International Journal of Research in Engineering, Science and Management
Volume-3, Issue-11, November-2020
https://www.ijresm.com | ISSN (Online): 2581-5792

A Comparative Study On Speech Emotion Recognition

Anushka Sandesara[1*], Shilpi Parikh[2], Pratyay Sapovadiya[3], Mrugendrasinh Rahevar[4]
[1,2,3]*UG Student, U & P U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Science and Technology, CHARUSAT, Changa, India*
[4]*Assistant Professor, U & P U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Science and Technology, CHARUSAT, Changa, India*
*Corresponding author: anushkasandesara@gmail.com

***Abstract*: Today's world has been "Chatting" with the machines for a long time. With the first known research paper by Daellert et. al on this topic, we can say that discussions on Speech Emotion Recognition technology have been there for such a long time and have been evolving and increasing its applications in our life. Although worthy of these many applications, speech emotion recognition is a challenging task as emotion is a subjective thing. Not all humans are the same, each human deals differently with emotions. There are no common criteria or steps to categorize emotions. Forget computers, even we humans at times fail to read the emotions behind the other person. This paper provides the list of some speech emotion recognition methods and a glimpse of method used.**

***Keywords*: Deep learning, Machine learning, Speech recognition.**

1. Introduction

Communication is the key to human existence. The most common and efficient method of communication in humans is by speech [1,2]. There are various methods for communication like text messages, email, etc. But somewhere along we feel that instead of communicating by text, communicating by speaking is way better. And the reason for the same is that many times we have to face ambiguous situations, for example, the phrase "This is good." can be said in a happy manner or in a sad manner. We humans can figure the meaning by figuring the emotion of the voice heard but machines cannot. Nowadays emojis have become so ordinary in text messages as these text messages can easily be misconceived without knowing the emotions behind it. And so, we would often send emojis with texts to pass emotion along with the text message as we do in speech so that we are not misunderstood.

From many years the research on this topic has been continued and until recently, emotion recognition from speech has acquired enormous recognition due to its growing application. The most common approach developed to extract emotion from speech was by understanding the relation between acoustic features and emotions. Different features of speech such as speaking rate, intonation, energy, formant frequencies, frequency (pitch), loudness, duration, and spectral characteristic (timbre) is used to encode emotion in speech [3], [4]. A bunch of machine learning algorithms have been developed to extract this information and build models for SER. The figure 1 shows the flow chart of traditional SER system.

Speech emotion recognition is inherently a multi-model process [5]. Speech is an integral part which conveys emotional information, but it isn't the only part of our body which conveys, emotional information. Visual and linguistic information such as facial expression, semantic, body language, are also necessary to convey information regarding emotions.

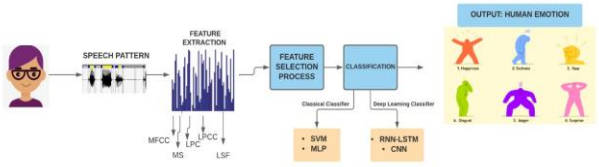

Fig. 1. Flow chart of traditional SER system

Speech emotion recognition is inherently a multi-model process [5]. Speech is an integral part which conveys emotional information, but it isn't the only part of our body which conveys, emotional information. Visual and linguistic information such as facial expression, semantic, body language, are also necessary to convey information regarding emotions. It is debated that when you communicate with one another, your 55% of the message is intended by your body language [6]. Mostly all of the exploration done on emotion recognition is only based on the speech modality databases which consist of emotions exaggerated by actors, and other modalities are yet to be explored.

In this paper, our aim is to gather and provide comprehensive information about various significant methods which have been developed and determined the best suited methods for speech emotion recognition. Information regarding particular technology, its techniques, their limitations, type of result generated is required before implementing any method which is provided in this paper.

2. Related Work

There have been previous works on Speech Emotion Recognition using various methodologies and datasets. Some of them make use of different kinds of neural networks and different types of classifiers for emotion classification. Depending upon the combination of feature extraction and type of classifier been used, we achieve different accuracies. The speech emotion recognition has been implemented using various classifiers as SVM, MLP, KNN, RNN, HMM, GMM, LSTM, CNN and various other. All of them have worked upon different feature extraction. The previous work can be broadly classified into two major parts, which are:

A. Classification schemes using machine learning approach

1) Feature classification using basic SVM classifier

The authors of this paper proposed a SER system using simple SVM [7]. In this system, three emotions have been acknowledged which are, sad, neutral and happy. The features extraction was – LPCC (Linear Predictive Spectrum Coding), pitch, MFCC (Mel-frequency Spectrum Coefficients), energy and MEDC (mel-energy spectrum dynamic coefficients). In this paper they have used combination of two different database: Berlin EmoDB and their self-made Chinese Emotional Database. Also, these two databases were used with different combination of features They created five different training models, each model working on different combination of features. The first model consisted of combination of Energy and Pitch. Likewise, second model was mixture of MFCC and MEDC features, third model consisted MFCC, LPCC and MEDC features in it, fourth model had Energy, MEDC and MFCC features in combination and finally the last- fifth model was a combination of pitch, energy, MFCC and MEDC features.

2) Feature classification using binary and multiclass SVM classifier

The authors in this particular paper proposed SER system using Binary SVM and Multiclass SVM [8]. In this proposed method, they have used seven different Binary Support Vector Machine [9] each for seven different emotions namely – Anger-Not Anger, Boredom-Not Boredom, Disgust-Not Disgust, Fear-Not Fear, Happy-Not Happy, Sad-Not Sad and Neutral-Not Neutral. These seven features are from the German Emotional Speech Corpus EmoDB [10]. Correlation based feature (CFS) [11] along with Sequential forward selection (SFS) [12] was used to select these features from binary SVMs. Along with these binary machines, another Multiclass SVM was also used. The below figure 2 shows the proposed SER architecture.

In the above proposed architecture, the Binary SVMs and one Multiclass SVM are not merged. They are kept separate only. Out of the 535 samples, 20% of the sample (107) are separated and kept for final evaluation of fused model and from the remaining 428 samples, 70% (300) of this data is used for the training of Binary SVMs and Multiclass SVM, while the remaining 30% (128) data is used for testing it. For labelling the samples in each Binary SVM and Multiclass SVM, of the 428 samples, Anger are labelled as positive label and all other remaining samples are labelled as negative label corresponding to Not-Anger feature. While evaluating the accuracy, the Happy-Not Happy samples were not used as it gave the highest negative examples.

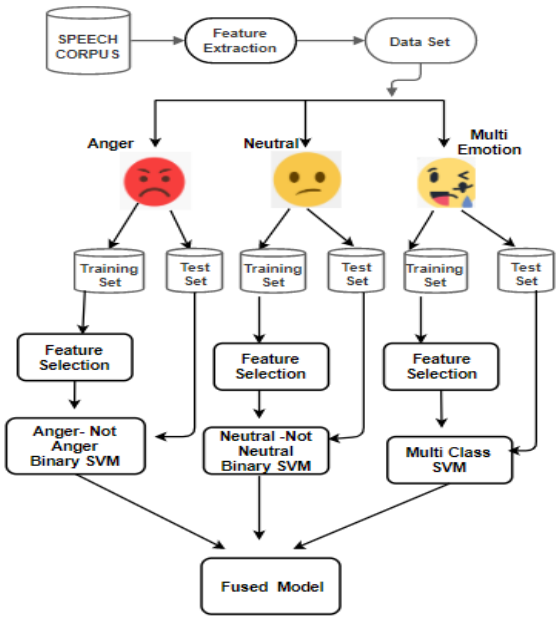

Fig. 2. SER architecture using Binary SVMs and Multiclass SVM

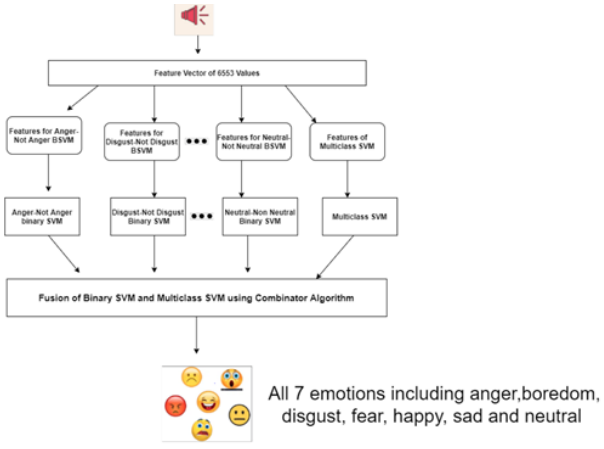

Fig. 3. Fusion Model

Another architecture of a Fusion Model, combining the seven Binary SVMs and one Multiclass SVM into one was also proposed which used the combinator algorithm. This model proved to show improvement in the overall accuracy.

3) Feature classification using three stage hierarchical SVM classifier

In this paper the authors used MFCC feature for constructing SER system with the help of SVM classifier [13]. In this proposed scheme, they made use of a 3-Stage Support Vector

Machine. Using this machine, they classified seven different emotions found in the Berlin Emotional Database [14]. They focused on the extraction of MFCC features [15] from the database. In the proposed system, in the first stage of the three stages of hierarchical SVM, emotions anger, disgust, fear and happiness are separated from emotions boredom, neutral and sadness. Further in the second stage, 2 SVMs are used. Of these 2 SVMs the first one classifies the emotions anger, happiness, disgust and fear whereas the second SVM classifies the boredom, neutral and sadness emotions. And finally, on the third stage, 3 SVM's are used. The first SVM isolates anger from happiness, the second one classifies disgust and fear and the last third one separates boredom form neutral. Thus, in all, 6 binary SVMs are required to create a 3-stage SVM in order to separate 7 emotions. The figure 4 shows the proposed system architecture.

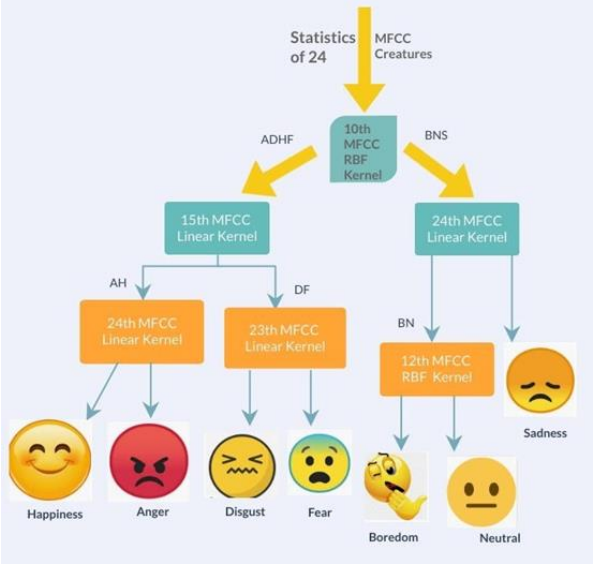

Fig. 4. Three Stage Hierarchical SVM System

4) SVM classier with Decision Tree

In this paper [16] the authors have proposed a new system in which they have made use of decision tree algorithm and applied it to the SVM classifier. For the feature selection process, they have used Fisher Feature selection process. Three notable highlights of this paper are: 1. Removing unnecessary features by making use of fisher criteria. This will eventually increase the recognition rate. 2. Invention of an algorithm to determination of the structure of tree at run-time. And at last they developed a genetic algorithm which will improve the features of SVM and merged the fisher criteria with the decision tree and SVM. The classifier is provided with data from two different datasets i.e. CASIA Chinese EmoDB and Berlin emotional dataset. They acquired five kinds of features which are combination of both prosodic and spectral features. The features acquired are as follows: pitch frequency, energy, zero-crossing rate, MFCC and Fourier Coefficient. Along with these 5 features another 5 analytical variables are calculated. These features are of multi-frame. Such features include, Median, Mean, Maximum, Minimum and Standard Deviation. Later on, these features are applied for emotion recognition process. The proposed SVM model can adopt two strategies for classification. The two strategies are: Injection (i.e. one-to-one) and One-To-Many. Comparing these two strategies, Injection strategy is found to be speedier than the other strategy. So, in the paper they have made use of Injection strategy for SVM. Along with it, they have used RBF kernel functions. The figure 5 shows the architecture of their proposed system.

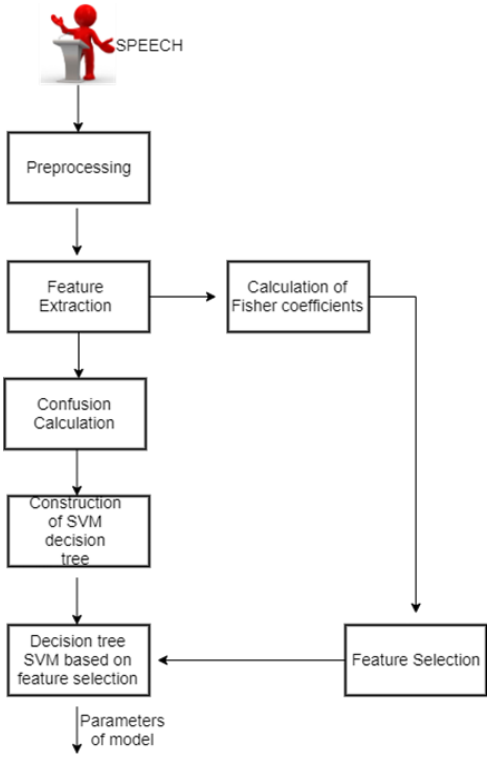

Fig. 5. Proposed system flowchart

One of the key problems solved by the proposed architecture was, reducing the confusion between the emotions. Doing so, the overall recognition rate got increased. This was possible because of the use of decision tree in classification problem.

5) Feature classification using MLP classifier

Another way to detect emotion from a speech is by feature extraction using MLP (Multi-Layer Perceptron) Classifier. Many works have been done using the MLP classifier. One of the methodologies proposed was using Neural Network along with MLP classifier to implement SER system [17]. In this paper [17], they focused on the underlying emotions in our speech, which are reflected through tone and pitch. In this paper they aimed to classify and identify major 8 types of emotions: happy, angry, sad, neutral, clam, disgust, fearful and surprised. They made use of the RAVDESS [18] dataset. The MLP Classifier was used for the classification of different emotions.

In the proposed methodology in paper, the MLP network had one input layer of 300 and 40,80,40 hidden layers within it

along with one output layer. The input layer takes the five features extracted from the audio file as input. These five extracted features are: MFCC, Mel Spectrogram Frequency, Chroma, Tonnetz and Contrast. A Logistic Activation function to act upon the input data is being used by the hidden layer and then it processes the data. The output layer brings out the information learned by the network as output. This layer classifies and gives output of the predicted emotion, according to the computation performed by the hidden layer. The MLP is made to train to the given dataset. The Multi-Layer Perceptron Classifier relies on an underlying Neural Network to perform classification. It trains the Neural Network using Back-Propagation. The dataset is spilt in the ratio of 75:25 for the training and testing.

Construction of MLP Classifier:

Step 1: Initialization of MLP Classifier by defining and initialization of required parameters.

Step 2: Train the Neural Network with the provided data.

Step 3: Prediction of output

Step 4: Calculating the accuracy

B. Classification Schemes Using Deep Learning Approach

1) Feature classification using time distributed CNN

Taking into consideration the flaws of traditional machine learning-based approaches, a new network was introduced which combined the use of a convolution neural network and one special circulation neural network. In this network LSTM and CNN, both are used to accomplish feature learning. From the experiments, it is proved that recognition of 7 emotions of Time Distribution CNN surpassed those of CNN and other LSTM networks. The idea of this network is to merge a deep CNN feature extraction architecture with a RNN model, which knows how to identify sequential dynamics in a speech signal. Unalike the SVM approach, this model will not work on international statistics obtained on attributes from domain of time and frequency.

2) Classification Using CNN based approach

The random forest classifier is an ensemble classifier. It falls in the category of ensemble algorithms because it is implemented using a combination of more than one algorithm of the same or different type. This model makes use of CNN as an uncommon characteristic extractor and all CNN layers are set with a multifarious-convolution nucleus to produce the obtained features more thorough. From the experimental results, the precision of the CNN-RF model is 3.25% greater than the convolutional model [19].

From the six emotions tested the two confusing were "surprise"' and "fear". The "neutral" emotion has the highest recognition rate in this model. This network model was successfully implemented on NAO robots and this system is performed on the Chinese speech emotion system. The average rate of recognition for all the emotions on the NAO robots is 75.57%. After a comprehensive test was conducted it was proposed that the CNN-RF model provides the NAO robots some cardinal responsibility of speech emotion recognition.

3) Feature classification using semi CNN based approach

In SER the main area of research is how discriminative, affect-salient aspects are drawn out from speech taken as input. The model of semi-CNN includes a layer of input, single convolutional layer, one layer which is thoroughly connected and lastly the important SVM classifier. By following the pecking order of semi-CNN, features extracted from each layer become highly invariant to any nuisance and also maintains the affect-salient goal of SER. The affect-salient learning method used in this approach is evaluated taking into consideration four common speech datasets that is Emo-DB, SAVEE, MES and lastly DES [20].

4) Feature classification using Deep CNN

The presented framework aims to employ attribute learning strategy for spectrograms obtained from speech. The model used here comprises three CNN layers, three fully joined layers, and a Softmax Layer. A spectrogram of 256 x 256 dimension is created from signals of speech that is taken as input. A spectrogram is considered to be a visual representation of the signals at various frequencies present in different waveforms [21].

5) Feature classification using Deep Neural Networks

Here two CNN neural networks and CNN LSTM networks namely 1 dimensional and 2-dimensional combination of CNN and LSTM network respectively are taken into consideration to acquire knowledge of native as well as general features correlated to emotions deduced from discourse and long-mel spectrogram. The two neural networks mentioned above comprises of similar architecture having four LFLBs and one LSTM layer. The LFLB present here consists of unique max-pooling layer and one convolutional layer respectively.

6) Feature classification using convolutional LSTM-RNN

In this model, a speech emotion recognition is proposed using convolutional long short-term memory (LSTM) and recurrent neural network (RNN) by using a phoneme-based feature extractor. This technique yields outputs phoneme-based emotion probabilities to all frames of an input utterance [22].

Now-a-days there is an increasing attentiveness to implement DL techniques to understand aspects from an emotional database. Deep Networks are implemented on top of standard classifiers such as SVM. In this method, an amalgam of biLSTM with a new pooling policy possessing mechanism for attention is taken into consideration. With this attention mechanism, this approach can simultaneously neglect silence frames as well as the extraneous chunks of speech which do not contain any psychological content. Experiments are performed on the interactive emotional dataset under this method. The feature extraction of this approach comprises of two steps. Firstly, the number of acoustic attributes which tend to be affected by feelings are pulled out from small frames normally 20 to 50 msec. They are recognized as LLD. Secondly, disparate analytical cluster functions are applied to all LLDs, and results are combined to create a lengthy feature vector called as HSF) at the utterance level. Below given table shows

common LLDs and HSFs for speech emotion recognition [23].

Table 1
List of common LLDs and HSFs

LLDs (Low-Level descriptors)	Pitch (F_0), voicing probability, energy, zero-crossing rate, MFCCs, formant locations/bandwidths, jitter, harmonics-to-noise ratio, Mel-filter bank features, etc.
HSFs (High-Level Statistical Functions)	Mean, variance, minimum, maximum, range, median, quartiles, linear regression coefficients, higher order moments (skewness, kurtosis) etc.

The performance is assessed on happiness, sadness, neutral and angriness. The experiments are applied to both raw spectral features and custom-made LLDs. The trained LLDs possess a softmax classifier that provides more precision in comparison to conventional LLDs that make use of SVM. The SVM model makes use of surfeit amount of HSFs to attain its best precision, the deep neural network (DNN) is lacking sensitivity to the array of second-hand HSFs.In certain circumstances test results are imbalanced so report generated contains weighted accuracy(WA) as well as unweighted accuracy(UA) that is the average recall over a variety of emotional categories.

7) Deep Neural Networks and Elm

A DNN is also referred to as a feed-forward neural network that contains more than contains numerous hidden layers midst of inputs and outputs. In this approach, the DNN grabs the common auditory attributes within the speech provided as input and generates fragments of emotional odds distributions from that vocal sound, are created which are later helpful to guess the emotional state. It is not necessary to implement the DNN model for the classification because fragments generated already provide an ample amount of data without any specific training. So instead of that ELM can be employed which includes one unique hidden layer network. The DNN that is trained priorly evaluates the emotional state accurately. Then from this emotion distributions, vocal features are created and passed to an ELM machine which determines the state with accuracy much higher than other conventional approaches. The training scheme of ELM is straightforward. While common neural networks require weights to be tuned by implementing the backpropagation algorithm, in ELM it works by assigning arbitrary values to the weights inserted between the input and hidden layer [24].

8) RNN: Bidirectional LSTM

The RNN has certain limitations which include the lack of covering long context information because of the gradient vanishing problem. To conquer this disadvantage, a LSTM network was established which comprises of recurrently connected memory blocks. Experiments are carried out on the IEMOCAP database. For evaluation criteria, four emotions (Angry, Happy, Sad, and Neutral) are taken into consideration. Two parameters are used for evaluation WA and un-weighted accuracy UA. The WA represents the accuracy obtained by classifying on the comprehensive test data set and on the other side UA is mean of all accuracies obtained during classification of all the emotions [25].

3. Results

The model is evaluated based on basic Support Vector Machine model. The dataset used here is EmoDB.

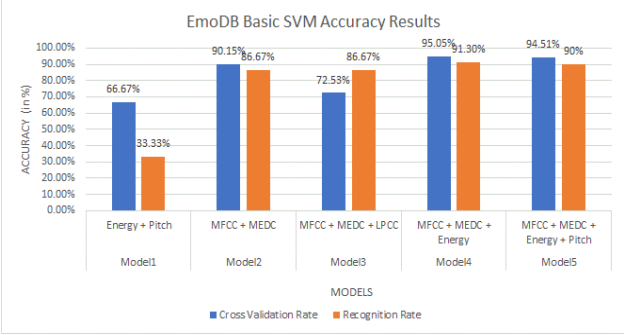

Fig. 6. Analysis of Cross Validation rate and Recognition rate of EmoDB features for Basic SVM model

And the same model trained using Chinese database gives the following result.

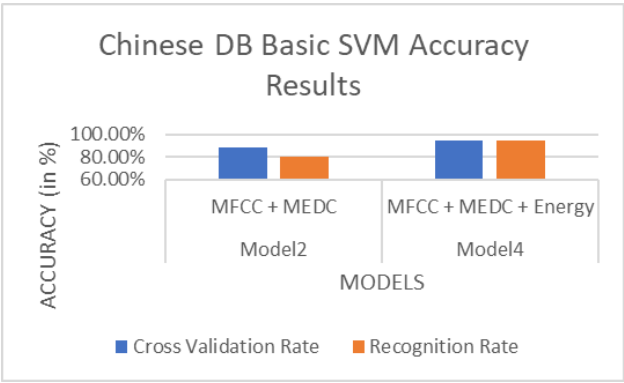

Fig. 7. Analysis of Cross Validation rate and Recognition rate of Chinese DB features for Basic SVM model

It can be observed from the above two tables that, different combination of features results in different recognition rate. For the Berlin EmoDB, the amalgam of MFCC + MEDC + Energy features gave the highest accuracy (95.1%). The reason for the same can be because it contained both prosodic features and spectrum features and these features has excellent characters. Similarly, for the Chinese Database, the same feature amalgam gave the highest accuracy (91.3%).

The method used here for training the model is Binary Support Vector Machine and Multiclass SVM. Dataset used here is EmoDB.

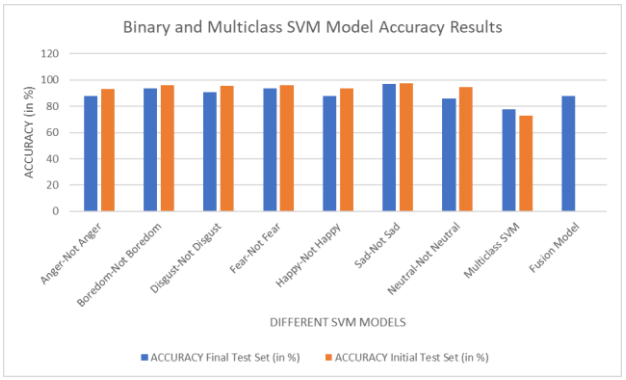

Fig. 8. Comparison between initial and final test accuracy of features for Binary and Multiclass SVM model

The model tested here is based on Three Stage Hierarchical Support Vector Machine and used on Berlin Emotional Database. The obtained accuracy is as below in the table.

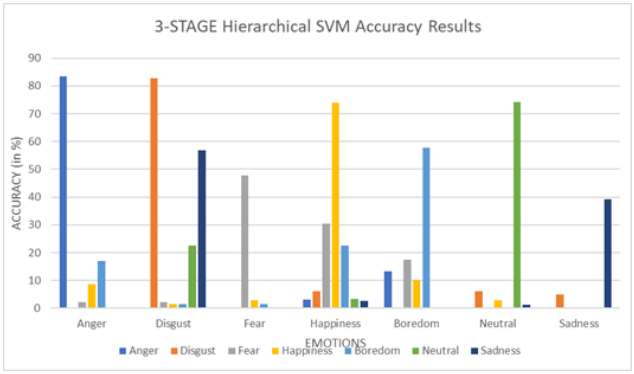

Fig. 9. Accuracy of various features for a 3-Stage Hierarchical SVM model

Here, 6 binary SVMs are required to create a 3-stage SVM in order to separate 7 emotions. Using 24 MFCC features an accuracy of 68% was obtained by this system.

Decision tree SVM model was used.

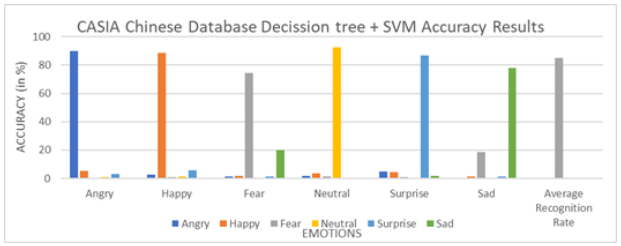

Fig. 10. Accuracy of various features in Chinese DB for Decision tree + SVM model

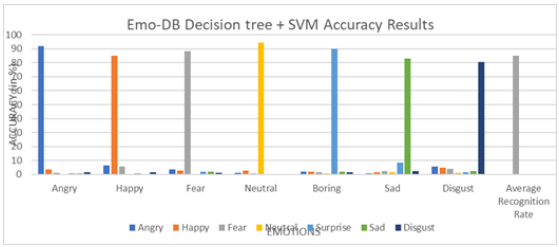

Fig. 11. Accuracy of various features in EmoDB for Decision tree + SVM model

The accuracy rate of this proposed model was obtained about 83.75%. Accuracy obtained by this model was 9% more than conventional SVM and for SVM not containing feature selection, the accuracy was about more 8.08%.

Thus, it depends a lot on the combination of features and the type of classifier that we select for prediction and recognition of emotion in speech.

On considering individual parameters with Logistic Activation Function and Relu Function, the obtained accuracy is as below in the table.

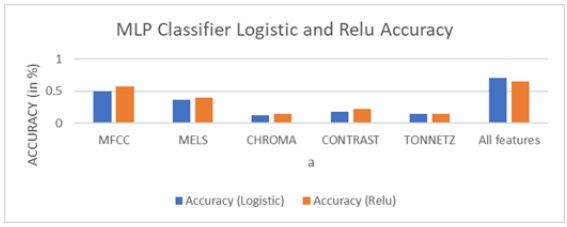

Fig. 12. Comparison of logistic accuracy and Relu accuracy for a MLP classifier

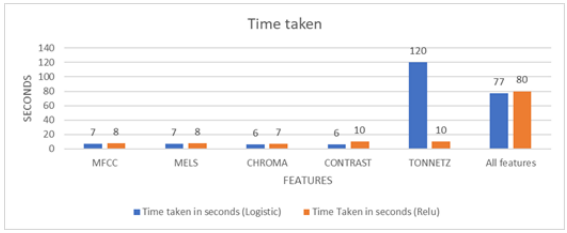

Fig. 13. Time taken for Logistic vs. Relu Activation function

On considering individual parameters with Logistic Activation Function and Relu Function, the obtained accuracy is as below in the table.

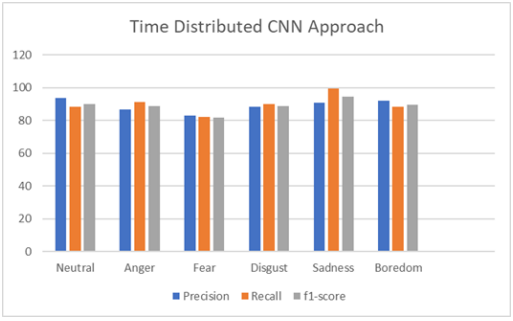

Fig. 14. Accuracy of Time Distributed CNN

This approach generated an extreme score of 74% on the verification set as well as 72% on the trial set. By implementing time distributed CNN ,10% higher accuracy is obtained if compared to the SVM approach.

By comparing the recognition rate of the semi-CNN approach with other models it is noted that this approach is robust to the variation of speakers. The table given below shows that all the trained features (LIF, semi-CNN (no_s), semi-CNN (no_or), and semi-CNN) perform better than fundamental features (RAW, TEO and other acoustic features) [20].

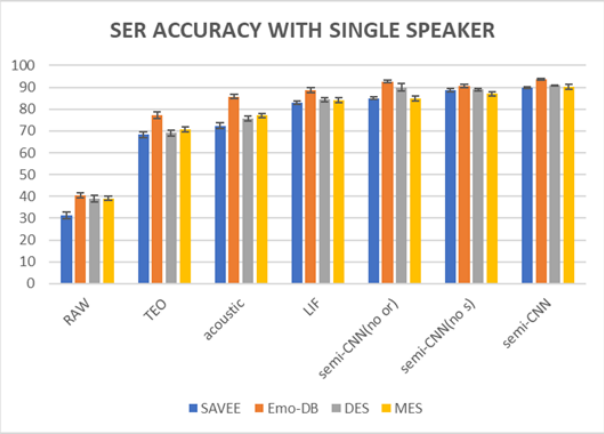

Fig. 15. Semi CNN Accuracy with Single speaker

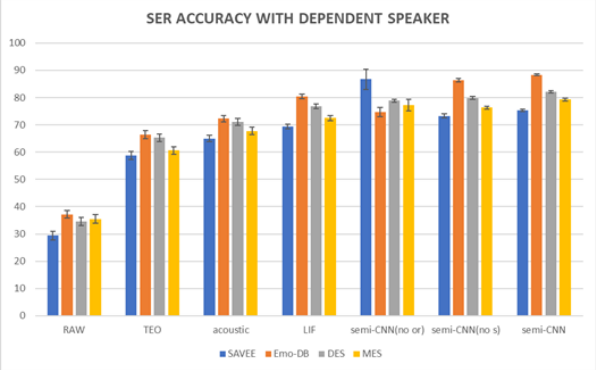

Fig. 16. Semi CNN Accuracy with Dependent speaker

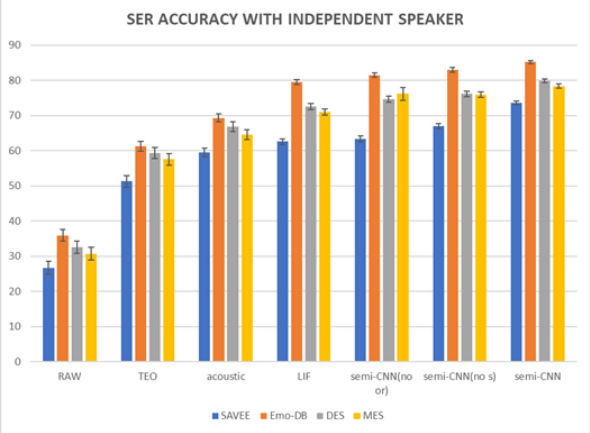

Fig. 17. Semi CNN Accuracy with Independent speaker

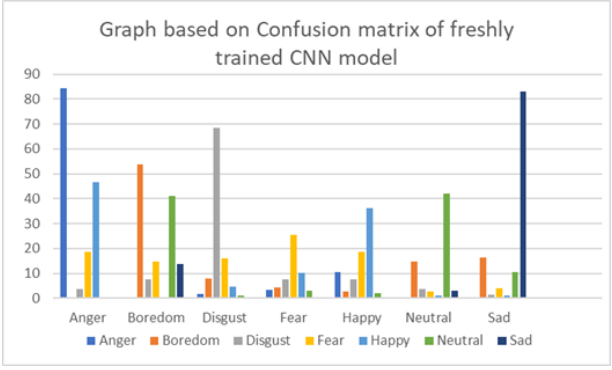

Fig. 18. Graph based on freshly trained CNN

The model mentioned was tuned on spectrograms from EmoDB by using fresh trained CNN approach. Results obtained from the model are displayed in graph above. From the observations made it is clear that the prediction performance for emotions like annoyance, boredom, disgust, and sadness is above 50 %. Nevertheless, the prediction accuracy for fear, happy and neutral emotions was nominal. The reason behind this is that emotions pertaining to fear are mostly confused with emotions related to annoyance, disgust, and happiness. Though, the rate of confusion claimed from the table (approximately 19% in all cases) is lower than correct predictions i.e. 25.33%. Moreover, emotions that relate neutrality are confounded with boredom. Also, happy emotion is confused with anger. Further works are needed to diminish confusions.

In this, the learned weights obtained from the priorly trained model are accustomed to set up the model before parameters are being tuned in relation to the novel dataset. The rate of learning in this model is often kept marginal (one-tenth of the conventional learning rate) so the weights are adjusted according to it.

Here Fine-Tuned CNN approach is used the confusion matrix of this finely tuned AlexNet model is represented in Table 2 below. This model enhances the precision accuracy of the emotions pertaining to annoyance, neutrality, and sadness. Nonetheless, the prediction accuracy of the other four emotions declined.

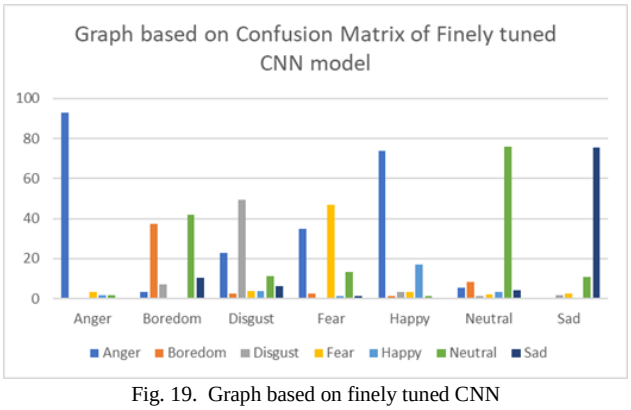

Fig. 19. Graph based on finely tuned CNN

From the experimental results of these two models, it is proved that excellent performance is achieved on the mission of recognizing speech emotions. Moreover, the 2-dimensional CNN LSTM network performs much better than the conventional approaches achieving correctness of 95.33% and 95.89% on the EmoDB of speaker-dependent as well as speaker-independent. Also, this model achieves higher recognition accuracy on the IEMOCAP database i.e. Approximately 89.16% and 52.14% on speaker-dependent and speaker-independent experiments respectively which is greater than accuracies obtained from traditional approaches [26].

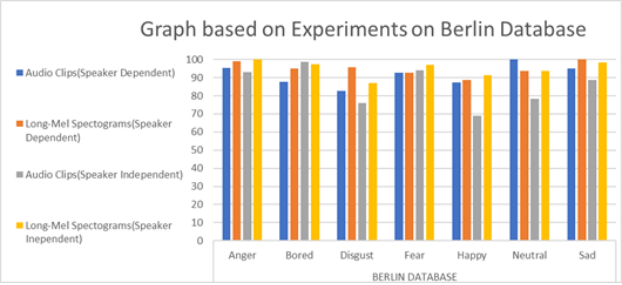

Fig. 20. Accuracy of 2D CNN LSTM Network on EmoDB

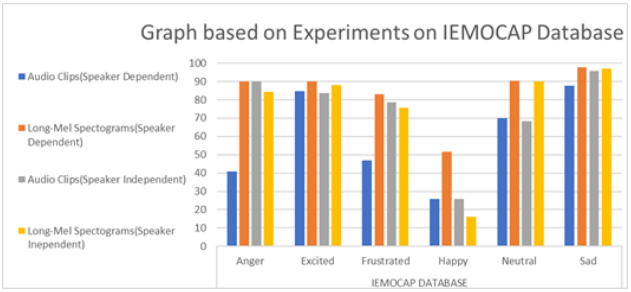

Fig. 21. Accuracy of 2D CNN LSTM Network on IEMOCAP

From the results, it is concluded that average recognition accuracies obtained from long-mel spectrograms possess more precision than raw-audio clips.

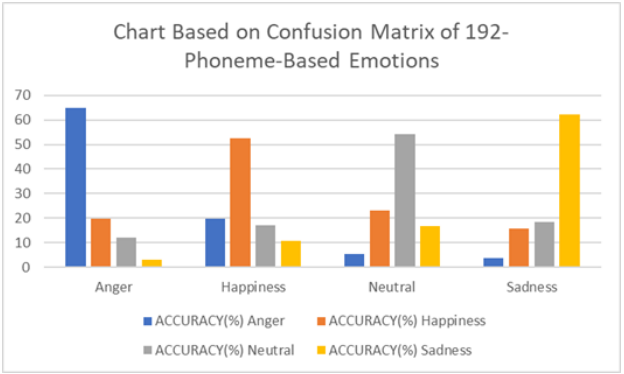

Fig. 22. Accuracy using 1D ConvLSTM-RNN

Experiments are performed on the IEMOCAP database and four emotions are classified namely anger, happiness, sadness, and neutral. Even though CNN based techniques generated a great result in SER, it was discovered that the performance is not so good in terms of the IEMOCAP database. This model has only one LSTM layer comprising of nodes including the sigmoid activation function. During the result phase, the softmax activation function was taken into consideration for obtaining odds of 4 emotions, 36 phoneme-class based emotions, and 192 phoneme-based emotions respectively [22].

The above graph represents the precision obtained from the proposed technique using,

1- dimensional ConvLSTM-RNN consisting of N = 4, 36 and 192 phoneme emotions.

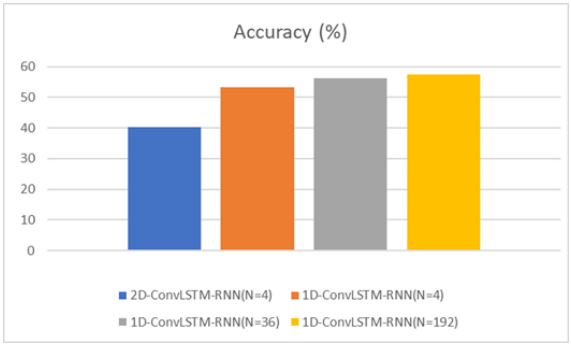

Fig. 23. Accuracy comparison

2D-ConvLSTM-RNN is considered as the baseline system for recognizing emotions. The figure represents the accuracies of both 1D-ConvLSTM-RNN as well as 2D-ConvLSTM-RNN respectively. It is clear from the results shown that with or without using phoneme-based emotions probabilities gave higher performance than baseline.

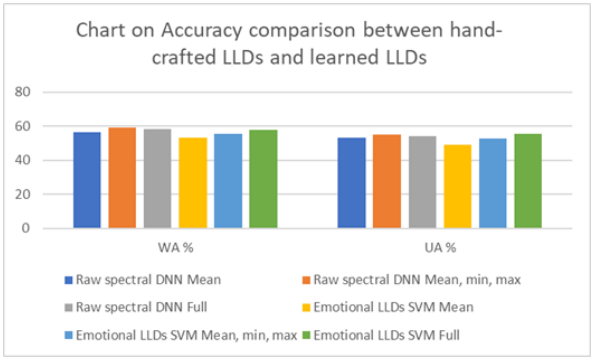

Fig. 24. Comparison between hand-crafted and learned LLDs

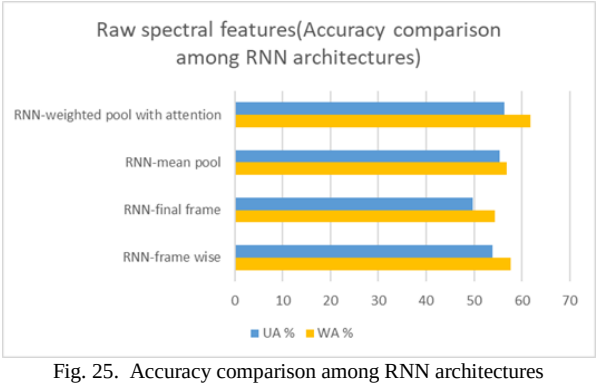

Fig. 25. Accuracy comparison among RNN architectures

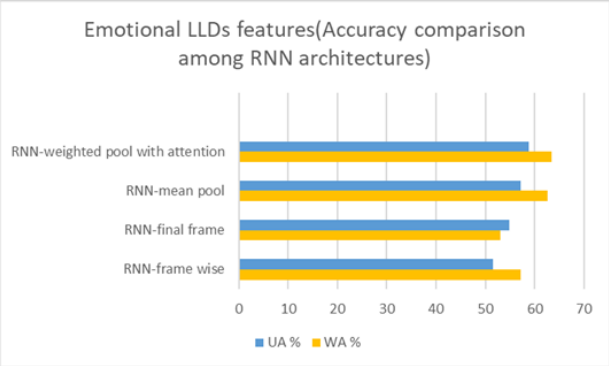

Fig. 26. Accuracy comparison among RNN architectures with LLDs feature

In table above, results are generated when custom-made LLDs central focus is on exploring temporal aggregation tasks with RNN layers. In certain situations, the results obtained from training done frame-wise and final-frame have low correctness. During such situations, Mean-pooling can provide significantly better results. In this attentive based approach, the accuracies obtained are +5.7% and +3.1% in WA and UA compared to the conventional SVM approach.

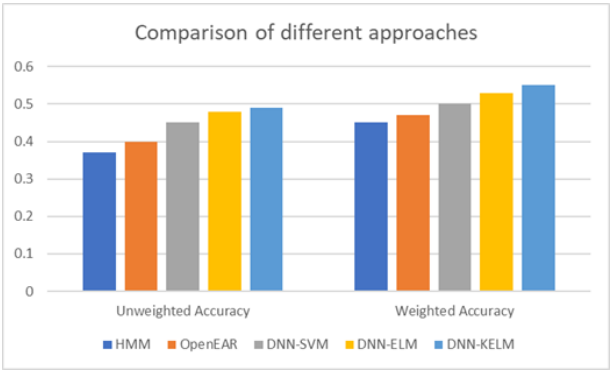

Fig. 27. Comparison among different approaches

The IEMOCAP dataset is used to evaluate the approach of dep neural network using extreme machine learning. This dataset consists of both audio and visual data from 10 actors. The DNN comprise of three hidden layers. The results obtained from this method are compared with the conventional HMM method, OpenEAR (uses SVM), DNN-SVM, DNN-ELM, DNN-KELM. The below table represents that comparison in two measures: WA and UA. Overall, the suggested method mentioned here outperforms the other approaches with 20% precision improvement for both weighted (0.451 -> 0.543) as well as unweighted (0.402 -> 0.482). In addition to this, the training time required for ELM is 10 times faster in comparison with SVM [24].

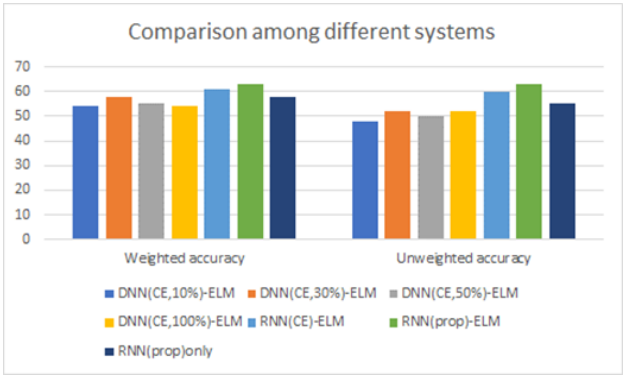

Fig. 28. Comparison between different systems of DNN and RNN

The above-given table shows the experimental results of the Bidirectional LSTM system. The first four bars represent the performance of DNN-based systems. For the following three bars the results of the proposed method are represented. RNN(CE)- ELM (Extreme Learning Machine) means that the DNN network is replaced with BLSTM-RNN with cross-entropy (CE). RNN (prop)-ELM in the table shows the system where the proposed method is implemented. By comparing the accuracies, it is clearly seen that there are improvements in the proposed system for both weighted accuracy (57.91% -> 62.85%) as well as un-weighted accuracy (52.13% -> 63.89%). Approximately the weighted accuracy shows increase of 5% and un-weighted accuracy shows around 12% [25].

Table 1
Comparative Table

Published Year	Name	Technology Used	Dataset	Strong Point	Weak Point
April 2012	Speech Emotion Recognition Using Support Vector Machine [7]	Basic SVM	Berlin Emotional Database (EmoDB and a self-taught Chinese Emotional Database	MFCC+MEDC+Energy features give the best accuracy among various spectral and prosodic features.	If this system applied to actual-time SER then accuracy is not satisfactory
Dec 2015	Efficient Speech Emotion Recognition using Binary support Vector Machines & Multiclass SVM [8]	Binary SVMs and Multi-Class SVMs	Berlin Emotional Database (EmoDB)	The N Binary SVM for N-Emotional states and Multiclass SVM combined in a fusion model has lowered the order to O(N) from the previous order of O(N^2).	If features selected are irrelevant and redundant then the classifier performs poorly.
May 2013	SVM Scheme for Speech Emotion Recognition using MFCC Feature [13]	Three Stage hierarchical SVM	Berlin Emotional Database (EmoDB)	Higher accuracy of MFCC features using Linear and RBF kernels with 3-stage hierarchical SVM	Only MFCC features are considered as well as the sigma value is also set to lowest i.e. 1

Jan 2019	Decision tree SVM model with Fisher feature selection for speech emotion recognition [16]	Decision Tree as well as Fisher Feature Selection	CASIA Chinese speech emotion collection and Berlin Emotional Dataset	It effectively decreases the emotional confusion problem and refines the overall emotion recognition rate	Confusion on Fear and Sad emotion still remains relatively high for CASIA
April 2020	Speech Emotion Recognition using Neural Network and MLP Classifier [17]	Neural Network and MLP Classifier	RAVDESS	It will ease up the process of feature extraction from audio script because the dataset used is being labelled according to decimal encoding.	Using neural networks into SER is sometimes computationally expensive and takes longer duration for development.
2016	Speech Emotion Recognition using Convolutional and Recurrent Neural Networks [27]	Time Distributed CNN	Berlin Emotional Dataset	It applies the same layer to several inputs. So, it is efficient and saves time.	It provides accuracy though when combined with other approach it would be more useful
June 2018	Speech emotion recognition based on convolution neural network combined with random forest [19]	CNN-Random Forest	CASIA EmoDB	It minimizes the interference of noise and provides more pure speech samples.	For data and attributes with multiple values, the weights generated by random forest classifier are not credible.
Nov 2014	Speech Emotion Recognition Using CNN [20]	Semi CNN Approach	SAVEE+EmoDB+MES+DES	This approach leads to robust performance in complex situations (environment distortion)	Sometimes problems persist if the image is positioned at different angles and background.
Feb 2017	Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network [21]	Deep CNN-LSTM	Berlin EmoDB	Short and discriminative features are learning automatically which is a basic step for SER.	In case of fear and happy emotions it often gets confused among the spectrograms
Dec 2017	Speech emotion recognition using convolutional long short-term memory neural network and support vector machines [22]	ConvLSTM-RNN	IEMOCAP Database	This approach focuses on phoneme-based feature extraction while conventional techniques do not focus on that.	Although it provides better results, for more accuracy to be obtained SVM parameters can be used.
Mar 17	Automatic speech emotion recognition using recurrent neural networks with local attention [23]	RNN with Local Attention	IEMOCAP Database	A combination model of bidirectional LSTM with new pooling strategy helps to focus on the salient features	This approach still possesses difficulty of existence of silence frames and non-emotional speech utterance. If these frames are present it will distort the orientation
Sept 14	Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine [24]	DNN-Extreme ML	IEMOCAP Database	It is effective and possess greater effectiveness when sample used for training is small	It often causes the problem of overfitting.
Sept 2015	High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition [25]	RNN-BLSTM	IEMOCAP Database	In conventional approaches all frames are mapped into the same label while this model gives the opportunity that all frames are classified individually.	It sometimes cannot process long sequences and the training task is very tedious.

4. Conclusion

In this paper we tried to discuss the different approaches through which speech emotion recognition for audio file type can be implemented and have provided the comparative analysis of these different approaches. We discussed basically 4 main classifiers i.e. SVM, MLP, RNN-LSTM and CNN, these classifiers are said to be most accurate than others. We later in this paper have made a comparative table based on the results of each model.

Acknowledgement

We are grateful to all who gave their invaluable feedback and comments in this. The work was performed in part of as academic contribution under guidance of Mrugendra Rahevar, an Assistance Prof., who is associated with CHARUSAT University, Changa.

References

[1] Dong Yu and Li Deng. Automatic Speech Recognition. Springer, 2016.
[2] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition volume 14. PTR Prentice Hall Englewood Clis, 1993.
[3] Louis Ten Bosch. Emotions, speech and the asr framework. Speech Communication, 40(1-2):213-225, 2003.
[4] Thomas S Polzin and Alex Waibel. Detecting emotions in speech. In Proceedings of the CMC, volume 16. Citeseer, 1998.
[5] Shahsavarani, Somayeh. (2018). Speech Emotion Recognition using Convolutional Neural Networks.
[6] Albert Mehrabian et al. Silent messages, volume 8. Wadsworth Belmont, CA, 1971.

[7] Speech Emotion Recognition Using Support Vector Machine Yixiong Pan, Peipei Shen and Liping Shen Department of Computer Technology Shanghai JiaoTong University, Shanghai, China.
[8] N. R. Kanth and S. Saraswathi, "Efficient speech emotion recognition using binary support vector machines & multiclass SVM," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, 2015, pp. 1-6.
[9] Chris. "Creating a Simple Binary SVM Classifier with Python and Scikit-Learn –." Machine Curve, 5 July 2020. www.machinecurve.com/index.php/2020/05/03/creating-a-simple-binary-SVM-classifier-with-python-and-scikit-learn.
[10] Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. (2005). A database of German emotional speech. 9th European Conference on Speech Communication and Technology, vol. 5. 1517-1520.
[11] Kowshalya, Meena & Madhumathi, R. & Gopika, N, Correlation Based Feature Selection Algorithms for Varying Datasets of Different Dimensionality. Wireless Personal Communications, vol. 108, 2019.
[12] Github.Io. http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/. Accessed 12 Aug. 2020.
[13] A., Milton & Roy, S. & Selvi, S. (2013). SVM Scheme for Speech Emotion Recognition using MFCC Feature. International Journal of Computer Applications, vol. 69. 34-39.
[14] Emo-DB. emodb.bilderbar.info. Accessed 5 Aug. 2020.
[15] Practical Cryptography. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.
[16] Sun, L., Fu, S. & Wang, F. Decision tree SVM model with Fisher feature selection for speech emotion recognition. J Audio Speech Music Proc. 2019, 2 (2019).
[17] Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama, "Speech Emotion Recognition using Neural Network and MLP Classifier."
[18] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).
[19] Zheng, Li & Li, Qiao & Ban, Hua & Liu, Shuhua. (2018). Speech emotion recognition based on convolution neural network combined with random forest. 4143-4147.
[20] Huang, Zhengwei & Dong, Ming & Mao, Qirong & Zhan, Yongzhao. (2014). Speech Emotion Recognition Using CNN. MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia. 801-804.
[21] Badshah, Abdul & Ahmad, Jamil & Rahim, Nasir & Baik, Sung. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. 1-5.
[22] N. Kurpukdee, T. Koriyama, T. Kobayashi, S. Kasuriya, C. Wutiwiwatchai and P. Lamsrichan, "Speech emotion recognition using convolutional long short-term memory neural network and support vector machines," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1744-1749.
[23] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2227-2231, doi:
[24] Han, Kun & Yu, Dong & Tashev, Ivan. (2014). Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.
[25] Lee, Jinkyu & Tashev, Ivan. (2015). High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition.
[26] Zhao, Jianfeng & Mao, Xia & Chen, Lijiang. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control. 47. 312-323.
[27] W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, 2016, pp. 1-4.