HEDCM: Human Emotions Detection and Classification Model from Speech using CNN

Anjali Tripathi, Upasana Singh, Garima Bansal, Rishabh Gupta, and Ashutosh Kumar Singh

National Institute of Technology, Kurukshetra, Haryana, India

Abstract

Emotion Detection and Classification using Speech is still in an emerging stage in the research community. In this paper, a speech-based classification model is proposed, categorizing speech into six basic emotion categories: anger, surprise, sadness, fear, happiness, and neutral. There are different techniques which are used in speaker discrimination and sentiment analysis task. Every method has its pros and cons. In the proposed model, the Mel Frequency Cepstrum Coefficient (MFCC) feature calculates the various attributes of the speech signal and Convolutional Neural Network (CNN) model to classify different types of emotions. The numerical evaluation has been performed with the RAVDESS dataset, which consists of recorded audio files by 12 Actors & 12 Actresses. For the emotion classification, prediction accuracy of 70.22% was obtained in our proposed model, along with a model accuracy of 96%. The accuracy is improved compared to other similar implementations as deep learning works better than old ML classification methods.

Keywords 1

MFCC, Deep neural network, Emotion recognition, Feature extraction, CNN

1. Introduction

Humans can predict the probability of emotions just by hearing, viewing, or speaking to a person; the need of the hour is to provide similar functionality to the machines. In such a case, post identification of humans' emotions, the device could act accordingly, keeping into consideration the user's requirements and priorities. Understanding human sentiments have always been a captivated area of research, and it is gaining a lot of concern from researchers and other disciplines.

Automatic emotion recognition has a natural application in this space since it can be used not only for automatic user feedback but also to construct more pleasant and natural conversation partners. Emotion recognition technology is essential to such assistants to become more seamlessly integrated into the users' daily lives. Detection of emotion from the audio is not an easy task given the accompanying reasons: In separating between different feelings which specific features of speech are more valuable isn't clear. Due to the various sentences, speakers' talking styles, speaking rates of speaker's greeting fluctuation were presented due to which speech signal features get straightforwardly influenced [1].

The study of individuals' feelings or frames of mind towards an occasion, discussion on themes, or general. The main distinguishing qualities that humans possess are their ability to demonstrate and understand emotions through various communication modes. Humans can comprehend even complex emotions in any way, and these emotions guide the understanding of their interpersonal relationships daily. As speech-based assistants' popularity surges, one of the most visibly

ACI'21: Workshop on Advances in Computational Intelligence at ISIC 2021, February 25–27, 2021, New Delhi, India
EMAIL: anjalitri65@gmail.com (A. Tripathi)
ORCID: 0000-0002-9860-7577 (A. Tripathi)

apparent deficiencies of such systems are their inability to understand their users' emotions and, further, to demonstrate any emotion in return [13] [7].

Numerous scientists have proposed useful features of speech that contain the information of emotion, for example, frequency of pitch, recurrence [3], energy [2], Linear Prediction Coefficients (LPC), MFCC, and Linear Prediction Cepstrum Coefficients (LPCC). Moreover, numerous analysts investigated a few classification methods, for example, Neural Networks (NN) [5], Kernel Regression and K- closest Neighbors (KNN), Hidden Markov model (HMM) [6], Gaussian Mixture Model (GMM), Maximum Likelihood Bayesian classifier (MLC) and Support vector machines (SVM). Although these systems can perform fundamental sentiment analysis, they do not inherently process the richness of emotion in everyday speech. To classify these emotions in various categories like Happy, Sad, Anger, surprise, neutral, and fearful. We propose a model that not only detects the sentiment, but it also classified that into different categories. In the HEDC model, the MFCC method and CNN classifier are used to classify the emotions. The system gives better accuracy compared to other pre-existing models. Work well in noisy data and classify sentiment into different categories like Happy, sad, anger, etc. Previously, all the work is done only on fixed-size input, but the HEDC model can also take dynamic size input. HEDC model can Easley detect the emotions in the case of Homophones words.

The paper is structured as Section 2 consists of the comparison in techniques of previous models. Section 3 contains the system model description. Section 4 includes the experimentation and results, and finally, Section 5 consists of the conclusion and some further future work.

2. Related Work

Understanding human sentiments have always been a captivated area of research, and it is gaining a lot of consideration from researchers and other disciplines. Communication over a long distance is very persistent in the current scenario, and the focus is on trying to communicate precisely what the person wants to say. But it’s hard to evaluate the actual emotions of a person while communicating over a long distance. Working on such an initiative will positively affect conveying the correct message [2] [4]. In the current years, knowledge of human behavior has obtained a lot of attention. Many techniques were used to understand human emotions and their polarity. Table 1 compares the various emotion detection methods using speech from some papers of the year 2005-2018.

Kagalkar et al. [2] worked on two types of datasets- training data and testing data. MFCC is used for feature extraction in both datasets. Using the extracted feature, the GMM and SVM classify the speaker's age into various age spaces. At that point, the feeling is predicted based on the trained data. Fung et al. [6] presented a model that uses a real-time CNN to detect emotions. This model can distinguish emotions into three classes- Happy, Sad, and Angry. The average accuracy gave by the model is-66.1%. The number of emotions detected is significantly less. Thus, it’s hard to foresee any feeling other than them. Chavhan et al. [9] proposed a model that classified emotion in 4 types using MEDC and MFCC for feature extraction technique with SVM classifier and applied this in three Kinds of input speech signal: gender independent, male and female. Yet, it just gives 100% accuracy for female speech.

Huang et al. [8] suggested another method for an emotion detection system. They utilized five layers of DBNs for feature extraction and classified emotion into four types with a non-direct SVM classifier's assistance. Yet, the downside of this new strategy was that the DBNs model's time cost was excessively more than other feature extraction techniques. Zheng et al. [11] utilized the PCA-DCNNs-SER approach for emotion detection and classification using the IEMOCAP database. This method was discovered in a way that is better than the SVM classification method. Because of the inappropriate appropriation of emotion-based data in the referenced database, the determined accuracy is less. Chen et al. [10] proposed an emotion recognition system using acoustic and linguistic features. Different feature representations are used for emotion detection in both acoustic and linguistic. Using these other representations, the accuracy of emotion classification is compared to a single database named USC-IEMOCAP.

Table 1
Comparison of classification methods

Classification Methods	Attributes
Support Vector Machine	Pros: • It doesn't have issues with local minima and over-preparing. • Ready to manage high dimensional input vectors Cons: • Doesn't work with variable length input. • The number of classes increments computational expense. • Not able to manage massive databases.
Hidden Markov Models	Pros: • Model time appropriation of a sound signal. • Development is easy. • Support vector length input. • Ready to display both discrete and nonstop signals. Cons: • It is accepting the likelihood of existing in an explicit state and autonomous of its past form.
Artificial Neural Network	Pros: • The capacity of self-association and self-learning. • Adjustability in various conditions. • Reasonable for design acknowledgment. Cons: • Requires broad training.
Naive Bayes Classifier	Pros: • Supervised classification • Highlights in a single class are thought to be autonomous of others. • The simplicity of straightforward execution and understanding. Cons: • Definite element autonomy suspicions. • Overfitting
CNN	Pros: • Automatically detect the import features of an image. • Speed is right on the short text.

- Ease to implement.
- Give the best results on an image-based data.

Cons:

- High computational cost.
- They used to need a lot of training data.

3. System Architecture

In this section, the proposed model, as shown in Fig. 1. will be discussed in detail. The whole system is divided into three phases:

3.1. Speaker discrimination

The objective of this phase is to determine whether the voice is of a male or a female. The input is first dividing into small chunks to make it easier and efficient because evaluating a whole audio file will take a bit more time than assessing the little chunks. Those chunks are compared with our training data to determine gender. This is important to evaluate first because there is a difference in the characteristics of voice of different genders for the same emotion, which can provide a more accurate result in the end. The output of this phase is a speaker id.

3.2. Speech recognition

Along with the first phase, this phase's process is also being executed. The audio file is going through a pre-processing part, which is filtration. The attempt to filter noise and remove background sound from the voice takes place to understand the voice and determine emotions more accurately clearly. After the filtration process, the MFCC graph of the file is being plotted. MFCC stands for Mel Frequency Cepstral coefficient. It is one of the standard strategies which is used for feature extraction of any speech signal.

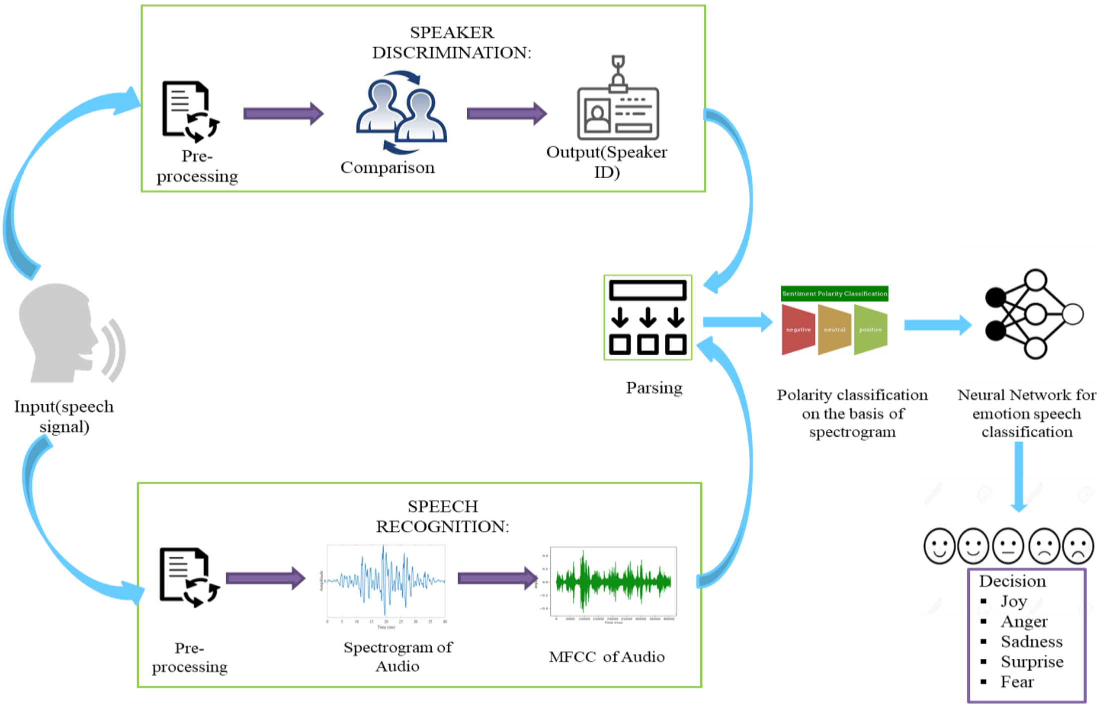


Figure 1: Implemented System Architecture

MFCC utilizes a non-linear recurrence unit to recreate the human sound-related framework. MFCC is the most used characteristic of voice in the emotion detection system. The next step is to plot the Spectrogram with the help of MFCC, in which ranges are defined for different polarity, and it is also the output of this phase. A spectrogram is a pictorial technique of representing the strength of a signal, or "loudness," of a signal over the long run at exclusive frequencies found in a particular waveform. Not exclusively might one be capable of seeing whether there may be quite plenty of strength at, for example, 2 Hz versus 10 Hz; however, you may likewise perceive how energy ranges fluctuate over the long haul. When implemented to a sound sign, spectrograms are called sonographs, voicegrams, or voiceprints. When the information is signified in 3D axes, it is probably called cascades [12].

When both phases execute, the output of the phases is combined. Depending on the Spectrogram, their polarity is determined whether the audio file is of either positive outlook or negative or neutral outlook.

3.3. Emotion classification

This final phase determines the audio's emotion or sentiment with the Convolutional Neural Network (CNN). The motive to use CNN is that it has high accuracy in image classification and recognition, as the output of the second phase is the image of a graph. In CNN, the polarity is further divided into different emotions based on their characteristics. For instance, the positive is also classified into happy, surprise. The negative polarity is further divided into anger and sad, and so on.

When implemented, this new proposed model proves to be more accurate compared to the previously proposed models.

4. Experiments and Results

The dataset used in this pro is The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This dataset includes:

1. Recorded speech and song versions of 12 Actors & 12 Actresses, respectively.
2. Song version data of Actor 18 is not present in the dataset.
3. Disgust Surprised and Neutral, emotions are not present in the song version of the dataset.

In this dataset, eight types of emotions say and sing two sentences and two instances perform by every 12 actors. Therefore, every actor would incite four examples for every emotion other than neutral, disgust, and surprise in view that there may be no making a song statistic for these emotions. Each sound wave is around 4 seconds, the first and last second are likely to quiet.

The input is an audio file that is passing through the early two phases simultaneously. So, in Speaker Discrimination Phase, speech is going through a pre-processing part in which it is further divided into small chunks. Then the comparison algorithm executes, and it determines whether it is the female voice or the male voice and then providing a speaker id, which is the output of this phase. Now, the question arises about how the comparison algorithm works? How is it differentiating that the audio file is in female voice or male voice? So, the answer is, in the dataset, every file is given a number, and it is arranged in such a way that every alternate file is female, so if the file number is even, then it is a female voice, and an odd file is of a male voice.

Simultaneously, in the second phase, i.e., the speech recognition phase is also being executed. In this phase, the audio file is again going through a pre-processing part, in which an attempt to remove noise from the audio file is being made. After this, a graph is plotted with the help of MFCC. MFCC is a state-of-art feature in Speech Recognition tasks because it turned into invented in the 1980s. The MFCC form, as shown in Fig. 2. governs what sound comes out. If anyone can find the shape correctly, this should deliver a genuine depiction of the produced phenomenon. With the MFCC graph's help, the next step is to plot the Spectrogram of audio, which is also the output of this phase, in which ranges are defined for different polarities. Loading audio data and altering it to MFCCs format can be speedily performed using the Python package deal Librosa [5].

After the execution of the first two phases, the output of both phases is parsed. Depending on the Spectrogram, their polarity is determined whether the particular audio file is of either a positive outlook or negative or neutral stance.

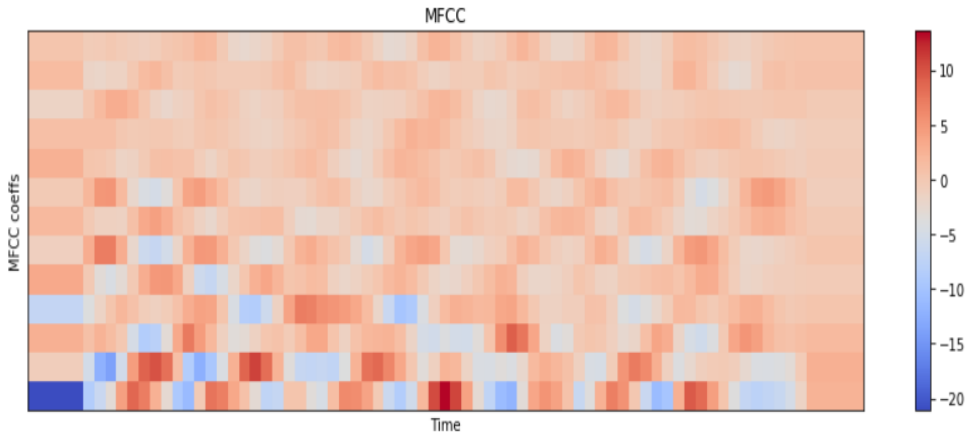
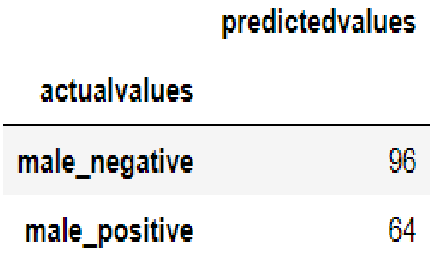


Figure 2: MFCC graph of a RAVDESS audio file

The final phase is the Sentiment Analysis phase. For this phase, Convolutional Neural Network (CNN) is used to further classify the polarity in different emotions. For instance, the positive is also classified into happy, surprise. The negative polarity is further divided into anger and sad, and so on [11].

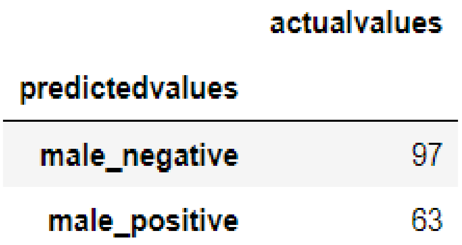
The CNN model is advanced with the assist of Keras and created with seven layers- 6 C0nv1D layers observed via a dense layer. The proposed model only trained with 700 epochs lacking any learning rate schedule etc. The accuracy of this model is comprised of its loss function and the evaluation metric.

The followed procedure maintains a model accuracy of 93.96%. The overall prediction accuracy is 75%. Fig. 5. shows the model loss graph, which perfectly illustrates the difference of prediction in the training and the testing data used. The difference is less as compared to the difference in other basic ML classifiers. Fig. 3 and Fig. 4 show the difference between the real and predicted values.



	predictedvalues
actualvalues	
male_negative	96
male_positive	64

Figure 3: Predicted Values



	actualvalues
predictedvalues	
male_negative	97
male_positive	63

Figure 4: Actual Values

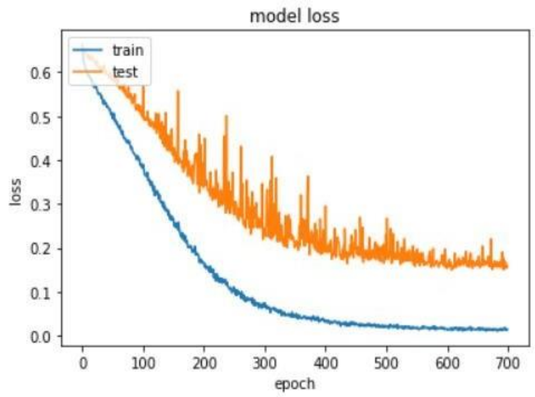


Figure 5: Model loss graph of testing and training data

The confusion matrix shows the comparison between the expected result and the predicted results of a model. The confusion matrix is shown in Fig. 6. shows that 107 times out of 160 times, the model predicts the correct outcome, whereas 53 times, it offers a different result in a particular audio file. In Fig. 7. the accuracy graph is shown for the same. The comparison of negative and positive emotions of a male is shown in the given confusion matrix. The CNN model needs a large data file though it requires more memory; the overall accuracy is better than all other techniques.

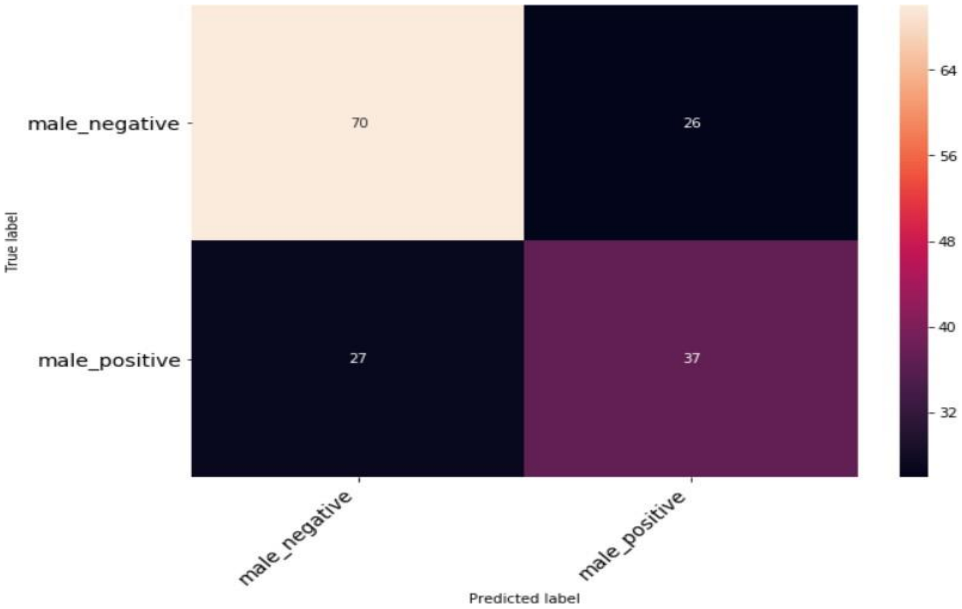


Figure 6: Confusion matrix of a male audio data

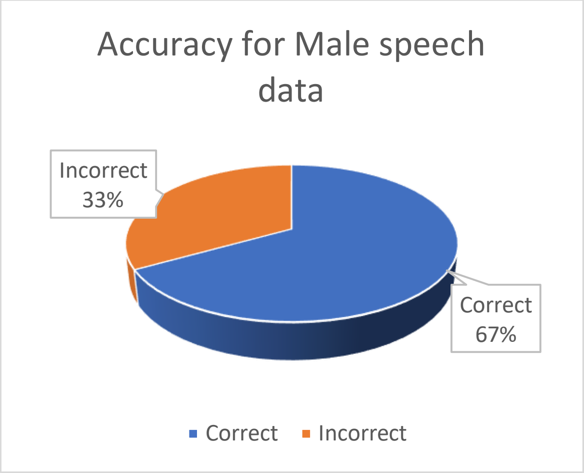


Figure 7: Accuracy for Male speech data

Prediction equation using Confusion Matrix= (TP+TN)/ (TP+TN+FP+FN) where,

1. True Positive (TP): Observation is positive and is predicted to be positive.
2. False Negative (FN): Observation is positive but is predicted negative.
3. True Negative (TN): Observation is negative and is predicted to be negative.
4. False Positive (FP): Observation is negative but is predicted positive.

5. Conclusion and Future Scope

In emotion classification, database choice played an essential role in good exactness in the result. Feature extraction from the audio signal is the second most crucial step in this field [3]. There are many methods for feature extraction, which is briefly covered in this paper: amongst them, the MFCC algorithm is widely used because it performs better than the other feature extraction techniques in the case of noise. The next important part of sentiment analysis is the use of Classifier. CNN is the most often used for solving Sentiment analysis and gives the best result in image-based data. As the confusion matrix shows, some emotions are easy to identify. Some are confused with other emotions and challenging to identify the model in which the speech belongs.

Many other issues need to be solved, like diversity in emotion, recognizing spontaneous emotion, and speaker recognition in simultaneous conversation. The future scope of this work is to investigate different strategies for ongoing issues in sentiment analysis. Also, extracting more useful features of a speech signal will enhance the model's accuracy and work better on a real-time system.

6. References

[1] Zhang, Shiqing, Shiliang Zhang, Tiejun Huang, and Wen Gao. "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching." IEEE Transactions on Multimedia 20 (2017)

[2] Chaudhari, Shivaji J., and Ramesh M. Kagalkar. "Automatic speaker age estimation and gender-dependent emotion recognition." International Journal of Computer Applications 117(2015).

[3] Souraya Ezzat, Neamat El Ghayar, and Moustafa M. Ghanem. “Investigating analysis of speech content through Text Classification.” International Conference of Soft Computing and Pattern Recognition(2010).

[4] Esraa Ali Hassan, Neamat El Gayar, and Moustafa M. Ghanem. “Emotions analysis of speech for call classification.” 10th International Conference on Intelligent Systems Design and Applications(2010).

[5] Abdul Malik Babshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. “Speech emotion recognition from spectrograms with the deep convolutional neural network." IEEE(2017).

[6] Bertero, Dario, and Pascale Fung. "A first look into a convolutional neural network for speech emotion detection." In 2017 IEEE international conference on acoustics, speech, and signal processing (ICASSP), IEEE (2017).

[7] Basu, Saikat, Jaybrata Chakraborty, Arnab Bag, and Md Aftabuddin. "A review on emotion recognition using speech." In 2017 International Conference on Inventive Communicationand Computational Technologies (ICICCT), IEEE (2017).

[8] Huang, Chenchen, Wei Gong, Wenlong Fu, and Dongyu Feng. "A research of speech emotion recognition based on deep belief network and SVM." Mathematical Problems in Engineering 2014 (2014)

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

[10] Jin, Qin, Chengxin Li, Shizhe Chen, and Huimin Wu. "Speech emotion recognition with acoustic and lexical features." In 2015 IEEE international conference on acoustics, speech, and signal processing (ICASSP), IEEE (2015).

[11] Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." In 2015 international conference on affective computing and intelligent interaction (ACII), IEEE (2015).

[12] Tripathi, Anjali, Upasana Singh, Garima Bansal, Rishabh Gupta, and Ashutosh Kumar Singh. "A Review on Emotion Detection and Classification using Speech." Available at SSRN 3601803 (2020).