

# Spectral features based speech emotion recognition using artificial neural network

Neha Dewangan<sup>1,\*</sup>

*School of Studies in Electronics & Photonics, Pt. Ravishankar Shukla University, Raipur-492010, India*  
[dewanganneha92@gmail.com](mailto:dewanganneha92@gmail.com)

Sunandan Mandal<sup>2</sup>

*School of Studies in Electronics & Photonics, Pt. Ravishankar Shukla University, Raipur-492010, India*  
[sunandan.mandal12@gmail.com](mailto:sunandan.mandal12@gmail.com)

Kavita Thakur<sup>3</sup>

*School of Studies in Electronics & Photonics, Pt. Ravishankar Shukla University, Raipur-492010, India*  
[kavithakur67@gmail.com](mailto:kavithakur67@gmail.com)

Bikesh Kumar Singh<sup>4</sup>

*Department of Biomedical Engineering  
National Institute of Technology,  
Raipur-492010, India*  
[bsingh.bme@nitrr.ac.in](mailto:bsingh.bme@nitrr.ac.in)

**Abstract**— Emotion is an essential part of human communication. People communicate with emotions through words and body language. Speech emotion recognition is a well-known technique to detect emotions from speech signals. Here, we have proposed binary and multiclass classification models that combine two cepstral coefficients, i.e., Mel-Frequency Cepstral Coefficient (MFCC) and Mel-Frequency Magnitude Coefficient (MFMC) to extract the spectral features from the speech signals and classify them using backpropagation artificial neural network (BPANN). In our study, it is found that when significant features of both spectral coefficients are combined it shows improvement in training and classification results. The proposed model achieved 85.24% accuracy for the multiclass classification of seven emotions using statistically significant features. The proposed model also achieved 100% accuracy for the binary classification of Happy versus Sad emotion and Sad versus Fear emotion.

**Keywords**—MFCC, Mel-Frequency Magnitude Coefficient, Neural Network, Speech Emotion Recognition, Multiclass classification model

## I. INTRODUCTION

Emotion is an internal state of feeling or agitation. One's mood or way of feeling about something is an affective state that facilitates communication. The expression of emotions in an identical situation varies from person to person [1]. Depending on the situation encountered, human convey their feeling using facial expressions and emotions within the voice. Emotions help us understand what other people think and feel through facial expressions, body language, and vocalizations. Most human communication involves conveying one's emotional state to another person. Thus, it serves as a communication tool. It is a natural, primary reaction of the human body. The person can recognize emotions while talking, working, watching movies, playing games, etc. Machines can learn these human emotional responses through physiological signals, speech signals, facial expressions (image or video signals), text signals, etc. Automatic speech emotion recognition (SER) is one of the state-of-the-art research topics in the human-computer interface (HCI). It has many applications in the medical field, call centers, crime investigation, telecommunication, robotics, computer games, and psychological assessment [2].

Speech is the primary medium of human communication;

therefore, detecting feelings or emotions is very easy and convenient compared to detecting emotions from the face or brain signal. Facial signals need high-quality video cameras, which are expensive, while brain signal acquisition is impossible at any time because it needs proper electrode setup, so speech emotion recognition is the easiest or less costly way to detect human emotions and is used for the human-computer interface.

Speech emotion recognition is a collection of methods that process and classify speech signals to detect embedded emotions [2]. To find the emotion embedded in a speech signal the first method is collecting data and then preprocessing it, the next step is to extract features from the utterance and then train the machine with these features with ground truth and finally trained machine will able to classify the inherent emotion in the speech. There are many types of features from which emotions can be identified from speech. These are prosodic features, spectral features, and Teager energy operators. Spectral features like Mel frequency cepstral coefficients (MFCC), log frequency power coefficients (LFPC), and linear predictive cepstral coefficients (LPC), and are the most used feature to extract emotion. Sato and Obuchi (2007) proved that spectral features give more accurate results than prosodic features [3]. The study by Zhang (2022) also shows the great result using Mel spectrograms for SER [14]. In this paper, we have used MFCC and MFMC based spectral coefficients to extract emotional features from a speech signal.

The forthcoming sections are arranged as follows: Section 2 discusses the work related to speech emotion recognition. Section 3 is the materials and methodology; it explains the dataset, feature extraction, and the proposed methodology. Sections 4 and 5 present the results and discussion, and conclusion, respectively.

## II. RELATED WORK

Ancilin and Milton (2021) [1] introduced a new spectral feature, the Mel frequency magnitude coefficient (MFMC), which is a modification of MFCC. There are two steps involved in extracting the MFMC in comparison to the Mel frequency cepstral coefficient: To begin with, the magnitude square is substituted by the magnitude of the fast Fourier transform. As a second step, the discrete cosine transform used in the decorrelation extraction of MFCC is excluded.

Their experiment results show improved accuracy in higher-order MFMCs to recognize speech emotions. Also, experimental results show that the Mel frequency magnitude coefficients classify emotions better than Mel frequency cepstral coefficients (MFCC), log frequency power coefficients (LFPC), and linear predictive cepstral coefficients (LPC). With the MFCCs alone, emotion was recognized with an accuracy of 81.50% for Berlin databases, 64.31% for RAVDESS databases, 75.63% for SAVEE databases, 73.30% for EMOVO databases, and 56.41% for eNTERFACE databases.

Sönmez and Varol (2017) [4] developed a lightweight, effective speech emotion recognition method called 1BTPDN. In this method, first, they applied 1D discrete wavelet transform (DWT) on raw speech signals. Then features were extracted from each filter by a one-dimensional local binary pattern and a one-dimensional local ternary pattern. By using neighborhood component analysis (NCA), 1024 features are selected from 7680 features. For RAVDESS, EmoDB, SAVEE, and EMOVO databases, they achieved 95.16%, 89.16%, 76.67%, and 74.31% success rates, respectively.

Nagarajan et al. (2020) [5] reported novel triangular filter banks based on bark and ERB frequency scales to recognize speech signals. In this work, MFCCs, and human-factor cepstral coefficients (HFCC) features with different types of triangular filter banks such as TFBCC-M (for MFCC), TFBCC-HF (for HFCC), TFB-B (for bark scale) and TFB-E (for ERB scale) were utilized. The experimental results reported that triangular filter banks were much more effective in extracting cepstral features for recognition and characterizing emotions than conventional triangular filter banks. For EmoDB database, the proposed method got the accuracies of 83.23% and 81.99% for the speaker-dependent (SD) scenario, 75% and 60.94% for the Speaker Independent (SI) scenario, and SAVEE database, 75% and 66.67% for the SD scenario, and 44.17% and 55% for SI scenario.

Choudhury et al. (2018) [6] combined several spectral features with excitation source features. In this work, time and frequency domain spectral features were extracted from the raw speech signals. In this work, Sequential Minimal Optimization (SMO) and Random Forest (RF) were utilized as classifiers. They achieved the accuracies of 75.5% and 75.5% for EmoDB, 55.5% and 55.3% for SAVEE, 99% and 97.7% for TESS older, and 99.1% and 99% for TESS younger for SMO and RF, respectively.

Kathiresan & Dellwo (2019) [7] proposed new dynamic features to improve emotion recognition ability. The proposed features were temporal dynamics (temporal delta and delta-deltas) and cepstral derivatives (cepstral delta and delta-deltas). Two different languages database, i.e. EmoDB for German and SAVEE for English were utilized for this work. By using these new features of different dimensions, they achieved 67.7% accuracy for the EmoDB dataset and 60.8% for the SAVEE dataset.

Langari et al. (2020) [8] proposed adaptive time-frequency features based on fractional Fourier transform fusion with a cepstral coefficient. They used two categories of features to recognize speech signals. The first one is the prosodic feature, i.e. pitch, energy, and duration, and the other one is related to the vocal tract, i.e. cepstral coefficient, formants, and DFT harmonics. This work reported that the

discrete fractional Fourier transform represents the angle on the time-frequency plane, and the rotation of the angle can restore the original data from the distorted signal in other space, which improves the accuracy. This work achieved accuracies of 97.57% for EmoDB, 80% for SAVEE, and 91.46% for the PDREC dataset.

Daneshfar et al. (2020) [9] proposed a three-stage hybrid system for speech emotion recognition, which concludes feature extraction, dimensionality reduction, and feature classification. They used perceptual-spectral features such as MFCC, PLPC, and PMVDR in combination with the prosodic feature like pitch. In their paper, they used a new pQPSO method (QPSO-based approach) and gaussian elliptical basis function (GEBF)-type neural network as a classifier for SER. This work utilized EmoDB, SAVEE, and IEMOCAP datasets and got 79.94%, 59.38%, and 65.71% accuracies, respectively.

Zhang et al. (2022) [14] proposed Mel-IMel dual-channel complementary structure with a convolutional neural network-stacked sparse autoencoder (CNN-SSAE). They focused on the low-frequency part and the high-frequency part of the speech signal using the Mel-spectrogram and the inverse Mel spectrogram, respectively to prove that the two spectrograms are complimentary. The experiment was conducted on the EMO-DB, SAVEE, and RAVDESS datasets, they achieved high accuracies for all the datasets.

### III. MATERIALS AND METHODOLOGY

This section describes the dataset, spectral feature extraction techniques, classification approach, and the proposed emotion recognition model.

#### A. Emotional database of the speech signal

In this paper, we have used SAVEE (Surrey Audio-Visual Expressed Emotion); it consists of acted audio-visual recordings of British English utterances. Four male actors in a visual media lab expressed seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Each subject utters 15 sentences for each emotion, except neutral, which has 30 utterances. It contains 480 audio files in .wav format with 16-bit encoding and a sampling rate of 44.1 kHz [10].

#### B. Pre-processing of the raw speech signal

The preprocessing begins with pre-emphasis, in which the signal is passed from a high pass filter to increase the amplitude of weak signals and reduce the noise in the background, enhancing the raw speech signal. The next step is silence removal. When an audio or speech signal is recorded, there are three parts: voiced, unvoiced, and silent [2]. As noise removal is necessary to remove unwanted signals, silence part removal is also essential because it contains no information, and there is no need to keep that part of the signal. Here, the speech signal is first framed in 20ms duration. For each frame, a zero cross rate is calculated. Then those frames that contain zero ZCR are eliminated, which is the silence part of the speech signal. A zero cross rate indicates the change of rate in signal from positive to negative or negative to positive. When there is no change in signal or the value of ZCR is zero, that means there is no change in signal, or there is no signal. Equation (1) shows the ZCR of the  $m^{\text{th}}$  frame

$$\text{ZCR}(m) = \frac{1}{2N} \sum_{i=1}^N | \text{sgn}(s(k)) - \text{sgn}(s(k-1)) | \quad (1)$$

Where  $\text{sgn}(s(k)) = \begin{cases} -1, & s(k) < 0 \\ +1, & s(k) \geq 0 \end{cases}$

$s(k)$  is the  $k^{\text{th}}$  sample amplitude of the  $m^{\text{th}}$  frame of length  $n$  [11].

Fig. 1 is depicted the raw speech signal and pre-processed speech signal after pre-emphasis and silent part removal using ZCR.

### C. Feature extraction

Features are the essential characteristics of finding emotions from the speech signal. There are various features exist for SER. Here, cepstral coefficients are used to extract spectral features. Spectral features are the short-duration frequency signals obtained using Fourier transform from the time domain signal.

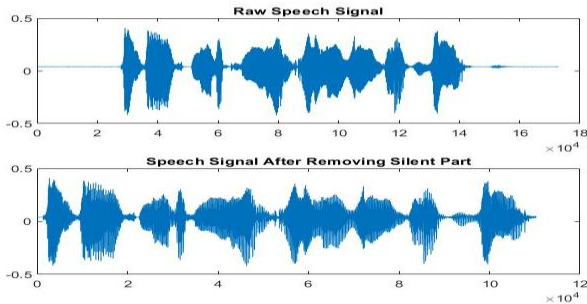


Fig.1. Speech signal before and after preprocessing

Cepstral coefficients are a type of spectral feature used extensively in speech recognition. Although they've been around for some time, they've recently received more attention as researchers have developed computer algorithms that use cepstral coefficients to perform speech emotion recognition.

1) *MFCC*: Mel frequency cepstral coefficient is one of the most used features in speech emotion recognition [11]. It is obtained by dividing the speech signal into frames of short duration, here 20 ms frames are taken with 50% overlapping; then, each time domain frame is converted into a frequency domain using FFT. Fourier representation (2) represents the speech frame as different frequencies, and the Fourier transform coefficients indicate the number of different frequencies contained in the speech frame.

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-i\frac{2\pi kn}{N}} \quad k=0,1,2,\dots,N-1 \quad (2)$$

Where  $N$  represents the number of samples in the speech frame.  $x(n)$  represents the  $n^{\text{th}}$  speech sample, and  $X(k)$  represents the  $k^{\text{th}}$  Fourier transform coefficient.

Next, the energy spectrum is calculated from the FFT signal and passed through Mel filter banks to calculate the subband energies. Then logarithm is applied to those energies. Lastly, a discrete cosine transform is taken to obtain MFCC.

2) *MFMC*: In [1], a new cepstral coefficient was introduced, which is a modification of the Mel frequency cepstral coefficient (MFCC), known as the Mel frequency

magnitude coefficient (MFMC). MFMC is obtained by dividing speech signal into frames of short duration (usually 20-30 ms) with 50% overlapping. Then each time domain frame is converted into a frequency domain using FFT. MFCC considers the energy spectrum, while MFMC uses the magnitude spectrum instead of the energy spectrum for further steps.

Next, the magnitude spectrum is converted into the Mel spectrum and divided into uniform bands. These uniform bands are converted into linear frequency scales and become non-uniform bands. Then the non-uniform bands will be passed through the triangular windows (3) with 50% overlapping to get Mel band magnitude. Lastly, the logarithm is applied to get MFMC.

$$H_m(k) = \begin{cases} \frac{k-fx}{fy-fz}, & fx \leq k \leq fy \\ \frac{fz-k}{fz-fy}, & fy \leq k \leq fz \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Where  $k$  is the frequency,  $fx$  is the lower frequency,  $fy$  is the middle frequency, and  $fz$  is the highest frequency of the triangular window.

$$Y(m) = \sum_{k=0}^{t-1} |X_m(k)| H_m(k), \quad 1 \leq m \leq M \quad (4)$$

Where  $t$  is the number of frequency components in the  $m^{\text{th}}$  band

$$\text{MFMC}(m) = \log_{10}(\sum_{k=0}^{t-1} |X_m(k)| H_m(k)), \quad 1 \leq m \leq M \quad (5)$$

### D. Cross-validation

Data division protocol or cross-validation is a technique used in machine learning to divide the dataset into two parts: the training set and the testing set. The model learns through the training set, and the remaining dataset, i.e. testing set, is used to test the model's performance [12]. Here, hold-out with dataset divided into a ratio of 67:33 and K-fold with  $K=5$  are used to classify emotion.

#### 1) Hold-out :

Hold-out cross-validation is the simplest and most common technique used for cross-validation (CV) [12]. The algorithm of the hold-out technique:

1. Divide the dataset into two parts: the training set and the test set. It can be set to 80:20, 70:30, 60:40, or another ratio as suitable. The first number represents the percentage of the dataset for training, and the other one represents the percentage of the dataset for testing purposes.
2. Training of model on the training set
3. Validation of the model on the test set

In this paper, hold-out is repeated 5 times to train the model and to get better accuracy in the testing phase.

#### 2) K-fold :

The K-fold method is another technique of cross-validation similar to the hold-out method. It splits the dataset into  $K$  equal parts known as folds.  $K-1$  folds of datasets are used for training, and one part of the dataset is used for testing; it repeats itself  $K$  times ( $K=5, 10$ , etc.). Each

validation tests a different set of datasets [12]. Here, 5-fold is selected for training the proposed model, the total dataset is divided into 5 equal parts and 4 parts were used for training and only one part is used in testing. This process repeats itself 5 times, each time new parts of the dataset were tested.

#### E. Backpropagation artificial neural network

A backpropagation artificial neural network is a supervised learning method that contains one input layer, one output layer, and some hidden layer. It enhances the output by changing the weight of neurons according to the errors between the actual output value and the expected output value. The training in this method is carried out by repeating the entire dataset until the errors are minimized [13].

#### F. Proposed emotion recognition model

The framework of speech emotion recognition is shown in fig. 2. Firstly, audio signals are pre-processed using a pre-emphasis filter. A ZCR-based algorithm is used to remove the silence part from the audio signals. The whole audio database is divided into two parts namely training and testing part using data division protocol. From fig. 2, it can be seen that the proposed model is also divided into two parts namely training and testing part using a vertical discontinued line. Except for ground truth, the training and testing part is almost the same. In the present work, spectral coefficients namely MFCCs and MFMCs are extracted as features in the next step. In the next step, significant feature selection is performed using an independent t-test and analysis of variance (ANOVA) for the binary and multiclass classification, respectively. Further, the classifier is trained with a feature vector and ground truth. After training of the classifier, learning parameters are generated. These parameters will help the classifier to emotion recognition during the validation and testing.

The basic steps are as follows:

1. Collecting raw data of speech signal
2. Pre-processing of the speech signal
3. Extracting and selecting the spectral features
4. Classification and validation

#### IV. RESULT AND DISCUSSION

This paper classified seven emotions from the SAVEE dataset, i.e. anger, disgust, fear, happy, neutral, sad, and surprise. Out of these seven emotions, four basic emotions namely anger, fear, happy, and sad are used to build binary classification problems (CP). In the present work, six binary CP (CP1 to CP6) and one multiclass CP (CP7) are used as a classification problem. The accuracies of six binary emotions CPs and one multiclass CP(all seven emotions) using MFCC, MFMC, and a combination of MFCC and MFMC features with hold-out and 5-fold are shown in Table I. Here, CP with numbers 1, 2, 3, 4, 5, and 6 are denoted by the binary emotion classification problems namely Fear versus Anger, Happy versus Anger, Happy versus Fear, Happy versus Sad, Sad versus Anger, and Sad versus Fear, respectively. CP7 is denoted the multiclass classification problem (all seven

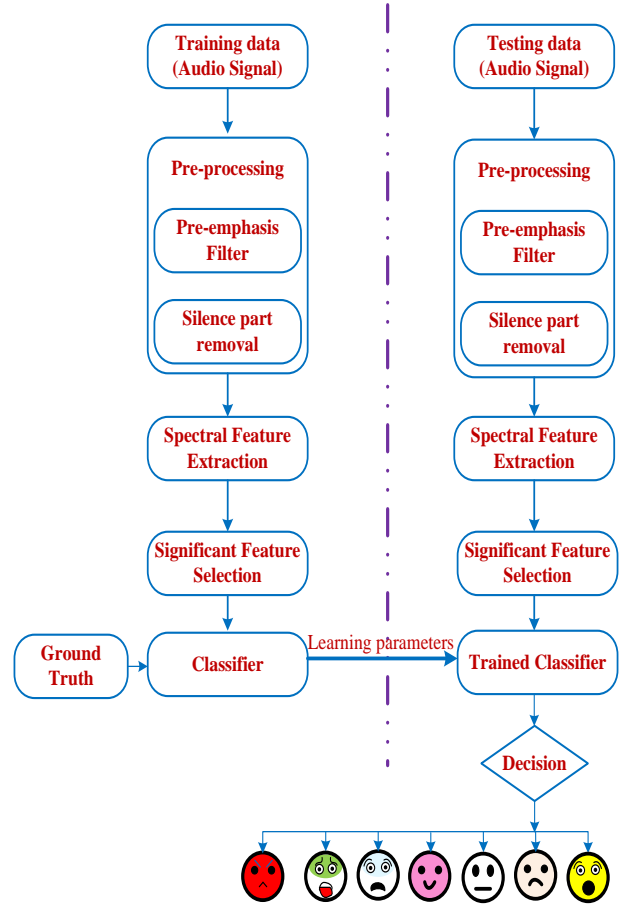


Fig. 2. Proposed model of speech emotion recognition

emotions). The alphanumeric code FV1, FV2, and FV3 are denoted the MFCC, MFMC, and combined MFCC+MFMC feature vectors, respectively.

The proposed model shows the 81.5%, 70.5%, 83.5%, 91%, 92%, 90.5%, and 81.5%, 66%, 83.5%, 94.5%, 93%, 93% accuracies for Fear versus Anger (CP1), Happy versus Anger (CP2), Happy versus Fear (CP3), Happy versus Sad (CP4), Sad versus Anger (CP5), and Sad versus Fear (CP6) for MFCC and MFMC features, respectively with hold-out CV and BPANN classifier. And 97.5%, 92.5%, 94.16%, 98.34%, 97.5%, 98.3%, and 84.98%, 74.16%, 85.82%, 98.32%, 97.5%, 95.82% accuracies for MFCC and MFMC features, respectively with 5-fold CV and BPANN classifier. When both features were combined, it shows the accuracy of 87.0% and 97.5% for CP1, 67.5% and 92.5% for CP2, 79.5% and 97.5% for CP3, 97.5% and 99.16% for CP4, 99.05 and 99.16% for CP5, and 95.0% and 98.33% for CP6 with hold-out and 5-fold CV, respectively. Here we also get 66.4% and 79.26% accuracies for all seven emotions (CP7) of SAVEE dataset for combined features (FV3) using hold-out and 5-Fold CV, respectively. We get the highest accuracy of 99.16% in 5-fold CV for Happy versus Sad (CP4) and Sad versus Anger (CP5) using FV3 feature vector with 5-fold CV and 80.74% accuracy for all seven emotions (CP7) using FV1 feature vector with 5-fold CV.

TABLE I. COMPARISON WITH COMBINED FEATURES ON SAVEE DATASET (%)

CV	Hold-Out			5-Fold		
Features Vector → Emotion CP↓	FV1	FV2	FV3	FV1	FV2	FV3
CP1	81.5	81.5	87.0	97.5	84.98	97.5
CP2	70.5	66.0	67.5	92.5	74.16	92.5
CP3	83.5	83.5	79.5	94.16	85.82	97.5
CP4	91.0	94.5	97.5	98.34	98.32	99.16
CP5	92.5	93.0	99.0	97.5	97.5	99.16
CP6	90.5	93.0	95.0	98.34	95.82	98.33
CP7	65.64	50.0	66.4	80.74	54.28	79.26

Table II shows the classification accuracies for statistically significant features of the FV3 feature vector. Here we get 100% accuracy for Happy versus Sad (CP4) and Sad versus Fear (CP6) in the 5-fold method for significant features of FV3 and get 85.24% accuracy for CP7. The results show that the 5-fold CV gives a better result compared to hold-out except for some cases. When MFCC and MFMC features are merged (FV3) the proposed method gives higher accuracy compared to MFCC (FV1) and MFMC (FV2) are used alone. From Table I and II, it can also see that multiclass CP (CP7) classification accuracy is improved by 5.98% using significant features from FV3 and 5-fold CV.

TABLE II. CLASSIFICATION ACCURACY FOR STATISTICALLY SIGNIFICANT FEATURES

Features Vector	Significant features from FV3	
CV→ Emotion CP↓	Hold-Out	5-Fold
CP1	80	95.84
CP2	65	80.84
CP3	85	96.6
CP4	97.5	100
CP5	97.5	88.34
CP6	97.5	100
CP7	57.98	85.24

A ROC (receiver operating characteristic curve) plot represents the performance of the classification model in graphical form. The ROC plot of CP2-FV3 for 5-Fold, CP6-FV3 for hold-out, and CP1-FV3 for 5-fold of Table I is shown in Fig. 3, Fig. 4, and Fig. 5, respectively.

Table III shows a comparison of the proposed work with some recently reported work. The results show that the proposed method gets a high accuracy of 85.24% for the SAVEE dataset compared to other works done on spectral features, as shown in Table III.

TABLE III. COMPARISON OF ACCURACIES FOR DIFFERENT WORK DONE ON SAVEE DATASET

Author(Year)	Features	Classifiers	Accuracy (%)
Ancilin and Milton ( 2021)	LPCC, LFPC, MFCC, MFMC	Support Vector Machine (SVM)	75.63
Sönmez and Varol (2017)	Time based, Frequency based, Cepstrum based, Wavelet transform based, Texture based, Deep features	DT, LDA, KNN, SVM	76.67
Nagarajan et al. (2020)	Cepstral coefficients	SVM	75, 66.67(SD) 44.17, 55(SI)
Choudhury et al. (2018)	Spectral features	SMO, Random Forest	55.5, 55.3
Kathiresan & Dellwo (2019)	MFCCs, 2TΔ, 2CΔ	SVM DNN	60.8
Langari et al.(2020)	MFCC, LPCC, Format	SVM	80
Daneshfar et al.(2020)	MFCC, PLPC, PMVDR, Pitch	GEBFNN	59.38
Proposed Method	MFCC, MFMC	BPANN	79.26
Proposed Method	Significant FV MFCC, MFMC	BPANN	85.24

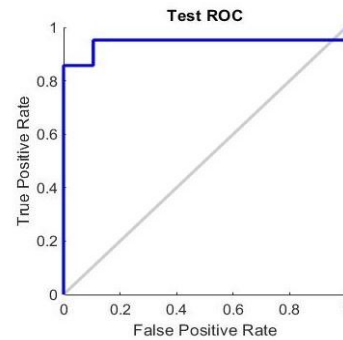


Fig. 3. ROC plot of CP2-FV3 for 5-fold CV (accuracy 92.5%)

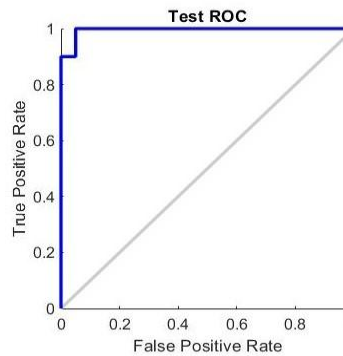


Fig. 4. ROC plot of CP6-FV3 for hold-out CV (accuracy 95 %)

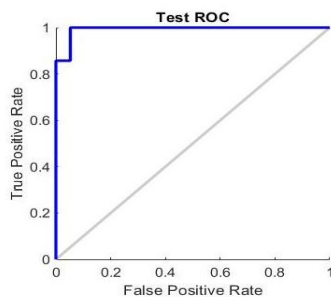


Fig. 5. ROC plot of CP1-FV3 for 5-fold CV (accuracy 97.5%)

## V. CONCLUSION

This paper used the combination of MFCC and MFMC to recognize emotion from the speech signal. MFCC and MFMC are both cepstral coefficients. MFMC feature has not been used in many works. Our work shows that both features, MFCC and MFMC, can stand alone and recognize the emotion from speech signals very well; also, when combined with MFCC, it gives better results. On binary classification, 100% accuracies are obtained for Happy versus Sad and Sad versus Fear classification problem. When MFCC and MFMC have used stand-alone, we get 80.74% accuracy using the MFCC feature for all seven emotions with a 5-fold CV and BPANN classifier. Compared to other work done on SAVEE dataset, we got the highest accuracy of 85.24% for all seven emotions when significant features are selected from a combination of MFCC and MFMC. Hence, the proposed model may help the machine recognize human emotion for further applications.

## REFERENCES

- [1] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, Aug. 2021, doi: 10.1016/j.apacoust.2021.108046.
- [2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, Elsevier B.V., pp. 56–76, Jan. 01, 2020, doi: 10.1016/j.specom.2019.12.001.
- [3] N. Sato and Y. Obuchi, "Emotion Recognition using Mel-Frequency Cepstral Coefficients," 2007.
- [4] Y. Ü. Sönmez and A. Varol, "A speech emotion recognition model based on multi-level local binary and local ternary patterns," *IEEE Access*, vol. 8, pp. 190784–190796, 2020, doi: 10.1109/ACCESS.2020.3031763.
- [5] S. Nagarajan, S. S. S. Netimi, L. S. Kumar, M. K. Nath, and A. Kanhe, "Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales," *Digital Signal Processing: A Review Journal*, vol. 104, Sep. 2020, doi: 10.1016/j.dsp.2020.102763.
- [6] Choudhury, Akash Roy, et al. "Emotion recognition from speech signals using excitation source and spectral features." 2018 IEEE Applied Signal Processing Conference (ASPCON). IEEE, 2018.
- [7] T. Kathiresan and V. Dellwo, "Cepstral Derivatives in MFCCs for Emotion Recognition," 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), 2019, pp. 56-60, doi: 10.1109/SIPROCESS.2019.8868573.
- [8] S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Inform Med Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100424.
- [9] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," *Applied Acoustics*, vol. 166, Sep. 2020, doi: 10.1016/j.apacoust.2020.107360.
- [10] S. Haq, P. J. B. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification."
- [11] K. Chauhan, K. K. Sharma, and T. Varma, "Improved Speech Emotion Recognition Using Modified Mean Cepstral Features," Dec. 2020, doi: 10.1109/INDICON49873.2020.9342495.
- [12] M. W. Browne, "Cross-Validation Methods," 2000. [Online]. Available: [www.idealibrary.com](http://www.idealibrary.com)
- [13] A. T. C. Goh, "Back-propagation neural networks for modeling complex systems," 1995.
- [14] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network", *Applied Science*, vol. 12, Sep. 2022, doi:10.3390/app12199518