2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566046

Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features

Hrithik Patni
Final Year B.Tech
Dept of Electronics and Telecommunication
College of Engineering
Pune,India
hrithikpatni31052000@gmail.com

Akash Jagtap
Final Year B.Tech
Dept of Electronics and Telecommunication
College of Engineering
Pune,India
akashbjagtap1234@gmail.com

Vaishali Bhoyar
Final Year B.Tech
Dept of Electronics and Telecommunication
College of Engineering
Pune,India
vaishubhoyar004@gmail.com

Dr Aditya Gupta
Adjunct Faculty
Dept of Electronics and Telecommunication
College of Engineering
Pune,India
adityagupta2590@gmail.com

***Abstract*—In recent years, increasing attention is given to the research of the emotions present in speech. Various systems are developed aiming to detect the emotions in the speaker's statements. One of the biggest differences between a machine and a human is understanding the emotions of others and behaving accordingly. Researchers are working on bridging this gap by recognizing emotions in speech or voice. This paper proposes a deep learning-based technique for speech emotion recognition (SER). The SER system is based on various techniques that use distinguished modules for emotion recognition. The model differentiates emotions such as neutral state, happiness, sadness, anger, surprise, etc. The performance of the classification system is based on features extracted and generated models. The features utilized in this include energy, pitch, chromagram, mel-frequency spectrum coefficients (MFCC), and Gammatone frequency spectrum coefficients (GFCC). The emotions are classified using a two dimentional Convolutional Neural Network (CNN). The classification model achieved an overall accuracy of 92.59% on the test data which is comparatively better than the previous algorithm. In future, the intention is to increase the system performance and detect more emotions.**

Keywords—Emotion recognition, Feature extraction, Cepstrum, Gammatone Filters, Glottal waveform, CNN.

I. INTRODUCTION

Understanding and detecting human emotions has always been an interesting task for data analytics specialists. Speech Emotion Recognition implies recognizing the emotional facet of speech regardless of the conceptual contents. It capitalizes on the truth that the speech signal reflects underlying emotion through pitch and tone. Dogs and horses can understand human emotion by this phenomenon. Research work in the Speech Emotion Recognition (SER) domain focuses on designing real-time methods of emotion detection of cellphone users, operators in call centres, etc. and also increase efficiency [1].

Embedding emotion into machines is crucial in making machines seem and work similar to humans [2]. Robots would be able to provide appropriate emotional responses and demonstrate personality with emotions. In some cases, it could replace people with computer-generated characters capable of conducting persuasive and natural conversations by requesting human emotions. Machines should understand the emotions conveyed through speech. Using this ability, a complete human-machine logical discussion with integrity and understanding can be achieved. Humans can do this function as a natural communication component of speech, the facility to automate it using programmed devices is still a matter of research. SER is difficult because the emotions are dependent on the speaker and, also defining audio is a challenge [3].

Machine learning algorithms include calculating feature parameters using raw speech data. Extracted features are used to train a model that produces the expected output labels [4]. The choice of features is a common problem with this method. The features which can efficiently cluster data into distinguished categories or classes is unknown. Insight can be given by examining many different aspects, combining unique features into a standard feature vector, or applying many features [5].

SER Applications include services for people with autism. The discovery of an angry caller in an automated call centre to transfer the call to a person [7]. Especially in persuasive communication, it requires special attention to what non-verbal clues the speaker conveys. Accurately measuring and analyzing the voice is a difficult task and entirely subjective.

II. BACKGROUND

Recent advancements in the field of data science and machine learning have attracted researchers to work on neural network-based techniques for speech emotion recognition. P. Harár [1] designed a Deep Neural Network (DNN) model which includes convolutional layers, and pooling layers. They worked on 3 emotion classes i.e. angry, sad, neutral of the German Corpus database which contains 271 labelled recordings. The model designed achieved 69.55% accuracy on emotion prediction. Badshah [2] designed a DNN model made up of 3 convolutional and 3 fully connected layers. Spectrograms of input signal were extracted and the model made predictions for the seven different emotion classes. The proposed method achieved an accuracy of 84.3% for the test data. Wootaek Lim [3] proposed an SER method using concatenated CNN and Recurrent Neural Networks (RNN) without manually creating features. The model build made predictions on 8 different emotions and achieved an average accuracy of

88.01%. S. An [4] used the LSTM-RNN model-based technique. The method was 96.67% accurate while predicting angry emotion, 100% accuracy for sad emotion prediction, and 86.67% accuracy achieved for the neutral case. The drawback with this system is that they have only designed it for only 2 emotions i.e sad and neutral. J. Zhao [5] extracted log-mel spectrograms from the IEMOCAP dataset and used it as input data to the designed 2-dimensional CNN-LSTM model. The model is 95% accurate in predicting only 5 emotion classes. Thiang [6] presented a speech emotion recognition system using Linear Predictive Coding (LPC) and ANN to control mobile robots. Sampled input signals, extracted features, and trained using LPC and ANN.

M.S. Likitha [7] used the standard deviation values and MFCC coefficients as features for emotions. The model was used to classify 3 emotion classes i.e happy, sad and angry. They achieved 80% accuracy even in a noisy environment. Kwon [8] worked on the RAVDESS dataset containing 1440 audio samples of 24 different actors. Extracted spectrogram of the speech signals and Implemented the CNN model and the final prediction accuracy was determined to be 81.75%. Issa [9] used RAVDESS, The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset and the EMO-DB dataset to extract MFCC, Chromagram, spectral contrast features. They used a deep CNN, with fully connected layers, the model achieved an overall accuracy of 71.61% on the RAVDESS dataset,64.3% on the IEMOCAP dataset and 86.1% on the EMO-DB dataset. They studied anger, calm, disgust, fear, happy emotion classes in this project. A. K. Samantaray [10] implemented Multilevel SVM to identify seven emotion classes. A different approach by combining zero-crossing rate, pitch, energy (i.e.prosody features), spectral features, formant frequencies etc. (i.e.quality features), MFCC, LPCC (i.e.derived features) and MEDC (i.e.dynamic features) for a powerful automatic emotion recognition system. The model achieved an 82.26% recognition rate.

S. Yildirim [12] looked into acoustic features of emotion classes like sad, anger, happy, and neutral. Acoustic features studied were duration, Fundamental and formant frequencies, root means square energy(RMSE), Spectral balance, Acoustic likelihood comparison. The model achieved an overall accuracy of 67%. C. H. Wu [13] used 2,033 speech signals for four emotional states and extracted pitch, intensity, formants, shimmer, MFCC features. They used multiple classifiers to recognize emotion classes like fear, sad, angry, happy, etc. By combining acoustic-prosodic information and semantic labels the model achieved an overall accuracy of 83.55%.

The earlier proposed system by [8], [9] has achieved quite a good accuracy but worked only on 2-3 output classes of emotion. The other proposed system by [14] works for 8 output classes and suffers from a lesser accuracy rate. The objective of this study is to propose an SER system using CNN and different features. The paper is arranged in the following way section 3 gives details of the utilized dataset in the study for SER. Section 4 gives details on the proposed methodology, whereas section 5 discusses the observed results. Based on observed results, the conclusion is drawn in section 6.

III. DATASET

Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS)[19] consists of voices of 24 different professionals (i.e 12 male, 12 female). 2 sentences are spoken by all professionals in an even North-American accent. Angry, happy, calm, fearful, sad, disgust and surprise are the various speech emotion expressions used. Every expression is generated in 2 levels of emotional intensity (light, bold), with a neutral expression. Every file out of 1440 files has a unique way of naming the file. The filename consists of 7 parts (e.g., 03-02-05-01-02-02-11.wav). The reason behind using the RAVDESS dataset is, each file is rated ten times on emotional genuineness, and intensity. Also, it has a high level of interrater reliability, emotional validity and test-retest reliability.

IV. PROPOSED METHODOLOGY

In speech signals, features indicate emotions present, and changes in the signal lead to changes in these features, which signify changes in existing emotions. Extracting these features from the speech signal is a major parameter in identifying the emotions of speech.

The speech features are categorised into two classes, long term and short-term features. It is mandatory to use a particular region of the audio signal for the process of feature extraction. Before extracting features of each segment of the complete utterance, the input speech signal is converted into frames using overlapping windows. The emotional state in which the speaker is present can be specified using prosodic features. According to studies, MFCC, duration, formant, energy, pitch and parameters of the speech signal are the factors on which emotions present in the speech are dependent. The different emotional states, parallel changes are observed in the speech rate, pitch, energy, and spectrum. Figure 1 gives details about the block diagram of the proposed algorithm.

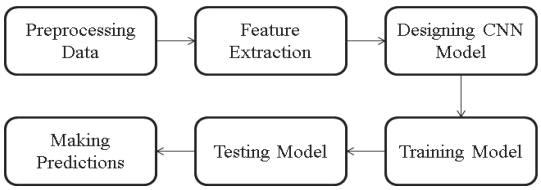

Fig. 1. Block Diagram of the proposed algorithm

Some general analysis about different emotions like in a happy state, the speech signal has a rise in variation range, mean value, pitch variation, and mean value of the energy. When a person is in an angry state it has a higher mean value, the mean value of energy and the variance of the pitch. However, for the sad state, the variation range, mean value, and pitch variation is decreased, the value of energy is reduced, the rate of speaking is slow, and the high-frequency components spectrum is shrunk. In the case of a fear speech signal, it has a varied range of pitch, a high mean value, and improvement in the spectrum of high-frequency

components. Hence emotions present in speech signals can be considered a function of the pitch, energy, and some spectrum feature.

A. Mel frequency cepstrum coefficients (MFCC)

MFCC is extensively used in voice recognition and SER based applications [9]. MFC is illustrating the short term power spectrum of sound. A high emotion recognition rate was achieved using MFCC. Due to MFCC filter bank characteristics, it achieves better frequency resolution and robustness to noise in the low-frequency region in comparison to the high-frequency region. MFCC feature extraction procedure is shown in Figure 2 [7].

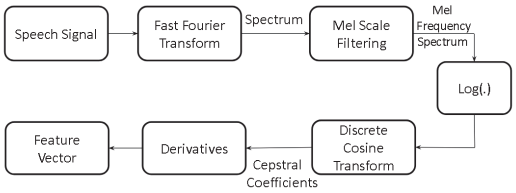

Fig. 2. MFCC feature extraction procedure

In the MFCC Extraction process, the signal is divided into frames. For every frame, the energy spectrum and the Fourier transform is calculated and mapped on the Mel-frequency scale. The discrete cosine transform(DCT) of the Mellogenergies is evaluated. A vector containing 40 MFCC coefficients is extracted after the DCT process. Out of 40 extracted coefficients, only the first 16 coefficients are used as the features in emotion recognition, because the later coefficients are redundant and also decrease the accuracy performance of the system. Equation 1 represents the FFT of the signal x[n].

$$X(k) = \sum_{n=0}^{N-1} \left(x[n] W_N^{nk} \right) \quad (1)$$

B. Gammatone Frequency cepstral coefficients (GFCC)

Features related to audio often have a variety of advances in processes related to speech. One such thing, also inspired by human hearing, is the Gammatone filtering bank [15]. The block diagram is shown in figure 3. This filter bank is the primary imitation of the cochlea. Gammatone Filters (GF) is structured in several mathematical frequency response methods and is used successfully in applications of speech processing. The Gammatone filter and the Cochleagram generation are used in the time domain. Figure 3 represents the block diagram of the GFCC feature extraction procedure.

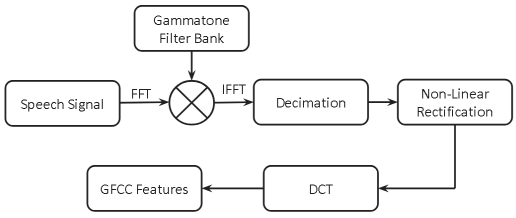

Fig. 3. GFCC feature extraction procedure

12 GFCC features are extracted from the input speech signal and these features are sufficient to describe an emotion class accurately. Equation 2 and 3 represents the transfer function of the filter and standard form of the Gammatone filter respectively.

$$H(z) = 1 + 4m*Z^{-1} + m*Z^{-2} \quad (2)$$

$$G(z) = \frac{1}{(1 - 4m*Z^{-1} + 6m^2*Z^{-2} - 4m^3*Z^{-3} + m^4*Z^{-4})} \quad (3)$$

C. Pitch and Chromagram

Pitch is an important characteristic of the speech signal. It is also known as the glottal waveform. Emotion depends on the subglottal air pressure and tension in vocal folds and this information can be extracted from the pitch signal. Chroma features [17] represent the harmonic content of a short-time window of the input speech signal. The feature vector is extracted from the magnitude spectrum by using a short-time Fourier transform(STFT), Constant-Q transforms (CQT), Chroma Energy Normalized (CENS). The mean value of variance, variation ranges, pitch and contours are different for eight emotions. 13 pitch and chromagram coefficients are extracted from the speech signal. STFT of the signal is given in equation 4 where x[n] is the input speech signal in discrete form and w[n] is the window function.

$$F[\tau,\omega] = \int_{-\infty}^{\infty} x[n] * w(n - \omega) * e^{-i\omega n} \quad (4)$$

D. Root mean square energy (RMSE)

The energy of a signal is considered to be equal to the total magnitude of the signal. Two methods for individually characterizing signal energy are short-time energy and root mean square energy measures. Computing the RMS value from audio samples is faster than alternative energy calculations requiring short-time Fourier transform (STFT) calculations [13]. Equation 4 and 5 represents the energy of signal and Root mean square energy(RMSE) of an energy signal respectively.

$$E = \sum_n |x(n)|^2 \quad (5)$$

$$E_{RMSE} = \sqrt{\frac{1}{N} \sum_n |x(n)|^2} \quad (6)$$

E. Convolutional Neural Networks

CNN has 3 key properties: pooling, weight sharing, and locality. All the properties are capable of boosting speech recognition performance. It is very useful in working with minute shifts in frequency which are very common in speech signals. Frequency shifts are the result of variations

in the length of the vocal tract among different speakers. Small changes in frequency may also occur for the same speaker. These changes cannot be handled properly with other models such as Support vector machines(SVM) and Gaussian Mixture Models (GMM).

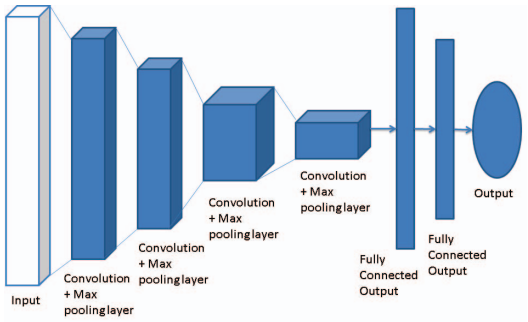

Fig. 4. Convolutional Neural Network(CNN) Architecture

Figure 4 represents CNN the architecture of the CNN model. All the features extracted are combined into a single 1-D vector, the vector consists of 16 MFCC, 12 GFCC, 13 pitch and chromagram and 1 RMSE feature. The model has five 2-D convolutional layers with different filter sizes, which are followed by two fully connected dense layers and an output layer. The input layer has 512 filters size, followed by four 2-D convolutional layers of filter size 256, 256, 128, 64 respectively. The 2 dense layers are of size 64 and 32 respectively. The output layer has 8 output nodes each representing a class of emotions. The activation function used in input and all the hidden layers is ReLU. In the output layer, the softmax activation function is used. Adam optimiser was used with a learning rate of 0.0007 because for a learning rate below it the model overfits and performs poorly on the test data. Figure 5 gives details about the CNN model designed.

Model: "sequential_4"

Layer (type)	Output Shape	Param #
conv2d_20 (Conv2D)	(None, 41, 824, 512)	2560
max_pooling2d_16 (MaxPooling	(None, 21, 412, 512)	0
batch_normalization_4 (Batch	(None, 21, 412, 512)	2048
conv2d_21 (Conv2D)	(None, 20, 411, 256)	524544
max_pooling2d_17 (MaxPooling	(None, 10, 206, 256)	0
batch_normalization_5 (Batch	(None, 10, 206, 256)	1024
conv2d_22 (Conv2D)	(None, 9, 205, 256)	262400
max_pooling2d_18 (MaxPooling	(None, 5, 103, 256)	0
batch_normalization_6 (Batch	(None, 5, 103, 256)	1024
conv2d_23 (Conv2D)	(None, 4, 102, 128)	131200
max_pooling2d_19 (MaxPooling	(None, 2, 51, 128)	0
batch_normalization_7 (Batch	(None, 2, 51, 128)	512
conv2d_24 (Conv2D)	(None, 1, 50, 64)	32832
max_pooling2d_20 (MaxPooling	(None, 1, 25, 64)	0
batch_normalization_8 (Batch	(None, 1, 25, 64)	256
flatten (Flatten)	(None, 1600)	0
dense (Dense)	(None, 64)	102464
dense_1 (Dense)	(None, 32)	2080
dense_2 (Dense)	(None, 8)	264

Fig. 5. Summary of the designed CNN model

V. EXPERIMENTAL RESULTS

Emotions of 288 unseen speech samples of the test dataset are predicted and classified into 8 emotion classes using the designed CNN model consisting of 1 input layer, 7 hidden layers and 1 output layer. The CNN model is trained with the 42 input features (16 MFCC, 12 GFCC, 13 Chromagram, 1 RMSE), batch size equal to 32 and epochs equal to 200. For 32 batch sizes, the model has the highest performance. The training, testing and validation dataset are divided as follows-

Table I: Dataset Classification of Training, Validation and Testing

	Sizes	No of Samples
Training Dataset	0.70	1008
Validation Dataset	0.15	216
Testing Dataset	0.15	216

Equation 7, 8, 9 gives the formula to calculate the precision, recall and F1 Score values respectively. Table 1 gives the details about the model's classification report, which shows the Precision, Recall and f1-score of the emotion classes.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + True\ Negative} \quad (7)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (8)$$

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (9)$$

Table II: CLASSIFICATION REPORT OF THE PREDICTIONS ON TEST DATA

Emotion	Precision	Recall	F1 Score
Angry	91.27	87.34	89.11
Calm	97.53	91.69	94.32
Disgust	93.08	84.21	89.85
fearful	99.16	98.03	98.72
Happy	90.54	93.89	92.68
Neutral	82.82	95.34	85.46
Sad	90.04	97.51	95.89
Surprised	97.18	90.29	93.68
Average(%)	89.90	92.287	91.213
Average Accuracy = 92.592%			

It can be observed that the precision of predicting emotions of each class is more than 82.0%. The Recall of each emotion is greater than 84.0. The f1 score for each of the classes is higher than 85.0. The trained models achieve an overall accuracy on the test data is 92.59%, whereas the earlier presented CNN model [3] achieved a limited accuracy of 81.75% on the RAVDESS dataset. Figure 5 represents the confusion matrix of the expected emotion vs predicted emotion, built using prediction of the emotion classes of the samples present in the test data. It is observed that out of 46 samples of angry emotion 45 are predicted correctly, out of 41 samples of calm emotion 36 are correctly predicted, out of 38 samples of disgust emotion 34 are correctly predicted, out of 36 samples of fearful emotion 34 were predicted, out of 40 samples of happy emotion 37 are correctly predicted, out of 16 samples of neutral emotion 14 are predicted correctly, out of 35 samples of sad emotion 33 are predicted correctly, and out of 36 samples of surprising emotions, 34 are predicted correctly.

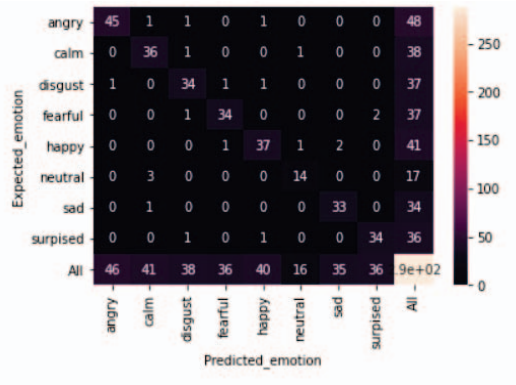

Fig. 6. Confusion Matrix of the predicted and expected emotions

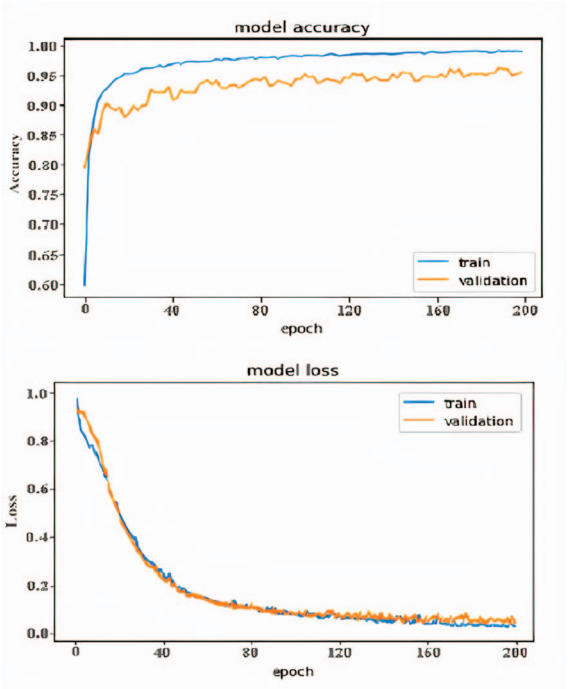

Fig. 7. Training and Testing accuracy graph

VI. CONCLUSION

Speech emotion recognition is increasing nowadays because it can lead to better human-machine interaction. This paper presents a salient features extraction mechanism and Convolutional neural network architecture for speech emotion recognition. By combinations of different features such as MFCC, GFCC, Chromagram and RMSE speech emotion recognition performance can be improved. The model is trained using Convolutional neural networks. The designed model achieved 97.18% accuracy for surprising emotions, 99.1% accuracy for fearful emotion, 97.5% accuracy for calm emotion, 91% for angry emotion, 93% for disgust emotion, 90% for happy emotion, 82% for neutral emotion, and 90% for sad emotion. The system archives an overall accuracy of 92.59% compared to an accuracy of 81.75% for the same data sets designed by [8]. The model classifies eight emotions whereas earlier proposed models have restricted themselves with 3-4 emotions only. Hence the proposed model provides better accuracy with detailed emotion classification. The convolutional neural network improves the accuracy and reduces the computational complexity of the overall SER system. In the future, the focus is on multimodal emotions. Also, testing the proposed model on different, real-time, larger available data sets can be seen as one of the future work.

VII. REFERENCES

[1] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning." 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, 2017, pp. 137- 140.

[2] Badshah, Abdul Malik, et al, "Speech emotion recognition from spectrograms with a deep convolutional neural network." *2017 international conference on platform technology and service (Plat Con)*. IEEE, 2017.

[3] W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks." 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea (South), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699.

[4] S. An, Z. Ling and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs." 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1613-1616.

[5] J. Zhao, X. Mao, L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomed. Signal Process." Control 47 (2019) 312–323

[6] Thiang and Suryo Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot." in Proceedings of International Conference on Information and Electronics Engineering (IPCSIT), Singapore, IACSIT Press, Vol.6, 2011, pp.179-183

[7] M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju, "Speech Based Human Emotion Recognition Using MFCC." IEEE WiSP- NET 2017 conference, 2017

[8] Kwon, Soonil, "A CNN-assisted enhanced audio signal processing for speech emotion recognition." *Sensors* 20.1 (2020): 183.

[9] Issa, Dias, M. Fatih Demirci, and Adnan Yazici, "Speech emotion recognition with deep convolutional neural networks." *Biomedical Signal Processing and Control* 59 (2020): 101894.

[10] A. K. Samantaray and K. Mahapatra, "A novel approach of speech emotion recognition with prosody, quality, and derived features using an SVM classifier for a class of North-Eastern Languages." Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on, pages 372–377, 2015

[11] Chavhan, Yashpalsing, M. L. Dhore, and Pallavi Yesaware, "Speech emotion recognition using support vector machine." *International Journal of Computer Applications* 1.20 (2010): 6-9.

[12] S. Yildirim, M. Bulut, and C. Lee, "An acoustic study of emotions expressed in speech." Proceedings of Inter Speech, pages 2193–2196, 2004.

[13] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels." IEEE Transactions on Affective Computing, 2(1):10–21, 2011

[14] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." Pattern Recognition 44, PP.572-587, 2011.

[15] N. Sugan, N. S. Sai Srinivas, N. Kar, L. S. Kumar, M. K. Nath and A. Kanhe, "Performance Comparison of Different Cepstral Features for Speech Emotion Recognition," 2018 International CET Conference on Control, Communication, and Computing (IC4), 2018, pp. 266-271, doi: 10.1109/CETIC4.2018.8531065.

[16] I. Chiriacescu, "Automatic Emotion Analysis Based On Speech." M.Sc. THESIS Delft University of Technology, 2009.

[17] M. Kattel, A. Nepal, A. K. Shah, D. Shrestha, "Chroma Feature Extraction" Department of Computer Science and Engineering, School of Engineering Kathmandu University, Nepal

[18] W.Q. Zheng, J.S. Yu, Y.X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks." in International Conference on Affective Computing and Intelligent Interaction IEEE, 2015, pp. 827–831

[19] Kerikeri, Leila, and Serrestou, Youssef and Raoof, Kosaiand CIMer, Catherine and Mahjoub, Mohamed andMbarki, Mohamed, "Automatic Speech Emotion Recognition Using Machine Learning." March 2019

[20] Murugan, Harini, "Speech Emotion Recognition Using CNN." International Journal of Psychosocial Rehabilitation. 24. 10.37200/IJPR/V24I8/PR280260 (2020).

[21] Lee, M.; Lee, Y.K, Lim, M.-T.; Kang, T.-K, " Emotion Recognition Using Convolutional Neural Network with Selected Statistical Photoplethysmogram Features." *Appl. Sci.* 2020, *10*, 3501. https://doi.org/10.3390/app10103501

[22] Shahsavarani, B. S, "Speech Emotion Recognition using Convolutional Neural Networks." (Master thesis, The University of Nebraska-Lincoln) (2018).

[23] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition." IEEE/ACM Transactions On Audio, Speech, and Language Processing, Vol. 22, No. 10, October 2014

[24] Y. Kuang and L. Li, "Speech emotion recognition of decision fusion based on DS evidence theory." Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, pages 795–798, 2013