Original Article

Speech Emotion Recognition And Analysis Using Mfcc & Cnn

Saurabh Aggarwal[1] and Saurav Kumar[2]

[1,2]Department of Information Technology, SRM Institute of Science and Technology, Modinagar, Ghaziabad (U.P.) India
[1]saurabhkumar933@gmail.com
[1]ORCID ID: 0000-0003-4416-6486, [2]ORCID ID: 0000-0001-5164-9852
DOI: 10.47750/pnr.2022.13.S07.906

Abstract

Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. Selection of suitable feature sets, design of a proper classification methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems. Detecting emotions is one of the most important marketing strategies in today's world. You could personalize different things for an individual specifically to suit their interest. Once machines have the capability of understanding the emotions of a person, it will greatly enhance the user experience. In this report, we will try to classify the emotion of a voice (audio clip) using different algorithms of Artificial Intelligence (AI). Based on the accuracy rates, a suitable choice could be made to make such application in the future.

Keywords: Audio Analysis; Speech Emotion; AI; MFCC; CNN; Voice Recognition

I. INTRODUCTION

This paper presents the implementation of emotion detection from voice. Speech is a fundamental means of communicating not only words, but also a vast range of human emotions. The speech emotion recognition involves analysis of the speech signal to identify the appropriate emotion based on training its features such as pitch. Detecting emotions is one of the most important marketing strategies in today's world. You could personalize different things for an individual specifically to suit their interest. For feature extraction and testing of a speech signal a good number of algorithms have been formulated. Few of them are Mel Frequency cepstrum coefficients (MFCC) and Convolutional Neural Networks (CNNs). In early 2000s, researchers exploring human-computer interaction discovered that people tend to interact with computers as if they were other people, respond to praise and criticism from computers the same way they respond to similar feedback from humans. "Emotions are a fundamental part of the human experience – but they've long been ignored by technology development because they seemed difficult to quantify and because the technology didn't really exist to read them. This has resulted in sometimes frustrating user experiences" [4].

The outline of this paper is as follows: Section II describes the literature survey, Section III illustrates about the algorithms of Convolutional Neural Networks (CNNs) & Mel Frequency cepstrum coefficients (MFCC). Section IV & V describes the final result and conclusion.

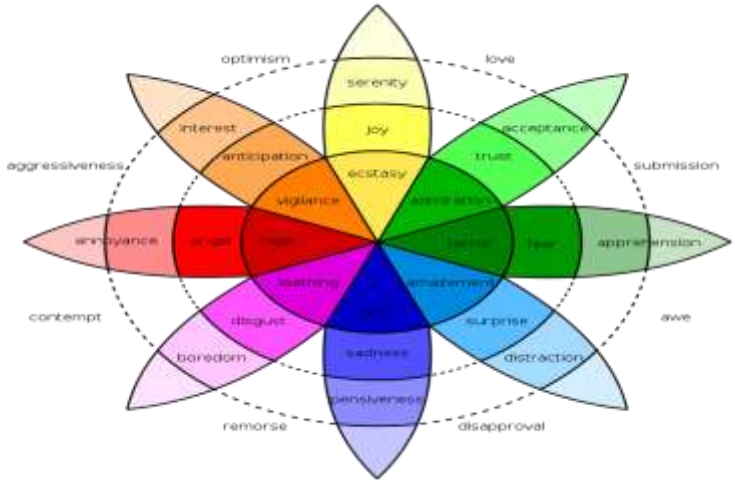

Fig.1: The emotion wheel

II. LITERATURE SURVEY

This section deals with the previously made works related with speech emotion recognition (SER), Mel-Frequency Cepstrum Coefficients (MFCC) Nicholson et al. [5] studied eight different emotions such as joy, teasing, fear, sadness, disgust, anger, surprise and neutral. Both prosodic and phonetic features were studied. Artificial neural network was used as the classifier. An overall recognition accuracy of 50% is reported for this study. Liqin Fu et al. [6] used a hybrid approach of spectral and prosodic features. The studied emotions were anger, disgust, happy, sad and neutral in their work. Hidden Markov model was the used classifier. An average recognition rate of 78% was obtained. Murugappan et al. [7] worked on the gender identification through vocal emotion classification with DWT and MFCC features. The emotions considered were angry, happy and sad. And SER technique based on an enhanced brain emotional learning (BEL) model, which is stimulated by emotional handling mechanism of limbic structure in the brain was proposed. However, BEL technique had its drawbacks and to overcome that, Genetic Algorithm (GA) was employed for updating the weights of BEL. The proposed technique had obtained a maximum average recognition accuracy of 90.28% in case of speaker-dependent speech emotion recognition while the highest average accuracy of 64.60% was obtained in case of speaker-independent speech emotion recognition [11]. In the same year, Sparse Hierarchical Coding (SC) approach was presented for emotion recognition systems. The proposed system comprised of motivated perceptual features (FPH) which resulted in a better and improved prediction of valence and arousal values compared to that of using only prosodic features (F200). Further, in the second stage of this technique, the proposed feature set was enhanced through an unsupervised feature learning method to automatically mine the non-linear relationship among the emotional speech data [12].

III. METHODOLOGY

3.1 **Convolutional Neural Network**

This section presents the deep learning convolutional neural network architecture that was implemented to classify emotions.

It's a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. In case of CNN the neuron in layer will only be connected to a small region of layer before it, instead of all neuron in fully connected manner. CNN compares the images piece by piece. The pieces it looks for are called features. By finding rough feature matches, in roughly the same position in two images, CNN get a lot better at seeing similarity than whole image matching scheme.

While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

3.2 **Mel-Frequency Cepstrum (MFCC)**

Mel-frequency cepstral coefficients(MFCC) is One of popular audio feature extraction method.

The key objectives are:

- Remove vocal fold excitation — the pitch information.
- Make the extracted features independent.
- Adjust to how humans perceive loudness and frequency of sound.
- Capture the dynamics of context.

It comprises of the following steps:

A. Frame Blocking

B. Windowing

C. Fast Fourier Transform

D. Mel Frequency Warping

The acoustic characteristic of the speech signal is Feature.

A small amount of data from the speech signal is extracted to analyse the signal without disturbing its acoustic properties.

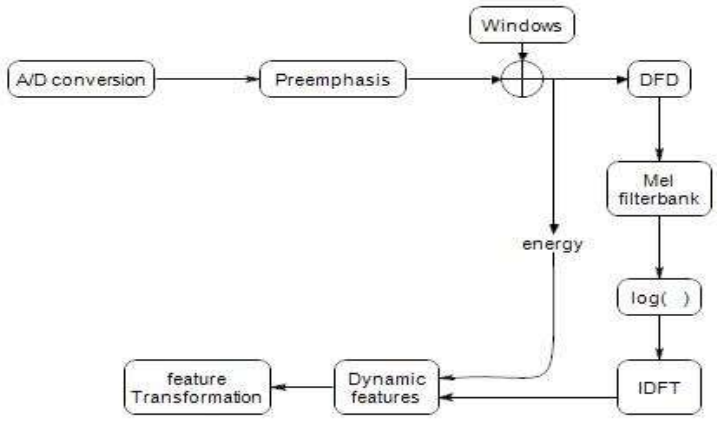

Fig.2: The MFCC flow-chart

IV. RESULT AND DISCUSSION

We first provide a brief description about the AI powered (CNN & MFCC) speech emotion detection and the implementation procedure for the classification. Corresponding classification statistics will then be provided in this section.

	0	1	2	3	4	5	6	7	8	9	label
0	-61.863117	-61.863117	-61.863117	-61.863117	-61.863117	-61.863117	-61.863117	-61.863117	-61.863117	-61.863117	male_surprised
1	-61.508526	-61.508526	-61.508526	-61.508526	-61.508526	-61.508526	-61.508526	-61.508526	-61.508526	-61.508526	male_surprised
2	-56.729975	-56.170952	-56.099735	-54.802708	-55.267311	-56.157734	-56.157734	-56.157734	-56.157734	-56.157734	male_angry
3	-63.524635	-63.524635	-63.524635	-63.524635	-63.524635	-63.524635	-61.551125	-63.393326	-63.524635	-63.524635	male_fearful
4	-43.072586	-44.302555	-44.097376	-44.237846	-44.318237	-44.431492	-44.208611	-44.297836	-43.697372	-43.814746	male_fearful
5	-42.992447	-42.992447	-42.871517	-42.288982	-40.687477	-40.202488	-40.153965	-42.670231	-41.780715	-41.448723	male_angry
6	-55.694332	-55.667202	-55.608786	-57.032059	-58.866116	-57.126228	-58.246738	-58.241409	-58.540890	-57.162392	male_disgust
7	-67.736809	-67.736809	-67.684508	-67.140762	-67.736809	-67.736808	-67.736809	-67.736809	-67.736809	-67.736809	male_sad
8	-61.581672	-61.581672	-61.581672	-61.581672	-61.581672	-61.581672	-61.581672	-61.581672	-61.581672	-61.581672	male_sad
9	-46.569328	-42.395744	-41.461449	-39.356552	-37.300949	-35.989304	-36.153702	-37.158661	-39.649632	-39.162895	male_disgust

Fig.3: Characteristic classification table

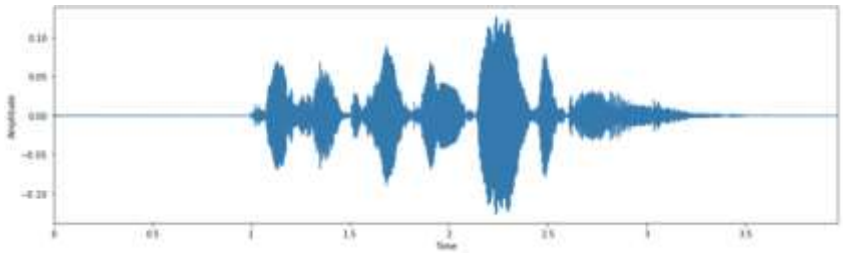

Fig.4: Speed waveform characteristic

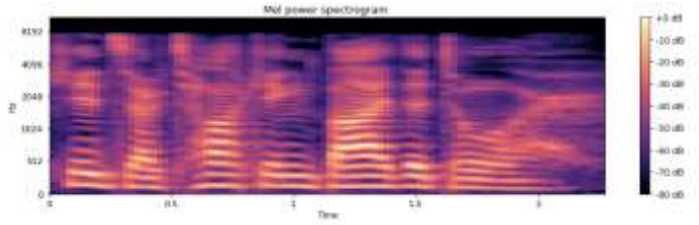

Fig.5: Mel power spectrogram

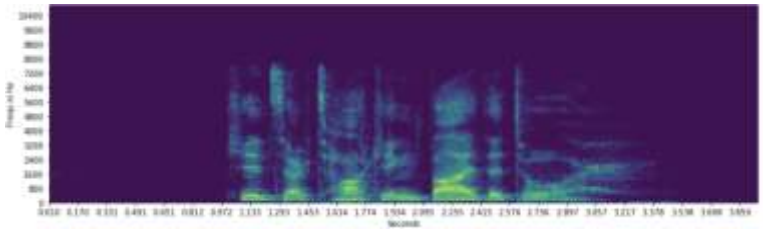

Fig.6: spectrogram 02

EXPECTED OUTCOMES
Neutral
Sad
Happy
Angry
Calm
Surprised
Fearful
Disgust
None

Fig.7: Expected outcomes

V. CONCLUSION

Speech emotion recognition is one of the rapidly developing techniques as it deals with the interaction between machine and human. In this proposed technique, an enhanced speech emotion recognition is carried out over eight basic emotions of angry, happy, sad, neutral, calm, disgust, surprise and fear. We have shown that it is possible to obtain a significant improvement using this method. The reason for evolution of all these technique is just for the advancement and reduction to time that assist people.

VI. REFERENCES

[1] Koteswara Rao Anne, Swarna Kuchibhotla, Hima Deepthi Vankayalapati: "Acoustic Modeling for Emotion Recognition", SpringerBriefs in Electrical and Computer Engineering 2019.

[2] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell, "On the Importance of Both Dimensional and Discrete Models of

Emotion" School of Psychology, The University of New South Wales, Sydney 2052, Australia.

[3] Seiichi Nakagawa, Member, IEEE, Longbiao Wang, Member, IEEE, and Shinji Ohtsuka "Speaker Identification and Verification by Combining MFCC and Phase Information" IEEE transactions on audio, speech, and language processing, vol. 20, no. 4, may 2018.

[4] R. Elbarougy and M. Akagi. Cross-lingual speech emotion recognition system based on a three-layer model for human perception. pages 1–10, 2019.

[5] Y. Wang and L. Guan. An investigation of speech-based human emotion recognition. pages 15–18, 2017.

[6] W. Han, C. F. Chan, C. S Choy, and K. P. Pun. An efficient MFCC extraction method in speech recognition. pages 145–148, 2016.

[7] Z. Huang, M. Dong, Q. Mao, and Y. Zhan. Speech Emotion Recognition Using CNN. pages 801–804, 2018.

[8] R. Hibare. Feature Extraction Techniques in Speech Processing : A Survey. International Journal of Computer Applications, 107(5):1–8, 2018.

[9] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech, pages 223–227, 2019.

[10] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller. Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks. Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), to be published, 2017.

[11] Jia Rong, Gang Li , Yi-Ping Phoebe Chen "Acoustic feature selection for automatic emotion recognition from speech", Elsevier , Information Processing and Management volume 45 (2016).

[12] Jaswinder Singh, Jaswinder Kaur "Proposed Security System to Embed Fingerprinting and Voice Recognition for ATMs" International Journal of Advanced Research in Computer Science and Software Engineering 5(5), May- 2015, pp. 256- 261.

[13] Chandrashekhar.S.PatilȦ and Gopal.N.DhootȦ "A Security System by using Face and Speech Detection" International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

[14] Seiichi Nakagawa, Member, IEEE, Longbiao Wang, Member, IEEE, and Shinji Ohtsuka "Speaker Identification and Verification by Combining MFCC and Phase Information" IEEE transactions on audio, speech, and language processing, vol. 20, no. 4, may 2019.

[15] Jia-Min Ren, Ming-Ju Wu, and Jyh-Shing Roger Jang, Member, IEEE "Automatic Music Mood Classification Based on Timbre and Modulation Features" IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6, NO. 3, JULY- SEPTEMBER 2015. pp. 236-246.