IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 2, April, 2022, pp. 603-609
Journal homepage: http://iieta.org/journals/ts

Emotion Recognition from Speech Utterances with Hybrid Spectral Features Using Machine Learning Algorithms

Kogila Raghu[1,2*], Manchala Sadanandam[1]

[1] Department of CSE, Kakatiya University, Warangal 506001, TS, India
[2] Department of CSE, Geethanjali College of Engineering and Technology, Hyderabad 501301, TS, India

Corresponding Author Email: kraghu.cse@gcet.edu.in

https://doi.org/10.18280/ts.390222

Received: 20 January 2022
Accepted: 16 March 2022

Keywords:
SER, speech prosody, feature extraction, SVM, MLP

ABSTRACT

Speech Emotion Recognition is always a complicated task in the domain of Speech Processing Research, though many research works have been done. The first and foremost challenge of SER is to selecting the Speech Emotion Database (Corpora), then extracting the related speech features and finally construct an appropriate Classification model. An effort is created during this work to discover the speech prosodies, spectral and combination of features with their dynamism to illustrate and classify the emotions of speech signal. The intrinsic or fine variations of speech samples are combined with the static delivery parameters within the Speech Emotion Recognition (SER) to refine the accuracy. The work in this paper, carried out the experiments on RAVDESS, IIITH IIITH-TEMD and our developed Database of native language DETL (Database for Emotions in Telugu Language) Speech Emotion Databases. This work extracted features like MFCC and Hybrid Features (MFCC+ΔMFCC+ΔΔMFCC) then finally applied those individual features and Combination of Features to different Classification models like SVM and MLP. We have got approximately 75%, 78% and 81% of accuracy for MLP with hybrid combination features on the above Databases respectively.

1. INTRODUCTION

Speech is very natural, normal, convenient and suitable way for human beings to communicate among them. In this information era, Automatic Speech Recognition (ASR) systems technologically advanced and became reachable as Human-Machine Interface. ASR takes Human Speech as the primary input. The speech signal contains features like Prosodic (Pitch, ZCR, F0), Spectral (MFCC, LPCC) and Hybrid Features used in various applications depending on the requirements. ASR model contains two components as acoustic model and language model. In ASR, Emotion Recognition is very complex. Emotion is a great sign to understand about individual of human being or mental state or developed product or any, So SER used in various applications like Medical Diagnosis, Call centers, Banking, Self driving cars, voiceless talk, etc.

ASR systems with their improvements enhances like the speed of hardware, parallel processing, and various technologies have led to several sensible speech applications like Siri, Ok Google, Alexa, etc.

Most of the current and existing ASR systems don't seem to be strong, once positioned in sensible environments. Performance and Working of those systems is exaggerated, once there's a gap between the training and testing phases. Making of ASR systems to robust is an important task. The primary objective of (robust) ASR is to handle the mismatch of training and testing phase's downside expeditiously. The sources of inconsistency that origin of mismatches represent in each intrinsic (speaker), extraneous (non-speaker) variation's, and they are presented in Figure 1.

The Gap Analysis in ASR system is presented in Figure 1. In ASR systems the mismatch is occurred because of gap between Training and Testing Phases. The different Sources of mismatches are a) Speaker related and b) Non-Speaker related mismatches. The Speaker related mismatches occurred due to Speakers pitch, stress, dialect, emotion, speaking rate etc. Where as in the case of Non-Speaker's related mismatches occurred due to microphone, background noise, drop out, echoes, etc. Because of the non-speaker connected variability's, mainly mismatch is produced.

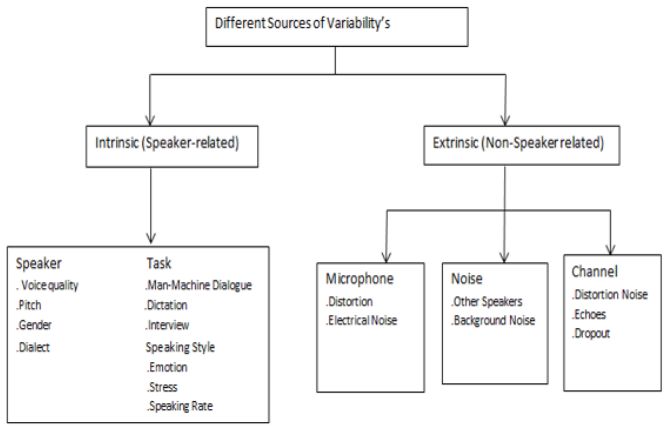

Figure 1. Gap analysis in early ASR system

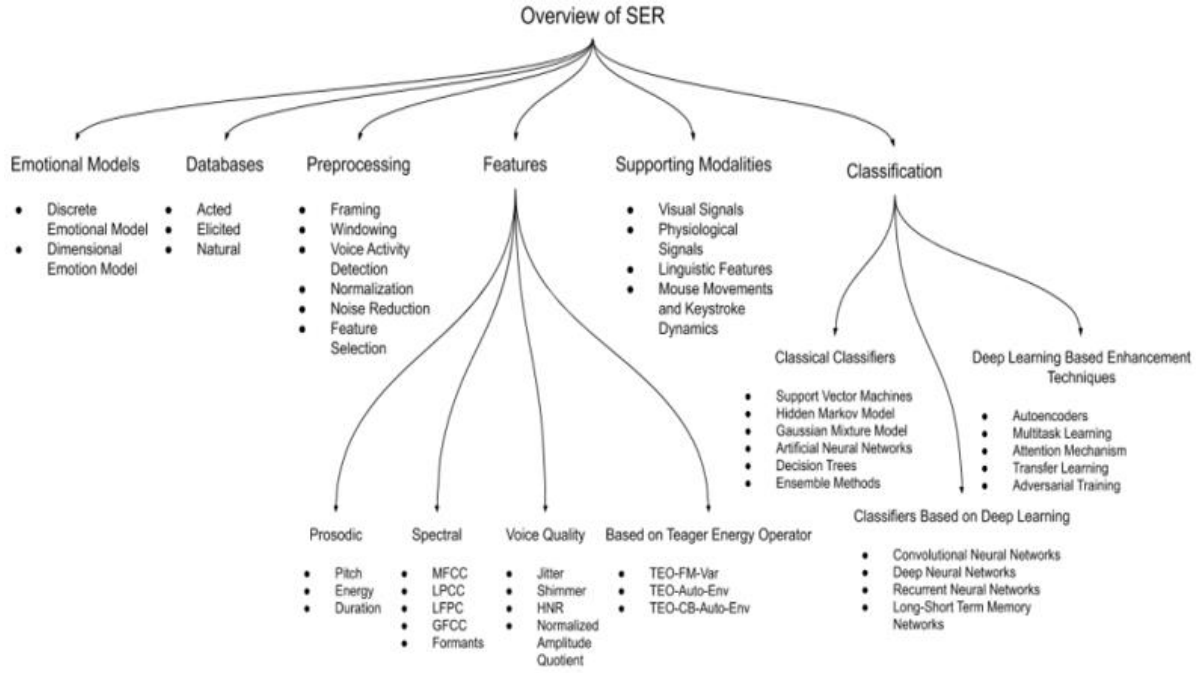

Figure 2. Bird's eye view of speech emotion recognition

Speech signal transmits phonological data (message), however it offers additional data from the speaker's Utterances like gender, age, accent, emotions, etc. These data sources of Speech Signals are accountable for variations in individualities, Characteristics, features and discriminations. Good performance of any SER system is established within the emotion settings because of convinced supply variations. The different sort of Speech that contains emotions of human beings comes into the category of speaker-related variations (intrinsic), and it is a massive task to work with the present natural, emotional condition ASR systems.

Emotion Recognition is a kind of well-known in each domain of human lifecycle and is manifested from face features, vocal translation, and traditional artifacts like photos or Speech files (music). Speech relics a natural mode, most suitable notably once the Speech Communication takes place via mobiles. SER has been around for over twenty years [1] and used applications of human-computer interaction [2], like as robots, call centers and psychological assessment. Though it's several applications, emotion recognition could be a challenging job, for the reason that emotions are individual and subjective. Generally Emotions are evaluated by sensitivity in different Human Beings.

SER system is a group of practices that processes speech signals and categorize those speech utterance signals to observe emotions fixed in them. Take a look on the Overview of SER, will isolated it into many individual domains, as illustrated in Figure 2. It might be useful to know emotions higher in order that the classification methods are often value-added. There are different styles and methods to categorize emotions, and its quiet and a complex open problem; but, the separate, high recognition rate models are mostly used. The assignment of recognizing emotions in speech remains troublesome because the performance of SER system models depends deeply on several factors [3]. They are

(a) Appropriate, consistent and robust features of emotions.

(b) A genuine information well-adjusted with samples, emotions etc., and

(c) An appropriate machine learning technique aimed at economical classifications related to emotions.

Among largely mentioned options, the spectral features used at frame level, whereas at the prosody features used at vocalization level are extracted. The present work during this area largely concentrates on prosodic speech that involves applied mathematics parameters in speech sample [4, 5]. Though, these explanations are static in the environments and don't take into account the dynamism as time passes throughout the spoken utterances which contains emotions. The native dissimilarities of the delivery information with time will give dynamic involvements in labeling emotional features and want to be discovered within the current state of affairs. This arranges the muse for this research work and inspires the researchers to portray and categorize the emotions of speech signal discrimination in each static and dynamics of speech. In the Introduction Part of this research work, ASR system and its Gap analysis explained clearly. The work mentioned with research objectives. And also clearly illustrated SER system in terms of various Databases, processing models, Features and Classification models. The Proposed work used various individual spectral features and robust hybrid features and different classifiers to improve the accuracy of the Speech Emotion Recognition.

The rest of the article is organized and directed below. Section 2 explained about Related Work. Section 3 of this paper describes the Proposed Methodology. Section 4 provides the Results and Discussions and in Section 5 conclusions given.

2. RELATED WORK

In the area of Speech Emotion Recognition (SER) Research, there are various Feature Extraction and Classification Techniques used from the last decade. SER identifies the emotions in Speech Utterances with shortest duration automatically. In the Literature Survey, different Feature Vectors and Process Models were used to analyze speech signal to identify the emotion in Speech of Human Beings.

Sahu et al. [6] reported a work on SER, the system used Feature Space Augment with GAN as Feature Extractor and

SVM as Classifier. The system got the accuracy of 60.29%.

Brochert and Dusterhoft [7] presented the research work on EMO-DB, extracted various features like formants, different frequency bands, spectral energy, jitter and shimmer. The work used SVM Classifier. The work got the accuracy of 70% for all emotions.

Han et al. [8] reviewed and presented a work on IEMOCAP-DB with extracted features like MFCC and delta features across time frames using Deep Neural Networks and got accuracy of 54.3% as recognition rate.

Mao et al. [9] contributed the research work on SAVEE DB, Berlin EMO-DB using CNN with their automated learning features, the research work got 73.6% for SAVEE.

Rong et al. [10] worked on their own created database of Mandarin i.e., 1 natural and 1 acted speech database. The work used Prosodic feature extraction techniques and k-NN Classifier. The work got 66.24% accuracy recognition rate.

Rao et al. [11] presented their work to recognize emotions from Telugu Speech database with prosodic feature extraction and SVM Classifiers and work got 66% average accuracy rate.

Kurpukdee et al. [12] worked on SER and the system extracted 1582 features using OpenSMILE tool. The system used SVM as classifier and got 65.13% as weighted accuracy.

Jiang et al. [13] reviewed and presented a work on SER using Deep Neural Networks with fusion systems to extract the features and SVM as the last phase classifier on IEMOCAP database, the system got accuracy of 64%.

Xia and Liu [14], reported a work on SER, the system extracted 1582 features from IS10 and applied SVM as classifier. The work go 60.9% as weighted accuracy and 62.4% unweighted accuracy.

3. PROPOSED METHODOLOGY

Recognition of Emotions from Speech is a two-step process [15]. Mainly it includes

a) Feature Extraction b) Classification

a) Feature Extraction: It is very important and foremost step in the SER system. The Features which are extracted must be capable to categorize the emotions of speaker. The Features must be more useful, reliable and effective for the process of SER. Obviously, the performance of the SER is varied due to the various Feature Extraction methods. In this work, Spectral Feature (MFCC) and combination of spectral features (MFCC+Δ MFCC+ΔΔMFCC)

b) Classification: It is the immediate next step of the Feature Extraction. Building a Classification Model takes two-steps. They are i) Training Phase ii) Testing Phase. In the Training step, we develop a Classifier reference model with the input speech features. In Testing step, we evaluate the testing samples feature vectors of speech against classification Model to predict the unknown speech utterance emotion. The Classification model establishes the class labels to various emotions like happy, sad, anger, fear and neutral. In our work, we used Classifiers like SVM (Support Vector Machines) and MLP (Multi-Layer Perceptron).

The chosen Speech Emotion databases, feature extraction methods and Classifiers employed in this proposed work are described during this section.

3.1 Feature extraction methods

In this section, used Feature Extraction techniques and their working procedure is given. In this work, Spectral features like MFCC, Delta1 MFCC and Delta2 MFCC are used and detailed description about each feature extraction is explained.

MFCC

Speech Processing related research work mostly uses MFCC (Mel Frequency Cepstral Coefficient) features, because it mostly resembles the human auditory system. Vocal track is exactly represented by MFCC. Every emotion contains its own spectrum, it can be identified by Spectral Features, so in this work MFCC features are used. Block Diagram of the MFCC process is in below Figure 3. The MFCC Process includes the following steps.

a) Apply Pre-emphasis to Speech signal to maintain high quality frequencies.
b) Apply Framing and Windowing to extract short duration and prominent characteristics of speech signal.
c) Apply DFT to convert speech signal from time domain to frequency domain
d) Apply FFT then pass the signal to filters like band-pass and apply logarithm to compute Mel.

$$Mel(f)=2595*\log_{10}(1+f/700) \quad (1)$$

where, f is frequency (Hz).

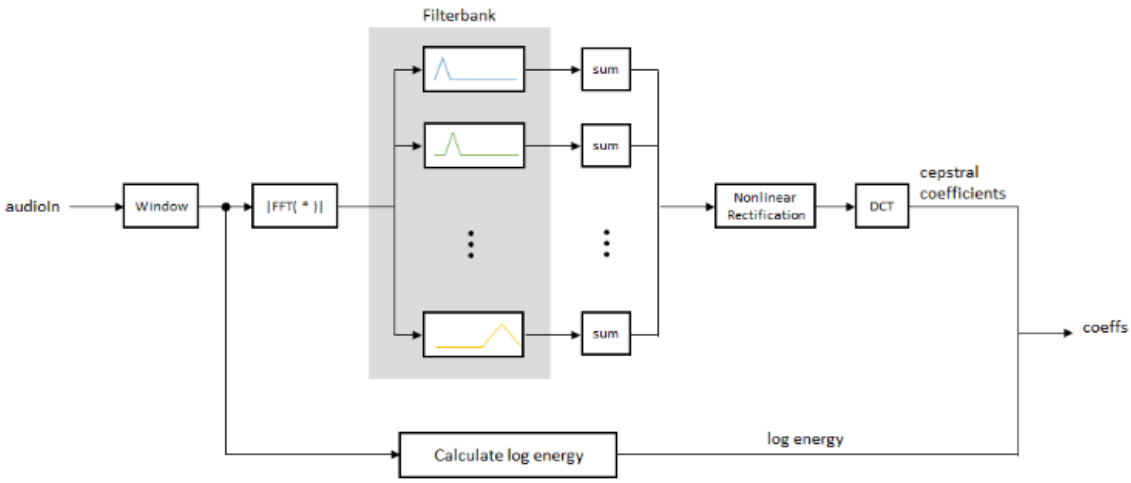

Figure 3. MFCC block diagram

Delta1 MFCC (or) Δ MFCC: It also same as the MFCC, but first order derivative of the MFCC is considered with respect to Time domain.

Delta2 MFCC (or) Δ Δ MFCC: It is also similar to MFCC, but second order derivative of the MFCC is calculated with respect to Time domain.

3.2 Databases

In this work, we used the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech Song) speech Database and IIITH-TEMD (Telugu Emotion Database).

RAVDESS: This is a simulated speech emotion database of English language speech [16]. It is created and developed by Ryerson University, Canada. There are 1440 emotional speech samples (.wav files) having sixty utterances for every class of emotions. There are 24 Actors in this Speech, Out of 24 Actors 12 Male and 12 Female Actors. The recordings are dispensed with a sampling rate of 44.1 kHz.

TEMD (Telugu Emotion Database): It is developed by IIIT-H (International Institute of Information Technology-Hyderabad). It is a simulated emotion Database, contains 38 native Telugu speakers aged between 21 years to 46 years. In TEMD19 speakers are professional actors (7 Male and 12 Female) and 19 speakers are Non-Actors (11 Male and 8 Female). On the whole Database is comprised of 5317 speech utterances but we used 458 speech files for the experimentation with general 5 emotions.

DETL (Database for Emotions in Telugu Language): It is our own created Database of native Telugu Language. It is also a simulated emotion speech Database, taken from mixed (male and female) Telugu speakers aged between 20 years to 70 years. The Speech Utterances are taken from different sources like YouTube, news, natural speech, movies etc. This Database contains mainly 5 categories of emotions namely Happy, Sad, Anger, Fear and Neutral; Each emotion category carries 100 speech files, overall 5*100 Speech files are created.

Different Telugu Speech Files with noise and without noise of various emotions as shown in Figures 4-9.

Speech File with Noise: "M.K College idhi, ikkadi gode kaadu, prathi adugu maa praanam. Inkokkadi cheyyi padithe irigi poddi."

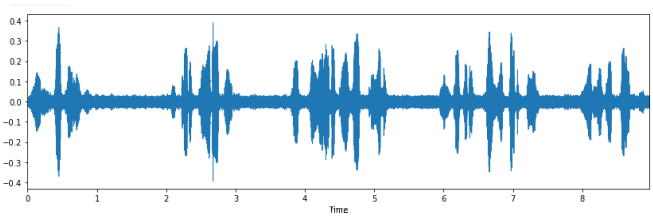

Figure 4. Wavelet plot for a sample speech file with noise

Speech File without Noise:

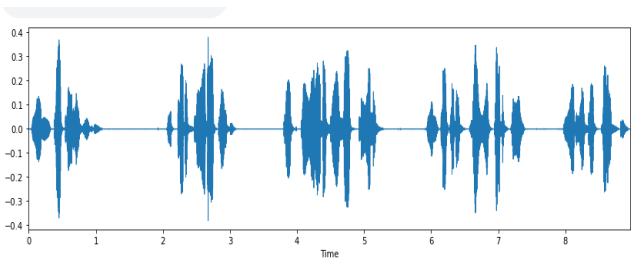

Figure 5. Wavelet plot for a sample speech file without noise

Angry Emotion: " Evarra Vaallu…"

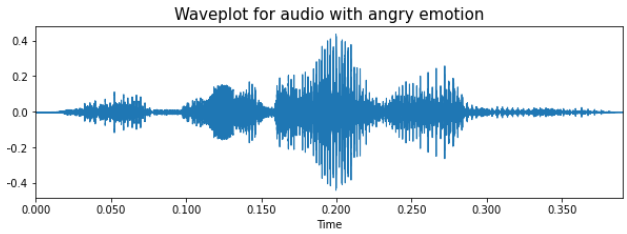

Figure 6. Wavelet plot for a sample angry speech emotion file

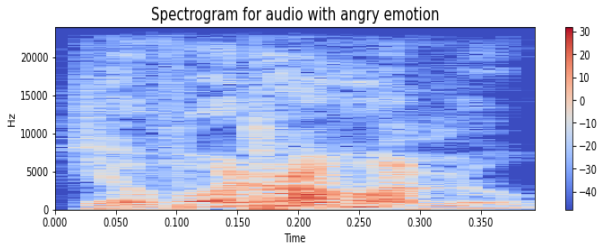

Figure 7. Spectrogram for a sample angry emotion speech file

Happy Emotion: "Hi Harsha, Anyhow good job, well-done."

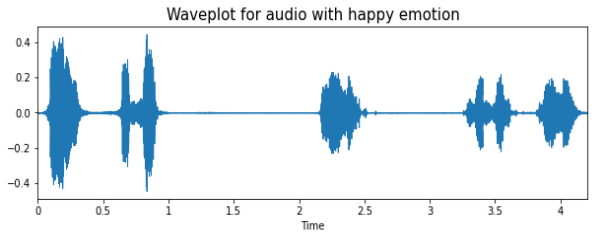

Figure 8. Wavelet plot for a sample Happy Emotion Speech File

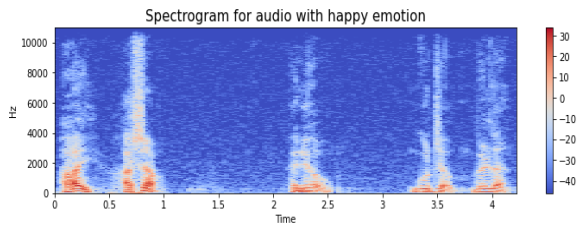

Figure 9. Spectrogram for a sample Happy Speech Emotion File

3.3 Classifiers

Classification model building takes two steps, comprising of Training and Testing. In Training Phase, a model is created as a reference model of the input speech file with feature vectors. In Testing Phase, the speech utterances of input are tested against training phase reference model to predict emotions of the unknown speech. In this work, we used SVM and MLP.

Support Vector Machine (SVM): SVM is a simple and well-organized Classification Algorithm, generally used in Pattern Recognition problems [17]. It was presented by

Vladimir Vapnik in 1995. The primary objective of this method is to find a function f(x), to identify Linear separable plane, Hyper plane or boundaries in Figure 10. From those planes it is very easy to differentiate the various categories of classes of any input data.

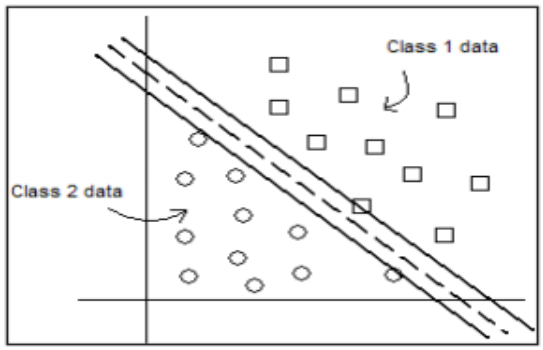

Figure 10. SVM classifier

However, SVM uses linear classification, it may also implement non-linear mapping of input data by kernel functions. Generally 3 kernel functions are used. They are i) linear kernel ii) Polynomial kernel and iii) RBF kernel.

SVM is one of the prominent algorithms that can be deployed as a classifier for emotion recognition. It is a higher dimensional vector supervised learning method. SVM can be considered as one of the best algorithms to be deployed in SER because we can easily assume the parameters of SVM. In addition to that not only training can be quickly done even on large databases but also its accuracy is very better when compared to other techniques. It is also termed as Maximum Margin classifiers. In this method we used RBF kernel, because it controls the input speech signal that fall within the boundaries as shown in Figure 11.

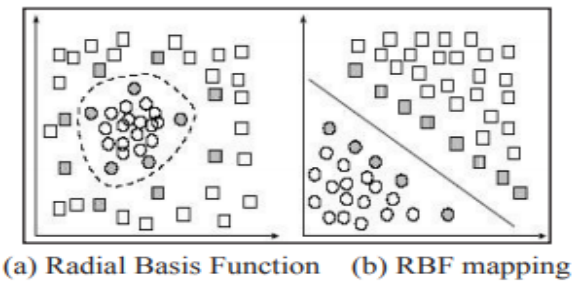

Figure 11. SVM kernel functions

The entire speech emotion classification process using SVM can be described as

1) The features are given as an input for SVM algorithm.
2) These features can be either linearly separable or non-linearly separable.
3) Based upon the features SVM will determine a function f(x) i.e. kernel using which boundaries can be determined by Kernel (x,y) = (x.y).
4) If the features are related to linearly separable then SVM takes a linear kernel function for mapping.
5) If the features are related to non-linearly separable then SVM uses Radial Basis Function (RBF) that maps non-linearly training patterns into a higher dimensional feature space where the features can be separable by using a hyper plane.

RBF kernel will restrict training features and make them lie in specific boundaries.

6) Using such kernels we can perform classification of features into classes of different emotions using SVM.

Multi-Layer Perceptron (MLP): MLP is a Neural Network based Classification Algorithm. MLP model classifying the emotions from the given speech utterances. The Architecture of the MLP model is showed in Figure 12.

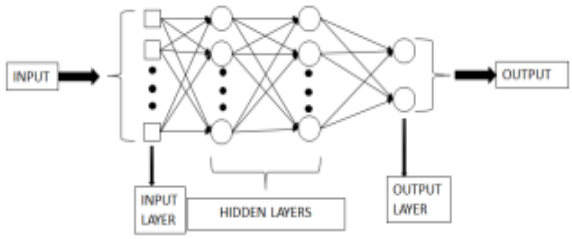

Figure 12. Architecture of multi-layer perceptron

MLP consists of mainly three parts, i) Input Layer ii) Hidden Layers and iii) Output Layer. Generally every MLP has one input layer, many number of hidden layers but one output layer. These hidden layers are defined depends upon Activation Function and the task.

The entire speech emotion classification process using MLP can be described as

1) MLP classifier will be adjusted by defining and initiating the parameters.
2) Then, Input data is trained by the Neural Network.
3) The trained model network is used to predict the emotions from the Speech.
4) Calculate the accuracy of the predictions.

4. RESULTS AND DISCUSSION

In this presented work, the experiments are carried on RAVDESS and IIITH-TEMD Databases with Python Language. For every Classification Model, 75% speech samples as training and 25% speech samples for testing. In Feature Extraction Process, individual spectral (MFCC) and combination of hybrid features (MFCC+ΔMFCC+ΔMFCC) are extracted. In the Classification process, SVM and MLP models are used.

Table 1. Speech emotion recognition accuracy rate of different algorithms along with various features for RAVDESS

Sl No	Features	Classifiers Accuracy (%) SVM	MLP
1.	MFCC	52.11	69.50
2.	MFCC+ ΔMFCC+ ΔΔMFCC	70.73	75.54

In the SVM Algorithm, the Speech utterances are plotted on a n-dimensional plane, here n is the number of features. Non-linear mapping of input data is done by kernels in SVM, in this work we applied RBF (Radial Basis Function) as kernel function. In the MLP model, the speech utterances features are extracted, those features are input to the Input Layer, then the

model uses hidden layers, here the count of hidden layers is 300, batch count is 32 and ReLu(Rectified Linear Unit) as Activation Function. The working procedure of MLP model is given in the section 3.3.

As results showed in Table 1, MLP gives better performance compared to the SVM on the given RAVDESS Speech Corpus.

Table 2. Speech emotion recognition accuracy rate of different algorithms along with various features for IIIT-TEMD

Sl No	Features	Classifiers Accuracy (%) SVM	MLP
1.	MFCC	54.41	73.87
2.	MFCC+ ΔMFCC+ ΔΔMFCC	72.13	78.84

As results showed in Table 2, MLP gives better performance compared to the SVM on the given IIITH-TEMD Speech Corpus with hybrid features.

As results showed in Table 3, MLP gives better performance compared to the SVM on the given DETL Speech Corpus with hybrid features.

Table 4 shows existing work and proposed work in SER.

Table 3. Speech emotion recognition accuracy rate of different algorithms along with various features for DETL

Sl No	Features	Classifiers Accuracy (%) SVM	MLP
1.	MFCC	56.24	74.47
2.	MFCC+ ΔMFCC+ ΔΔMFCC	74.36	81.32

Table 4. Performance comparison - existing work and proposed work

Sl No	Feature Extraction Method	Data Base	Classification Model	Accuracy (%)
1.	LPCC+Formants (Rao et al. 2013 [11])	IITKGP-SESC	GMM AANN	70 61
2.	Acoustic Features (Pitch, ZCR, MFCC) (Avots et al. 2019 [18])	EmoDB	HMM(5 Folds)	78
3.	MFCC+ZCR+HNR+TEO (Aouni et al. 2020 [19])	RML	SVM-Polynomial SVM-RBF	64 65
4.	MFCC+ Spectral Contrast Features (Parikh et al. 2021 [20])	RAVDESS	Deep CNN	71
5.	MFCC+ ΔMFCC + ΔΔMFCC (Proposed Work)	RAVDESS	SVM MLP	71 75
6.	MFCC+ ΔMFCC + ΔΔMFCC (Proposed Work)	IIITH-TEMD	SVM MLP	72 78
7.	MFCC+ ΔMFCC + ΔΔMFCC (Proposed Work)	DETL	SVM MLP	74 81

The comparison of Accuracy report used Classifiers and the Databases is shown in the below Figure 13. From the graph, it is observed that MLP produces better results than SVM on the two Databases.

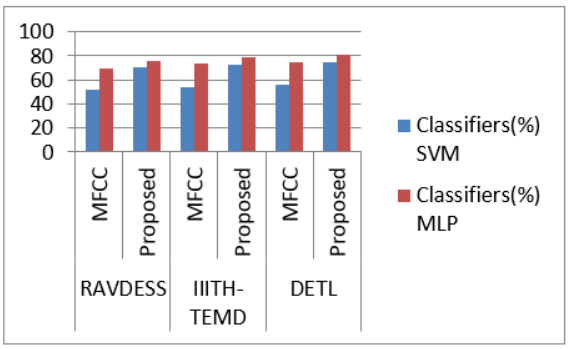

Figure 13. Comparison of different databases and classifiers

5. CONCLUSION

Many exploration works are going on to recognize emotions from speech, so they are locating different Automatic Recognition Systems with improved strategies, however those examination works are limited to natural languages and dialects. Most of the researchers took Spectral features are subjective and important, in that MFCC features resembles the human auditory system. Difficulties of the current research which have no standard Database to the specific region spoken language. This work utilizes Hybrid Feature extraction technique (Spectral) to take out the prominent features from the speech input. The separated significant features (MFCC+ ΔMFCC + ΔΔMFCC) are used in the Classification Model. Use of Feature Extraction Techniques for recognition of Emotion from Speech is an effective phase towards designing a standard SER. From the experimentation, it is noticed that among the two classifiers SVM and MLP for the RAVDESS and IIITH-TEMD Speech Emotion Databases the acknowledgment accuracy is better for MLP only. It is noticed that a good prominent feature subset will always give the better results.

REFERENCES

[1] Palo, H.K., Mohanty, M.N. (2018). Comparative analysis of neural networks for speech emotion recognition. International Journal of Engineering & Technology, 7(4): 111-126.

[2] Rao, K.S., Reddy, R., Maity, S., Koolagudi, S.G. (2010). Characterization of emotions using the dynamics of prosodic features. In Speech Prosody 2010-Fifth International Conference, pp. 1-4. https://www.isca-speech.org/archive_v0/sp2010/papers/sp10_941.pdf.

[3] Palo, H.K., Chandra, M., Mohanty, M.N. (2018). Recognition of human speech emotion using variants of Mel-Frequency cepstral coefficients. In Advances in

Systems, Control and Automation, pp. 491-498. https://doi.org/10.1007/978-981-10-4762-6_47

[4] Huahu, X., Jue, G., Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. In 2010 International Conference on Artificial Intelligence and Computational Intelligence, 1: 537-541. https://doi.org/10.1109/AICI.2010.118

[5] Boco, P.F.O., Tercias, D.K.B., Cruz, K.R.D., Raquel, C.R., Guevara, R.C.L., Naval Jr, P.C. (2010). EMSys: An emotion monitoring system for call center agents. https://www.researchgate.net/profile/Prospero_Naval/publication/268276414_EMSys_An_Emotion_Monitoring_System_for_Call_Center_Agents/links/54d470650cf25013d02978e7/EMSys-An-Emotion-Monitoring-System-for-Call-Center-Agents.pdf.

[6] Sahu, S., Gupta, R., Espy-Wilson, C. (2018). On enhancing speech emotion recognition using generative adversarial networks. arXiv preprint arXiv:1806.06626.

[7] Borchert, M., Dusterhoft, A. (2005). Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In 2005 International Conference on Natural Language Processing and Knowledge Engineering, pp. 147-151. https://doi.org/10.1109/NLPKE.2005.1598724

[8] Han, K., Yu, D., Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech.

[9] Mao, Q., Dong, M., Huang, Z., Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 16(8): 2203-2213. https://doi.org/10.1109/TMM.2014.2360798

[10] Rong, J., Li, G., Chen, Y.P.P. (2009). Acoustic feature selection for automatic emotion recognition from speech. Information Processing & Management, 45(3): 315-328. https://doi.org/10.1016/j.ipm.2008.09.003

[11] Rao, K.S., Koolagudi, S.G., Vempada, R.R. (2013). Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology, 16(2): 143-160. https://doi.org/10.1007/s10772-012-9172-2

[12] Kurpukdee, N., Koriyama, T., Kobayashi, T., Kasuriya, S., Wutiwiwatchai, C., Lamsrichan, P. (2017). Speech emotion recognition using convolutional long short-term memory neural network and support vector machines. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1744-1749. https://doi.org/10.1109/APSIPA.2017.8282315

[13] Jiang, W., Huang, L., Liu, Q., Lü, Y. (2008). A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In Proceedings of ACL-08: HLT, pp. 897-904.

[14] Xia, R., Liu, Y. (2015). A multi-task learning framework for emotion recognition using 2D continuous space. IEEE Transactions on Affective Computing, 8(1): 3-14. https://doi.org/10.1109/TAFFC.2015.2512598

[15] Basharirad, B., Moradhaseli, M. (2017). Speech emotion recognition methods: A literature review. In AIP Conference Proceedings, 1891(1): 020105. https://doi.org/10.1063/1.5005438

[16] Livingstone, S.R., Peck, K., Russo, F.A. (2012). Ravdess: The Ryerson audio-visual database of emotional speech and song. In Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science, pp. 205-211.

[17] Le, B.V., Lee, S. (2014). Adaptive hierarchical emotion recognition from speech signal for human-robot communication. In 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 807-810. https://doi.org/10.1109/IIH-MSP.2014.204

[18] Avots, E., Sapiński, T., Bachmann, M., Kamińska, D. (2019). Audiovisual emotion recognition in wild. Machine Vision and Applications, 30: 975-985. https://doi.org/10.1007/s00138-018-0960-9

[19] Aouani, H., Ayed, Y.B. (2020). Speech emotion recognition with deep learning. Procedia Computer Science, 176: 251-260. https://doi.org/10.1016/j.procs.2020.08.027

[20] Parikh, N., Mistry, K., Bhavsar, Y., Hakimi, A., Magare, A. (2021). Real time speech emotion recognition using machine learning. International Research Journal of Engineering and Technology (IRJET), 8(6): 2786-2792.