Speech Emotion Recognition System using SVD algorithm with HMM Model

Divya Sharma, Amarjeet Pal Cheema, Koushik Reddy K, Kusalanatha Reddy C, Badri Ram G, Avinash G, Pavan Kumar Reddy
Department of Electronics and Communication Engineering, New Horizon College of Engineering, Bangalore
E-mail : er.divyasharma@gmail.com, ece.amarjeet@gmail.com, koushikreddy949@gmail.com, kushalnathreddy2002@gmail.com, galijerlabadriram5978@gmail.com, avinashroyal2002@gmail.com, ankepallipavankumar138@gmail.com

Abstract— From many years there has been gaining interest in the field of SER using Matlab. SER states the emotional state by analyzing the input speech. SER has a simple pattern and also including feature extraction, feature matching, classification and database. Here, from the input speech by using algorithm features are extracted by using some models the feature matching takes place. By this process we use to analyze the characteristics of the input speech signal. Hence, the system recognize the state of emotion. The system states some the of emotions: Angry, Boredom, Anxiety, Disgust ,Happiness, Neutral, Sadness. The main purpose of this paper is to give survey on two of the algorithms using HMM model with different speech emotion databases. There are several audio features for extracting are available. And also various classifiers are available. The most popular models are Hidden Markov model(HMM), Vector Quantization(VQ), Gaussian Mixture model(GMM), Deep Neural Networks (DNN), Neural Networks(NN) and Artificial Neural Networks(ANN). There are several speech emotion databases included Berlin database. Hence, we reviewed some of the models will be discussed in this paper.

Keywords— Hidden Markov model(HMM), Modelling, MATLAB, feature extraction, feature matching, classification, databases.

I. INTRODUCTION

Now a days Speech Emotion Recognition(SER) has been gaining interest and is one of the topics which has been continuously researched in speech processing. In the field of Human Computer Interaction (HCI) it is the most important topic. Recent technologies like artificial intelligence, IOT and Machine learning have scaled up communication and automation industries [1][2][3][4]**.** It research was started from the late fifties. From late fifties it indicated that the growth of publication papers are increasing every year. It was widely used in many applications and also applied in so many fields such as security, education, human computer interaction, teaching, entertainment and so on. It show the desired information about the emotional/mental state of an individual which had a way to new research field called automatic emotion recognition. The majority researchers are interested in SER because when compare to other signal speech signals that are more readily acquired and a good source for affective computing. As now SER was emerging across various fields such as artificial intelligence and also an important topic in signal processing and pattern recognition. To detect the state of emotion of the individual from his input speech signal is the main objective of SER system. SER has three main objectives for getting successful output one is to use a good database with many utterances and second one is choice of algorithm for extraction of features and the last one is classification for feature matching. In database there are several individual's speeches and each has their own emotional state. So, many researchers have utilized database with categorized emotions.

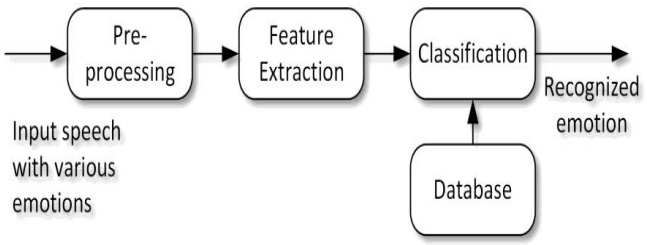

Fig.1. Speech Emotion Recognition System

With more utterances then more accuracy we be will getting for attaining emotional state of the unknown input speech. The other objective is feature extraction. It plays a crucial role in the SER why because when we extract features the features contain information about speech signal such as pitch, frequency formant. Most researchers prefer to use best feature sets for storing more information. If the features increases then the dimensions also increases. The feature selection model is also important to reduce the dimensions. The last objective is the classification. It classifies the raw data in the form of utterances into a particular emotion from the data. In recent years, so many classification algorithms have been proposed and also different types of classifiers are also merged for SER such as Hidden Markov Model(HMM), GMM, VQ, DNN, NN and ANN. By both the feature extraction, feature selection helps in improving learning performance and less complexity with better models helps in decreasing storage.

II. LITERATURE SURVEY

Several research has be done natural language processing for various applications and numerous classifiers have implemented for the accuracy and also design of the system for recognizing the emotional state. The review based on classifiers, databases has been presented in [5][6][7].

There are so many databases that are available which some contains large numbers of audio files and ravadness is one of the high quality formats which a review based on the that database is presented in [8].

There are so many approaches that are published for ser system using different techniques in which the feature extraction process plays main role in ser system. Every year new techniques are published in which one of those is by TEKO model which is presented in [9].

The study [10][11] proposes the different types of models implemented in ser system and the general architecture of the ser system for implementing using different models. In this they presented several models like HMM, neural networks and other models. And review of some classifiers.

There are so many models for the implementation of ser systems also for classification. There are some techniques for extracting features based on some feature decomposition method which can increase the accuracy using less acoustic features in which some of them are svd, lda and also combination with different types of classifiers and their accuracies are presented in [12].

Now a days so many hybrid models have been implemented like LSTM- CNN, LSTM-RNN, CNN-LSTM-DNN etc are some hybrid classifiers for the rate of recognition and accuracy with different databases as study [13] proposes DNN model.

This study [14] proposes the different types of classifiers and the their accuracies using different types of databases to get accuracies. In those CNN_LSTM with IEMOCAP has got accuracy of 95.89% and the SVM eith EMOdb has got accuracy of 95.3% and DNN with EMOdb has got accuracy of 96.97%.

In SER system fature extraction is main task. Which means the data we are extracting from the audio files. There are so many algorithms that are used in feature extraction. This study [15] proposes the Unsupervised Feature Extraction Using Singular Value Decomposition. SVD technique is matrix factorization method used in feature extraction.

Study [16] proposes about feature selection on ser from the speech and not only this and so many papers [17] published on speech emotion recognition based on factors that affect and also about the reduction of unwanted noise that increases the computational cost.

Study [18] proposes a survey on ser in natural environment also the models that are used in that environment and their disadvantages in the context of the speaker, type of environment the speech signal is recorded.

III. PROPOSED MODEL

This section presents the proposed system. The SER flow/Block diagram of proposed model is shown in Fig. 2. The steps involved are discussed below.

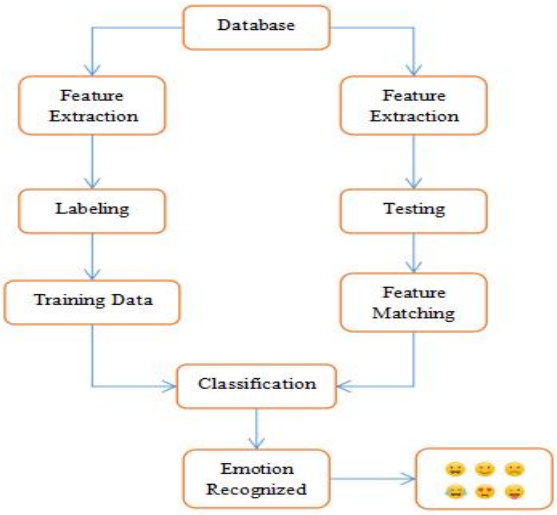

Fig.2. SER flow

A. *Selection of databases:*

Database is a collection of raw data which is stored in file formats.Raw audio data is stored in the form of wav file format. In this paper methodology we use some databases for SER and showed the accuracy of each database.For example we use Berlin database for SER in the form of wav files. It contains more than 500 utterances by different speakers with different texts and emotions.In that there are 10 different speakers with different genders and ages. And also it contains 10 different texts and with seven emotional states that are shown in below table.

TABLE 1: Databases Preview

Database	**Emotions**
English database	Angry, Sadness, Neutral
Urdu database	Angry,Anxiety,Disgust,Happiness, Neutral, Sadness
Berlin database	Angry,Anxiety,Boredom,Disgust,Happiness, Neutral, Sadness

B. *Creating an audio file:*

A format used for storing audio data on system is called as audio file format. There are several types of audio file formats that are determined by their file extension. The audio file can be compressed or uncompressed to lessen the size of the file. The data can be embedded in a container format though it can be in raw audio data. It is important to differentiate between the container and audio coding format. Waveform audio file format is a format of an uncompressed audio file format. It mostly contains of uncompressed audio in PCM format which is compatible with windows. It stores raw audio data and also it does not requires processing and retains its original data. Wav file format is a Lossy compressed audio format (LCAF). LCAF is a file format which use audio data compression to

lessen greater portions in size of file by removing some inaudible sounds in audio file. Data gathering is one of the most important aspects of creating a model, that is because the quality and quantity of the dataset impact the output of the model to a great extent.

C. *Feature Extraction:*

We use wav file and svd algorithm used for feature extraction. At first we load audio files in the form of wav files and extract features from audio files also using svd algorithm for feature extraction.

i) SVD algorithm:

Singular Value Decomposition (SVD) is a method used for decomposition of data matrix. It is a matrix factorization method expresses a data matrix as into orthogonal matrices based on singular values on its diagonal in decreasing order. To compute both singular values and vectors of original matrix \

$$[U,S,V] = \text{svd (original matrix)}$$

The SVD theorem states that

$$Amxn = UmxmSmxnV'nxn$$

Let the matrix A is mxn matrix. It expresses an mxn matrix as

$$A = U*S*V'$$

Here, U, V are mxm and nxn orthogonal matrices which are left and right singular vectors for corresponding singular values. It is a eigenvalue method. This function removes extra rows or columns of zeroes from the diagonal matrix of singular values S, along with the columns in U or V that multiply those zeroes in the expression A = U*S*V'. By doing this removing of zeroes and columns can decrease storage necessity without comprising the accuracy also decreases execution time.

D. *Classification:*

Generally, all the events that are upto the classification are not visible,there will be some hidden events. HMM is generally a Markov chain whose internal states are observed only through some probabilistic functions. The hidden states of the model helps in capturing the attributes of the data. The goal of HMM is to adapt a Markov chain by observing its hidden states. At the time of classification,speech signal is taken and the probability of every single speech signal in the model are calculated and compared it to the test sample.For better understanding of HMM model. Let's say that there are two friends M and N. M mood changes according to the breakfast he eats in the morning. N looks at M and observes how his mood changes daily. N does he not know the breakfast taken by M, but he predicts it by the mood of the M. Breakfast of M is a hidden event but it can be predicted by known events. This implies that it allows us to predict a sequence of hidden variables from the set of observable variables. It describe about sequence of events. It is also called as sequential generating probabilistic model. HMM model tells about both observable events and hidden events which we don't observe them directly. This proposed methodology states that a database has to be selected first and then the utterances has to be converted into wav file format. Then we train the data and using SVD algorithm the features are to be extracted and every file will be labeled. Now using HMM we classify the test samples.

IV. RESULTS & DISCUSSION

As now we discussed about the models and algorithms as for feature extraction we wav files, SVD algorithm and for classification we HMM model. SVD algorithm and HMM model are designed and implemented on matlab. We extract features from audio files in database and train data. After labeling the trained data, A test speech sample is loaded from database which we trained and extract features from speech signal which will be stored in the workspace then compare it with the trained data by HMM model as classifier.

A. *Data Sheet*

Now, let us check the recognition rate of emotions of databases by using svd and hmm we get accuracy of 65% from Berlin database

TABLE 2: Recognition Rate of Emotions of Different Databases
(A:Angry, F:Anxiety, B:Boredom, D:Disgust, H:Happy, S:Sad, N:Neutral)

Recognition Rate							
Database	**A**	**F**	**B**	**D**	**H**	**S**	**N**
English	65%	-	-	-	-	55%	60%
Urdu	70%	-	-	-	65%	60%	70%
Berlin	85%	70%	60%	60%	75%	65%	80%

From the above table we can say that the Berlin database achieve high accuracy than the these two databases. Here, we applied different databases with different languages and utterances with different emotions. When compared to Berlin database the other database have more utterances with different emotion. By this we can say that the higher the utterances the more the accuracy increases.

B. *Graphical Analysis*

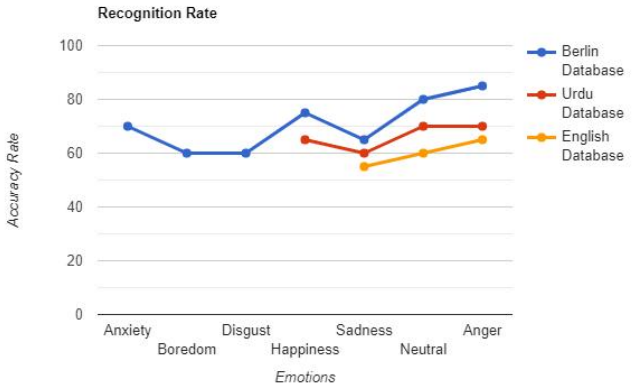

Fig.3 Recognition Rate of Emotions in each database

Now, we discuss about the accuracy of classifiers and databases. We report the recognition accuracy of using SVD and PCA classifiers with some databases. Apart of that SVD and PCA are both same but by some numerical controversies SVD is more stable than PCA due to SVD doesn't use co-variance matrix. But that is not the main thing PCA uses SVD in that process. By using the Berlin database to achieve results. By using Berlin database with the SVD we achieve results of about more than by using PCA when compared to SVD we get accuracy of 72%. By the above results by using SVD we get better results when compared to other.

C. *Frequencies of Different emotions*

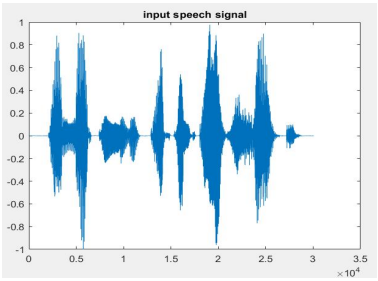

Fig. 4 Speech signal of Anger

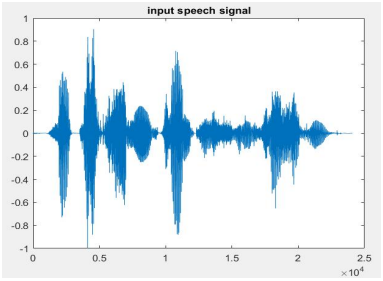

Fig. 5 Speech signal of Anxiety

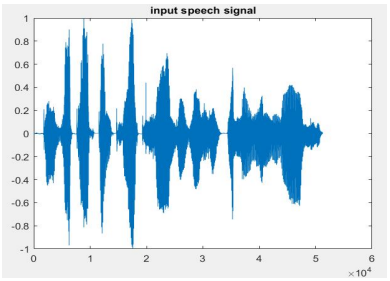

Fig. 6 Speech signal of Disgust

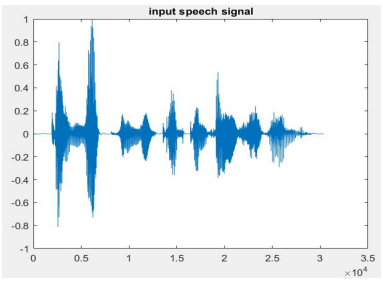

Fig. 7 Speech signal of Happiness

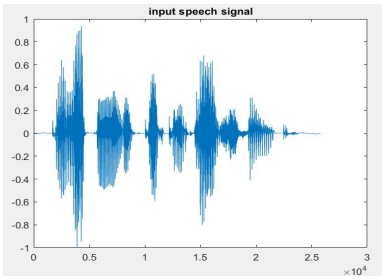

Fig. 8 Speech signal of Neutral

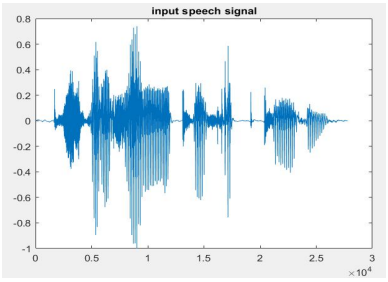

Fig. 9 Speech signal of Sadness

The above graphs were frequency speech signals of different emotions from the Berlin database. The features extracted in speech signal are pitch, frequency from the input speech signals. As now we discussed about how the results displayed after the program.

V. CONCLUSION

In this study, we presented an speech emotion recognition (SER) system using SVD algorithm with HMM model using three databases with seven emotions. By using SVD we extracted features from three types of databases and by using classifier we do feature matching. By this study we know how the classifiers and features impact on the accuracy of SER system. In feature selection highly discriminant features are selected and compared with trained data features by test sample features. SER reported that Berlin database has the high accuracy of emotion recognition when compared to other. Hence, this study proposes using the SVD method to extract features from the audio file and HMM model is used for the design of classification. Therefore, this paper presents an easy way to approach for classifying features from files for distinguishing features by SVD. This study gives the improved results when compared to old methods and it shows that our proposed model can be applied to recognition of emotion.

REFERENCES

[1] D. Sharma, S. Jain and V. Maik, "Energy efficient clustering and optimized loadng protocol for iot," Intelligent Automation & Soft Computing, vol. 34, no.1, pp. 357–370, 2022.

[2] Application of artificial intelligence in energy efficient hvac System design: a case study P Adhikary, S Bandyopadhyay, S Kundu - ARPN journal of Engineering and Applied sciences, 2017.

[3] D Sharma, I Mishra, S Jain, A Comprehensive Analysis of Security Requirements and Approaches For Internet of Things, International Journal of Advanced Computer Technology, vol 4, issue 6, pp 105-109.

[4] M. Dhivya and T. Parameswaran, “Smart scheduling on cloud for IoT-based sprinkler irrigation,” International Journal of Pervasive Computing and Communications, 2020, http://dx.doi.org/10.1108/IJPCC-03-2020-0013.

[5] Mehmet Berkeehan Akcay, Kaya Oguz (2020). “Speech emotion recognition: Emotional models databases, features, prepossessing methods, supporting modalities and classifiers”.

[6] Teddy Surya Gunawan, Muhammad Fahreza Alghifari, Arman Morshidi, Mira Kartiwi. (2017). A Review on Emotion Recognition Algorithms using Speech Analysis.

[7] M. Maithri, U. Ragavendra, Anjan Gudigar, Jyothi Samanth, Prabal Datta Barua, Murugappan Murugappan, (2022). “Automated emotion recognition: Current trends and future perspectives”.

[8] Livingstone SR, Russo FA (2018). “The Ryerson Audio-Visual Database of Emotional speech emotion and Song (RAVDESS)”.

[9] Leila Kerkeni, Youseef Serrestou, Mbarki, Raoof, Ali Mahjoub, Cleder (2019).“ Automatic Speech Emotion Recognition”.

[10] Saliha Benkerzaz, Youssef Elmir, Abdeslem Dennai (2019). “A Study on Automatic speech emotion Recognition”.

[11] Harshavardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey, Siddharth Swarup Rautaray, (2020). “A comprehensive survey and analysis of generative models in machine Learning”.

[12] Palani Thanaraj Krishnan, Alex Noel, Vijayarajan (2021). “Emotion classification from speech signal based on empirical mode decomposition and non-linear features”.

[13] Nithya roopa S, Prabhakaran M, Betty.P,(2018). “Speech Emotion Recogntion using Deep Learning”.

[14] Raoudha YAHIA CHERIF, Abdelouahab MOUSSAOUI, Nabila FRAHTA, Mohamed BERRIMI(2021).”Effective speech emotion recognition using deep learning approaches for Algerian dialect”.
[15] Kourosh Modarresi (2015) “Unsupervised Feature Extraction Using Singular Value Decomposition”.
[16] J.Rong,Y.P.P.Chen, “Acoustic feature selection for automatic Emotion Recognition from speech”.
[17] Zhe Chen, Yanmei Zhang, Junbo Zhang, Rui Zhou, Zhen Zhong, Chaogang Wei, Yuhe Liu Jing Chen,(2021).”Cochlear Synaptopathy: A Primary Factor Affecting Speech Recognition Performance in Presbycusis”.
[18] Shah Fahad, Ashish Ranjan, Jainath, Akshay Deepak,(2021). “A survey of speech emotion recognition in natural environment”.
[19] S. S. Narayanan, (2005). “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process.