Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary & Peer Reviewed
Volume: 04, Number: 01, Pages: 124- 130, Year: 2021

Automatic Speech Emotion Recognition using Mel Frequency Cepstrum Co-efficient and Machine Learning Technique

Shoaib Mustafa[1], Akmal Khan[2], Shabir Hussain[3], M. Zeeshan Jhandir[2], Rafaqat Kazmi[2], and Imran Sarwar Bajwa[2]

[1]National College of Business Administration & Economics Rahim Yar Khan, Pakistan
[2]Department of Computer Science, The Islamia University of Bahawalpur, Pakistan
[3]School of Information Engineering, Zhengzhou University, Henan, China

Corresponding author: Akmal Khan (e-mail: akmal.shahbaz@iub.edu.pk).

Abstract- **An audio speech signal is the quickest and accepted way of communication between humans. This fact provoked researchers and scientists to use the speech signal to communicate between humans and machines to make machines work more efficiently. In the study of human-robot interaction (HRI), emotion can be helpful in many applications. The most significant difference between humans and machines is an emotion; if the machine acts with emotion, more people can acknowledge the machine. Every day through conversation: emotion in speech is a significant key to meaning out the speaker's underlying intention to identify several emotional states, which help people who have a problem in understanding and recognition emotion. Automatic speech emotion is a difficult task that depends on the capability of the Speech feature. In this work, study an algorithm involving MFCC computation and Support Vector Machine (SVM) is used to perform the task of Speech emotion recognition system of collectively five emotions named Angry, Happy, Neutral, Pleasant Surprise and Sadness. Two databases are used for this purpose (Toronto University speech dataset and IEMOCAP speech dataset) with 97% and 86% accuracy. This work can be enhanced by adding preprocessing steps before feature extraction and considering more artifact features (like pitch, time domain) and the current features. Moreover, along with these databases, other popular databases like the Berlin speech database can enhance accuracy.**

Index Terms—**SER, MFCC, Cepstral coefficients, SVM, Toronto Database, IEMOCAP Database.**

I. INTRODUCTION

The audio speech signal is the quickest and natural way of communication between humans. This fact provoked researchers and scientists to use the speech signal to interact between humans and machines to make machines work more efficiently [1]. The automatic speech emotion recognition (SER) system is one of the most commonly used methods for bettering efficiency, normalizing the human-machine interface, and improving artificial intelligence. Recently human-computer interaction (HCI) is a very famous area in the field of research. The primary purpose of HCI is to facilitate communication between the user and the machine using several parallel channels [2-10]. In SER, the speech signal articulates emotions such as anger, neutral speech, sadness, happiness, excitement, fear, and many others.

The speech signal is processed on a paralinguistic basis [2]. SER has many applications, by many modalities Emotion recognition can be utilized [11-15,16], it can be used in vehicle board system to identify the mental state of the driver for his/her safety moreover, it can be used in the call control room to detect the emotional behavior of the clients [3]. Emotion recognition has implication in psychiatry [11] medicine [12], psychology [13] and design of human-machine interaction system [14]. Many research works have been done in SER, but the problem of accuracy is always there. Three main issues for a successful speech emotion recognition system are selecting a better speech emotion dataset, more operative extracting feature, and planning a reliable classifier. The main issue in SER is feature selection: a good SER system depends on better feature selection. Many researchers offered an imperative speech feature that held personal data, such as pitch, energy, formant frequency, Linear Prediction, Cepstrum Coefficients (LPCC), and Mel recurrence Cepstrum coefficients (MFCC). So, most researchers like to utilize a list of features that make out various kinds of features with more personal data. In any case, utilizing a joining list of the features may offer high dimension and speech feature redundancy in this manner; it prepares for the learning procedure confounded freehand machine efficiency of learning and improved the probability overfitting. This way, the selection of features is imperative toward the decrease dimensional repetition of the feature. An analysis of techniques and collection features is introduced. Mutually component feature selection and feature extraction remain fit for successful learning routines.

Classification is the final stage of the speech emotion for recognition (SER) system. It includes organization the rare

information as utterance or utterance frames into a specific class of emotion based on feature extraction from the information. The researcher offered many classifications like hidden Markov Model, Gaussian mixture model, support vector machine, recurrent neural network, and neural network.

We proposed a system for emotional recognition, such as (Angry, Happy, Neutral, Pleasant Surprise, and Sadness.). We extracted the Mel-Frequency cepstrum coefficient (MFCCs) and utilized them for training three different classifiers (SVM LDA and DECISION TREE); we confirmed that the mixture of both features has a high accuracy rate 97% and 86% on the Toronto University speech dataset and IEMOCAP speech dataset. The training and testing emotion points is shown in FIGURE 8.

We proposed the maximum accuracy of automatic speech emotion recognition (SER) using different classifiers. The current SER system cannot be achieved accurately for some reason. By emphasizing these problems, the following are the research questions that have been answered in this paper. We also compare our results with previously work done of different researches as shown in Tab. I.

- How to extract speech emotion using Mel-frequency Cepstral Coefficient (MFCC)?

MFCC is a classical efficient and successful approach used for speech-related tasks such as gender identification through speech recognition and speech emotion recognition. The primary purpose of using the Mel-frequency cepstral coefficient (MFCC) to extract features from speech because they are derived from the human speech patterns MFCC tries to mimic the human ear where frequencies are nonlinearly determined on the audio spectrum.

- Which Tools were used to solve the accuracy Problem?

There are many tools used for speech emotion recognition, but MATLAB is a better tool to solve this accuracy problem.

- Which Techniques were used to solve the accuracy problem?

Several machine learning techniques have been utilized for separate emotion classification. This machine learning techniques' objective is to study from the training samples and then utilize them to classify new observations. Similarly, there is still no clear answer to selecting learning machine algorithms; every technique has its specific advantages and restrictions. For this purpose, here we select to analyze the performance of three different classifiers SVM, LDA, and DT.

A. CONTRIBUTION

- We used two database IEMOCAP speech dataset and Toronto university speech dataset for proposed model.
- We utilized three algorithms (LDA, SVM, and DECISION TREE) to build our model to recognize five emotions named Angry, Happy, Neutral, Pleasant Surprise, and Sadness.
- Our proposed model described the better emotion recognition rate of 97% on the IEMOCAP dataset utilizing an SVM classifier. For the Toronto college discourse dataset, 86% exactness was accomplished by all classifiers as shown in Table II, III, and IV.

II. LITERATURE REVIEW

Amongst the cognitive senses, speech is the most effective channel. For the sack of transferring data, thought, and emotion, speech is an initially communicated medium. Moreover, it is also a unique medium of passing human intellect from one member to another. Bill Gates (co-founder of Microsoft Corp.) in his book. The Road Ahead "describes a speech recognition process is an essential tool or innovation for future computing. In human-computer interaction, speech identification is always looked like an allure. Speech signal identification changes acoustic speech signals taken through a medium or telephone in sentences. An analysis of several speech emotion detection procedures, features available, as detailed in the paper [17-20, 21]. Feature extraction, such as the Mel frequency cepstral coefficient (MFCC), was described entirely [22, 23]. The identified arguments may be the last output for different tenders like control and signals, raw material or signal, document, and penetrate work preparation. Identifying the speech signal has been a vital area goal for researchers for more than four decades. Like a machine, by allocators in the whole environment, the spoken digits are particular to human beings. Classification and recognition of signals together involve a man's perception. An excellent example of the human disposition may be the arrival of language to classify inherent and identical sequences. Discovering and recognizing patterns in the speech signal is the most challenging task for the sequence recognition machine [1, 2]. Speech recognition is a lengthy process in this procedure; it involves speech, language, gesture, destination capability, and the subject. The functioning of a voice recognizer framework relies upon its assignment and its design. Humans can speak by machine and the PC through a different medium or other device that is slow in performance.

The maximum fundamental feature of MFCCs and formants is energy. This decreases the methodology's computational problem and can be prompt equally energy and data transmission saving funds when the voice is captured on cell phones. They utilized classifiers for human speech emotion, such as Hidden Markov Model (HMM), Neural Network (NN), Kernel deterioration and K-nearest Neighbors approach (KNN), support vector machine (SVM), and Naive Bayes classifier [24], Gaussian Mixture Model (GMM) [25]. They picked SVM as a fundamental classifier because of its detailed planning and effort by quite a few attributes. SVM part capacities are utilized to outline to an advanced dimensional element space externally defeated the innovation. That regular strategy for utilizing bit works in SVM run to models at making groups and discover the kernel capacity that achieves the extreme average of correctness for the specific problem.

For some time, the classification of emotions has been a vital subject to discussion in various psychology areas, affecting science and emotion study. Firstly, the method of emotions is labeled by a discrete quantity of classes. Numerous scholars have led studies to figure out which emotions are essential [24]. A most well-known model is Ekman [25], who proposed a rundown of 6 fundamental emotions named (disgust, happiness, fear, sadness,

anger, and surprise). He clarifies that every emotion goes about as a discrete classification instead of a person's emotional state. In the second methodology, emotions mix a few mental measurements and are recognized by axes. Different analysts characterize emotions as indicated by at least one measurement.

In 1897, Wilhelm Max Wundt proposed three measurements that could depict this emotion: (1) strain versus easiness, (2) pleasurable versus unpleasable, and (3) exciting versus calming [25]. Not at all like the prior three-dimensional models, Russell's model features just two measurements, which incorporate (1) excitement (or activation) and (2) valence [21]. The precise methodology is usually utilized in SER [20]. It describes emotions utilized in common emotion words, for example, anger and joyfulness. In that work, many six raw emotions (neutral, anger, fear, disgust, sadness, and surprise) and the six emotions of Ekman's model were utilized to identify emotion from speech utilizing the clear-cut methodology. It was speech signal working difficulty so that the researcher can read these difficulties. Therefore, this method cannot give the result that was needed and failed to provide the learning data. Recently approach numerous instructional papers that give us relevant detail to begin research work utilizing different models and techniques for speech emotion recognition.

TABLE 1
COMPARISONS OF RELATED WORK WITH OUR PROPOSED METHODOLOGY

Proposed Methodology	Methodology	DB	CNN	SVM	KNN	DT	ANN	DNN	ACCR %	EMO.C
Semiye Demircan et all 2016 [21]	MFCC LPC	EMOCAP DB	✕	✓	✕	✕	✓	✕	92.9	A,B,D,F,H,S,N
Eyben et al. 2016 [23]	MFCCVQC	EGEMAPS	✕	✓	✕	✕	✕	✕	86	A,S,N,H,D
Quan et al. 2017 [22]	Correlation cepstral MFCC	EMO-DB	✕	✓	✕	✕	✕	✕	80	A,H
Abdul Malik Bad shah et al. 2017 [22]	MFCC	EMO-DB	✓	✓	✕	✓	✕	✕	89.46 86.56	A,B,D,F,H,N,S
Palo et al. 2018 [14]	MFCCVQC PLPVQC	EMO-DB	✕	✕	✕	✕	✕	✓	83	A,B,D,F,H
Shuiyang Mao et al. 2019 [23]	MFCC GMM HMM	EMO-DB EMOCAP	✕	✕	✕	✕	✕	✓	91.32	A,H,N,S,S
Suraj Tripathi et al. 2019 [24]	MFCC	EMOCAP	✕	✕	✓	✕	✕	✕	74.6	A,S,N,H
Pengxu Jiang et al 2019 [25]	PCRN PCRN PCRN PCRN	CASIA EMO-DB ABC SAVEE	LSTM	✕	✕	✕	✕	✕	61.5 90 80 84	A,F,H,N,S,S
Huang-	MFCC	EMO-	✕	✕	✕	✕	✕	✓	61.	
Cheng et al. 2019 [26]		DB							48	N,A,H,S
Yang, Ningning et al 2020 [35]	MFCC	EMO-DB	✕	✓	✕	✕	✕	✕	92..4	H,A,S,N
PROPOSED WORK	MFCC	TORONTO UNIVERSITY, EMOCAP	✕	✓	✕	✓	✕	✕	9786	A,H,N,P,S

III. METHODOLOGY

A. TORONTO EMOTIONAL SPEECH DATASET

Two databases are used to enhance the accuracy of the speech Emotion Recognition system. Audios in this database are recorded at the University of Toronto with 2 actresses (aged 27 and 65). We have included four emotions (angry, happy, disgust, and surprise) for speech emotion recognition purposes.

B. MOCAP DATABASE

This dataset acted Multi speaker and multi-model datasets recently currently conduct from SAIL Lab at USC. IMOCAP database used in [17,18] from this database. It is most commonly used in SER to plan new methods and analyze acoustic features in speech emotion recognition [19]. We have included three emotions (Angry, Natural, and sad). The task of Emotion Recognition involves 2 phases, classification and Feature Extraction. We have used MFCCs of speech signal as extracted features as shown in Fig. 2, 6 and then classified those extracted features (MFCCs) using SVM classifier to perform Speech Emotion Recognition.

C. FEATURE EXTRACTION

As we know that feature extraction refers to the method utilized to remove dimensionality from a given database and hence minimize the time and space of the system. Feature vector contains discriminating information that distinguishes one object/instance from another object/instance [8], so it is a critical task to select an appropriate algorithm. For feature extraction, in this paper, we have used MFCCs as the extracted features as shown in Fig. 5 for the further classification of emotions.

D. SPEECH FEATURES OF EMOTION DETECTION

In contempt of the fact that the sound is a singular duct, which is extracted and examined. These are two types of features extracted for the sack of prosodic, spectral and for speech quality feature the paralinguistic feature is classified, for this our proposed model shown in Fig. 1.

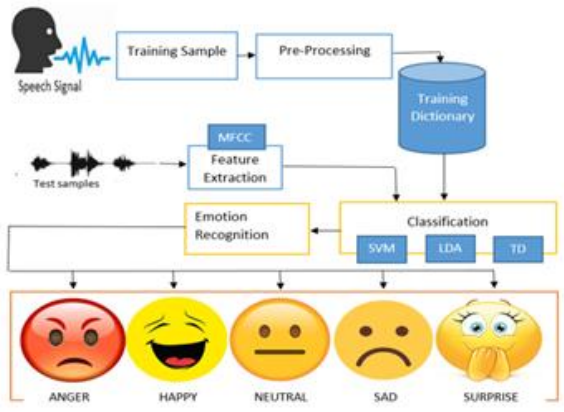

FIGURE 1: BLOCK DIAGRAM FOR SPEECH EMOTION RECOGNITION SYSTEM

E. MEL FREQUENCY CESTRUM CO-EFFICIENT

MFCC is a classical efficient and successful approach used for speech-related tasks such as gender identification by voice, speech recognition, speech emotion recognition, and many others. The Mel-frequency cepstral coefficients (MFCC) is one of the most famous audio features [23, 24] MFCC coefficients are successful because they are derived from human speech patterns [2]. The MFCC is a prestigious technique utilized in speaker recognition, which is focused on the speaker's concrete vocal tract properties [25]. MFCC tries to mimic the human ear where frequencies are nonlinearly determined on the audio spectrum [6]. The Mel scale is a perceptual size of pitches that decided to be equivalent in useful ways with the audience. Mel originates from the term tune show scale depends on pitch relationship. The orientation fact between that scale and typical rate estimation is categorized by comparing a 1000 Hz tone, 40 dB over the audience's edge, by a pitch of 1000 Mel's. Mel scale corresponds linearly less than 1 KHz and logarithmically over 1 kHz [6]. Computation of Mel-frequency Cepstral coefficients involves different steps as shown Fig. 2 and Fig. 6. Computation of Mel Frequency Cepstral coefficients involves the following steps.

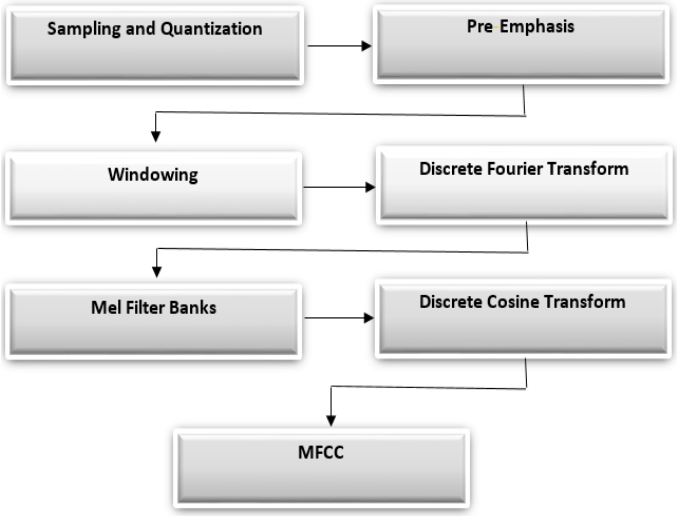

FIGURE 2: BLOCK DIAGRAMS FOR MFCC FEATURE EXTRACTION

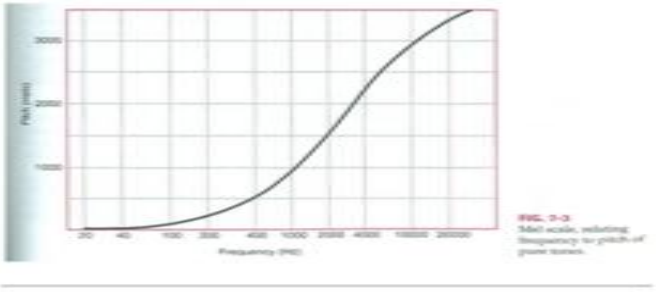

FIGURE 3: COURTESY WEB

F. SAMPLING AND QUANTIZATION

The audio wave is a continuous signal, while the computer is a digital machine that cannot represent the continuous signals directly, so we must have to convert these continuous signals into a discrete finite set of information; this is known as sampling. In comparison, quantization is the process of representing real-valued numbers as the set of integers [4].

G. PRE-EMPHASIS

The speech spectrum has higher energy at low frequencies as compared to high frequency. [7] Through pre-emphasis, energies are boosted at high-frequency levels, which balance the spectrum's level of energies. Pre-emphasis is considered a noise reduction tool as it reduces the noise power without affecting the rest of the signal [4].

H. WINDOWING

The speech signal varies over time but is stationary in short segments. Thus, in this step, the speech signal is divided into short chunks of rectangular segments of approximately 10 ms for further analysis by minimizing the signal's discontinuities using the Hamming window, which shrinks the magnitude of the signal towards zero at the edges of the window hence avoiding discontinuity. As shown in Fig. 4, below are plots of windowing and hamming window effect [9]:

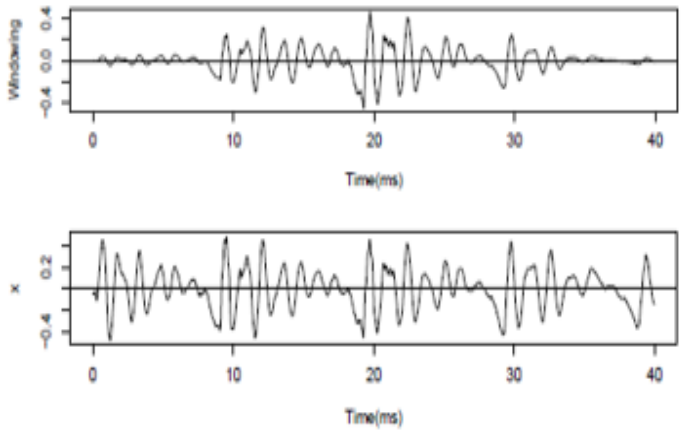

FIGURE 4: PLOT OF WINDOWING AND HAMMING WINDOW EFFECT

I. DISCRETE FOURIER TRANSFORM

After windowing, the next step is DFT; this step is based on the spectral analysis to extract the speech feature based on magnitude spectrum computation [4]. At this stage, the energy level /

spectral information of every window is collected. The input to DCT is the windowed signal, and the output is the complex numbers representing the magnitude and the phase of the frequency. The commonly used algorithm to find out DCT is Fast Fourier Transform or FFT.

J. MEL FILTER BANK

Referring to the human hearing ability that we are more sensitive to sound between 20 and1000 Hz, computation of Mel Frequency Spectrum is performed after the DFT bypassing the spectrum through Mel filters to obtain Mel Spectrum. It is used to match the human ear's frequency resolution by linearly spacing the energies of frequencies below 1000 Hz and logarithmically spacing them after that. Graphically the Mel filter is shown in Fig. 5.

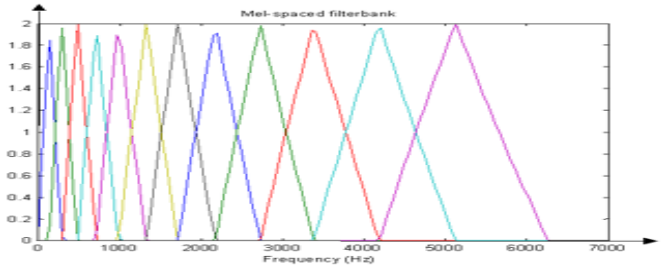

FIGURE 5: MEL FILTER BANK WITH OVERLAPPING FILTERS

K. LOG OF MEL SPECTRUM VALUES

We take the Mel spectrum log because the human hearing ability is less sensitive to the slight change of amplitude and more sensitive to amplitude.

L. DISCRETE FOURIER TRANSFORM

To find out the Cepstrum coefficients, we calculate the Discrete Cosine Transform of each Mel output. The first K coefficients are the MFCC (Usually, K=13). Also, notice that the MFCCs form an n × q matrix for each signal where n is the number of window frames, and q is the number of MFCCs. If we pass the MFCC matrices to a vector-based pattern recognition technique, these matrices must be transformed or summarized to vectors. The simplest way of doing this is to take each of the n column vectors [9].

M. EXTRACTION OF MEL-FREQUENCY CESTRUM COEFFICIENTS (MFCC)

In the field of artificial speech recognition, MFCCs is regularly using the feature. Its function is to mimic human speech. In the music-related field, MFCC has proven to be one of the most popular and successful spectrum features in speech. MFCCs resource of required is lo-moderate and easiest way of implementation and also Multi speaker Multilanguage.

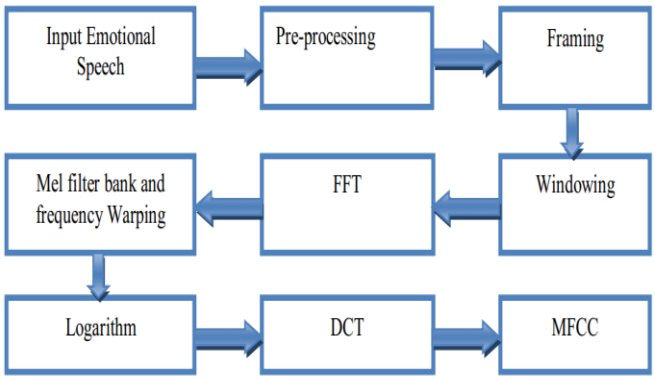

FIGURE 6: BLOCK DIAGRAM FOR MFCC FEATURE EXTRACTION PROCESS

N. CLASSIFICATION

After feature extraction, the very next step is emotion classification. This work has used the following two SVMs for the Toronto speech dataset and IEMOCAP dataset, respectively: the Gaussian support vector machine and the quadratic support vector machine. The developed model is shown in Fig. 7.

Support vector machine (SVM) is high dimensional vector supervised learning strategy that depends on emotive suppositions. It calculates the appearance (or non-appearance) of a predefined feature of a class that was no identified with every other feature's appearance (or non-appearance). It is easy to compute and implement, its parameters are easy to expect, even on exceptionally enormous

Databases learning or preparing is extremely quick and successful, and its exactness is relatively better in contrast with different procedures. The dynamic recognition process by preparing and testing stages is shown in Fig. 6.

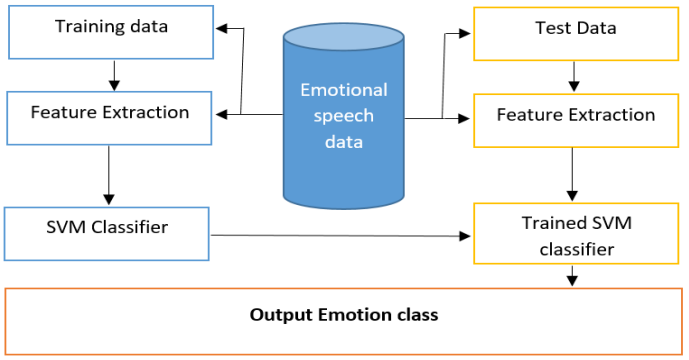

FIGURE 7: BLOCK DIAGRAM OF SPEECH EMOTION RECOGNITION CLASSIFICATION

To explore classification models automatically, utilize the Classification Learner application. In our work, we utilize three classifiers for speaker/voice recognition, which is separate a speaker is a Male or Female.

O. LINEAR DISCRIMINANT ANALYSIS (LDA)

Linear Discriminant Analysis (LDA) speculates Fisher's direct discriminant; a technique utilized in statistics, design recognition, and AI to locate a direct mixture of the features that describe or divorces at least two classes of articles or cases. LDA proceeds

multi-dimensional information, utilizes previous class data (Supervised Learning), and declares the information in a structure that enhances the various classes' parting. It takes covariance of a class (of information) with itself, means of the whole information, means of each class, and prior probabilities of the class. LDA additionally utilizes the scatter inside a class, scatter in the middle of classes, and attempts to best divorce 2 classes of information. Linear Discriminant Analysis (LDA) is one of the most mechanical experienced classification systems. The LDA basic idea is straightforward: for each class to be recognized, analyze (different) linear capacity. There are many linear models of classification, and they diverge to a great extent in how the coefficients are built up. One pleasant value of LDA is different nearly of the replacements, and it doesn't require many passes the information for enhancement. Moreover, it usually handles issues with naturally two or more classes, and it can give chance evaluations to every candidate's classes. LDA looks to some amount like principal components analysis (PCA), in that many functions are delivered (utilizing every variable), which are proposed, in some sense, to give information decrease through the improvement of data.

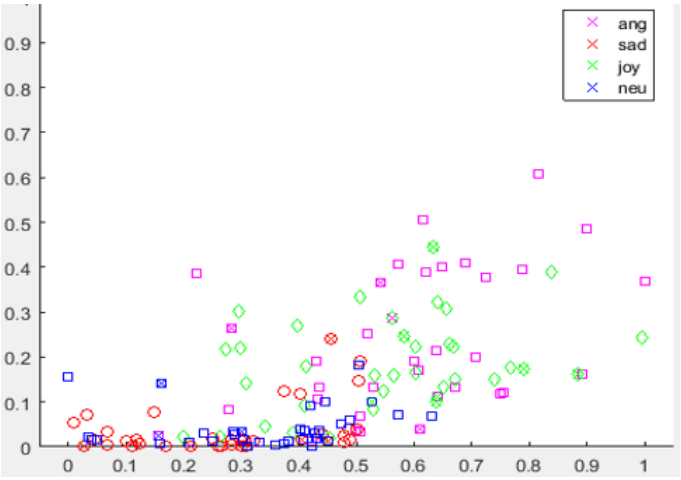

FIGURE 8: TRAINING AND TESTING POINTS OF EMOTION CLASSIFICATION

IV. EXPERIMENTS AND RESULTS

As mentioned before, in this paper, we have experimented with SER on 2 databases Toronto University speech Dataset, IEMOCAP speech dataset. In the Toronto University speech dataset, we have divided the dataset into training and testing datasets. For the training dataset, 200 samples (50 samples of each category) of emotions (Angry, Happy, Disgust, and Pleasant Surprise) were separated). In contrast, 24 samples (6 from each category) were separated for testing data. The MFFCC vector of this training data was passed to the classifier, and it gave an accuracy of 97%.

In the IEMOCAP speech, dataset 150 samples (50 for each category) of emotions (Angry, Neutral, and sad) were passed to the classifier as the training dataset. At the same time, 30 samples (10 from each category) were treated as testing data. The accuracy of this dataset after training was 86%. The confusion matrix results and accuracy are shown in table 2 and table 3, and we also compare our whole model with previous work done as shown in Tab. IV.

TABLE II
CONFUSION MATRIX FOR EMOTION PREDICTION USING SVM WITH TORONTO UNIVERSITY SPEECH DATASET

Emotion Class	Angry %	Happy %	Neutral %	Surprise %	Sadness %	Result %
Angry	97	1	0	1	1	97
Happy	1	97	0	2	0	97
Neutral	0	0	3	96	1	96
Surprise	0	0	3	96	1	96
Sadness	0	0	2	1	97	97

TABLE III
CONFUSION MATRIX FOR EMOTION PREDICTION USING SVM WITH IEMOCAP

Emotion Class	Angry %	Happy %	Neutral %	Surprise %	Sadness %	Result %
Angry	89	0	3	2	6	89
Happy	5	85	4	5	1	85
Neutral	1	4	86	6	3	86
Surprise	2	5	3	84	6	84
Sadness	3	3.	4	4	86	86

TABLE IV
RESULT COMPARISON WITH PRIOR STUDY AND PROPOSED ACOUSTIC FEATURE SET USED DIFFERENT DATABASE

Literature	Feature	Classifier	Accuracy	Database
Quan et al. [23]	Correlation, Cepstral MFCC	SVM	80%	EMO-DB
Palo et al. (2018) [24]	LPCCVQC MFCCVQC PLPVQC	MLP RBFN PNN, DNN	83% 79% 76%	EMO-DB
Eyben et al. [25]	EGEMAPS Compare	SVM SVM	78.1% 86.7% 97%	EMO-DB TORONTO
Proposed Work	MFCC	LDA DC	86%	UNIVERSITY EMOCAP

V. CONCLUSION

One method has been experimented with in this paper using two databases with two different types of SVM. In total, 5 distinct emotions have been classified. The results are achieved by putting cepstral coefficients in a feature vector where each row represents each signal, and the matrix was formed by adding labels to the interval of 50 rows (one label for 50 samples of each category).

We introduced an automatic speech emotion recognition (SER) system utilizing three algorithms (LDA, SVM, and DECISION TREE) to order collectively five emotions named Angry, Happy, Neutral, Pleasant Surprise, and Sadness. In this manner, sorts of features (MFCC) were extracted from two databases used for this purpose (IEMOCAP speech dataset and Toronto university speech dataset), and a mix of these features was displayed. We work on how features and classifiers affect the correctness in emotion from speech. A subsection of extremely discriminant features is chosen. Feature selected methods prove that more data isn't in every time better in applications. The models of machine

learning were prepared and calculated to identify the emotion from features.

SER described the better emotion recognition rate of 97% on the IEMOCAP dataset utilizing an SVM classifier. For the Toronto college discourse dataset, 86% exactness was accomplished by all classifiers. We can see that SVM is consistently performed by additional information and encounters broad exercise occasions from this result. Along these lines, we set up an SVM model with an excellent approach for practical use for controlled information in the LDA relationship. Improvement of the strength of speech emotion recognition (SER) system is yet conceivable with merging datasets and with combination of classifications.

The impact of preparing different emotion indicators may be researched with utilizing them in a particular detection framework. To utilize another component of feature selection (FS) technique is proposed. The speech emotion recognition (SER) system rate depends on the quality of good feature selection: a better feature selection (FS) technique can provide choice features selection of emotion rapidly.

VI. FUTURE WORK

This work can be enhanced by adding preprocessing steps before feature extraction and considering more artifact features (like pitch, time domain) and the current features. Moreover, along with these databases, other popular databases like the Berlin speech database can enhance accuracy.

REFERENCES

[1] Shambhavi, S. S., & Nitnaware, V. N., Emotion speech recognition using MFCC and SVM. *International Journal of Engineering Research and Technology*, vol. *4*, *no.*6, pp.1067-1070, 2015.

[2] SS Shambhavi, VN Nitnaware Emotion Speech Recognition using MFCC and SVM (International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS060932 www.ijert.org

[3] Lalitha, S., Geyasruti, D., Narayanan, R., & Shravani, M. Emotion detection using MFCC and cepstrum features. Procedia Computer Science, vol. 70, pp.29-35, 2015.

[4] Palo, H. K., Mohanty, M. N., and Chandra, M., Use of different features for emotion recognition using MLP network. In Computational Vision and Robotics (pp. 7-15), 2015. Springer, New Delhi.

[5] Watile, A., Alagdeve, V., & Jain, S., Emotion Recognition in Speech by MFCC and SVM. International Journal of Science, Engineering and Technology Research, vol. 6, 2017.

[6] Madan, A., & Gupta, D., Speech feature extraction and classification: A comparative review. International Journal of computer applications, vol. 90, no. 9, 2014.

[7] Saste, S. T., & Jagdale, S. M., Emotion recognition from speech using MFCC and DWT for security system. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (vol. 1, pp. 701-704), 2018. IEEE.

[8] D Tacconi, O, Mavora p Lukowicez B Amrich c setz G Troster and c Haring Activity and emotion recognition to support early diagnosis of psychiatric diseases in pervasive computing technologies for healthcare,2008 pervasive health 2008. Second international conference on IEEE, 2008, PP 100-102.

[9] B Maier and W, A Shiblis Emotion in medicine in the philosophy and practice of medicine and bioethics pp 137-159 springer, 2011.

[10] Zuroff, D. C., & Colussy, S. A., Emotion recognition in schizophrenic and depressed inpatients. Journal of clinical psychology, vol. 42, no.3, pp.411-417, 1968.

[11] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G., Emotion recognition in human-computer interaction. IEEE Signal processing magazine, vol. 18, no. 1, pp.32-80, 2001.

[12] Guo, J., Lei, Z., Wan, J., Avots, E., Hajarolasvadi, N., Knyazev, B., & Escalera, S., Dominant and complementary emotion recognition from still images of faces. IEEE Access, vol. 6, pp. 26391-26403, 2018.

[13] RE. Haamer K. Kulkarni, N imanpour, M.A. Haque E, E Avots, M. Breisch K. Nasrollahi, S.E. Guerreo, C. Ozcinar, X. Baro et al, "Changes in facial expression as biometric: a database and benchmarks of identification", in IEEE Conf. on Automatic Face and Gesture Recognition Workshops. IEEE, 2018.

[14] Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., & Espy-Wilson, C. Adversarial auto-encoders for speech based emotion recognition, 2018. arXiv preprint arXiv:1806.02146.

[15] Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. In Interspeech (pp. 3097-3101), 2018.

[16] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, vol. 42, no.4, pp.335, 2008.

[17] T.S. Gunawan N.A.M Saleh, and M. Kartiwi, Development of Quranic Recier identification system using MFCC and GMM Classifier, International journal of electrical and computer Engineering (IJECE), vol, 8, 2018.

[18] Gunawan, T. S., Alghifari, M. F., Morshidi, M. A., & Kartiwi, M. (2018). A review on emotion recognition algorithms using speech analysis. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 6(1), 12-20.

[19] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A,. A review on speech emotion recognition: Case of pedagogical interaction in classroom. In 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-7), 2017. IEEE.

[20] Matilda S. Emotion recognition: A survey. International journal of advanced computer research, vol. 3, no. 1, pp. 14-19, 2015.

[21] Koolagudi SG, Rao KS. Emotion recognition from speech: A review international journal of speech technology, vol.15, no. 2, pp. 99-117, 2012.

[22] N.J.Nalini, S. Palanivel. "Music Emotion Recognition: The Combined Evidence of MFCC and Residual Phase. Egyptian Information Journal, 2017.

[23] Quan, C, Zhang, B, Sun, X, Ren, F.A combined cepstral distance method for emotional speech recognition. Int. J.Adv. Robot. Syst.2017, 14.

[24] Eyben F Scherer, K R, Schuller, B W Sandburg J Andre E Basso C Deville's L,Y, Epps, Laukka P, Narayanan, S, S. et al, The Geneva minimalistic acoustic parameter set for voice research and affective computing. IEEE trans. Affect computer, vol. 7, pp.190-202, 2012.

[25] Yang, Ningning, Dey Nilanjan, Sherratt, R. Simon, Shi Fuqian. Recognize basic emotional statesin speech by machine learning techniques using mel-frequency cepstral coefficient features. Journal of intelligent and fuzzy system, vol, 39 no, 2 pp.1925-1936, 2020.