



Convolution Neural Network for Speech Emotion Recognition

¹Shruti Garg, ²Gaurav Kumar

¹Assistant Professor, ²Student

¹Birla Institute of Technology, Mesra, Ranchi, India

Abstract: Automatic Emotion Recognition from audio signals is a challenging task for machines however it is quite easy for human. The audio signals vary according to different features such as pitch, loudness, timbre, speech rate and pauses. In this work, speech emotions were classified seven different emotions namely happy, sad, surprised, neural, angry, disgust, fear. A convolution neural network of nine layers has been applied to classify above emotions. Feature extraction is done through Mel-Frequency Cepstral Coefficient (MFCC).

Index Terms - MFCC, CNN, SER, SAVEE.

I. INTRODUCTION

The speech emotion recognition (SER) comes under category of human computer interaction [2]. These systems can be used in call centers, criminal identification, neuroscience, psychological problem identification etc. [3]. The speech signal varies in term of pitch, frequency, intensity, speaking rate and voice quality for different emotions [4].

Speech emotion recognition has been done for only for two sentiments i.e. happy and sad. On the other hand, different researcher classified multiple emotions such as happy, angry, sad, disgust, fear, surprise and neutral [7]. Categorizing in multiple emotions in more challenging than only two emotions. Convolution neural network has been used to categorized multiple emotions in this work. Because the classification using machine learning models requires lots of preprocessing and won't give much accuracy [8]. The emotions in speech signals are classified in two phases 1. Feature extraction and 2. Emotion classification. A block diagram for tradition SER system is shown in figure 1.

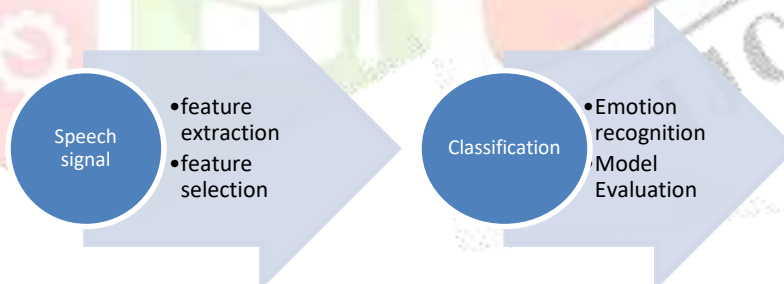


Figure 1: Tradition speech emotion recognition system

Whereas in SER using convolution neural network (CNN) required only classification step because the feature extraction in CNN is done within the layers.

Several researchers have been classified speech signals using statistical, machine learning or deep learning methods. SER using hidden markov model (HMM), gaussian mixture model (GMM) and k-nearest neighbor (KNN) has been done in [5][8]. SER using different machine learning models has been done in [9]. A survey of CNN for SER system has been presented in [3][4]. The SER using multilayer perceptron, radial basis neural network, probabilistic neural network and deep neural network has been done in [10]. The classification of speech has been done by many for different emotions. The neural networks were outperformed for emotion classification. The convolution neural networks give high accuracy for many classification problems because they convolve features in data at many levels also learns by experience.

The research papers [15]-[22] shows work in voice classification using traditional techniques as well as deep learning techniques having similarity in work.

The further subsections of paper describe methodology in section II, results and discussions in section III and conclusion remarks are given in section IV.

II. METHODOLOGY

2.1 Dataset

The dataset used in this research paper is SAVEE (Survey Audio Visual Express Emotions). The SAVEE dataset has been designed for automating the recognition feature of emotions.

The database consists of recording of four male actors in seven different emotions. The audio visual has been recorded from the standard TMIT corpus and phonetically balanced emotions.

The audio-visual dataset consisting of six basic emotions and one neutral. The TMIT sentences being uttered by four English actors with total 480 utterances. Ten subjects has been used to evaluate the database[11].

The files are audio and are named in alike that the prefix word describes the emotion different classes such as:

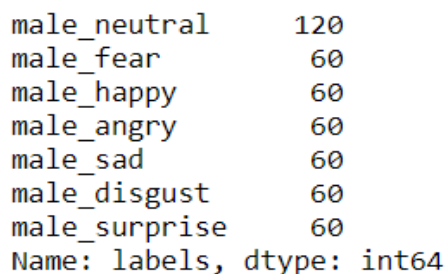
- 'anger'='a'
- 'disgust'='d'
- 'fear'='f'
- 'happiness'='h'
- 'neutral'='n'
- 'sadness'='sa'
- 'surprise'='su'

As a rule, a SER(Speech Emotion Recognition) framework is includes two sections: first is the preprocessing part that concentrates appropriate highlights or data and a classifier that apply those highlights to perform on emotion recognition. All the audio files are arranged as first two alphabets are actor initials followed by underscore then first character of emotion. For example DC_d03.wav is for 3rd disgust sentence uttered by actor DC.

2.2 Machine Intelligence Library

2.2.1 Pre-Processing

Pre – processing has been an important step, the length of the audio signal needs to be consistent length, the discourse length of the audio signal has been maintained for each data by auditing the edges. Normalization has been performed along with missing value replacement. The different types of Emotion present in the SAVEE dataset are shown in figure 2:



```

male_neutral      120
male_fear         60
male_happy        60
male_angry        60
male_sad          60
male_disgust      60
male_surprise     60
Name: labels, dtype: int64

```

Figure 2: Number of instances of each class

2.2.2 Feature Extraction

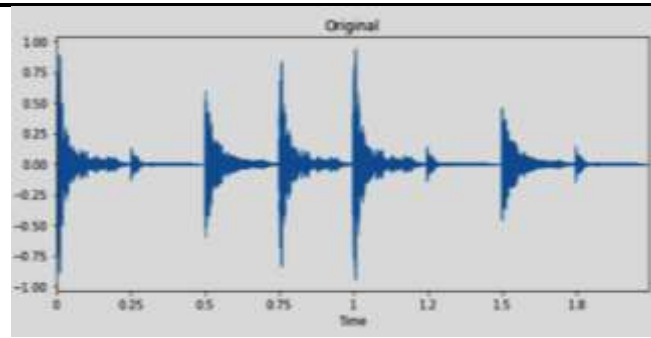
Feature Extraction is an important step to make machines learn through feature categorization. In this paper the feature extraction has been done through MFCC.

MFCC, short for Mel-Frequency Cepstral Coefficient. MFCC is a sentence, is an "image" of the vocal tract that delivers the sound. The initial phase in any programmed is speech acknowledgment framework is to remove the valuable component that recognize the pieces of the sound sign that are useful for distinguishing the etymological substance and disposing of the various stuff which conveys data like foundation commotion, feeling and so on. Mel Frequency Cepstral Coefficients (MFCCs) are an element broadly utilized in programmed discourse and speaker acknowledgment [12].

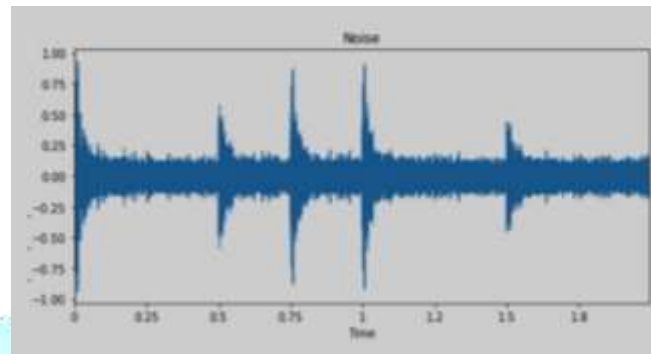
Librosa function has been used to convert the audio signals into two tuples of array to generate features.

2.2.3 Data Augmentation

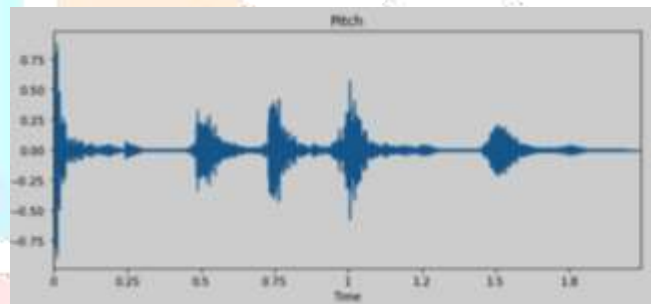
To generate the copies or syntactic data for increment in dataset, augmentation has been applied through injection of noise, pitch change, time and speed change to syntactic data. The augmentation with injection of noise and pitch change is done in this work. Signal after adding white noise and pitch change is shown in figure 3.



(a)



(b)



(c)

Figure 3: (a) original signal (b) Signal after adding white noise (c) Signal after pitch tuning

2.3.4 Convolution Neural Network (CNN)

Convolutional Neural Network is a type deep learning algorithm used as a feature extractor as well in end to end learning. In this paper CNN has been used along with Softmax classifier. The CNN used here consists of 9 layers, in which 8 layers of 1D CNN and 1 layer of dense layers [13]. The general architecture of CNN used in present work is shown in figure 4.

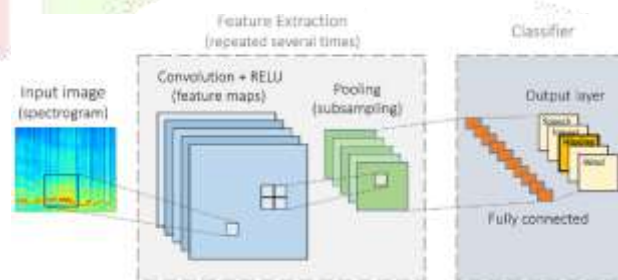


Figure 4: Convolutional neural network framework

For the development of eight CNN layers, ReLU activation function has been used along with 8×8 filters, along with the dropout layer to reduce the dimensions.

Dense layer has been used in last in order to flatten the output from 2D to 1D with Softmax Classifier to achieve the classification. The learning algorithm used here is Root Mean Square Propagation (RMSProp) [14].

The following experiment has been performed on intel i5 8th generation processor with 32 GB RAM with python 3 on jupyter notebook platform. The standard functions used in the Keras package.

III. RESULTS and DISCUSSIONS

The training and test set has been divided into 80:20. There were seven classes available in labelled dataset. The confusion matrix achieved after experiments is shown in figure 5:

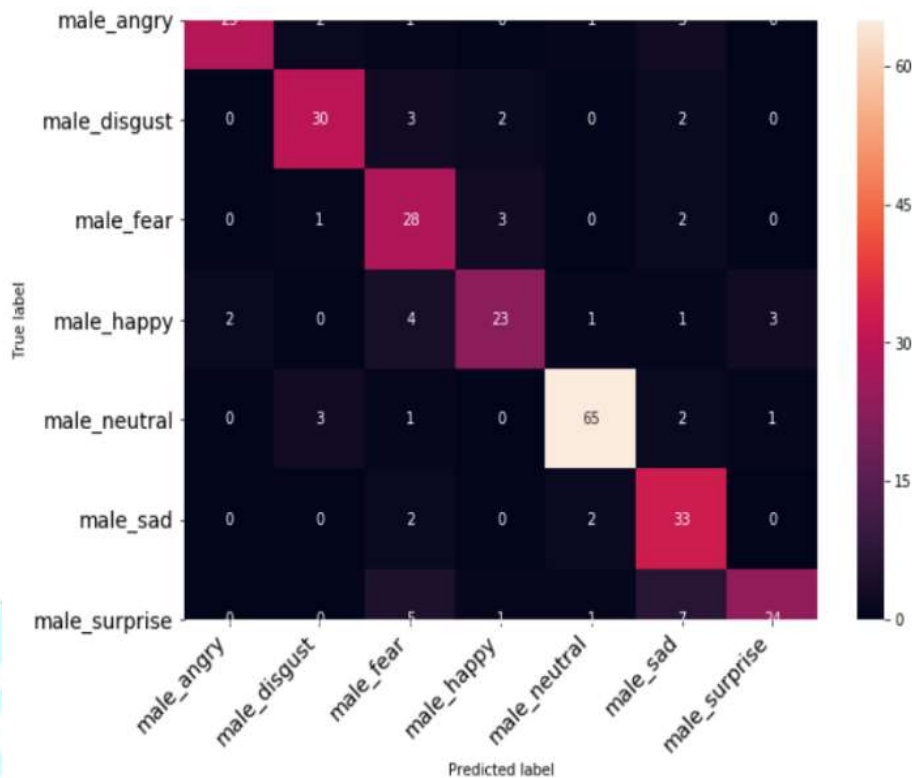


Figure 5: Confusion matrix obtained after experiments

The average accuracy obtained here is 84.31% after evolving the neural network for 50 epochs.

The dataset has been evaluated on the basis of confusion matrix. The precision, recall, F1-score and support found by experiments are shown in table 1.

	Precision	Recall	F1Score	Support
angry	.76	.68	.68	36
disgust	.58	.70	.63	37
fear	.32	.85	.46	34
happy	.68	.38	.49	34
neutral	.90	.62	.74	72
sad	.83	.65	.73	37
surprise	.58	.37	.45	38

Table1: Showing statistical metrics based on confusion matrix.

Contrasting present work with the previous work in speaker independent accuracy found in emotion classification is in between 50-70% shown in table 2.

REF NO.	Feature/Dataset	Accuracy
[14]	MEL+SAVEE	70.00%
[15]	SAVEE	44.18%
[16]	MFCC	57.2%
[21]	MFCC+SAVEE	51%
[22]	SAVEE	78.44%
Present work	MFCC+SAVEE	84.31%

Table 2: Accuracy comparison from previous work

CNN works efficiently in many classification problems but suffers with the problem of overfitting as well. In order to stop overfitting data augmentation is applied here.

IV. CONCLUSION

The present work shows the implementation of convolution neural network for speech emotion recognition. It is difficult to classify emotions because they subjective in nature. The SAVEE dataset has been used to classify seven emotions in speech signals namely anger, disgust, fear, happiness, neutral, sadness and surprised. Experiment shows that the performance of CNN is better than traditional statistical and machine learning methods shown in table 2. The experiments also been performed with and without data augmentation. The results with data augmentation was found better than without augmentation.

REFERENCES

- [1] S. Haq and P.J.B. Jackson. "Speaker-Dependent Audio-Visual Emotion Recognition", In Proc. Int'l Conf. on Auditory-Visual Speech Processing, pages 53-58, 2009.
- [2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communication ACM*, vol. 61, no. 5, pp. 90–99, 2018.
- [3] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.
- [4] Khalil, Ruhul Amin, et al. "Speech emotion recognition using deep learning techniques: A review." *IEEE Access*, vol. 7, pp. 117327-117345, 2019.
- [5] Abdulbasit Al-Talabani, Harin Sellahewa, Sabah A. Jassim, "Emotion recognition from speech: tools and challenges," *Proc. SPIE 9497, Mobile Multimedia/Image Processing, Security, and Applications 2015*, 94970N (21 May 2015); <https://doi.org/10.1117/12.2191623>
- [6] Lanjewar, Rahul B., Swarup Mathurkar, and Nilesh Patel. "Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k-nearest neighbor (k-nn) techniques." *Procedia computer science* 49 (2015): 50-57.
- [7] O. Kwon, K. Chan, J. Hao, T. Lee, "Emotion recognition by speech signal," in Proc. EUROSPEECH, Geneva, Switzerland, 2003, pp. 125–128.
- [8] Mao, Shuiyang, et al. "Revisiting Hidden Markov Models for Speech Emotion Recognition." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- [9] Kerkeni, Leila, et al. "Automatic Speech Emotion Recognition Using Machine Learning." *Social Media and Machine Learning*. IntechOpen, 2019.
- [10] Mohanty, Mihir Narayan, and Hemanta Kumar Palo. "Segment based emotion recognition using combined reduced features." *International Journal of Speech Technology* 22.4 (2019): 865-884.
- [11] Jackson, P., and S. Haq. "Surrey audio-visual expressed emotion (savee) database." University of Surrey: Guildford, UK (2014).
- [12] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." *arXiv preprint arXiv:1003.4083* (2010).
- [13] Chua, Leon O., and Tamas Roska. "The CNN paradigm." *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 40.3 (1993): 147-156.
- [14] Chenchah, Farah, and Zied Lachiri. "Speech emotion recognition in acted and spontaneous context." *Procedia Computer Science* 39 (2014): 139-145.
- [15] Liu, Zhen-Tao, et al. "Speaker-Independent Speech Emotion Recognition Based on CNN-BLSTM and Multiple SVMs." *International Conference on Intelligent Robotics and Applications*. Springer, Cham, 2019.
- [16] Kim, Eun Ho, et al. "Improved emotion recognition with a novel speaker-independent feature." *IEEE/ASME transactions on mechatronics* 14.3 (2009): 317-325.
- [17] A. Batliner, B. Schuller, D. Seppi, S. Steidl, Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in *Emotion-Oriented Systems*. Springer, 2011, pp. 71–99.
- [18] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.
- [19] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5005–5009.
- [20] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-errorbased learning for continuous emotion recognition in speech," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2367–2371.
- [21] Kishore, KV Krishna, and P. Krishna Satish. "Emotion recognition in speech using MFCC and wavelet features." *2013 3rd IEEE International Advance Computing Conference (IACC)*. IEEE, 2013.
- [22] Yogesh CK, Hariharan M, Ngadiran R, Adom AH, Yaacob S, Berkai C, Polat K. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*. 2017 Mar 1;69:149-58.