SPEECH BASED EMOTION DETECTION SYSTEM USING MFCC

Vemula Yakub Reddy[1], Mangipudi Pavan Kumar[2], Mankala Sushma[3],Gurindagunta Kiran[4],Vijaya Kumar Gurrala[5]

[1-4]Dept. of Electronics and Communication Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India.
[5]Assistant Professor, Dept. of Electronics and Communication Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India.

---***---

Abstract - *Generally, people express their emotions through speech, facial expressions and body pose. But, estimating the state of his emotion can be found out easier through speech only. Recent studies says that harmony features of speech signal help in recognizing emotions easily. Because speech is a major channel for communicating emotion. Here we developed a speech based emotion detection system considering German emotional corpus database (EMODB) using Neural Network approach. It comprises 10 sentences which covers 7 classes of emotion from everyday communication. Using Fourier parameters of speech signal i.e. when speech signal is Fourier transformed, harmonies can be calculated by extracting features using Mel Frequency Cepstrum Coefficient (MFCC). Thus, by extricating the emotional conditions of speaker from speech, we can improve the exhibition of speech based emotion detection system and subsequently extremely valuable for criminal investigations, smart assistance surveillance and the location of dangerous events in health care systems too.*

Key Words: Mel-Frequency Cepstral Coefficients(MFCC), Speech Recognition, Cepstrum, Speech analysis, Neural Networks.

1. INTRODUCTION

Speech recognition [1] is the way toward changing over an acoustic signal, caught by a Microphone to a lot of words. These words can be utilized for applications such as orders and control, information passage, and record planning. Speech is acoustic signal which contains data about the perspectives on the speaker and furthermore the thoughts that pivoting in the brain of a speaker. Automatic Speech Recognition [2] (ASR) is just based on acoustic data in audio signal. But in an uproarious situation, its precision level is less. Along these lines, rather than Audio Speech Recognition (ASR) [3], we can utilize Audio-Visual Speech Recognition [4] (AVSR) which utilize both speech and visual data moreover. Audio is one an antiquated approach for communication. In present days these speech signal are utilized in man-machine communication also. When inspected in an adequately brief timeframe (5-100 m sec), its attributes are fixed. Speech based emotion detection plays a major role in machine learning platform by improving man-machine interaction. Emotions plays a major role in human environment, we can find the emotion of a person by seeing his/her facial expressions or by noticing his/her actions. Here, this system deals with the detecting emotions of a person from his speech. By recording the a speech of a person and extracting features from those speech and by performing specific actions on them, emotion behind those speech can be extracted.

To improve machine man interface speech based emotion detection system gives some different applications, for example this system can be utilized in Airplane cockpits to give examination of Mental condition of pilot to stay away from calamities, for example, mishaps. Speech emotion detection system also uses to recognize worry in speech for better execution lie recognition, in Call centre conversation to break down lead examination of the customers which helps with improving nature of nature of a call systematic and in like manner in clinical field for Mental determination. Emotion detection in criminal investigations also helps in finding criminals who hides emotions behind their facial expressions. If machine will prepared to understand individuals like emotion conversation with programmed robot toys would be dynamically reasonable and pleasant. In vehicle board system where information of the mental state of the driver may be given for the system in keeping in mind of his/her security.

2. SPEECH EMOTION DETECTION

Generally, speech emotion [7] can be recognized by deeply analyzing the speech signal. Here the speech signal is divided into frames and separate frames are analyzed and features such as pitch or fundamental frequencies, energy, MFCC values are obtained and using neural network mechanism emotion is classified. The assessment of the speech emotion detection [8] system depends on nature of speech/audio. In event that the substandard speech is utilized as a contribution to the system, at that point we might have wrong conclusions. The audio signal as a contribution to the emotion recognition [9] system might have this present reality emotions. The main aim of this model is to detect the emotion of the speaker from his voice with help of feature extraction with a popular technique called MFCC feature extraction and neural network classifier to modify its emotion detection accuracy. The speech emotion detection have 5 primary sections which are shown in the below block diagram.

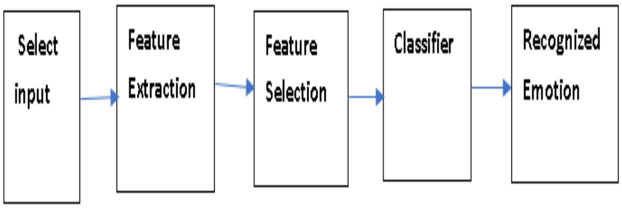

Fig -1: Emotion detection system

2.1 Feature Extraction in Input Speech

In this step, feature [10] extraction must have a lot of consideration since emotion detection relies intensely upon this stage. The principle objective of feature[11] extraction step is to process a miserly succession of feature vectors giving minimized portrayal of provided input speech signal. This step is generally acted in 3 phases. The principal phase is known as the pre-emphasis. In this primary step speech signal is digitized and sampled into several frames." The subsequent stage involves in Fourier transform of signal. At last, the final step changes the all-encompassing vector into progressively minimized & powerful vector then they are given to the detector.

2.2 Mel-Frequency Cepstral Coefficients(MFCC)

In Mel-Frequency Cepstral Coefficients[13] (MFCC) first step involves in the windowing of the speech signal which is obtained from the Fast Fourier Transform(FFT)in the pre emphasis step. "This windowing technique used is called hamming window which is used to eliminate discontinuous edges to eliminate noise added the signal. Then the Mel-scale filter bank is applied and its Discrete Cosine Transform (DCT) is calculated and then these cepstral coefficients are calculated. Then its DFT is measured to perform its power spectral analysis, where its real values are converted into time domain. The process MFCC is a feature extraction procedure which extracts features from the audio like where human used for hearing audio and eliminates all other data.

3. CLASSIFICATION

3.1 Neural Network Model

Neural Networks are the mathematical models that have been utilized in data processing. On a very basic level, Neural Network models are an interconnected system of nodes, corresponding to the huge system of neurons in the human brain. In an Artificial Neural Network [16](ANN), every node is allotted to the system represents to a neuron. Particularly neurons get the input from the other corresponding neurons by means of neurotransmitter association called synapse. A neuron commonly associates with an individual processing element, which is called perceptron. These Neural Network models have the intermediate layers where the whole data is processed called intermediate layers. These intermediate layers may be single layer or the multiple layers. The number of intermediate layers used is mainly depends on the complexity of input data given to the input layer.

3.2 Multi-Layer Perceptron (MLP)

Multi Layer Perceptron(MLP)[17] comprises more than one hidden layer of perceptron in a system. A typical arrangement of layers in a MLP has input, output, and hidden layers. In an Artificial Neural Networks, the input layer is the primary layer acts a channel for entering the information. The subsequent layer is a hidden layer, these hidden layers is used to process the data, the number of hidden layers is mainly depends on the complexity of the input data. The final layer is the output layer where the processed data is presented.

3.3 Feed-Forward Neural Network (FFNN)

The Feed-Forward Neural Network are the metamorphosis model which transforms input layers to the output layers within the forward direction algorithm is one of the most significant standard strategies for chemical characterization of sediments utilizing hyperspectral information.

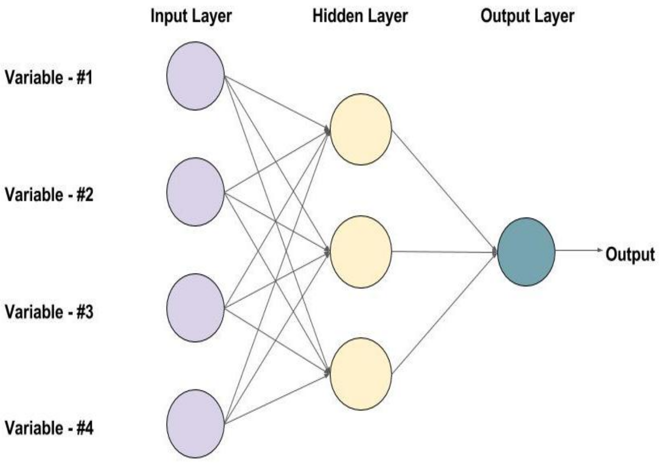

Fig -2: An example of Feed –forward Neural Network with 3 neurons (1 bias) and 1 hidden layer.

4. RESULTS

In the speech emotion detection system after extracting features, they are given to the classifier section. There are various methods of classifier are proposed for the task of speech emotion detection. Here, we used artificial neural network model. The calculated features are given to the neural network model then it processes the data and classifies the output by showing the as emoji of an emotion based on the respective input signal and also it provides an audio of the input signal at the time of displaying emoji. Here are some of the resulting waveforms obtained by giving a particular speech signal of angry emotion and the duration of that signal is 3 sec.

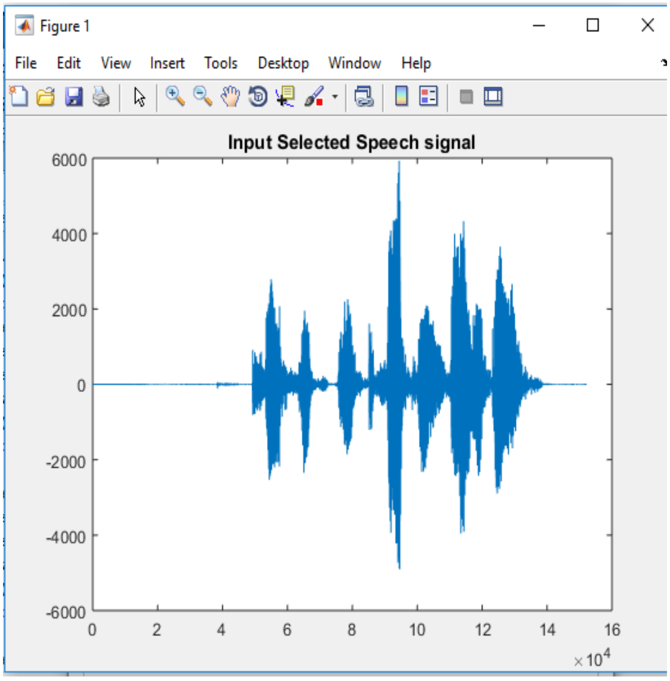
Figure 1
File Edit View Insert Tools Desktop Window Help
Input Selected Speech signal
6000
4000
2000
0
-2000
-4000
-6000
0
2
4
6
8
10
12
14
16
× 10^4

Fig -3: Input speech signal

This was the input speech signal which was given as input to the speech based emotion detection system.

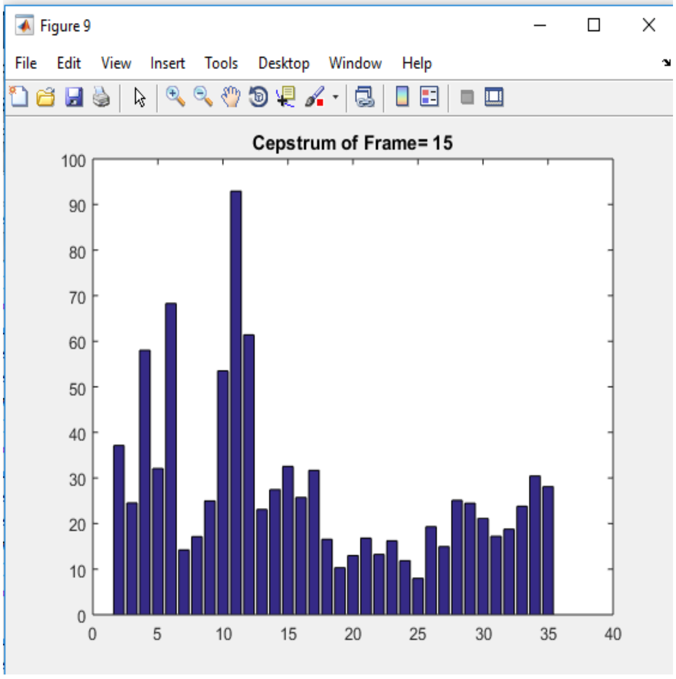

Fig -4: Cepstrum of one of the frame

These were cepstrums for divided frames from the input speech signal. Here we have presented cepstrum of one frame in Fig 4. Similarly, we will have cepstrums of all other frames.

Fig -5: Detected Emotion

This was the final detected emotion emoji based on the input speech signal along with we will have audio output of speech signal.

5. CONCLUSION AND FUTURE WORK

The proposed system clearly shows the variations between different speech signals based on their nature, for better precisions we have to take the speech input signals from good and silent environments otherwise speech input may contain noise so that we cannot detect emotion from the speech as expected. The significant issues in speech emotion recognition system are the signal processing unit in which proper features are separated from accessible speech signal and another is a classifier which recognizes emotions from the speech signal. The normal precision of the vast majority of the classifiers for speaker independent system is not as much as that for the speaker dependent. Emotion recognitions from the human speech are expanding now daily since it brings about the better collaborations among human and machine. To improve the emotion recognition process, combinations of the given strategies can be determined. Additionally by separating more effective features of speech, precision of the speech emotion recognition system can be upgraded. This work can be further researched in many ways. Neural network systems and hidden markov model for automatic emotion detection are the most regular devices in such ways. By taking the video input along with speech signal input we can improve the emotion detection system with more accuracy. Similarly the mixture of various techniques like video based or manual association will be inspected further.

REFERENCES

[1] M. A. Anusuya and S. K. Katti, "[2009] Speech Recognition by Machine: A Review," IJCSIS) Int. J. Comput. Sci. Inf. Secur., vol. 6, no. 3, pp. 181–205, 2009.

[2] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef, "Comparative study of automatic speech recognition techniques," IET Signal Process., vol. 7, no. 1, pp. 25–46, 2013, doi: 10.1049/iet-spr.2012.0151.

[3] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically- inspired binary cascade classification schema," Int. J. Speech Technol., vol. 15, no. 2, pp. 131–150, 2012, doi: 10.1007/s10772-012-9127-7.

[4] M. Anusuya and S. Katti, “Front end analysis of speech recognition: A review,” Int. J. Speech Technol., vol. 14, pp. 99–145, Jun. 2011, doi: 10.1007/s10772-010-9088-7.

[5] D. O’Shaughnessy, “Interacting with computers by voice: Automatic speech recognition and synthesis,” Proc. IEEE, vol. 91, pp. 1272–1305, Oct. 2003, doi: 10.1109/JPROC.2003.817117.

[6] S. Zahid, F. Hussain, M. Rashid, M. H. Yousaf, and H. A. Habib, “Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods,” Math. Probl. Eng., vol. 2015, p. 209814, 2015, doi: 10.1155/2015/209814.

[7] S. Saksamudre, P. P. Shrishrimal, and R. Deshmukh, “A Review on Different Approaches for Speech Recognition System,” Int. J. Comput. Appl., vol. 115, pp. 23–28, Apr. 2015, doi: 10.5120/20284-2839.

[8] C. Maaoui and A. Pruski, “Emotion Recognition through Physiological Signals for Human-Machine Communication,” 2010.

[9] R. Cowie et al., “Emotion recognition in human-computer interaction,” Signal Process. Mag. IEEE, vol. 18, pp. 32–80, Feb. 2001, doi: 10.1109/79.911197.

[10] S.-T. Pan and T.-P. Hong, “Robust Speech Recognition by DHMM with A Codebook Trained by Genetic Algorithm,” J. Inf. Hiding Multimed. Signal Process., vol. 3, Jan. 2012.

[11] C. Kotropoulos and D. Ververidis, “Emotional speech recognition: Resources, features, and methods,” Speech Commun., vol. 48, no. 9, pp. 1162–1181, 2006.

[12] I. Luengo, E. Navas, and I. Hernaez, “Feature analysis and evaluation for automatic emotion identification in speech,” IEEE Trans. Multimed., vol. 12, no. 6, pp. 490–501, 2010, doi: 10.1109/TMM.2010.2051872.

[13] I. Patel and Y. S. Rao, “Speech recognition using hidden markov model with MFCC-subband technique,” ITC 2010 - 2010 Int. Conf. Recent Trends Information, Telecommun. Comput., vol. 1, no. 2, pp. 168–172, 2010, doi: 10.1109/ITC.2010.45.

[14] V. K. Kale, P. D. Deshmukh, and H. R. Gite, “Voice Based Biometric System Feature Extraction Using MFCC and LPC Technique,” Int. J. Adv. Eng. Res. Sci., no. 5, pp. 4–8, 2016.

[15] V. Krishnan and B. Anto, “Features of Wavelet Packet Decomposition and Discrete Wavelet Transform for Malayalam Speech Recognition,” vol. 1, Jan. 2009.

[16] N. Pushpa, R. Revathi, C. Ramya, and S. S. Hameed, “Speech Processing Of Tamil Language With Back Propagation Neural Network And Semi- Supervised Traning,” vol. 2, no. 1, pp. 2718–2723, 2014.

[17] G. Sivaram and H. Hermansky, Multilayer perceptron with sparse hidden outputs for phoneme recognition. 2011.