Study on Feature Extraction of Speech Emotion Recognition

Sweta Bhadra[1], Uzzal Sharma[2], Alok Choudhury[3]
Don Bosco College of Engineering and Technology, Assam Don Bosco University
Airport Road, Azara, Guwahati - 781017, Assam. INDIA.
[1]sweta.bhadra4@gmail.com, [2]uzzal.sharma@dbuniversity.ac.com, [3]alok.choudhury@dbuniversity.ac.com

Abstract:*Speech emotion recognition system aims at automatically identifying the emotion of the speaker from the speech. It is a modification of the speech recognition system which only identifies the speech. In this paper, we study the feature extraction algorithm such as pitch, formant frequency and MFCC.*

Keywords: Feature extraction, pitch, formant frequency, MFCC.

1. Introduction

Speech is a form of vocalized communication for people to interact with each other [1]. It is the most natural way of exchanging information [2]. A system which can recognize speech without requiring any technical proficiency from its users was desired by the people. Therefore people's desire was the motivation behind the development of speech recognition system. Speech recognition provides the interface to recognize speech. The process of converting speech signals into sequence of words by means of an algorithm is called speech recognition. Speech recognition process has made human voice understandable by a computer [3]. The speech recognition program consists of a set of grammatical rules which enables the computer to recognize an input and form sentences from it.

Speech signals not only convey words and meanings but also the emotion of the speaker. Emotions influence processes such as perception, learning, decision making and therefore it has a great impact onhuman behavior [4]. Emotion colours the understanding of what is going around, interpretation of the motives of people and responses one thinks to be appropriate. Enabling emotion to speech recognition is known as speech emotion recognition which not only recognizes speech but also helps to detect the emotional state of the speaker. Thus implementing emotion in speech recognition system improves machine interaction with its users and adds naturalness to the system. An efficient speech emotion recognition system provides natural and friendly interaction between a computer and a human [5]. Speech emotion recognition system helps to avoid misunderstandings and therefore finds its application in medical field, automatic dialog systems, e-learning etc.

Detecting human emotion from speech is a very challenging issue and that is why it has been an area of huge interest for many researchers in the past few years. Recognizing emotion from human speech is a very tedious task due to the ambiguity related between classifying the acted and the natural emotions [4]. Studies has been carried out to explore the acoustic indicators to detect emotions in human speech. The acoustic features helps to identify the speaker and the nature of utterances to achieve the maximum natural interaction in an automated system [6].

Human emotions can be classified as happy, joy, sad, angry, neutral etc. Extracting distinct features from the speech helps in accurate recognition of emotions. The use of feature sets increases the rate of recognition in the system. In this paper we discuss the methods to extract the features from a recorded speech sample to detect the emotion from the speech. We study few of the feature extraction method such as pitch, formant frequency, MFCC and compare the results to find which method provides the maximum accuracy.

2. Feature Extraction

Feature extraction is the most important stage of recognition. There are many methods to extract the features of a speech emotion recognition system. Some of them are discussed below.

2.1 Pitch

Pitch is a fundamental property of speech. It is well known that speech is driven by noise which is produced by the vibration of the vocal folds that varies at a rate between 50 Hz to about 400 Hz, which is known as fundamental frequency. Pitch is known as the fundamental frequency of the speech signal. The characteristics of pitch are considered widely in the field of stress evaluation which includes pitch frequency assessment and mean, variance and distribution analysis. There are many ways for extraction of pitch, one of which is the 'Cepstral method' which has been discussed.

In the Cepstral method, the analog signal is first sampled at a suitable rate and then it is quantized to convert the signal into digital form [7]. The digital signal is then converted into frames of suitable size by passing the signal through a hamming window and then by using Fast Fourier Transform the signal is converted to the frequency domain. After the signal is converted into its frequency domain, the absolute values of the signal are considered and the signal logarithm is obtained. Using Inverse Fast Fourier Transform the signal is transformed into Cepstral domain in which the peak signal represents the pitch frequency. Cepstrum of a signal is calculated by the formula s[n] = c[n] + Θ[n], where

$$c[n] = \frac{1}{2\pi}\int_{-\pi} \log|S(w)|e^{jnw}\, dw \quad (1)$$

$$S(w) = \sum_{n=-\infty}^{\infty} s[n]e^{-jnw} \quad (2)$$

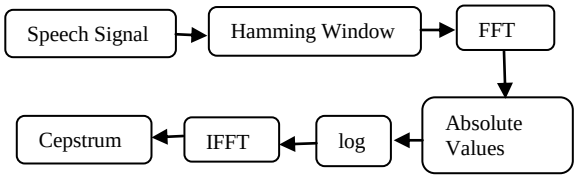

Figure1: Block Diagram for Cepstral Analysis.

2.2 Formant Frequency

Formant Frequency plays a very important role in analyzing the emotional state of a person. It is defined as the spectral peak of the sound spectrum. In phonetics, formant basically means the acoustic resonance of the human vocal tract. They are measured as amplitude peaks in the frequency spectrum of the sound.

For extracting the formant frequencies the Linear Predictive Coding (LPC) is used in which the analog signal is converted to .wav format [7]. Using Fast Fourier Transform the signal is transformed into its frequency domain and the power spectrum is calculated. The signal is then passed through a Linear Predictive Filter with coefficients and the roots of the polynomial are obtained.

2.3 MFCC

Mel frequency cepstral coefficients was introduced in 1980 by Davis and Mermelstein [8]. It is based on the human peripheral auditory system. MFCC is less susceptible to noise and provides better recognition performance [9]. MFCC due to its good performance is used widely in audio classification experiments. It is used to extract features from speech signal. Mel is a unit to measure the perception of speech or frequency of a tone [6]. The Mel-scale is a mapping of real frequency scale to the perceived frequency scale which is linear below 1000 Hz and logarithmically above.

In this algorithm the speech signal is segmented into frames by using the sliding window method. To obtain the magnitude spectrum Discrete Fourier Transform of each framed signal is considered. Later to transform the spectrum into Mel scale frequency wrapping is done in which triangular uniformly spaced filter banks are obtained. To obtain the MFCC the filter banks are multipliedwith the magnitude spectra [6].

The steps in MFCC are as follows [10]

Step 1: Frame Blocking

The process of segmenting the speech samples into a small frame with length ranging between 20 msec to 40 msec. The voice signal is divided into frames of N samples and adjacent frames are separated by M (M<N). The values used for M is 100 and N is 256.

Step 2: Windowing

The speech signal is segmented into frames by this technique. Hamming window is used as the window shape and represented as W (n) and $0 \leq n \leq N-1$ where N is the number of samples in each frame. The result is given by $Y(n)= X(n) \times W(n)$ where X(n) is the input signal and Y(n) is the output signal.

$$W(n)=0.54-0.46\cos 2\pi nN-1 \quad 0\leq n \leq N-1. \qquad (3)$$

Step 3: Fast Fourier Transform

Each frame of N samples is converted to frequency domain from time domain using the Fast Fourier Transform (FFT).

$Y(w) = FFT[h(t) * X(t)]= H(w)* X(w)$ (4) where X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Step 4: Mel frequency wrapping

In FFT spectrum voice signal does not follow the linear scale. It uses set of triangular filters to compute a weighted sum of filter spectral components so that the output approximates to a Mel scale. The magnitude frequency response of each filter is triangular in shape and at the centre frequency equals to unity and decrease linearly to zero between two adjacent filters. The sum of its filtered spectral components is the output of each filter.

$$F \text{ (el)} = [2595 * log10[1 + f] 700] \qquad (5)$$

The equation is used to calculate the Mel for frequency f in Hz.

Step 5: Cepstrum (Discrete Cosine Transform)

Here the log Mel spectrum is converted into time domain using Discrete Cosine Transform (DCT). The output of the conversion is known Mel Frequency Cepstrum Coefficient.

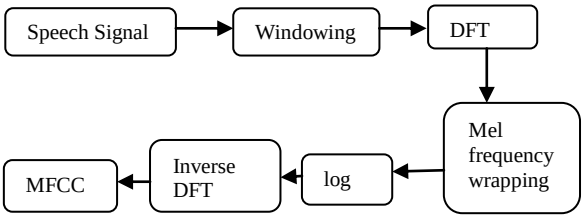

Figure2: Block Diagram for MFCC

3. Comparison

Extracting speech features helps in the detection of human emotion i.e. happy, angry, sad, neutral etc. The use of feature set helps to increase the recognition rate of the system. Anger is the emotion which has the highest level of pitch and energy. In angry state males present higher levels of energy than the females. It has been observed that males express anger with a slow speech rate but females expresses fast speech rate under similar circumstances. Disgust state on the other hand is expressed with a low pitch of level, low intensity level, and a slow speech rate than the neutral state. The fear is related with a high pitch level and high intensity level. Sadness is related to low levels of the mean intensity and pitch. The speech rate is slower than the neutral state in similar circumstances. The pitch is an important parameter because it helps to separates fear from joy.

From the experiments performed, we first identify the three emotional states happy, neutral, and angry. It has been found that pitch provides the best feature to identify the emotional states such as neutral and angry. It has been observed that the pitch frequency for both neutral and angry states are quite different after considering the mean of pitch frequency for eight different speakers. The mean of pitch frequency for angry state is found to be 336 Hz which is much higher than the mean of neutral state which is 130 Hz. Thus for identifying the angry and neutral emotions from speech pitch is considered the best indicator [7].

Formant Frequency method helps to recognize the happy emotion satisfactorily. In the experiment formant frequencies of five different speakers were considered which lied in a particular range. The range of formant frequencies mean was found to be 338 Hz, 802 Hz, and 1628 Hz. Thus formant frequency analysis easily identifies the happy emotion but it is not able to indicate the neutral and angry emotion [7].

The MFCC approach does not require any calculation of other acoustic features. The Mel frequency cepstral coefficients provide a better representation of the signal. They are less susceptible to noise and provides better recognition. MFCC on Berlin Database which consists of 7 emotional states that is anger, boredom, fear, happiness, sadness and neutral was used for feature extraction. It was observed that 16 mel cepstrum coefficients was extracted from data and each data size was reduced by certain amount and window size equals to 128 [11].

4. Conclusion

It is not possible to build a perfect recognition system and compare the performance of the machine in relation to human performance. In this paper, we studied few of the

feature extraction algorithm for speech emotion recognition. The extraction of the best parametric representation of acoustic signals is an important task since it affects the recognition performance. After studying the methods we have found that MFCC produces the most efficient output.

References

[1] Santosh K.Gaikwad, Bharti W. Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications, November 2010, Volume 10- No.3.

[2] Om Prakash Prabhakar, Navneet Kumar Sahu, "A Survey On: Voice Command Recognition Technique", International Journal of Advanced Research in Computer Science and Software Engineering, May 2013, ISSN: 2277 128X, Volume 3, Issue 5.

[3] Miss Himanshu, Sarbjit Kaur, Vikas Chaudhary, "Literature Survey on Automatic Speech Recognition System", International Journal of Emerging Technology and Advanced Engineering, July 2014, ISSN: 2277 128X, Volume 4, Issue 7.

[4] Farah Chenchah, Zied Lachiri, "Speech Emotion Recognition in acted and spontaneous context", 6th International conference on Intelligent Human Computer Interaction, IHCI 2014, Procedia Computer Science 39(2014) 139-145.

[5] V. V. Nanavare, S.K. Jagtap, "Recognition of Human Emotion from Speech Processing", ICAC3'15, Procedia Computer Science 49(2015) 24-32.

[6] Rahul B. Lanjewar, Swarup Mathurkar, Nilesh Patel, "Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) technique", ICAC3'15, Procedia Computer Science 49(2015) 50-57.

[7] Bageshree V. Sathe-Pathak, Ashish R. Panat, "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person",International Journal of Computer Science Issues, July 2012, ISSN (Online): 1694-0814, Vol. 9, Issue 4.

[8] Purnima Chandrasekar, Santosh Chapaneri, Deepak Jayaswal, "Emotion Recognition from Speech using Discriminative Features", International Journal of Computer Applications, September 2014, Volume 101– No.16.

[9] Mandar Gilke , Pramod Kachare , Rohit Kothalikar , Varun Pius Rodrigues, Madhavi Pednekar, "MFCC-based Vocal Emotion Recognition Using ANN", 2012 International Conference on Electronics Engineering and Informatics, IPCSIT vol. 49 (2012) © (2012) IACSIT Press, Singapore,DOI: 10.7763/IPCSIT.2012.V49.27.

[10] Jagvir Kaur, Abhilash Sharma, "Emotion Detection Independent of User Using MFCC Feature Extraction", International Journal of Advanced Research in Computer Science and Software Engineering, June 2014, ISSN: 2277 128X, Volume 4, Issue 6.

[11] S. Demircan, H. Kahramanlı, "Feature Extraction from Speech Data for Emotion Recognition", Journal of Advances in Computer Networks, March 2014, Vol. 2, No.1.

[12] Ankur Sapra, Nikhil Panwar, Sohan Panwar, "Emotion Recognition from Speech",International Journal of Emerging Technology and Advanced Engineering, February 2013,ISSN: 2250-2459, Volume 3, Issue 2.

[13] Varsha Singh, Vinay Kumar Jain, Dr. Neeta Tripathi, "A Comparative Study on Feature Extraction Technique for Language Identification", International Journal of Engineering Research and General Science, April-May 2014, ISSN:2091-2730, Volume 2, Issue 3.

Authors Profile

Sweta Bhadra, BTech, is currently pursuing MTech in Artificial Intelligence from Assam Don Bosco University (INDIA). She received her BTech degree in Computer Science from Assam Don Bosco University in 2015.

Dr Uzzal Sharma,is currently working as an Assistant Professor - Stage II, in the Dept. of CSE & IT in Assam Don Bosco University (INDIA). His area of interest include Speech Signal Processing, Computer Architecture, Computer Networks and holds many publications for the same. He is also a member of International Association of Computer Science and Information Technology and International Association of Engineers (IACSIT).

Alok Choudhury,is currently working as an Assistant Professor in the Dept. of CSE & IT in Assam Don Bosco University (INDIA). His area of interest include Data Mining, Computer Networks, Cryptography and Network Security and holds many publications for the same. He is also a member of International Association of Engineers (IAENG).