Machine Learning Based Speech Emotions Recognition System

Dr. Yogesh Kumar, Dr. Manish Mahajan

Abstract: The speech signal is one of the most natural and fastest methods of communication between humans. Many systems have been developed by various researchers to identify the emotions from the speech signal. In differentiating between various emotions particularly speech features are more useful and if not clear is the reason that makes emotion recognition from speaker's speech very difficult. There are a number of the dataset available for speech emotions, it's modelling, and types that helps in knowing the type of speech. After feature extraction, another important part is the classification of speech emotions so the paper has compared and reviewed the different classifiers that are used to differentiate emotions such as sadness, neutral, happiness, surprise, anger, etc. The research also shows the improvement in emotion recognition system by making automatic emotion recognition system adding a deep neural network. The analysis has also been performed using different ML techniques for Speech emotions recognition accuracy in different languages.

Index Terms: Emotion recognition, Feature extraction, Emotions, Modeling, Machine Learning , deep neural network, Dataset

1 INTRODUCTION

One of the fastest and natural methods of communication between humans is a speech signal. For interaction between human and machine use of speech signal is the fastest and most efficient method [1]. To maximum awareness of received message, all available senses are used by human's natural ability. For machine emotional detection is a very difficult task, on the other hand, it is natural for humans. So, knowledge related to emotion is used by an emotion recognition system in such a way that there is an improvement in communication between machine and human [3]. The female or make speakers emotions find out through speech in speech emotion recognition. Linear prediction cepstrum coefficient (LPCC), Fundamental frequencies and Mel+ frequency cepstrum coefficient (MFCC) are some of the studied speech features. These features make a base for speech processing. In differentiating between various emotions particularly speech features are more useful is not clear is the reason that makes emotion recognition from speakers' speech very difficult [29]. There is an introduction of accosting variability due to the existence of different speaking rates, styles, sentences and speakers that affects the features of speech. Different emotions may be shown by the same utterance and there are different portions of the spoken utterance of each correspond emotion that makes it difficult to differentiate these portions of utterance [25]. The emotion expression depends on the culture and environment of the speaker that creates another problem as there is variation in the style of speaking by the variations in environment and culture. Transient and long terms emotion are two types of emotions and it is not clear about the type of emotion detected by recognizer [21]. Speech information recognized emotions may be speaker independent or speaker dependent.

Various classifiers like K-nearest neighbors (KNN), Support vector machine (SVM), CNN, etc are available for classification [7]. In this paper brief introduction about speech emotion recognition is given along with the speech emotion recognition system block diagram description in the second section of the paper. Various work has been done on different datasets so some of the existing datasets are covered in the third section along with modeling of emotions speech and different types of speech. The fourth section gives brief details about various feature extraction mechanism for speech emotion recognition and then focus is given on review on the classification part. In this section, we have covered KNN, SVM, CNN, recurrent neural network, etc. The sixth section gives brief about the use of deep learning for speech emotion recognition.

2 SPEECH EMOTION RECOGNITION SYSTEM

There is a pattern recognition system stage in speech emotion recognition system that makes them both same [22]. Energy, MFCC, Pitch like derived speech features patterns are mapped using various classifiers. It consists of five main modules are:

- **Speech input:** Input to the system is speech taken with the help of microphone audio. Then equivalent digital representation of received audio is produced through pc sound card.

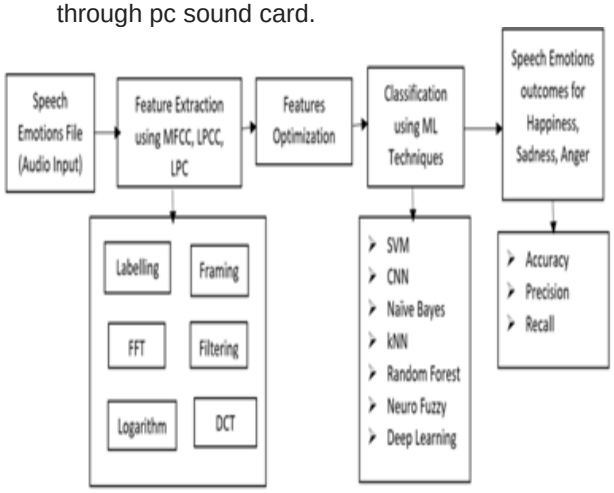

Fig. 1 Framework for Speech Emotions Recognition

- *Dr. Yogesh Kumar is currently working as an associate professor CSE in CEC, Landran, Mohali.*
  Email: Yogesh.arora10744@gmail.com@gmail.com
- *Dr. Manish Mahajan, Professor and Head of CSE in CEC, Landran, Mohali. Email: Manishmahajan4u@gmail.com*

- **Feature extraction and selection:** There are 300 emotional states of emotion and emotion relevance is used to select the extracted speech features [22]. For speech feature extraction to selection corresponding to emotions all procedure revolves around the speech signal.
- **Classification:** Finding a set of significant emotions for classification is the main concern in speech emotion recognition system. There are 300 emotional states contains in a typical set of emotions that makes classification a complicated task [9].
- **Recognized emotional output:** Fear, surprise, anger, joy, disgust and sadness are primary emotions and naturalness of database level is the basis for speech emotion recognition system evaluation.

3 DATASET FOR SPEECH EMOTIONS

In the field of affect detection, a very important role is played by suitable choice of speech database. Three databases are used for good emotion recognition the system as given below [8]:

1. Elicited emotional speech database: In this case emotional situation is created artificially by collecting data from the speaker.

- Advantage: This type of database is similar to a natural database
- Problem: There is unavailability of all emotions and if the speaker
  knows about it that they are being recorded then artificial emotion can be expressed by them [9].

2. Actor based speech database: Trained and professional artists collect this type of speech dataset.

- Advantage: In this database-wide variety of emotions are present and collecting it is also very easy.
- Problem: It is very much artificial and periodic in nature

3. Natural speech database: Real world data is used to create this database

- Advantage: For real world emotion recognition use of natural speech database is very useful.
- Problem: It consist of background noise and all emotions may not be present in it.

4 MODELLING EMOTIONS

From overview of 64 emotional speech data collection most represented emotions are sadness, surprise, anger, joy, fear and are known as big n category of emotions [11]. Further emotions are modelled into categories that are separated from main categories such as neutral, positive, negative and anger, joy, sadness, fear, etc like big n emotions. These are separated into surprised, irritated and bored, etc shades of main categories. Dimension aspects are another way to model emotions in which how many dimensions taken into account is the main concern. "Arousal is the individual's global feeling of dynamism or lethargy", is a three-dimensional space to represent the states of human emotion. Both physical and mental activity is subsuming, preparedness to act and overt activity. Control and power are two related concepts subsumes by power dimension on other hand people's sense of their own power is the central issue about emotion. In this valence is an individual overall sense of woe or weal and is related to what is faced by them [19].

4.1 TYPES OF SPEECH:

On the basis of ability they have to recognize a speech recognition systems can be separated in different classes [12]. Following are the classification:

- **Isolated words:** In this type of recognizers sample window both sides contains low pitch utterance. At a time only single word or utterance is accepted by it and there is need to wait between utterances by speaker as these systems have listen/non-listen states. For this class isolated utterance is a better name.
- **Connected words:** In this separate utterance can run together with minimal pause between them otherwise it is similar to isolated words.
- **Continuous words:** It allows users to speak naturally and content are determined by computer. Creation of recognizers that have continuous speech capabilities are difficult due to determination of utterance boundaries by utilizing a special method.
- **Spontaneous words:** It can be thought of as speech at basic level that is natural sounding and not rehearsed. Variety of natural speech features are handle is the ability of spontaneous speech with ASR system.

5 FEATURE EXTRACTION MECHANISM FOR SPEECH EMOTION RECOGNITION

Relevant emotional features extraction from speech is the second important step in emotions recognition. To classify features there is no unique way but preferably acoustic and linguistic features taxonomy is considered separately. Due to extreme difference concerning thwie extraction methods and database used is another distinction. An importance is gain by linguistic features in case of spontaneous or real life on other hand their features lose their value in vase of acted speech. Earlier only small set of features were used but now larger number of functional and acoustic features are in use for extraction of very large feature vectors [32]. In this section explanations of acoustic, linguistic features and functional are discussed.

- **Acoustic features:** Large statics measures of energy, duration and pitch is used to characterized acoustic features that are derived from speech processing [33]. In order to mask particular items in speech of humans a involuntary and voluntary acoustic variation is basic used for emotion recognition using acoustic features. Measurement of energy, pitch or voiced and unvoiced segments duration is in seconds that can represent duration features by applying different types of normalisation. Words, utterance, syllables or pauses like phonemes unit's higher phonological parameter duration is exclusively represented [26].
- **Linguistic features:** In reaction of our emotional state an important role is played to grammatical alternations or words chosen by us. Bag-of-Words and N-Grams are two prime methods from number of existing techniques used for analysis. To predict next given sequence a probabilistic base language model is used and N-grams is a numerical representation form of texts in automatic document categorisation. Reduction of speech complexity by elimination of irrelevant words and stopping words that do not

increase a general minimum frequency of occurrence is useful before applying this technique. Cries, laughs, sighs, etc non-linguistic vocalisations can be integrated into vocabulary.

- **Functional:** After extraction of Low-level descriptor (LLD) a functional are applied and number of functional and operators. Out of each base contour equal size feature vector is obtained [18]. To obtain constant number of elements a feature vector is used per word to provide normalization over time that are ready to be model by static classifier. Before applying functional a LLD can be transformed or altered as for linguistic features. Example of functional features is peak (distance, number), four first moments (curtosis, standard deviation, mean and skewness), segments (duration, number) or extremes values (max, min, range).
- **Feature selection:** To describe phenomenon from a larger set of redundant or irrelevant features is a subset of features selected by feature selection. Feature selection is done to improve the accuracy and performance of classifier [20]. Wrapper based selection methods are generally used approaches that employ an accuracy of target classifier as optimization criterion in a closed loop fashion [26]. In this features with poor performance are neglected.
  Hill climbing, sequential forward search is commonly chosen procedure with a sequentially adding and empty set. These features give performances improvement. Selected subset of features effects are ignored by use of filter methods which is a second general approach. Reduced features sets obtained from the acted and non-acted emotions difference is very less.

6 FEATURE EXTRACTION FOR SPEECH EMOTIONS RECOGNITON

There are number of methods for feature extraction like Linear predictive cepstral coefficients (LPCC), Power spectral analysis (FFT), First order derivative (DELTA), Linear predictive analysis (LPC), Mel scale cepstral analysis (MEL), perceptual linear predictive coefficients (PLP) and Relative spectra filtering of log domain coefficients (RASTA) [30]; [5].

- **Linear predictive coding (LPC):** In encoding quality speech at a low bit rate LPC method is useful that is one of the most powerful techniques of speech analysis. At current time specific speech sample can be approximated as a linear combination of past speech samples is the basic idea behind linear predictive analysis. It is a human speech production base model that utilizes a conventional source filter model. Vocal tract acoustics are simulated by Lip radiation, vocal tract and glottal transfer functions that are integrated into one all pole filter. Over a finite duration the sum of squared differences between estimated and original speech signal is minimized using LPC that helps in having unique sets of predictor coefficients.
  In real recognition actual predictor coefficients are not used as a high variance is shown by it. There is transformation of predictor coefficient to a cepstral coefficients more robust set of parameters. Some of the types of LPC are residual excitation, regular pulse excited, pitch excitation, voice excitation and coded excited LPC.
- **Mel frequency cepstral coefficients (MFCC):** It is considered as one of the standard method for feature extraction and in ASR most common is the use of 20 MFCC coefficients. Although for coding speech use of 10-12 coefficients are sufficient and it depend on the spectral form due to which it is more sensitive to noise. This problem can be overcome by using more information in speech signals periodicity although aperiodic content is also present in speech. Real cesptal of windowed short time fast Fourier transform (FFT) signal is represent by MFCC [21]. Non linear frequency is use. The parameters similar to humans used for hearing speech are used to extracts parameters using audio feature extraction MFCC technique. Other information is deemphasizes and arbitrary number of samples contain time frames are used to divide speech signals. Overlapping from frame to frame is used to smooth the transition in most systems and then hamming window is used to eliminate the discontinuities from each time frame.
- **Perceptual linear prediction (PLP):** Hermansky developed a PLP model that uses psychophysics concept of hearing to model a human speech. The speech recognition rate gets improved by discarding irrelevant information by PLP. Spectral characteristics are transformed to human auditory system match is the only thing that makes PLP different from LPC.
  The intensity-loudness power-law relation, equal-loudness curve and critical-band resolution curves are three main perceptual aspects approximates by PLP.
- **Mel scale cepstral analysis (MEL):** PLP analysis and MEL analysis is similar to each other in which psychophysically based spectral transformations is used to modify the spectrum. According to the scale of MEL a spectrum is wrapped in this method on other hand according to bark scale a spectrum is warped in PLP. So output cepstral coefficients are the main different between scale cepstral analysis of PLP and MEL. The modified power spectrum is smooth using all pole model in PLP and then on the basis of this model a output cepstral coefficients are computed. On other hand modified power spectrum is smooth using cepstral smoothing in MEL scale cesptral analysis. In this Discrete Fourier Transform (DFT) is used to convert log power spectrum is directly transform into capstal domain.
- **Relative Spectra filtering (RASTA):** The ability to perform RASTA filtering is provided by analysis library to compensate for linear channel distortions. It can be used either in cepstral or log spectral domains and in both of them linear channel distortions is appear as an additive constant. Each feature coefficient is band passes by RASTA filter and convolutional introduced noise in the channel effect is alleviated by band pass filter equivalent high pass portion. Then frame to frame spectral changes are smoothened with the help of low pass filtering.
- **Power spectral analysis (FFT):** This is the more common techniques of studying speech signal and over the frequency content of the signal over time is described by speech signal power spectrum. Discrete

Fourier Transform (DFT) of the speech signal is the first step to compute power spectrum that computes time domain signal equivalent frequency information. Real point values consist speech signal can use Fast Fourier Transform (FFT) to increase the efficiency.

7 CLASSIFICATION FOR SPEECH EMOTIONS RECOGNITION SYSTEM

The best features come after features calculation is provided to the classifier. In expression of speaker's speech an emotion is recognizes by classifier and for speech emotion recognition number of classifiers have been proposed by various researchers. In this section review of some of the classifier has been given.

- **K-Nearest Neighbours (KNN):** Automated speech services such as interactive voice recognition systems have used speech based emotion recognition. In mental depression like medical applications, lie detectors like investigative application use of speech services play great implications. Renjith S, et.al, (2017), have worked on Telugu and Tamil languages to detect emotions happiness, sadness and anger using speech recordings [23]. In their work they have pre-processed to separate disturbances from speech waveforms and raw speech signals. They have extracted Hurst and Linear Predictive Cepstral Coefficients (LPCC) features and then classification is done on the basis of statistical parameters obtained from these features. Both KNN and ANN is used to identify the reactive emotions and then accuracy, precision and recall parameter is used to compare their performance for both features individually and in combination. As compared to LPCC when use of Hurst gives better results when tested for individual features in terms of recall, precision and accuracy.
  In developing and existing technology security plays an important role and in order to avoid criminal activities cyber security has gone beyond the password. Public and personal information can be access by implementing the parallel human emotion and biometric based recognition framework. Steven A. Rieger Jr, et.al, (2014), have focused on collection of KNN and spectral feature extraction with pattern recognition paradigm. adaptive component weighted cepstrum (ACW), linear predictive cepstrum (CEP), line spectral frequencies (LSF), postfilter cepstrum (PFL) and mel frequency cepstrum (MFCC) are five spectral features [27]. In their work they have trained ensemble of kNNs using bagging algorithm and by ensemble classification fusion is accomplished implicitly. In all the experiments a LDC emotional prosody speech database is used and results show that performance of single kNN is inferior compared to two kNNs.
- **Naive Bayes classifier:** In human communication an important role is played by emotion as feelings can be easily convey through it. In speech processing domain, emotion detection from speech has become a challenging and major area of research. This task has become even more challenging due to categorization of several emotional classes from extracted suitable features from speech. Naive Bayes classifier is used by Atreyee Khan, et.al, (2017), along with both spectral and prosodic features for emotion detection [15]. As spectral features a Mel-Frequency Cepstral Coefficients (MFCC) has been used and pitch is used as prosodic feature. Naïve Bayes Classifier is used to perform classification and they have considered seven emotional classes to develop both gender independent and dependent system. Berlin Emo-db popular speech database speech samples are used to test accuracy of the system after performing classification. Classification of audio signal into four basic emotional state is implemented by S. K. Bhakre, et.al, (2016) by considering MFCC, picth, ZCR and energy statistical features from 2000 utterances of the created audio signal database [2]. Average magnitude difference method (AMDF) is used to extract pitch features and magnitude spectrum sum of square absolute value is used to calculate energy. Energies spectrum of Discrete cosine transform (DCT) is used to calculate MFCC in which they have considered only 1-14 coefficients of DCT and rest is discarded. For variables approximate calculation is done using regression analysis of statistical process in statistical modelling. Audio signal is classified into four different emotions using Naive Bayes classifier. It is totally a probability based classifier so it gives accurate prediction in speech analysis as there is need to predict future sample because speech signal is a random signal. There is need of millions of dataset by signal classifier for signal recognition in speech signal but use of Naive Bayes classifier requires minimum dataset.
- **Support Vector Machine (SVM) classifier:** Human computer Interface (HCI) subset automatic emotion and speech recognition has become widely researched topic with the advent of digitization of every possible avenue. As we understand machines, machine has also understood us as menial jobs are taken with machines. From given sample amplitude, pitch and MFCC features are extracted and it run across growing and existing database of training samples. Ashwini Rajasekhar, et.al, (2018), have distinguished the given sample using SVM and speaker utterance is detected using MFCC [24]. In the end SVM classifier differentiates between fear, anger, sadness, happiness and updates the database accordingly. Amiya Kumar, et.al, (2015), have introduced a novel approach by combining MFCC, LPCC derived features, energy, ZCR, pitch prosody features, MEDC dynamic features for automatic recognition of speaker's emotion state [13]. Then happy, surprise, anger, sad, disgust, neutral and fear are seven discrete emotional states identified using multilevel SVM classifier in five native assamese languages. The proposed approach is evaluated for combination of features in terms of accuracy that shows a good result for speaker independent cases as compared to individual features.
- **Convolution Neural network (CNN) classifier:** Extraction of speech emotion features are main part of speech emotion recognition so Li Zheng, et.al, (2018), have proposed a random forest and CNN based new network model (CNN-RF) (Zheng, L., 2018). From normalized spectrogram a speech

emotion features are extracted using CNN and then speech emotion features are classify using RF classification algorithm. From results it has been predicted that as compared to traditional CNN model use of CNN-RF model gives improved results and it also improves the Nao record sound command box. Finally, Nao robot can "try to figure out" a human's psychology through speech emotion recognition and also know about people's happiness, anger, sadness, and joy, achieving a more intelligent human-computer interaction.

In the community, a lot is attention has been by cross model approaches like CNN. This is mainly developed for images analysis and can be used in speech processing. Norman Weibkirchen, et.al, (2017), have represented the emotion afflicted speech input using spectrograms adapting CNN based classification architecture (Weißkirchen, N., 2017). On SUSAS, eNTERFACE and EmoDB benchmark corpora proposed network architecture is applied and in Leave-One-Speaker-Out setting proposed classification ability is investigated. The results show a close to real life corpus for SUSAS.

- **Recurrent neural network:** Transfer of each utterance categorical label into a label sequence is the challenge which needs to be considered while modeling the categorical speech emotion recognition tasks in a sequential approach. So, Xiaomin Chen, et.al, (2018), have to make a hypothesis. In which both non-emotional and emotional segments consist of utterance alternatively (Chen, X., 2018). On the basis of that hypothesis, they have treated an utterance label sequence as a chain of nulls denoting non-emotional frames and emotional states denoting emotional frames are two kinds of states. For automatically alignment and label an utterance emotional segments with emotional labels a connectionist temporal classification based recurrent neural network (CTC-RNN) is exploited and the same is used for non-emotional segments with non-emotional labels. The proposed method is tested on IEMOCAP corpus that shows the effectiveness of it as compared to state of the art emotion recognition algorithms. Effectiveness of speech features used for classification make automatic emotion recognition from speech as a challenging task. For automatically discovering of emotionally relevant features from a speech deep learning is used by Seyedmahdad Mirsamadi, et.al, (2017). Both the emotionally relevant short-term frame-level acoustic features can be learned using a deep recurrent neural network (Mirsamadi, S., 2017). It also helps in appropriate temporal aggregation of those features into a compact utterance level representation. Then feature pooling over time is done using new novel strategy in which local attention is used to focus on regions of the emotionally silent speech signal. They have also used IEMOCAP corpus to evaluate the proposed solution that gives more accuracy in terms of prediction as compared to existing emotion recognition algorithms.

TABLE 1: COMPARISON TABLE OF DIFFERENT CLASSIFIER

S. No.	Author name	Classifier	Database
1	Renjith S, et.al, (2017),	kNN and ANN	Amritaemo
2	Steven A. Rieger Jr, et.al, (2014),	KNN	LDC emotional prosody speech database
3	Atreyee Khan, et.al, (2017),	Naive Bayes	Berlin Emo-db
4	Sagar K. Bhakre, et.al, (2016),	Naive Bayes	They have made dataset by considering 2000 sentences audio signal from 20 different speakers.
5	Ashwini Rajasekhar, et.al, (2018),	SVM	They have used computerized voice dataset
6	Amiya Kumar, et.al, (2015),	Multilevel SVM classifier	Utterances of "Multilingual Emotional Speech Database of North East India" (MESDNEI).
7	Li Zheng, et.al, (2018),	Convolution Neural Network combined with Random Forest (CNN-RF)	RECOLA natural emotion database
8	Weibkirchen, et.al, (2017),	CNN	we utilised three data sets Berlin Emotional Speech Database (EmoDB), eNTERFACE and Speech Under Simulated and Actual Stress (SUSAS)
9	Xiaomin Chen, et.al, (2018),	Connectionist temporal classification based recurrent neural network (CTC-RNN)	IEMOCAP corpus

8 DEEP LEARNING FOR SPEECH EMOTION RECOGNITION SYSTEM

Pipeline of pre-processing, feature extraction, dimensionality reduction, and classification are the basis of traditional speech emotion recognition. On the basis of professional classifier and engineering accuracy of recognition, performance may vary that is more difficult in case of big data. Now a day's number of researchers has started working on automatic emotion recognition from raw signal due to the finding of final result automatically by learning representation from a neural network. So, Huijuan Zhao, et.al, (2018), have given a review on current progress on end-to-end speech emotion recognition problems [35]. They have covered network model requirement, process procedures, current achievements and some of the potential future issues in speech emotion recognition. Towards better interaction between machine and human effect, recognition is considered as an important component. In call centres and human-computer interaction and other several areas use of emotion recognition in speech have been found. In emotion recognition use of deep neural networks (DNN) has given great success. For continuous emotion recognition from speech Panagiotis Tzirakis, et.al, (2018), has presented a new model in which Convolutional Neural Network (CNN) is used to

train a model end to end [28]. In these features are extracted from the raw signal using CNN and stacked on top of it a 2-layer Long Short-Term Memory (LSTM) is considered as contextual information in the data. To test the proposed model RECOLA database is used and tested it in terms of concordance correlation coefficient that shows good results. In academia also use of speech emotion recognition (SER) is a hot topic and choice of speech emotion features using SER systems is one of the key issues. So, feature selection is necessary to establish a speech emotion recognition system that gives a perfect representation of speech emotion attributes. Most of the methods are based on a single type as each kind of features is effective. A deep learning-based feature fusion is done in the proposed method by Gang Liu, et.al, (2018), in which both hyper-prosodic based pitch features and features based on spectral are combined [18]. The use of proposed methods results in improvement of SER system. A new method has been described by Pavol Harár, et.al, (2017), in which they have used DNN architecture with pooling, convolutional and fully connected layers [10]. Three class subset of emotional speech Berlin database (German Corpus) is used by them are:

- Neutral
- Angry
- Sad

The database contains 271 labelled recording with a total length of 783 seconds. Unit variance and mean of every audio file is zero due to standardized of raw audio data and without overlap, with the segments of 20 millisecond, every file is split. Then further to divide all data into Testing (10%), Validation (10%) and Train (80%) sets and eliminated silent segments using Voice Activity Detection (VAD) algorithm. Stochastic Gradient Descent is used to optimize DNN and without feature selection, raw data is used as input. The proposed trained model is tested in terms of test accuracy that gives 96.97% on whole file classification. From the last few decades increasing attention on emotion recognition and application value of emotion has been considered. In 2016, Xi Zhou, et.al, have worked on making a feasible SER system in which two deep learning methods are used to make effective models [36]. Deep belief network and stacked autoencoder network are twp deep learning methods used for emotion states classification and automatic emotion feature extraction respectively. German Berlin Emotional Speech Database is used for testing the proposed model that gives 65% of accuracy in the best case. Along with this, they have also validated the influence of different emotion categories and speakers in recognition accuracy.

9 COMPARATIVE ALNALYSIS FOR SPEECH EMOTIONS RECOGNITON USING DIFFERENT MACHINE LEARNING TECHNIQUES

Performance analysis based on different machine learning techniques for different languages. The comparison shows that the different machine learning methods have been used for recognizing speech emotions system for numerous languages. Based on that, the accuracy has been computed for the best case. Although emotions are of different types, in the paper, the accuracy is taken for the best case using different feature extraction techniques and machine learning methods.

TABLE 2: PERFORMANCE ANALYSIS BASED ON ACCURACY FOR DIFFERENT ML TECHNIQUES

S. No.	Authors	Feature Extraction Technique	Machine Learning Techniq-ues	Accuracy (%)	Speech Emotions Language
1	Li Zheng et al. (2018)	CNN	CNN	81.4%	English
2	Li Zheng et al. (2018)	CNN	CNN-Random Forest	84.6%	English
3	Aishah Abdul Razak et al. (2015)	LPC	Neuro-Fuzzy	60%	English
4	Aishah Abdul Razak et al. (2015)	LPC	Neuro-Fuzzy	60%	Malay
5	Aishah Abdul Razak et al. (2015)	MFCC	SVM	87%%	English
6	Renjith S, Manju K G (2017)	LPCC	Hybrid kNN	61.29%	Tamil
7	Renjith S, Manju K G (2017)	LPCC	Hybrid-kNN	62.05%	Telugu
8	Renjith S, Manju K G (2017)	LPCC	ANN	62.37%	Tamil
9	Renjith S, Manju K G (2017)	LPCC	ANN	67.18%	Telugu
10	Sagar K. Bhakre et al. (2016)	MFCC	Naïve Bayes	81%	English

10 CONCLUSION

In the paper brief introduction about speech emotion recognition is given along with the speech emotion recognition system block diagram description. In the field of affect detection, a very important role is played by a suitable choice of speech database. For good emotion recognition system mainly three databases are used. On the basis of ability, they have to recognize a speech recognition system can be separated in different classes are isolated, connected, spontaneous and continuous words. Relevant emotional

features extraction from the speech is the second important step in emotions recognition. To classify features there is no unique way but preferably acoustic and linguistic features taxonomy is considered separately. There are a number of methods for feature extraction like Linear predictive cepstral coefficients (LPCC), Power spectral analysis (FFT), First order derivative (DELTA), Linear predictive analysis (LPC), Mel scale cepstral analysis (MEL), perceptual linear predictive coefficients (PLP) and Relative spectra filtering of log domain coefficients (RASTA) and some of them are briefly covered in this paper. Another important part of speech emotion recognition system is the use of classifier. In the paper, the detailed review on KNN, SVM, CNN, Naive Bayes, and recurrent neural network classifier for speech emotion recognition system. The last section of the paper covers the review on the use of the deep neural network to make speech emotion recognition system. To further improve the efficiency of system combination of more effective features can be used that enhances the accuracy of speech emotion recognition system.

REFERENCES

[1] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition, vol. 44, pp. 572-587, 2011.

[2] S. K. Bhakre, A. Bang, "Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier", 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2363-2367, 2016.

[3] I. Chiriacescu, "Automatic Emotion Analysis Based On Speech", M.Sc. THESIS Delft University of Technology, 2009.

[4] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, D. Jiang, "Sequence-to-sequence Modelling for Categorical Speech Emotion Recognition Using Recurrent Neural Network", 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1-6, 2018.

[5] P. Cunningham, J. Loughrey, "Over fitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets Research and development in intelligent systems", XXI, 33-43, 2005.

[6] C. O. Dumitru, I. Gavat, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", International Symphosium ELMAR, Zadar, Croatia, 2006.

[7] S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009.

[8] R. Elbarougy, M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception", 2013 AsiaPacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–10, 2013.

[9] D. J. France, R. G. Shiavi, "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Transactions on Biomedical Engineering, pp. 829–837, 2000.

[10] P. Harár, R. Burget, M. K. Dutta, "Speech Emotion Recognition with Deep Learning", 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 137-140, 2017.

[11] Q. Jin, C. Li, S. Chen, "Speech emotion recognition with acoustic and lexical features", PhD Proposal, pp. 4749–4753, 2015.

[12] Y. Kumar, N. Singh, "An Automatic Spontaneous Live Speech Recognition System for Punjabi Language Corpus", I J C T A, pp. 259-266, 2016.

[13] Y. Kumar, N. Singh, "A First Step towards an Automatic Spontaneous Speech Recognition System for Punjabi Language", International Journal of Statistics and Reliability Engineering, pp. 81-93, 2015.

[14] Y. Kumar, N. Singh, "An automatic speech recognition system for spontaneous Punjabi speech corpus", International Journal of Speech Technology, pp. 1-9, 2017.

[15] A. Khan, U. Kumar Roy, "Emotion Recognition Using Prosodic and Spectral Features of Speech and Naïve Bayes Classifier", 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1017-1021, 2017.

[16] A. Kumar, K. Mahapatra, B. Kabi, A. Routray, "A novel approach of Speech Emotion Recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages", 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 372-377, 2015.

[17] Y. Kumar, N. Singh, "Automatic Spontaneous Speech Recognition for Punjabi Language Interview Speech Corpus", I.J. Education and Management Engineering, pp. 64-73, 2016.

[18] G. Liu, W. He, B. Jin, "Feature fusion of speech emotion recognition based on deep Learning", 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 193-197, 2018.

[19] C. M. Lee, S. S. Narayanan, "Toward detecting emotions in spoken dialogs", IEEE Transactions on Speech and Audio Processing, pp. 293-303, 2005.

[20] S. Mirsamadi, E. Barsoum, C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention", 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227-2231, 2017.

[21] A. Nogueiras, A. Moreno, A. Bonafonte, J. B. Marino, "Speech Emotion Recognition Using Hidden Markov Model", Eurospeech, 2001.

[22] J .Pohjalainen, P. Alku, "Multi-scale modulation filtering in automatic detection of emotions in telephone speech", International Conference on Acoustic, Speech and Signal Processing, pp. 980-984, 2014.

[23] S. Renjith, K. G. Manju, "Speech Based Emotion Recognition in Tamil and Telugu using LPCC and Hurst Parameters", 2017 International Conference on circuits Power and Computing Technologies (ICCPCT), pp. 1-6, 2017.

[24] A. Rajasekhar, M. K. Hota, "A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines", International Conference on

Communication and Signal Processing, pp. 0114-0118, 2018.

[25] M. Shrivastava, A. Agarwal, “Classification of emotions from speech using implicit features”, In 9th International Conference on Industrial and Information Systems, pp. 1-6, 2014.

[26] B. Schuller, A. Batliner, S. Steidl, D. Seppi, “Recognising realistic emotions and a_ect in speech: State of the art and lessons learnt from the first challenge”, Speech Communication, vol. 53, pp. 1062-1087, 2011.

[27] A. Steven, J. Rieger, R. Muraleedharan, R. P. Ramachandran, “Speech Based Emotion Recognition Using Spectral Feature Extraction and an Ensemble of kNN Classifiers”, 2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 589-593, 2014.

[28] P. Tzirakis, J. Zhang, B. W. Schuller, “End-to-end speech emotion recognition using deep neural networks”, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089-5093, 2018.

[29] T. Vogt, E. Andre, J. Wagner, “Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization”, LNCS 4868, pp. 75-91, 2008.

[30] D. Ververidis, C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods Speech Communication”, vol. 48, pp. 1162-1181, 2006.

[31] N. Weißkirchen, R. Bock, A. Wendemuth, “Recognition of Emotional Speech with Convolutional Neural Networks by Means of Spectral Estimates”, 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 50-55, 2017.

[32] S. Yildirim, M. Bulut, C. Lee, “An acoustic study of emotions expressed in speech”, Proceedings of InterSpeech, pp. 2193–2196, 2004.

[33] J. Yuan, L. Shen, F. Chen, “The acoustic realization of anger, fear, joy and sadness in Chinese”, Proceedings of ICSLP, pp. 2025–2028, 2002.

[34] L. Zheng, Q. Li, H. Ban, S. Liu, “Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest”, The 30th Chinese Control and Decision Conference (2018 CCDC), pp. 4143-4147, 2018.

[35] H. Zhao, N. Ye, R. Wang, “A Survey on Automatic Emotion Recognition Using Audio Big Data and Deep Learning Architectures”, 2018 4th IEEE International Conference on Big Data Security on Cloud, pp. 139-142, 2018.

[36] X. Zhou, J. Guo, R. Bie, “Deep learning based Affective Model for Speech Emotion Recognition”, 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, pp. 841-846, 2016.