2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377192

Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks

Shambhavi Sharma
Amity School of Engineering and
Technology, CSE
Amity University Uttar Pradesh
Noida, India
shambhavisharma110800@gmail.com

Abstract—This paper presents a comparative study on two classifiers created for speech emotion recognition. Perceiving a person's feeling has consistently been an intriguing task for everyone. These feelings can be expressed through facial expressions, speech, actions, and so forth. The most widely used form of communication is through speech. Speech is an elaborated form of communication constituting various details. These details provide several information such as the abstract of the message, tone of the speaker, language used, background noise, any form of musical sound, emotions, etc. The significance of speech emotion recognition technology is getting mainstream with the advancement of "Voice User Interface" technology. This technology makes it possible for computers to interact with humans by applying speech analysis to understand the instructions given by a person and perform the required tasks and commands. There is always an emotion attached to a piece of speech while communicating but recognizing this emotion is a complex job in the research field. This is mainly because the way emotions are perceived from an audio differs from person to person. I have created two models for speech emotion recognition. I have used Mel Frequency Cepstral Coefficient (MFCC) for feature extraction from the audio files. The first model has been created using Multi-Layer Perceptron (MLP) classifier which gave an accuracy 57.29 percent. The second model was created Long Short-Term Memory (LSTM) and gave a good accuracy of 92.88. I have made use of RAVDESS dataset for classification purpose.

Keywords—Speech emotion recognition, Mel Frequency Cepstral Coefficient, Long Short-Term Memory, Multi- Layer Perceptron, Recurrent neural network, Artificial neural networks, Deep learning

I. Introduction

Feelings are a crucial part of our day to day lives. It has a fundamental role in all the decisions that we take. It encourages us to coordinate and comprehend the sentiments of others by passing on our emotions to other people. In today's time we can easily communicate with our computers, laptops, and smart phones through voice operated technologies such as Siri, Google assistant, Alexa, Cortana, etc. But the question arises that do these technologies actually know what we are feeling. There is a remarkable advancement in recognizing the speech, but still, we are a little away from building the emotional understanding between a human and a computer.

Paralinguistic features such as voice tone, voice modulation and voice pitch are some essential part of our speech. These features highlight the parts of our speech. They are not marked by the words or sentences verbally spoken. On the other hand, they add weight and emphasis to the words and sentences that are verbally expressed by a person. Paralinguistic features have the ability to completely change the meaning of a sentence. Prosodic features such as rhythm, intonation and stress also highlight the features of our speech. it plays an important role in assessing the emotion being delivered in a speech. Therefore, it is important to train our machines to understand these features and interpret the emotion attached to a speech.

There have been numerous hypothesis and plenty of research work done in this field. Emotions are considered to be momentary states, on the other hand attitude of a person, behavior, demeanors or characters are something that is considered to hold a long-term value. Emotional states are corresponded with specific physiological states which affects the attributes of our speech like pitch, tone, etc. For example, the blood flow and pressure are very high in a situation where person is feeling joy, dread, anger, anxiety or bliss,. In these situations, it is also seen that the mouth gets dry and sometimes jaws are also clenched. Trembling of muscles in the gut is also seen. A high energy voice is seen in such cases. The person speaks is an uproarious, quick and disorderly manner. When a person is exhausted, tired out, unhappy or in dismal, a whole lot of different symptoms are seen. The nervous activity of the parasympathetic sensory structure increased. Lower pressures in blood are seen which is exactly different from the situation in which a person is in joyous mood. It is also seen that there is escalation in the salivation rates. As a result, the speech of the person is lack of vitality and enthusiasm. These physiological impacts are ubiquitous. There are propensities in the relationship between some acoustical characteristic and essential feelings across various societies [1]. Speech emotion recognition has wide range of applications such as human-computer interaction, Telecommunication, Business analytics, Security, Medicine, Psychological therapies, etc. There are three major principals that are considered in making of an artificial framework that is fit for perceiving feelings from a speech and recognizing it correctly. These are signal processing, extraction of features, emotion recognition and categorization [2]. Bringing out the essential attributes is the foremost and the most important step to get the accurate output. It is important to segregate the important and useless attributes from the speech. Important attributes are pitch, tone, etc. Useless features are background noise, disturbances, etc.

978-1-6654-1451-7/21/$31.00 ©2021 IEEE

Human mouth organs like vocal tract, teeth, gums, tongue play an important role in changing the sounds of our speech. Different forms of sound are created because of different shapes of these organs. To get an accurate data of a speech it is important to study the impact of the varying shapes of these human organs. Various algorithms have been used for feature extraction from speech signals such as linear prediction cepstrum coefficients (LPCC), Artificial neural networks (ANN), Mel Frequency cepstrum coefficients (MFCC), amalgamation of Linear Prediction coefficients and Mel cepstrum coefficients (LPCMCC), the Support Vector Machine (SVM) [3][4]. My proposed work is based on feature extraction using MFCC and decision making using standard deviation. Several different models are presented for speech emotion analysis [11] [12]. All these models proposed differ in various aspects such as the datasets being used, nature of features used, and type of classifiers and neural networks built for emotion analysis. Some of the most commonly used datasets are Berlin Database [5], CREMA-D [6], RAVDESS, Mandarin Affective Speech Database etc. In this paper I have made use of Ravdess database [7].

It is believed that this field has a large scope of research. Current research work mostly concentrates on emotion detection for either a small piece of speech or a few words, where as a speech generally comprises of a blend of emotions. There can be various degrees or extents to which an emotion can be reflected in a speech. In the past years different methodologies have been used to create different models such as, K-nearest Neighbors (KNN). This methodology was concluded to have a satisfactory generalization capability [8]. Gaussian mixture models were made that improved the biased intensity of the wavelet parcel entropy highlights. This idea was proposed for recognition of multiclass emotions [9]. Naïve Bayes classification method was used that could deal with self-assertive number of autonomous factors whether persistent or unqualified [10]. Some other methodologies adopted are Neural Networks, Kernel Regression, Hidden Markov Model and support vector machines.

II. METHODOLOGY

A. Speech Corpus

I have used Speech sound only records (16bit, 48kHz .wav) from the "RAVDESS" dataset. This bit of the RAVDESS contains 1440 records: 60 preliminaries for every entertainer x 24 on-screen characters = 1440. The RAVDESS contains 24 expert on-screen characters "12 female" and "12 male", vocalizing two lexically-coordinated explanations in a nonpartisan "North American" inflection. The emotions were labelled as follows: 01-'neutral', 02-'calm', 03-'happy', 04-'sad', 05-'angry', 06-'fearful', 07-'disgust', 08 -'surprised'. Every articulation is created at two degrees of enthusiastic force (ordinary, solid), with an extra nonpartisan articulation.

B. Extraction of features using MFCC

Bringing out the essential attributes is the foremost and the most important step to get the accurate output. It is important to segregate the important and useless attributes from the speech. Important attributes are pitch, tone, etc. Useless features are background noise, disturbances, etc.
Human mouth organs like vocal tract, teeth, gums, tongue play an important role in changing the sounds of our speech. Different forms of sound are created because of different shapes of these organs. To get an accurate data of a speech it is important to study the impact of the varying shapes of these human organs. We can interpret the shape of the vocal tract by marking the frequencies and various other attributes. This is slightly challenging because this task is to be carried out in a small timespan. The task of MFCCs is to precisely and correctly give this representation.

Steps involved in working of MFCC: [4][11]

- Energy increment: This is the first step. In this process the energy of the input voice signal is increased by passing it through a filter bank.
- Framework: The signals of a sound alter at a very rapid rate when a person is speaking. It is presumed that on a brief time span there is no change in the signals. For this reason, we break the signals into shorter time intervals like 20-40 milliseconds. An extremely shorter time interval is undesirable because then we will require more examples to get a good output. If the time interval is extended that is also undesirable because then the signal will vary a lot in the time frame.
- Overlapping: After building up of frameworks, the next task is to overlap the consecutive frameworks. This is done in order to make the signal continuous in nature.
- Computation: Computation or calculation of power range is the next task. The idea for this mechanism is taken from the working of an ear part that is, cochlea. As a person speaks this organ starts vibrating at different spots. The frequencies approaching sound waves decide these spots. The periodogram gauge it carries out a similar job that is extremely important. It makes out the amount of vitality present around in various frequencies approaching towards it. It is helpful in studying the various frequency regions. This step often makes use of Fast Fourier transformations.
- Mel scaling: The "mel scale" is used to get an idea about the building of the gauze banks. These gauzes are of different sizes. The initial one is very narrow that gives the idea about the vitality near to 0 Hertz. As the frequencies are escalated the gauzes get broader to give accurate information about energy at higher frequencies.
- Applying Logarithmic function: The next step is to take out their logarithm. In order to understand the audio correctly we need to double the recognized volume of audio. This implies huge variations in energy may not be all that varying if the sound is loud enough.
- Discrete Cosine Transformation (DCT): The last step is to process the DCT of the log filter bank energies. The range of its coefficients is between 3-14. There are various other coefficients that are not required. They are disposed off.

C. Classifier

a) Artificial Neural Network – Multilayer Perceptron (MLP)

Artificial Neural Network is equipped for learning any nonlinear function. Henceforth, these organizations are prominently known as Universal Function Approximators. ANNs have the ability to learn loads that map any contribution to the yield. One of the principal purposes for general guess is the enactment work. Enactment capacities acquaint nonlinear properties with the organization. This enables the organization to gain proficiency with any unpredictable connection among input and output [13]. The Expression MLP is applied equivocally allude to a feedforward artificial neural network.

These neural networks have the ability to learn through training and observation. Through training, the network builds up the relationship between the set of inputs and outputs. To minimize the error the parameters like weights and biases are attuned. At times they mark reference to networks that are made out of numerous layers of perceptron. Multilayer perceptron is informally alluded to as "vanilla" neural network systems, particularly when they have a solitary hidden layer. An MLP comprises a minimum of three layers junctions. These layers are Input layer, Hidden layer and Output layer. The output from every neuron is connected to the input of neurons of the subsequent layer. They are generally used for classification problems.

All the nodes of an MLP uses certain type of activation functions. The very first node uses a linear function whereas rest of the nodes use a nonlinear function. While educating the multilayer perceptron a superintended plan is used that is called “backpropagation”. The quality that makes it different from other neural networks is the type of activations function it uses. It has the ability to distinguish different types of data. For complex tasks a network with multiple layers is used. A straight line is used to categorize the inputs in this algorithm. The input is a constituent vector which is further multiplied by a load ‘w’ and added to a constant (called bias) ‘b’. Mathematical equation for the same is given below.

$$\boxed{y = \psi(\textstyle\sum_{i=1}^{n} w_i x_i + b) = \psi(w^T x + b)} \qquad (1)$$

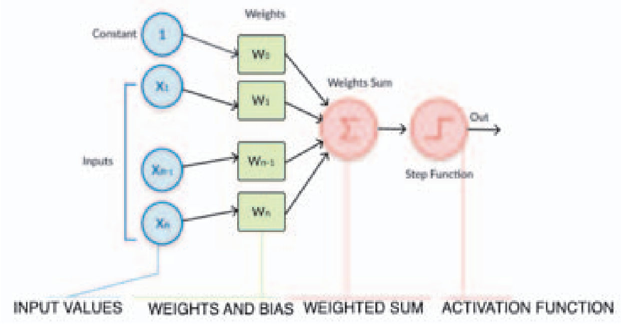

Fig. 1. Structure of a perceptron

b) Recurrent Neural Network – Long Short Term Memory (LSTM)

Recurrent Neural Network (RNN) are used to model tasks where the output is contingent on multiple inputs and dependent on these inputs. Unlike traditional neural networks, RNN operates on a sequence of data inputs and stores the information of the previous inputs. For example, while you are reading a text you don’t process individual words, rather you try to remember the previous words and sentences to update the incoming information. In this way you gather a better understanding of the text. RNN works in a similar way. In a speech emotion recognition model, a sequence of voice signals is input in the model and a function is executed at each time step, thus it maintains a state all through.

There is an inner loop in a recurrent neural network, the data points are processed in this loop. These loops provide the network activations from a past time venture as inputs to the network organization to impact forecasts at the current time step. Keras holds up three RNN layers: Simple RNN, LSTM and GRU. One of the major obstacles faced by RNN is gradient vanishing problem. RNN is unsuccessful to learn within the sight of delays more than five to ten discrete time ventures between applicable information functions and target signals. The learnings from the previous advances becomes unimportant in the gradient plunge step. Long Short-Term Memory (LSTM) is able to overcome this problem. LSTM can figure out how to connect negligible delays more than 1000 discrete time gaps by upholding steady error course through consistent blunder carrousels inside unique units, called cells. LSTM was the first architecture that made use of a separate memory cell to train the network. For an input vector of length “T” the equations corresponding to memory cell are described below. “i”, “f”, “o” corresponds to input gate, forget gate, and output gate respectively. “a” represents cell input activation function, “h” is the hidden layer and “c” is state vector [14].

$$\boxed{i^t = \sigma(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i)} \qquad (2)$$

$$\boxed{f^t = \sigma(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f)} \qquad (3)$$

$$\boxed{o^t = \sigma(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^{t-1} + b_o)} \qquad (4)$$

$$\boxed{a^t = tanh(W_{xc}x^t + W_{hc}h^{t-1} + b_c)} \qquad (5)$$

$$\boxed{h^t = o^t \odot \tanh{(c^t)}} \qquad (6)$$

$$\boxed{c^t = f^t \odot c^{t-1} + i^t \odot a^t} \qquad (7)$$

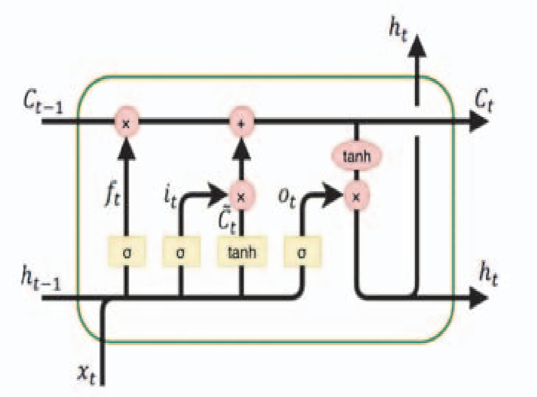

Fig. 2. Architecture of LSTM Network

D. Proposed Architecture

Software Requirements: Anaconda 64-bit, Python 3.6.6

Python Libraries Used: TensorFlow GPU v2.1.0, Keras v.2.3.1, Numpy v1.18.1, MatPlotLib v2.2.3, Pandas v1.0.3, Os, Librosa v0.7.2, Wave, Scikit-learn v0.23.1.

In my proposed model the above-mentioned python libraries were used. The “Ryerson Audio Visual Database of Emotional Speech and Song” dataset was downloaded that contained about 1440 audio records. The MFCC features were extracted from the audio files to obtain the mean of each dimension that depicts the short-term power range of the audio files. This was saved in a NumPy array which was further used as inputs to the model. There were in total eight speech labels: neutral, calm, happy, sad, angry, fearful, disgust and surprised. A function was defined to split the dataset into training and testing set. The samples in the training set structures the knowledge that the algorithm uses to learn. The samples in the test set assess the performance of the model on the basis of the learnings it has gained. 80% of the data was kept as training set and 20% was kept as testing set. The random state which ensures the similar arrangement of arbitrary numbers is created each time you run the code was kept as 9.

The first model was made using the MLP Classifier, it uses log-loss function using LBFGS or stochastic slope drop. The other parameters of the MLP classifier were kept as follows: alpha, also called as L2 parameter regularization was set to 0.01. Batch size was set as 256. Epsilon defines the numerical stability in adam optimizer and it was set to 1e-08. Hidden layer size was set to 300. Learning rate was kept as ‘adaptive’ and max iteration was set to 500. ‘Adaptive’ retains the learning rate consistent insofar as training loss keeps decreasing. Each time two continuous epochs are unsuccessful in diminishing training loss, or unsuccessful in expanding validation score, the current learning rate is divided by 5.

Then the second model was made using the LSTM classifier. Sequential classification was done because here the predictive model is fed with inputs that is distributed over equally spaced time intervals and my aim was to forecast the classification group it belongs to. Dense layers were added in the model, dropout layers were also added to reduce overfitting of my prediction model. Keras library supports regularizations of dropout layers. Dropout works by eliminating inputs to a layer, which might be input factors in the information test or enactments from a past layer. It has the impact of recreating countless organizations with totally different organization structure and thus, making hubs in the organization for the most part heartier to the data sources. While adding drop out layers “Relu” activation function was used. An activation function is in charge of modifying the added weight contribution from the hub into the actuation of the hub or yield for that input. “Relu” activation function is a piecewise linear function that will yield the information straightforwardly in the event that it is positive, else, it will yield zero. It looks and acts like a linear function, yet is a nonlinear function permitting complex connections in the information to be acquired by the model. This function utilizes stochastic angle plunge with backpropagation of mistakes to prepare profound neural organizations. After the last dense layer “softmax” activation function was used. After that configuration of the model was done for training. All the prepared layers were fed into the LSTM classifier to make the prediction.

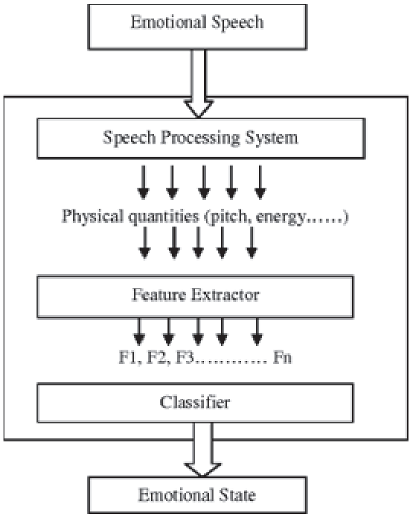

Fig. 3. Mechanism for speech emotion recognition

III. Results and discussions

I was successful in creating two Speech Emotion Recognition models. The first model was created using MLP classifier and gave an accuracy of 57.29 percent. The second model was created LSTM and gave a good accuracy of 92.88 percent. So, it can be concluded that the LSTM model gave a higher accuracy.

MLP can inexact any capacity, to a self-assertive accuracy, along these lines there is no requirement for RNN. Anyway, that doesn't mean it is usable. Accepting we are discussing time arrangement input, course book answer would be that you can take care of your time arrangement in a feed forward system, by having an information layer contain additionally contributions from past time focuses. Subsequently viably changing time arrangement issue, into feed forward issue. Anyway, you should pick length of your information in advance, and you won't have the option to learn capacities that relies upon the data sources quite a while prior. You can tackle this issue by having a RNN, that can hypothetically, store data from quite a while prior, in its setting layer. However, you will have a gradient exploding/vanishing problem. So, can be utilized further for feelings acknowledgment, yet there is one significant thing that we need to remember is, for the better execution and appropriate rumbustiousness of our framework, we should prepare our framework emotionally and appropriately with the assistance of different legitimate datasets which are principally worked for testing and preparing of a feeling acknowledgment framework. So as to utilize and process our framework in better manners, effectively and with more comfort there is a fundamental requirement for appropriate datasets and expressions for the correct training of our framework.

So, in order to use and process our system in more efficient ways, easily and with more convenience there is an essential requirement for suitable datasets and phrases for the proper and efficient training of our system. Basically, for the training of our system we mainly focus on the use of different datasets which are listed below:

A. function sentiments

B. impulsive emotion

C. extract response

D. standard response

In the created model the focus has been kept on the use of function sentiments database. This is because of their packages stacked with strong sentimental expressions which can be easily rendered and can be easily used as a means of input of speech signals. For arrangement of various emotion in my framework I have utilized the RAVDESS dataset. It was utilized in light of the discourse bundles that were prepared to use for framework at whatever point are any place I wanted them to utilize.

The bundles hold discourse sign and vocals of around 24 on-screen characters with 7356 records which are evaluated by around 247 coders from around the world to use as a contribution for a feeling acknowledgment framework. The addresses were partitioned into fundamental feeling type like neutral, calm, happy, sad, angry, fearful, disgust and surprised. This paper proposes a new feature enhancement method for improving the multiclass emotion recognition based on LSTM model. The main contribution of this paper was the method of extracting features for a signal within a short period of time.

Graphs were plotted as follows. Accuracy = Number of correct predictions divided by (/) Total number of predictions.

A. Training and validation accuracy

Training accuracy is the exactness of a model on data samples it was built upon. Validation accuracy is the exactness of the model on the data samples it hasn't been built upon and is new to the model. From my training and validation graphs I can see that the training curve is closely tracked with the validation curve. Hence, I conclude that I got a model with good accuracy of 92.88 percent.

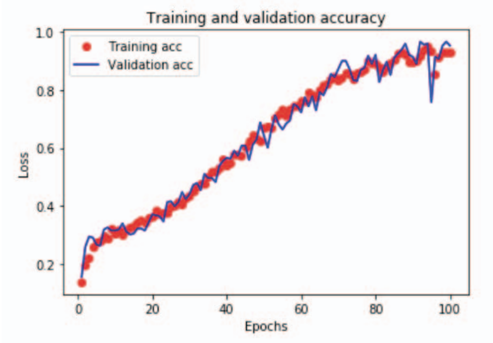

Fig. 4. Training and validation accuracy curves for the proposed LSTM model

B. Training and validation loss

Training loss is the inaccuracy on the training data samples. Validation loss represents the inaccuracy after the validation dataset has been run through the trained model layers. If validation loss is much higher than training loss, it is called over fitting. If validation loss is much less than training loss, it is called under fitting. To have a good accuracy of the model our aim should be to have an extremely low validation loss.

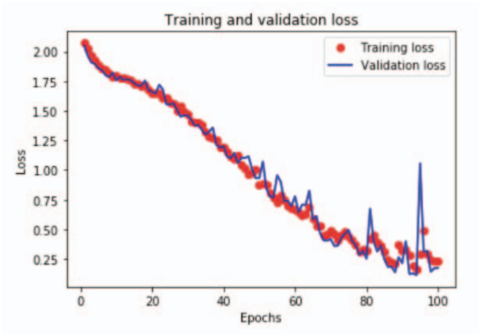

Fig. 5. Training and validation loss curves for the proposed LSTM model

IV. CONCLUSION

SER frameworks can be improved in strategies for their preparation time speculations or exactness with the assistance of blend of databases and classifiers, however it will be hard to for a person to deal with the intricacy which be brought about the perplexing utilization of databases, AI calculations and classifiers. A superior determination of removed highlights can be utilized so as to improve the feelings acknowledgment precision. Different development techniques can be utilized to extricate the quality highlights. Subsequent studies include adding more experimental data in different environments and improving the practicability of front-end algorithms. In the future work, more low-level and high-level speech features will be derived and tested by the proposed methods. Other filter, wrapper, and embedded based feature selection algorithms will be explored and the results will be compared. The proposed methods will be tested under noisy environment and also in multimodal emotion recognition experiments.

REFERENCES

[1] Wu, Dongrui & Parsons, Thomas & Narayanan, Shrikanth. (2010). Acoustic feature analysis in speech emotion primitives estimation. Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. 785-788.

[2] Zhao, Li, and C-H. Jiang. "A study on emotional feature analysis and recognition in speech." Acta Electronica Sinica 32.4 (2004): 606-609.

[3] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases.

[4] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 2257-2260, doi: 10.1109/WiSPNET.2017.8300161

[5] Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. (2005). A database of German emotional speech. 9th European Conference on Speech Communication and Technology. 5. 1517-1520.

[6] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," in IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 377-390, 1 Oct.-Dec. 2014, doi: 10.1109/TAFFC.2014.2336244.

[7] Livingstone, Steven & Russo, Frank. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE. 13. e0196391. 10.1371/journal.pone.0196391.

[8] Bombatkar, A., Bhoyar, G., Morjani, K., Gautam, S., & Gupta, V.S. (2014). Emotion recognition using Speech Processing Using k-nearest neighbor algorithm.

[9] Bombatkar, A., Bhoyar, G., Morjani, K., Gautam, S., & Gupta, V.S. (2014). Emotion recognition using Speech Processing Using k-nearest neighbor algorithm.

[10] M. A. Tischler, C. Peter, M. Wimmer and J. Voskamp, “Application of emotion recognition methods in,” 2007.

[11] S. Basu, J. Chakraborty and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 333-336, doi: 10.1109/CESYS.2017.8321292.

[12] Zheng W, Xin M, Wang X, Wang B (2014) A novel speech emotion recognition method via incomplete sparse least square regression. IEEE Signal Process Lett 21(5):569–572

[13] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.

[14] An, S., Ling, Z., & Dai, L. (2017). Emotional statistical parametric speech synthesis using LSTM-RNNs. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). doi:10.1109/apsipa.2017.8282282