IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

A Perspective Study on Speech Recognition

Sana Fatema N. Ali[1] and Prof. S. T. Khandare[2]
ME Scholer, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India[1]
Associate Professor, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India[2]

Abstract: *Emotions play an extremely important role in human mental life. It is a medium of expression of one's perspective or one's mental state to others. Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions including Neutral, Anger, Happiness, and Sadness in which any intelligent system with finite computational resources can be trained to identify or synthesize as required. In this work spectral and prosodic features are used for speech emotion recognition because both of these features contain the emotional information. Mel-Frequency Cepstral Coefficients (MFCC) is one of the spectral features. Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features which are used to model different emotions. The potential features are extracted from each utterance for the computational mapping between emotions and speech patterns. Pitch can be detected from the selected features, using which gender can be classified. The audio signal is filtered using a method known as feature extraction technique. In this article, the feature extraction technique for speech recognition and voice classification is analyzed and also centered to comparative analysis of different types of Mel frequency cepstral coefficients (MFCC) feature extraction method. The MFCC technique is used for deduction of noise in voice signals and also used for voice classification and speaker identification. The statistical results of the different MFCC techniques are discussed and finally concluded that the delta-delta MFCC feature extraction technique is better than the other feature extraction techniques..*

Keywords: Digital Signal Processing, MATLAB, Machine Learning, voice modulation

I. INTRODUCTION

Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI).

Speech Emotional Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. It has potentially wide applications, such as the interface with robots, banking, call centers, car board systems, computer games etc. For classroom orchestration or E-learning, information about the emotional state of students can provide focus on enhancement of teaching quality. For example, teachers can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why the learner's emotional state should be considered in the

classroom. In general, the SER is a computational task consisting of two major parts: feature extraction and emotion machine classification. The questions that arise here: What is the optimal feature set? What combination of acoustic features for a most robust automatic recognition of a speaker's emotion? Which method is most appropriate for classification? Thus came the idea to compare a Recurrent Neural Network (RNN) method with the basic method Middle Latency Response (MLR) and the most widely used method Support Vector Machine (SVM). And also all previously published works generally use the Berlin database. To our knowledge the Spanish emotional database has never been used before. For this reason we have chosen to compare them. In fact, the emotional feature extraction is a main issue in the SER system.

For the audio signal analysis used a feature extraction technique known as MFCC. The objective of this paper is to transform the audio waveform to frequency domain representation, for advanced signal processing and analysis. Here also discuss the comparative analysis of different MFCC methods. The important parameter of speech signal in feature extraction method is Cepstral coefficients and pitch frequency.

It is used for speech recognition, speech synthesis and speaker verification, etc. Here extract the highlights of the discourse section, for example, essential frequency, groups, Cepstral coefficient line spectral pairs, MFCC and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

spectrogram. Here we are mainly discussing Cepstral coefficients for the speech recognition and different types of the MFCC. The process of the feature extraction technique is first the speech is analyzed over a short frame window and then each short frame window, obtained by Fast Fourier Transform (FFT). The Mel spectrum is obtained when the output of the FFT is passed through a Mel-filter. For the MFCC the Mel spectrum is performed by Cepstral coefficients. Therefore, the audio signal is signified as a sequence of the Cepstral vector.

II. TYPES OF FEATURE EXTRACTION TECHNIQUES

2.1 Linear Prediction Coding (LPC)

LPC is one of the good signal analysis methods for linear prediction in the speech recognition process. The feature extraction techniques find out the basic parameters of speech.

LPC is the most powerful method for determining the basic parameter and computational model of speech. The idea behind LPC is the Speech sample can be approximated as a linear combination of past speech samples.

A. Advantages

1. The main advantage of linear predictive coding is to reduce the bitrates of the speech i.e. reduces the size of the transmitting signal.
2. The signal transmitted through LPC required less bandwidth and hence number of users can be increased
3. This method of coding uses the encryption of data so the data is secured until the destination.

B. Disadvantages

1. Due to reduction in the bitrates of the speech signal, the quality of voice signal is reduced.
2. This technique is lossy compression technique, hence data gets faded if transmitted to the long distance.

2.2 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is the most popular feature extraction technique. Frequency bands are placed logarithmically here so it approximates the human system response more closely than any other system. Due to its advantage of less complexity in implementation of feature extraction algorithm, only sixteen coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstral are extracted from spoken word samples in database As shown in below figure the first step is pre-processing in which the signals are pre-processed before feature extraction. In framing the signal splits into a number of frames in time domain, then on each individual frame the hamming window is applied. Discrete Fourier Transform (DFT) is used to convert each frame from time domain to frequency domain. The filter bank is created by calculating the number of picks spaced on Mel-scale and again transforming back to the normal frequency scale. Discrete Cosine Transformation (DCT) is used to convert the Mel spectrum coefficient to the time domain.

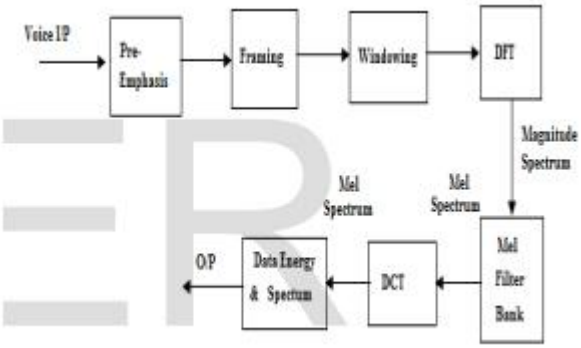

Fig 1: Mel Frequency Cepstral Coefficient (MFCC)

2.3 Linear Prediction Cepstral Coefficient (LPCC)

The feature extraction is used to demonstrate the speech signals by finite number of measures of the signals. To obtain LPCC coefficients the LPC coding is used. LPCC implemented using the autocorrelation method. The main drawback of LPCC is that the LPCC are highly sensitive to quantization noise.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

A. Advantages

1. Provides good discrimination
2. Low correlation between coefficients
3. Not based on linear characteristics; hence, similar to the human auditory perception system Important phonetic characteristics can be captured

B. Disadvantages

1. Low robustness to noise
2. In a continuous speech environment, a frame may not contain information of only one phoneme, but of two consecutive phonemes
3. Not flexible since the same basic wavelets have to be used for all speech signals.

2.4 Discrete Wavelet Transform (DWT)

DWT can be considered as a filtering process achieved by a low pass scaling filter and a high pass wavelet filter. The transform decomposition separates the lower frequency contents and higher frequency contents of the original signals. The lower frequency contents provide a sufficient j sampling process of overall coefficients is still the same and there is no redundancy.

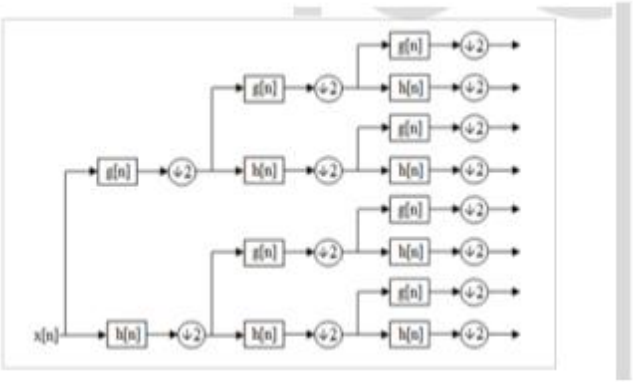

Fig 2: Wavelet Packet Decomposition (WPD)

A. Advantages

1. Successfully used for denoising tasks.
2. Capable of compromising a signal without major Degradation**.**

B. Disadvantages

1. Not Flexible since the same basic wavelets are useful for all speech signals.

2.5 Perceptual Linear Prediction (PLP)

The Perceptual Linear prediction (PLP) technique was developed by Hermansky. PLP removes the unwanted information of the speech and thus improves speech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of the human auditory system.

A. Advantages

1. Reduction in the discrepancy between voice and invoice speech.
2. Resultant feature vector is low dimensional

B. Disadvantages

1. Resultant feature vector dependent on the whole spectral balance of the formanant aptitude.
2. Spectral balance is easily altered by communication channels and noise and equipment used.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

III. LITERATURE REVIEW

Thiang, presented speech recognition using Linear Predictive Coding (LPC) and Artificial Neural Network (ANN) for controlling movement of mobile robots. Input signals were sampled directly from the microphone and then the extraction was done by LPC and ANN [1].

Ms.Vimala. C and Dr.V.Radha proposed an independent isolated speech recognition system for Tamil language. Feature extraction, acoustic model, pronunciation dictionary and language model were implemented usingHidden Markov Model(HMM) which produced 88% of accuracy in 2500 words [2].

Cini Kurian and Kannan Balakrishnan found development and evaluation of different acoustic models for Malayalam continuous speech recognition. In this paper HMM is used to compare and evaluate the Context Dependent (CD), Context Independent (CI) models and Context Dependent tied (CD tied) models from this CI model 21%. The database consists of 21 speakers including 10 males and 11 females [3].

Suma Swamy introduced an efficient speech recognition system which was experimented with Mel Frequency Cepstral Coefficients (MFCC), Vector Quantization (VQ), HMM which recognize the speech by 98% accuracy. The database consists of five words spoken by 4 speakers ten times [4].

Annu Chaudhary proposed speech recognition system for isolated and connected words of Hindi language by using Hidden Markov Model Toolkit (HTK). Hindi words are used for dataset extracted by MFCC and the recognition system achieved 95% accuracy in isolated words and 90% in connected words [5].

IV. CONVOLUTIONAL NEURAL NETWORK (CNN)

A sort of feed-ahead artificial network in which the joining sequence among its nodes is motivated by presenting an animal visual-cortex. Single cortical neurons give response to the stimuli at a prohibited area of the region known as the receptive areas. The receptive areas of various nodes semi-overlap so that they can match the visual area. The reply of a single node for stimuli among its receptive area could be mathematically through the convolution operations. Convolutional networks were motivated by natural procedures and are varieties of multi-layer perceptions formulated to use the least quantity of pre-processing. They have broad use in image and video recognition, recommendation systems and Natural Language Processing (NLP). The dimensions of the Characteristics Map (Convolved Features) is regulated by following parameters:

- Depth: Representing the filter count we used in the convolution operation.
- Stride refers to size of the filter, if the size of the filter is5x5 then stride is equal to 5.
- Zero-padding: Padding the input matrix with 0s was often convenient around the border, in order to apply filter to ‘Input Audio’ matrix’s bordering elements. Using zero padding size of the characteristics map can be governed.

V. FEATURE EXTRACTION

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The Feature extraction is usually performed in three stages.

The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectral temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals.

The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage (which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer.

Although there is no real consensus as to what the optimal feature sets should look like, one usually would like them to have the following properties:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

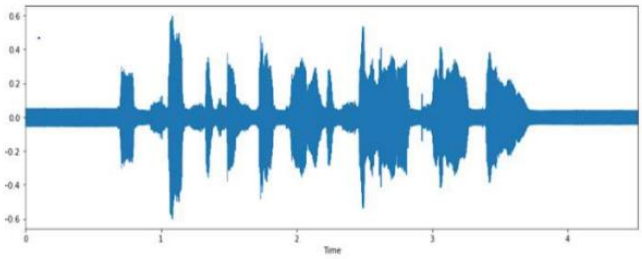

Fig 3: Feature Extraction

They should allow an automatic system to discriminate between different through similar sounding speech sounds, they should allow for the automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environment. Automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environments.

VI. METHODOLOGY

Speech samples are first passed through a gender reference database which is maintained for recognition of gender before it gets into the process. Statistical approach is followed by taking pitch as a feature for gender recognition. A lower and upper bound pitch for both male and female samples could be found using the reference database. Input human voice samples were first broken down into frames of frame size 16 ms each. This was done for frame level classification in further steps. For each frame MFCC (Mel Frequency Cepstral Coefficient)was calculated as the main feature for emotion recognition. Reference database is maintained which contains the MFCCs of emotions i.e. of Sad, Anger, Neutral and Happy. MFCC of the frames were compared with the MFCCs stored in the reference database and the distance was calculated between the comparable frames. Based on the distance of the analysis frame from the reference database, one can classify the frame as anger, happy or normal. The output is displayed in terms of emotional frame count Y.

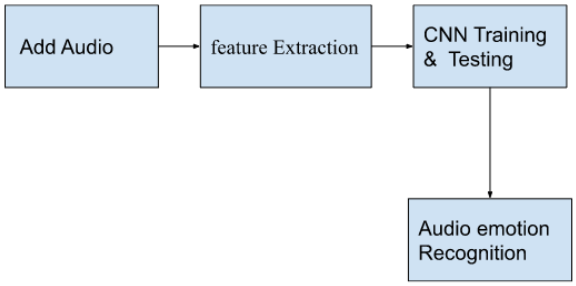

Fig 4: Block Diagram of Speech Emotion Recognition

VII. CONCLUSION

After constructing various models, we got the better CNN model for the emotion distinction task. We expected to get 85% accuracy from the previously available model. Our model would've performed better with more data. Also our mode will perform very well when distinguishing among a masculine and feminine voice. Weanalyse different types feature extraction technique and observe their Accuracy and recognition rate which is given below.

Table 1: Feature Extraction Techniques their Accuracy and Recognition Rate.

Method Type	Accuracy	Word Recognition Rate
CNN	90%	89.49%
LPC	87.1%	84%
MFC	89%	80%
HMM	88%	80%
DWT	87.8%	…

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

Above table is a summary of the literature we study for our project and analysis that CNN gives the 90 % accuracy and 89.49% Word Recognition Rate.

REFERENCES

[1]. Thiang and Suryo Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robots", in Proceedings of International Conference on Information and Electronics Engineering (IPCSIT).

[2]. Ms.Vimala.C and Dr.V.Radha, "Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM", in Proceedings International Conference on Communication Technology and System Design 2020, Procedia Engineering 30 ISSN: 1877-7058, 13March 2020, pp.1097-1102.

[3]. Cini Kuriana and Kannan Balakrishnan, "Development & evaluation of different acoustic models for Malayalam continuous speech recognition", in Proceedings of International Conference on Communication Technology and System Design 2020 Published by Elsevier Ltd, December 2020, pp.1081-1088.

[4]. Suma Swamy and K.V Ramakrishnan, "An Efficient Speech Recognition System". Computer Science & Engineering: An International Journal (CSEIJ), Vol.3, No.4, and DOI: 10.512 1/cseij.2019.3403 August 2021, pp.21-27.

[5]. Annu Choudhary, Mr. R.S. Chauhan and Mr. Gautam Gupta et.al. "Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK)", in Proceedings of International Conference on Emerging Trends in Engineering and Technology, 03.AETS.2013.3.234, 22-24th February 2020, pp.244– 252.

[6]. P. Sharma, V. Abrol, A. Sachdev and A. D. Dileep, et.al. "Speech emotion recognition using kernel sparse representation based classifier," in 2021 24th European Signal Processing Conference (EUSIPCO), pp. 374-377, 2021.

[7]. Linhui Sun, Yiqing Huang, Qiu Li and Pingan Li, et.al.Multi-classification speech emotion recognition based on two-stage bottleneck features selection andMCJD algorithm, Signal Image and Video Processing,10.1007/s11760-021-02076-0, 2022.

[8]. Yu Wang, Research on the Construction of Human-Computer Interaction System Based on a Machine Learning Algorithm, Journal of Sensors,10.1155/2022/3817226 2022, Vol 2022, pp. 1-11.

[9]. Sandeep Kumar Pandey, Hanumant Singh Shekhawat and S.R.M Prasanna, et.al. Attention gated tensor neural network architectures for speech emotion recognition, Biomedical Signal Processing and Control ▫ 10.1016/j.bspc.2021.103173, 2022, Vol 71pp. 103173.

[10]. Jason C. Hung and Jin-Che Chen, Construction and Research of E-sports Speech Emotion Recognition Model, Lecture Notes in Electrical Engineering - Innovative Computing, 10.1007/978-981-16-4258-6_4,2022, pp. 23-31.