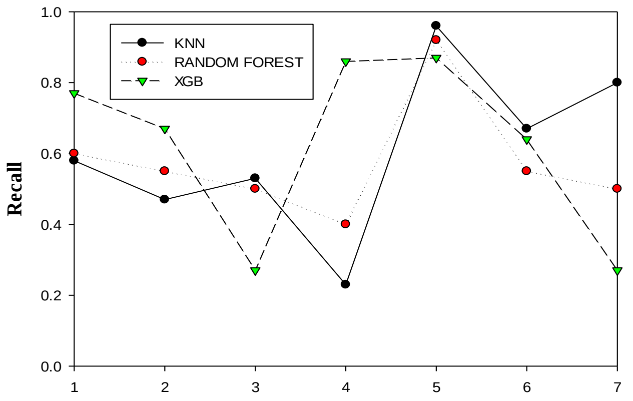
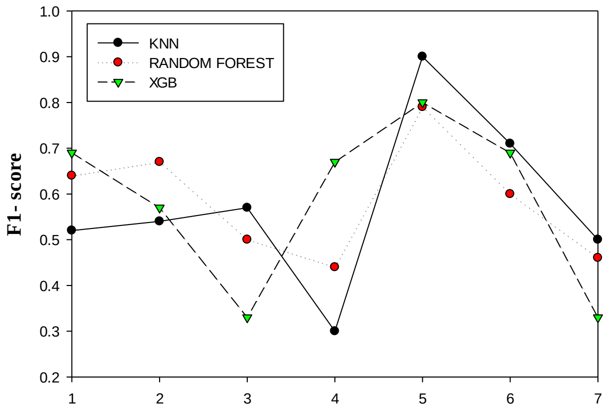


Fig.8 (a). Emotion Vs Precision score for first 13 MFCC using KNN, Random Forest and XGB on SAVEE.



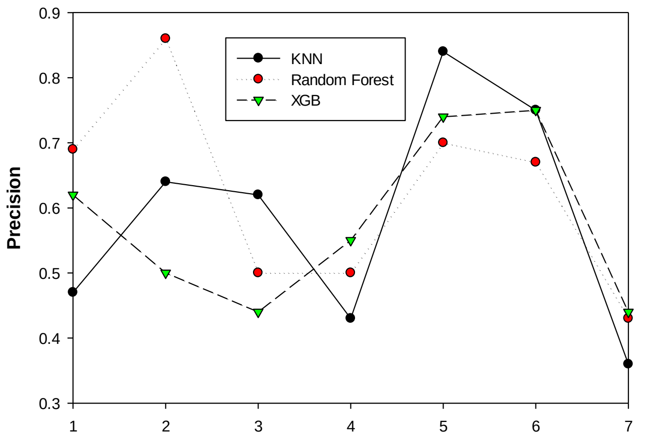
Emotion

Fig.8 (b). Emotion Vs Recall score for first 13 MFCC using KNN, Random Forest and XGB on SAVEE.



Emotion

Fig.8 (c). Emotion Vs F1 score for first 13 MFCC KNN, Random Forest and XGB on SAVEE.



Emotion

Fig.8 (d). Emotion Vs Recall score for 13 MFCC using three decision tree classifiers on EMODB.

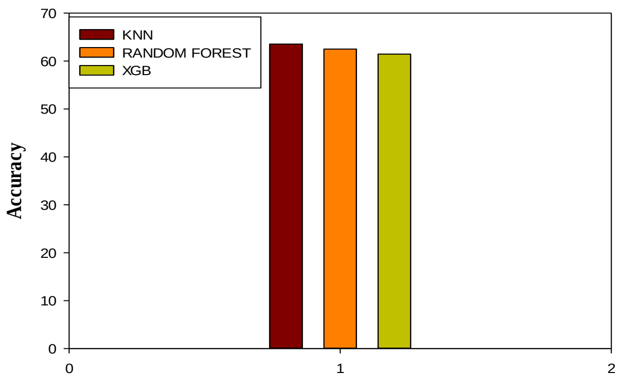


Fig.9. Accuracy score obtained with first 13 MFCC using KNN, Random Forest and XGB.

5. Conclusion

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 utilized 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 has 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 utilized for predicting emotions and personality of the speaker. The results can be integrated with various application pertaining to human psychology and medical treatments.

References

[1] S. G., K. Koolagudi, and K. S. Rao, ‘Emotion recognition from speech: A review’, in International Journal of Speech Technology, 2012

[2] C. Marechal et al., ‘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), 2019.

[3] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, ‘Emotion recognition from speech using global and local prosodic features’, International Journal of Speech Technology, 2013.

[4] S. G. Koolagudi, A. Barthwal, S. Devliyal, and K. Sreenivasa Rao, ‘Real life emotion classification from speech using gaussian mixture models’, in Communications in Computer and Information Science, 2012.

[5] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, ‘Transfer learning for improving speech emotion classification accuracy’, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2018-Septe, no. January, pp. 257–261, 2018.

[6] C. M. Lee and S. S. Narayanan, ‘Toward detecting emotions in spoken dialogs’, IEEE Transactions on Speech and Audio Processing, 2005.

[7] R. Banse and K. R. Scherer, ‘Acoustic profiles in vocal emotion expression.’, Journal of Personality and Social Psychology, vol. 70, no. 3, pp. 614–636, 1996.

[8] V. Hozjan and Z. Kačič, ‘Context-independent multilingual emotion recognition from speech signals’, International Journal of Speech Technology, 2003.

[9] S. Ramakrishnan, ‘Recognition of Emotion from Speech: A Review’, in Speech Enhancement, Modeling and Recognition- Algorithms and Applications, 2012.

[10] N. Sebe, I. Cohen, and T. S. Huang, ‘Multimodal emotion recognition’, in Handbook of Pattern Recognition and Computer Vision, 3rd Edition, 2005.

[11] Q. Zhang, Y. Wang, L. Wang, and G. Wang, ‘Research on speech emotion recognition in E-learning by using neural networks method’, in 2007 IEEE International Conference on Control and Automation, ICCA, 2007.

[12] S. Jing, X. Mao, and L. Chen, ‘Prominence features: Effective emotional features for speech emotion recognition’, Digital Signal Processing: A Review Journal, vol. 72, no. October, pp. 216–231, 2018.

[13] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, ‘Spoken emotion recognition using hierarchical classifiers’, Computer Speech and Language, 2011.

[14] A. Özseven, T.; Düğenci, M.; Durmuşoğlu, ‘A Content Analysis of The Research Approaches in Speech Emotion’, International Journal of Engineering Sciences & Research Technology, 2018.

[15] K. V. Krishna Kishore and P. Krishna Satish, ‘Emotion recognition in speech using MFCC and wavelet features’, in Proceedings of the 2013 3rd IEEE International Advance Computing Conference, IACC 2013, 2013.

[16] A. Yousefpour, R. Ibrahim, and H. N. A. Hamed, ‘Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis’, Expert Systems with Applications, 2017.

[17] L. Shu et al., ‘A review of emotion recognition using physiological signals’, Sensors (Switzerland). 2018.

[18] S. Oosterwijk, K. A. Lindquist, E. Anderson, R. Dautoff, Y. Moriguchi, and L. F. Barrett, ‘States of mind: Emotions, body feelings, and thoughts share distributed neural networks’, NeuroImage, 2012.

[19] L. Pessoa, ‘Emotion and cognition and the amygdala: From ``what is it?{’’} to ``what’s to be done?{’’} (Reprinted from Neuropsychologia, vol 48, pg

[20] S. G., K. Koolagudi, and K. S. Rao, ‘Emotion recognition from speech: A review’, in International Journal of Speech Technology, 2012.

[21] P. Winkielman, P. Niedenthal, J. Wielgosz, J. Eelen, and L. C. Kavanagh, 'Embodiment of cognition and emotion, in APA handbook of personality and social psychology, Volume 1: Attitudes and social cognition., 2014.

[22] A. Fernández-Caballero et al., 'Smart environment architecture for emotion detection and regulation', Journal of Biomedical Informatics, 2016.

[23] H. Guan, Z. Liu, L. Wang, J. Dang, and R. Yu, '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), 2018.

[24] L. Cen, F. Wu, Z. L. Yu, and F. Hu, 'A Real-Time Speech Emotion Recognition System and its Application in Online Learning', in Emotions, Technology, Design, and Learning, 2016.

[25] V. Shuman and K. R. Scherer, 'Emotions, Psychological Structure of', in International Encyclopedia of the Social & Behavioral Sciences: Second Edition, 2015.

[26] P. Ekman, 'Basic Emotions', in Handbook of Cognition and Emotion, 2005.

[27] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, 'Presentation and validation of the radboud faces database', Cognition and Emotion, 2010.

[28] P. Ekman, 'Facial expression and emotion', American Psychologist, 1993.

[29] C. Bourke, K. Douglas, and R. Porter, 'Processing of facial emotion expression in major depression: A review', Australian and New Zealand Journal of Psychiatry. 2010.

[30] J. Van den Stock, R. Righart, and B. de Gelder, 'Body Expressions Influence Recognition of Emotions in the Face and Voice', Emotion, 2007.

[31] R. Banse and K. R. Scherer, 'Acoustic Profiles in Vocal Emotion Expression', Journal of Personality and Social Psychology, 1996.

[32] T. Gulzar, A. Singh, and S. Sharma, 'Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks', International Journal of Computer Applications, 2014.

[33] U. Shrawankar and V. M. Thakare, 'Techniques for Feature Extraction In Speech Recognition System : A Comparative Study', 2013.

[34] R. E. Haamer, E. Rusadze, I. Lüsi, T. Ahmed, S. Escalera, and G. Anbarjafari, 'Review on Emotion Recognition Databases', in Human-Robot Interaction - Theory and Application, 2018.

[35] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, 'Emotion Detection Using MFCC and Cepstrum Features', Procedia Computer Science, vol. 70, pp. 29–35, 2015.

[36] P. Jackson and S. Haq, 'Surrey audio-visual expressed emotion (savee) database', University of Surrey: Guildford, UK, 2014.

[37] Z. T. Liu, Q. Xie, M. Wu, W. H. Cao, Y. Mei, and J. W. Mao, 'Speech emotion recognition based on an improved brain emotion learning model', Neurocomputing, 2018.

[38] P. Ekman et al., 'Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion', Journal of Personality and Social Psychology, 1987.

[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, 'A survey of affect recognition methods: Audio, visual, and spontaneous expressions', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

[40] A. Koduru, H. B. Valiveti, and A. K. Budati, 'Feature extraction algorithms to improve the speech emotion recognition rate', International Journal of Speech Technology, 2020.

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

[42] V. Tiwari, 'MFCC and its applications in speaker recognition', International Journal on Emerging Technologies, 2010.

[43] N. Dave, 'Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition', International Journal for Advance Research in Engineering and Technology, 2013.

[44] M. Yankayi, 'Feature Extraction Mel Frequency Cepstral Coefficients (Mfcc)', pp. 1–6, 2016.

[45] S. Ananthakrishnan and S. S. Narayanan, 'Automatic prosodic event detection using acoustic, lexical, and syntactic evidence', IEEE Transactions on Audio, Speech and Language Processing, 2008.

[46] T. Kinnunen and H. Li, 'An overview of text-independent speaker recognition: From features to supervectors', Speech Communication, 2010.

[47] W. Y. Wang, F. Biadsy, A. Rosenberg, and J. Hirschberg, 'Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification', Computer Speech and Language, 2013.

[48] J. Lyons, 'Mel Frequency Cepstral Coefficient', Practical Cryptography. 2014.

[49] H. K. Palo, M. Chandra, and M. N. Mohanty, 'Recognition of Human Speech Emotion Using Variants of Mel-Frequency Cepstral Coefficients', Lecture Notes in Electrical Engineering, vol. 442, pp. 491–498, 2018.

[50] M. Yazici, S. Basurra, and M. Gaber, 'Edge Machine Learning: Enabling Smart Internet of Things Applications', Big Data and Cognitive Computing, 2018.

[51] Xia Wang, Yuan Dong, J. Hakkinen, and O. Viikki, 'Noise robust Chinese speech recognition using feature vector normalization and higher-order cepstral coefficients', 2002.

[52] S. B. DAVIS and P. MERMELSTEIN, 'Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences', in Readings in Speech Recognition, 1990.

[53] D. Palaz, M. Magimai-Doss, and R. Collobert, 'End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition', Speech Communication, 2019.

[54] V. Passricha and R. K. Aggarwal, 'A comparative analysis of pooling strategies for convolutional neural network

based Hindi ASR', Journal of Ambient Intelligence and Humanized Computing, 2020.

[55] C. Vimala and V. Radha, 'Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words', International Journal of Computer Science and Information Technologies, 2014.

[56] C. P. Dalmiya, V. S. Dharun, and K. P. Rajesh, '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, 2013.

[57] A. NithyaKalyani and S. Jothilakshmi, 'Speech summarization for tamil language', in Intelligent Speech Signal Processing, 2019.

[58] S. S. Stevens, J. Volkmann, and E. B. Newman, 'A Scale for the Measurement of the Psychological Magnitude Pitch', Journal of the Acoustical Society of America, 1937.

[59] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, 'Features for Content-Based Audio Retrieval', 2010.

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

[61] S. B. Kotsiantis, 'Supervised machine learning: A review of classification techniques', Informatica (Ljubljana). 2007.

[62] M. Luckner, B. Topolski, and M. Mazurek, '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), 2017.

[63] O. Sutton, 'Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction', Introduction to k Nearest Neighbour Classification, 2012.

[64] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, 'Efficient kNN classification algorithm for big data', Neurocomputing, 2016.

[65] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, '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, 2018.

[66] Ruslan V. Skuratovskii. The timer compression of data and information Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP 2020, pp. 455-459..

[67] Skuratovskii, R. V. Employment of Minimal Generating Sets and Structure of Sylow 2-Subgroups Alternating Groups in Block Ciphers. Advances in Computer Communication and Computational Sciences, Springer, pp. 351–364, 2019.

[68] Gnatyuk, V. A. Mechanism of laser damage of transparent semiconductors.Physica B: Condensed Matter,. pp. 308-310, 2001.

[69] Mikhail Z. Zgurovsky, N.D. Pankratova. System Analysis: Theory and Applications. Springer Verlag. Berlin. 2007. P. 446.