International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2015): 3.791

Speech Emotion Recognition using Artificial Neural Networks

P. Prithvi[1], Dr. T. Kishore Kumar[2]

[1, 2]National Institute of Technology, Warangal, Telangana – 506 004, India

Abstract: ***Emotion recognition in speech is a topic in which little research has been done to date. Emotion recognition in speech is an interesting and applicable research topic. In this paper, we present a system for emotion recognition using neural networks. By using a database of words, our system will be speaker independent. The classifiers will used to distinguish emotions such as neutral, anger, happy, sorrow etc. Emotional speech samples will used as database for emotion recognition from speech and extracted features from speech samples are prosodic features like pitch, energy, formats and spectral features like mel frequency cepstral coefficients for speech are used. Further the classifiers will be trained by using these features for classifying emotions accurately. Thus, many components like pre-processing of speech, MFCC computations, classifiers come together in the implementation of emotion recognition system using speech.***

Keywords: MFCC, Prosodic Features, Emotions, Neural Networks

1. Introduction

The interaction between humans and computers has received a lot of attention off late. It is one of the most popular areas of research with great potential. Teaching a computer the understanding of human emotions is an important aspect of this interaction. A lot of successful applications related to voice recognition are available in the market. People can use their voice to give commands to car, cell-phones, computer, television and many electrical devices. Thus, to make a computer understand human emotions and give a better interaction experience becomes a very interesting challenge. [1]

The most common way to recognize any speech emotion is extracting important features that are related to various emotional states from the voice signal (i.e. energy is an important feature to distinguish happiness from sadness), feed these features to the input end of a classifier and obtain different emotions at the output end. This process is shown in the figure below.

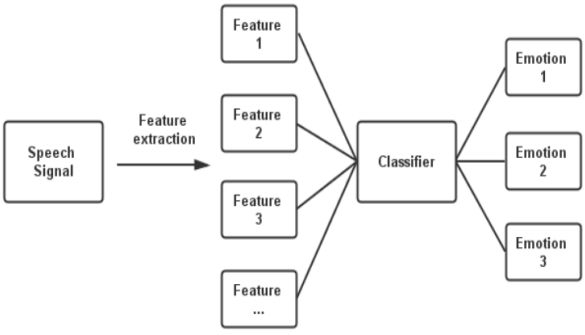

Figure 1: Basic flow of emotion recognition using speech

2. Emotion Recognition using Speech

The aim is to classify a batch of recorded speech signal into four categories, namely: happy, sad, angry, nature, using MATLAB. Before extraction, these speech signals need pre-processing. Samples are taken from the speech to convert analog signal to digital signal. Then the normalization makes sure each sentence is in the same volume range. At last, segmentation separates signal in frames so that speech signal can maintain its characteristics in short duration. Four commonly used features are chosen to study and extracted using MATLAB. Energy is the most basic feature of speech signal but it still has significant difference between emotions such as angry and sad. Pitch is frequently used in this topic and autocorrelation method is used to detect the pitch in each frame. After that statistical values such as mean, variance, max value of pitch are calculated for speech signals. Formant is another important feature. Linear predictive coding (LPC) method is used to extract the first formant. Similar to pitch, statistical values are calculated for first formant. Mel frequency cepstral coefficient (MFCC) is a representation of short-term power spectrum on a human like mel scale of frequency. First three coefficients of MFCCs are taken to derive means and variances. All 15 features of words are put into an input matrix along with a target matrix, which indicate the emotion state for each sentence composed the input of neural network. MATLAB neural network pattern recognition tool is used to train and test the data and perform the classification, in the end figures of mean square error and confusion will be given to show how good the performance is. [3]

3. Pre-processing

Prior to feature extraction, some necessary steps are taken to manipulate speech signal. Pre-processing mainly includes sampling, normalization and segmentation. [2]

Figure 2: Speech pre-processing

According to sampling theorem, when the sampling frequency is larger or equal to 2 times of the maximum analog signal frequency, the discrete time signal is able to reconstruct the original analog signal. As indicated in figure, sampling is performed by collecting points from analog signal at a certain rate T_s. Generally the sampling frequencies for speech signals are 8000Hz, 16000Hz and 44100Hz. In

MATLAB sampling is applied automatically after the recording function.

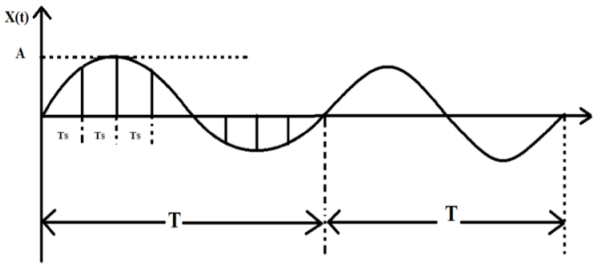

Figure 3: Sampling Process

Normalization process uses the signal sequence divided by maximum value of the signal to ensure each sentence has a comparable volume level.

Segmentation process divides the signal sequence into multiple frames with overlap as shown in figure. Overlapping is utilised to avoid loss of data due to aliasing. The signal s(n) becomes si(n) once framed, where i indicates the number of frames. After pre-processing, characteristics of the whole speech signal can be studied from statistical values.

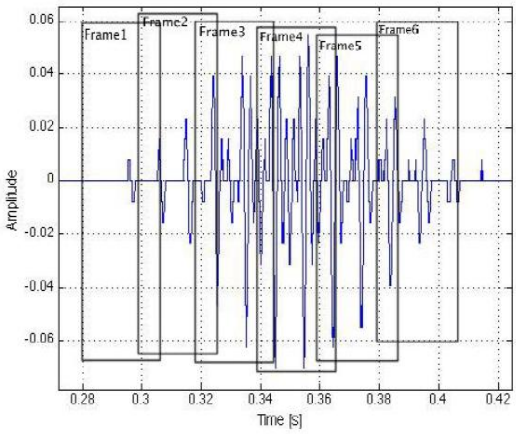

Figure 4: Segmentation Process

4. Features

4.1 Energy

Energy is the most basic feature in speech signal processing and it plays an important role in emotion recognition, e.g. speech signals corresponding to happiness and anger have much higher energy than those of sadness. The record function used in MATLAB has a threshold level for minimum voice, which highly reduces the effect of noise.

4.2 Pitch

Pitch known as the perceived rising and falling of voice tone, is the perceptual form of fundamental frequency. It represents the vibration frequency of vocal folds during speaking. It's called fundamental frequency because it sets the periodic baseline for all higher- frequency harmonics contributed by the pharyngeal and oral resonance cavities above. It is the source of speech model. Besides, it is the most frequently used feature in speech emotion recognition. There are many ways to estimate pitch from a speech signal. Auto-correlation method is used for it is a commonly used method and is easy to practice. This method uses short-term analysis technique to maintain characteristic for each frame, which means pre-processing should be fully applied before pitch extraction. Since autocorrelation can decide the period of a periodic signal, for each frame autocorrelation is applied.

4.3 Formant

Formant frequencies are defined as resonances in vocal tract and they determine characteristic timbre of vowel. It is also a very useful feature for speech recognition and could be found in many speech emotion studies. The peaks of the frequency response from a linear prediction filter are the formants. This illustration also provides a way to obtain the formants. That is to compute the roots of a linear prediction coding (LPC) polynomial. This should be done at frame level as well. The linear prediction coding as its name indicate that it predicts current sample as a linear combination of past samples.

4.4 MFCC

The Mel-Frequency Cepstrum Coefficients (MFCC) is an accurate representation of short time power spectrum of a sound. The advantage of MFCC is that it imitates the reaction of human ear to sounds using a mel scale instead of linearly spaced frequency bands.

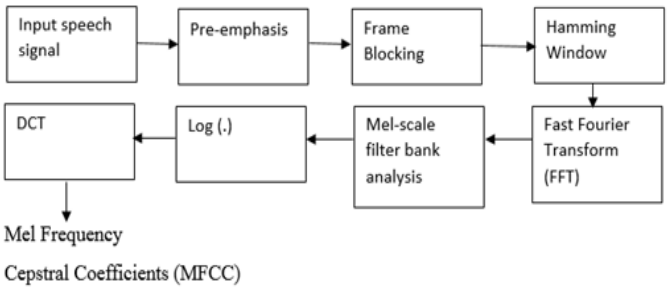

Figure 5: MFCC block diagram

5. Results and Conclusions

Energy of the speech signal is extracted after normalization. Sad emotion has the least energy while angry emotion has the highest emotion.

The pitches of speech signal vary with time. Mean value and variances are calculated for each speech signal. All the pitches are in the same range. The formants are in the range of 350 – 1000 Hz range. Mean and variance values of formants are calculated.

Sad speech are slow, depressing and powerless. Reasonable ranges for pitch and formant have to be defined to avoid erroneous values. Accuracy is of utmost importance in feature extraction process. The accuracy depends on the number of times the neural network is trained. [4][5][6]

MATLAB is a good tool with a good user interface. It can be used for speech recognition and classification. Energy, pitch, formant, MFCC prove to be best representations of emotion for speech signal.

References

[1] Lawrence Rabiner, Ronald Schafer, Introduction to digital speech processing, Prentice Hall.
[2] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Upper Saddle River, NJ: Prentice-Hall, 1993.
[3] F. Dellaert, T. Polzin, & A. Waibel, Recognizing Emotion in Speech, *Proc. ICSLP*, Philadelphia, PA, USA, 1996, 1970-1973.
[4] S. Yacoub, S. Simske, X. Lin, & J. Burns, Recognition of Emotions in Interactive Voice Response Systems, *Proc. European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, 729-732.
[5] J. Liscombe, Detecting Emotion in Speech: Experiments in Three Domains. *Proc. HLT/NAACL*, New York, NY, USA, 2006, 231-234.
[6] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, & A. Stolcke, Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog, *Proc. ICSLP*, Denver, Colorado, USA, 2002, 2037-2040.

Author Profile

Smt. P. Prithvi is an Assistant Professor at the Dept. of ECE, National Institute of Technology, Warangal. Her major areas of interest include Speech Processing, Electronic Instrumentation.

Dr. T. Kishore Kumar is an Associate Professor at the Dept. of ECE, National Institute of Technology, Warangal. His areas of interest include Speech Processing, Radar Signal Processing, Real Time Embedded Systems, Advanced Digital System Design, VLSI systems, Real Time Processing using Embedded Systems.