JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SPEECH-BASED EMOTION RECOGNITION

Valandas Sai Shashank[1], Nuthanakanti Bhaskar [2], Dr.K. Srujan Raju[3], A.Raji Reddy4

[1]Research Scholar, Department of Computer Science and Engineering, CMR Technical Campus, Hyderabad, India.
[2]Associate Professor, Department of Computer Science and Engineering, CMR Technical Campus, Hyderabad, India.
[3]Professor, Department of Computer Science and Engineering, CMR Technical Campus, Hyderabad, India.
[4]Professor, Department of Mechanical Engineering, CMR Technical Campus, Hyderabad, India.

***Abstract*— Speech is the primary form of communication for humans, followed by language. Social contact depends heavily on emotion. As we are dealing with human-machine interaction, identifying the emotion in a speech is both crucial and difficult. People's emotions differ from one another, and the same person may experience several different emotions at once and show them in various ways. When a person expresses his or her emotions, each will have a different energy level, pitch fluctuation, and tone variation are grouped together depending on the issue. As a result, computer vision's long-term objective is to recognise speech emotions. Our project's goal is to build a convolutional neural network-based system for intelligent emotion recognition in voice. which employs many modules for the classifier's recognition of emotions.**

1. INTRODUCTION

Speech emotion recognition (SER), a component of naturalistic human-computer interaction (HCI), is becoming more crucial in a variety of contexts. Speech emotion identification is now a developing discipline that straddles artificial intelligence and artificial psychology, in addition to being a well-liked research area in signal processing and pattern recognition. The research is extensively used in sectors such as interactive teaching, entertainment, security, and human-computer interaction. The three components of a speech emotion processing and recognition system are typically speech signal acquisition, feature extraction, and emotion recognition. The neural network-based approach to speech recognition is the most advantageous method. Artificial neural networks (ANN) are information processing techniques with biological inspiration. Artificial neural networks (ANN) are a fast-growing alternative to HMM for speech recognition modelling since they don't require any prior understanding of the speech process. RNN can model time-dependent phonemes and learn the sublunary link between speech and data. The use of multi-layer perceptron (MLP) type conventional neural networks for speech recognition and other speech processing tasks has been growing. An acoustic signal that was recorded by a microphone or a phone is converted into a collection of characters during the speech recognition process. These can also be used as the input for additional linguistic processing, which is addressed in section, to produce speech understanding. It is well known, voice recognition software performs functions comparable to those of the human brain.

The objective of this research is to categorise these emotions from a given speech sample in the best way possible. We intend to employ the multilayer perceptron with several techniques for emotion prediction. Support Vector Machine (SVM) and Multi Layer Perceptron Classifier are the two classifiers that we compare (MLP Classifier).

Goals & Issue Identification :

Speech The process of attempting to identify human emotion and the related affective states from speech is known as emotion recognition, or SER. This makes use of the fact that tone and pitch in the voice frequently convey underlying emotion. Recent years have seen a fast expansion in the scientific field of emotion recognition. Machines can neither sense nor display emotion, unlike humans. Yet, automating emotion recognition can enhance human-computer connection and minimise the need for human intervention.

2. Literature survey

[1] Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. The research provides a potential method for audio signal emotion identification using deep learning methods. The authors present examples of feature extraction using CNNs and classification using Decision Trees. The study emphasises how these systems could be applied in practical ways to enhance human-machine interactions and create intelligent systems for mental health screening.

[2] Jianfeng Zhao, Xia Mao, Lijiang Chen. Learning Deep features to Recognise Speech Emotion using Merged Deep CNN. In the paper, a novel method for deep CNN-based speech emotion recognition is presented. The authors use two benchmark datasets to show how their merged deep CNN architecture captures discriminative characteristics from voice signals and achieves cutting-edge performance. The study emphasises how deep learning methods could enhance the reliability and accuracy of voice emotion identification systems.

[3] A key contribution to the field of speech recognition is the 1980 paper "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences" by Davis and Mermelstein. The authors of the research evaluate various parametric models for identifying monosyllabic words in continuous speech. They contrast three distinct representations: auditory filter banks (AFB), perceptual linear predictive (PLP), and linear predictive coding (LPC). A linear filter is used in the common speech processing technique known as LPC to simulate the vocal tract. AFB models the frequency response of the auditory system using a filter bank, whereas PLP modifies LPC to integrate perceptual information.

[4] Emotion recognition in speech is an interesting and applicable research topic and present a system for emotion recognition using one-class-in-one neural networks. By using a large database of phoneme balanced words, our system is speaker and context independent. We achieve a recognition rate of approximately 50% when testing eight emotions

3. Methodology

use the provided input to predict emotions using the MLP classifier. Using the five retrieved features, we obtain results. We provide the model the varying five features. While a single feature parameter is insufficient for creating an effective forecast, using the characteristics independently and passing it all together results in a significant variance of the prediction emotion. In order to train the model, the MLP Classifier is given the Ravdess dataset, which is divided into the training and testing datasets in a ratio of 75:25. These eight emotional categories are covered. As it is effective for time series-based data, such as the audio from which we will be predicting the emotion, the classifier is being used. In order to forecast the mood, we then test the remaining 25% of the data set.

Block Diagram

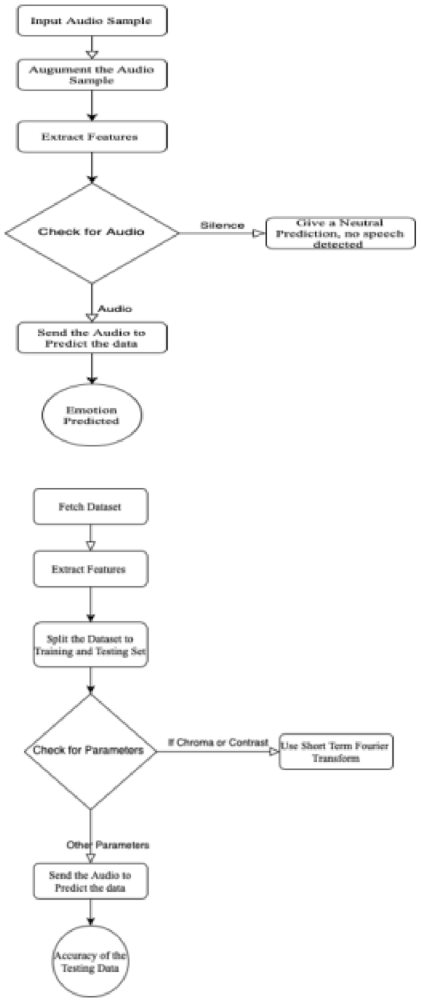

Input Audio Sample
Augument the Audio Sample
Extract Features
Check for Audio
Silence
Give a Neutral Prediction, no speech detected
Audio
Send the Audio to Predict the data
Emotion Predicted
Fetch Dataset
Extract Features
Split the Dataset to Training and Testing Set
Check for Parameters
If Chroma or Contrast
Use Short Term Fourier Transform
Other Parameters
Send the Audio to Predict the data
Accuracy of the Testing Data

Fig 1 : Training process

In order to gain a solid understanding of the audio, the audio is captured for 15 seconds with a 0.5 second break at the beginning and finish of the audio file. The volume of the audio will vary, which will make it difficult to extract the characteristics. To overcome this, we normalise to average the volume across the whole sample.

The range of the audio in terms of DB will be from -758 to 770 DB because the length of the audio will be represented as a 32-bit number in float format with a ten-to-the-power expression, which has higher relevance because it can represent larger and smaller values. A list of hyper-parameters are provided to the MLP-Classifier. The logistic activation function is a differential function that aids in determining the slope of a curve at any two locations.

5. Implementation

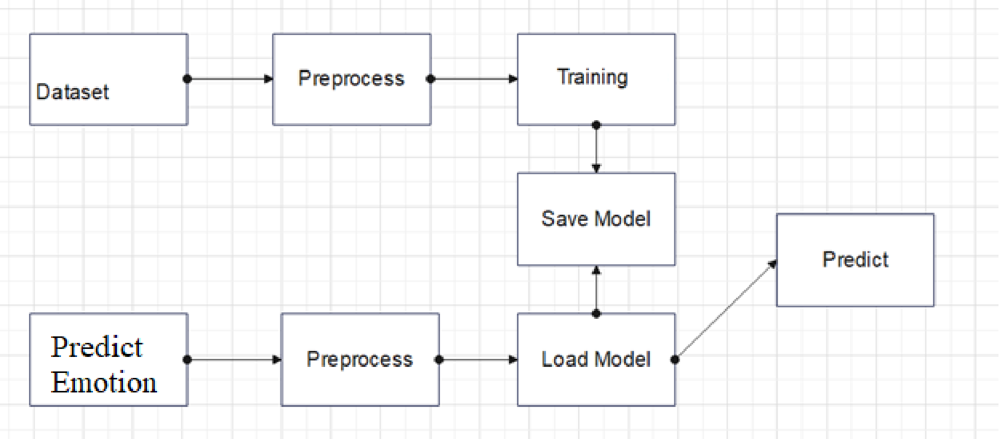


Fig 2 : System Architecture

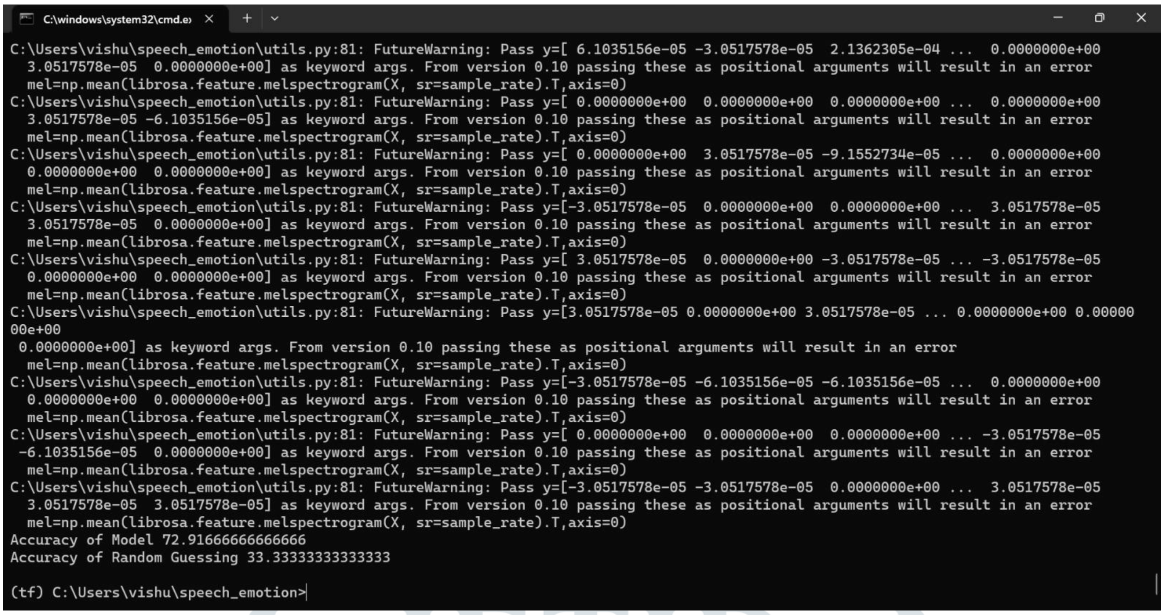
A system architecture, sometimes known as a systems architecture, is a conceptual model that describes a system's behaviour, structure, and other aspects. A formal description and representation of a system is an architecture description. arranged to aid in understanding of the system's structures and actions.

3-Tier Architecture:

In order to get around the restrictions of the two-tier design, the three-tier software architecture (also known as a three-layer architecture) was developed in the 1990s. Between the user interface (client) and data management (server) components is the third tier, sometimes known as the middle tier server. By offering services like queuing, application execution, and database staging, this middle tier provides process management where business logic and rules are put into action and can support hundreds of users (as opposed to only 100 users with the two tier architecture).

When a distributed client/server design is required that offers (in comparison to the two tier) better performance, flexibility, maintainability, reusability, and scalability, while concealing the complexities of distributed processing from the user, the three tier architecture is utilised. Three layer architectures are a common option for Internet applications and net-centric information systems as a result of these qualities.

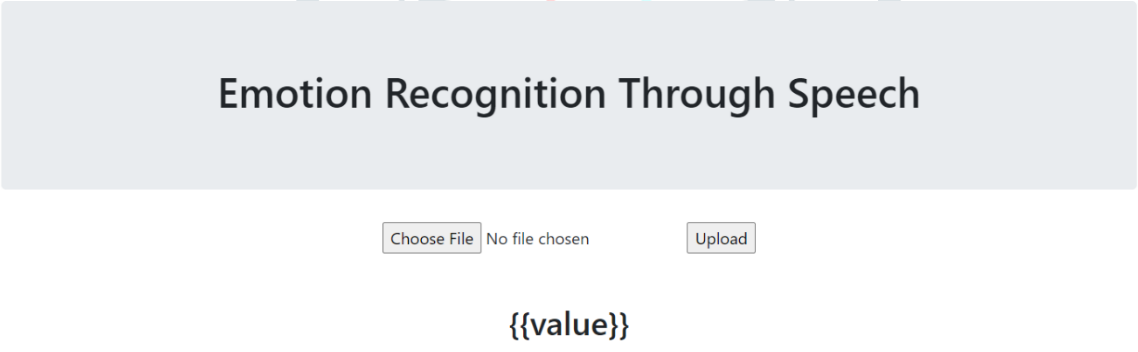
6. Result



```
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[ 6.1035156e-05 -3.0517578e-05  2.1362305e-04 ...  0.0000000e+00
  3.0517578e-05  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ...  0.0000000e+00
  3.0517578e-05 -6.1035156e-05] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[ 0.0000000e+00  3.0517578e-05 -9.1552734e-05 ...  0.0000000e+00
  0.0000000e+00  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[-3.0517578e-05  0.0000000e+00  0.0000000e+00 ...  3.0517578e-05
  3.0517578e-05  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[ 3.0517578e-05  0.0000000e+00 -3.0517578e-05 ... -3.0517578e-05
  0.0000000e+00  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[3.0517578e-05 0.0000000e+00 3.0517578e-05 ... 0.0000000e+00 0.00000
00e+00
 0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[-3.0517578e-05 -6.1035156e-05 -6.1035156e-05 ...  0.0000000e+00
  0.0000000e+00  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ... -3.0517578e-05
 -6.1035156e-05  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
C:\Users\vishu\speech_emotion\utils.py:81: FutureWarning: Pass y=[-3.0517578e-05 -3.0517578e-05  0.0000000e+00 ...  3.0517578e-05
  3.0517578e-05  3.0517578e-05] as keyword args. From version 0.10 passing these as positional arguments will result in an error
  mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
Accuracy of Model 72.91666666666666
Accuracy of Random Guessing 33.33333333333333

(tf) C:\Users\vishu\speech_emotion>
```

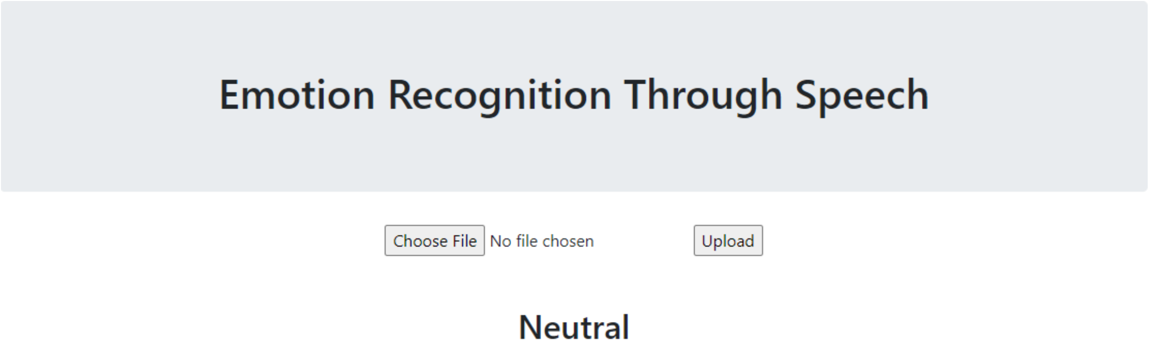
Fig 3 : Dataset Screen



Emotion Recognition Through Speech

{{value}}

Fig 4 : Home Screen



Emotion Recognition Through Speech

Neutral

Fig 5 : Emotion Detected

To obtain the results, the training process was initiated by executing command prompts and running specific commands in the command-line interface (CLI)[Fig 3]. The model was trained using the established methodology outlined in the research paper. Subsequently, a web page was created to present the output to the users[Fig 4]. The web page was linked to the trained data by configuring the Flask framework and setting the appropriate paths. Upon successful configuration, the web page was accessible through a generated IP address. Users were able to access the web page using a browser and select a specific WAV file to retrieve the corresponding output[Fig 5].

7. CONCLUSION

This project demonstrates how effective MLPs are in categorising voice signals. Even with reduced models, it is simple to recognise a small number of characters. In comparison to other methodologies, we have achieved higher accuracy for individual emotions. The effectiveness of the pre-processing has a significant impact on a module's performance. The reliability of Mel Frequency Cestrum Coefficients is excellent. Every human emotion has been carefully examined, dissected, and its correctness verified. The findings of this study show that speech recognition is possible and that MLPs can be utilised for any task involving voice recognition and identifying each emotion expressed in speech.

8. REFERENCES

[1].Hadhami Aouani, Yassine Ben Ayed," Speech Emotion Recognition with deep learning" in Procedia Computer Science,Volume 176, 2020, Pages-251-260 DOI:https://doi.org/10.1016/j.procs.2020.08.027
[2]. Zijiang Zhu,Weihuang Dai,Yi Hu,Junshan Li,"Speech emotion recognition model based on Bi-GRU and Focal Loss" in Pattern Recognition Letters,Volume 140, December 2020, Pages 358-365. DOI:https://doi.org/10.1016/j.patrec.2020.11.009
[3].Dias Issa,M. Fatih Demirci,Adnan Yazici,"Speech emotion recognition with deep convolutional neural networks" in Biomedical Signal Processing and Control,Volume 59, May 2020,101894. DOI:https://doi.org/10.1016/j.bspc.2020.101894
[4]. Mixiao Hou,Jinxing Li,Guangming Lu,"A supervised non-negative matrix factorization model for speech emotion recognition" in Speech Communication,Volume 124, November 2020, Pages13-20.DOI:https://doi.org/10.1016/j.specom.2020.08.002
[5]. Shadi Langari, Hossein Marvi, Morteza Zahedi,"Efficient speech emotion recognition using modified feature extraction",Informatics in Medicine Unlocked,Volume 20, 2020, 100424.DOI:https://doi.org/10.1016/j.imu.2020.100424

[6]. Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019. DOI:https://www.ijitee.org/portfolio-item/L26981081219/

[7]. Jianfeng Zhao, Xia Mao, Lijiang Chen. Learning Deep features to Recognise Speech Emotion using Merged Deep CNN. IET Signal Process., 2018.DOI:https://doi.org/10.1049/iet-spr.2017.0320
[8]. H.K. Palo, Mihir Narayana Mohanty and Mahesh Chandra. Use of different features for Emotion Recognition using MLP network. Springer India 2015, Computational Vision and Robotics, Advances in Intelligent Systems and Computing DOI:http://dx.doi.org/10.1007/978-81-322-2196-8_2
[9]. Ayush Kumar Shah, Mansi Kattel,Araju Nepal. Chroma Feature Extraction using Fourier Transform. Chroma_ Feature_ xtraction. January 2019.DOI:https://www.academia.edu/42216949/Chroma_Feature_Extraction
[10]. Sabur Ajibola Alim and Nahrul Khair Alang Rashid Some Commonly Used Speech FeatureExtractionAlgorithms.DOI:10.5772/intechopen.80419. DOI:http://dx.doi.org/10.5772/intechopen.80419
[11] Keshi Dai1, Harriet J. Fell1, and Joel MacAuslan2"Recognizing Emotion in Speech Using Neural Networks", IEEE Conference on "Neural Networks and Emotion Recognition" in 2013. DOI:https://doi.org/10.1109/ICONIP.1999.845644
[12] Margarita Kotti and Constantine Kouropoulos "Gender Classification in Two Emotional Speech Databases" IEEE Conference on2004. DOI:http://dx.doi.org/10.1109/ICPR.2008.4761624

[13] Mohammed E. Hoque1, Mohammed Yeasin1, Max M. Louwerse2 "Robust Recognition of Emotion from Speech", International Journal on October 2011, Volume 2, pp. 221-225. DOI:https://doi.org/10.1007/s10772-018-9546-1

[14] Nobuo Sato and Yasunari Obuchi. "Emotion Recognition using MFCC"s" Information and Media Technologies 2(3):835-848 (2007) reprinted from: Journal of Natural Language Processing 14(4): 83-96 (2007)DOI:https://doi.org/10.11185/imt.2.835

[15] T L Nwe'; S W Foo L C De Silva, "Detection of Stress and Emotion in speech Using Traditional and FFT Based Log Energy Features" 0-7803-8185-8/03 2003 IEEE (2003) DOI:10.1109/ICICS.2003.1292741

[16] Yixing Pan, Peiwei Shen and Lipping Shen, “Speech Emotion Recognition Using Support Vector Machine” International Journal on 2012, Issue 3, pp. 654 DOI:http://dx.doi.org/10.1007/978-3-642-21402-8_35

[17] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Transactions on Multimedia, vol. 16, pp. 2203-2213, 2014. DOI:10.1109/TMM.2014.2360798

[18] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8614-8618. DOI:10.1109/ICASSP.2013.6639347

[19] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," arXiv preprint arXiv:1301.3605, 2013. DOI:https://doi.org/10.48550/arXiv.1301.3605

[20] Deep Learning Technique for Speech Emotion Recognition “Jonnadula Narasimharao” 2022 Interanational Conference on Futuristic Technologies,INCOFT 2022,Nov 2022 DOI: https://doi.org/10.1109/INCOFT55651.2022.10094534