Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 10, October 2015, pg.295 – 302

RESEARCH ARTICLE

An Emotion Based Speech Analysis

Mr. Prabal Deep Das, Prof. Mrs. Sharmila Sengupta

M.E. Student, Electronics and Telecommunication, Vivekanand Education Society Institute of Technology, Maharashtra, India

Professor, Computer Science, Vivekanand Education Society Institute of Technology, Maharashtra, India

Abstract: In a real world, when two human beings are having a conversation between them then they are able to identify the mental state of the speaker by hearing there voice (when speaking on a telephone) or both by seeing there facial expression as well as the way they speaking. Whereas, when a human being is having a conversation with a robot, then the robot is not able to understand the emotion of the speaker spoken in real time. This paper mainly focuses on how a robot can be trained by only using the speech voice spoken by a human being, so that the robot will be able understand and recognize the mental state of the speaker. Such type of work will prove to make a human robot interaction possible in real world. By having such type of system it will also provide a number of application which are possible by having human robot interactions.

Keywords: Feature Extraction system, formant, intensity, Cepstral analysis, Feature Recognition system.

1. INTRODUCTION

An Emotional based speech analysis has became a very important topic for ensuring a reliable interaction between a human and a robot. For last two decades many researchers have worked for implementing such type of a system in which the different feature of a speech voice will be extracted and are used to train a robot, so that when a person utters a sentence in some emotion then the robot should be able to analyze and recognize the emotional or mental state of the same person. This type of application will give rise to implementation of a dynamic programming in which the robot will work according to the emotional state of the speaker.

Many researches have made to examine the total number of emotions, later it was concluded that there are total 300 emotions among which the there are eight basic emotions such as neutral, joy, disgust, fear, sadness, stressed and surprise. Rests of the emotions are generated from the combinations of these primary emotions. These emotions are clearly reflected in a human speech. For example if a person is angry then he will utter the words at a faster rate with varying intensity, etc.

A speech analysis is mainly considered as a system used to study the different characteristics of a sound signal which have been recorded. The system mainly consists of two subsystems which would help in analyzing the speech emotions in a more reliable manner.

The two subsystems are:

1. Feature Extraction Subsystem: This subsystem mainly emphasizes on the method which can be used to extract or take the various features from a speech signal which can be used to analyze the emotional content of a speech signal. The features are broadly classified as:
   a) Prosodic Features: These features are the time domain features which are needed to be extracted for distinguishing different emotions. The prosodic features which are usually extracted are Pitch, Intensity, Formants, and Speech Rate etc.
   b) Spectral Features: These features are the time domain features which are needed to be extracted from a speech signal. The spectral features are mainly associated with Evaluation of cepstral coefficients and from these cepstral coefficients various other features can also be extracted are skewness, kurtosis, jitter, and shimmer etc.

2. Emotion Recognition Subsystem: It the final part of the speech analysis system which gives an opportunity to recognize the emotion from the recorded speech signal by using the feature comparison of the present speech signal with the features which has already been extracted from a set of speech signal. These are implemented by using the following steps: [11]

i. Extracting the features from the speech signal in real time.
ii. Summarization of the features into a set of features or the features is being selected from a group of features which has been extracted from the speech signal.
iii. These reduced feature sets are used to supervise the system which will help in to classify the emotions.

This paper is organized as follows: We first summarized the features which are going to be extracted from a segment of a speech signal. Secondly, we summarized the various types of emotion recognition methods which are traditionally used to classify the different emotions. Third, we have implemented the MATLAB programs for extracting the features and the observations from the programs output, followed by the results. This paper has also highlighted the various applications of implementing such type of system.

2. FEATURE EXTRACTION SYSTEM

The feature extraction process is the most important part of an emotion based speech analysis system. The speech signals which are required for extracting the features so that they can be used for training purpose should be acquired from a well defined speech corpus or a database. These databases are a collection of .wav files which have been recorded in a closed vacuum chamber in order to avoid the occurrence of noise. The database can be made by having a collection of speech recorded in a real life situation or it may be a collection of speech recorded by the actors and actresses belonging to different age groups. There are various commercially used databases which are: [17]

1. Berlin Emotional Speech Database.
2. Danish Emotional Speech Database.
3. The Speech Under Simulated and Actual Stress (SUSAS) Database.
4. Texas Instrument and Massachusetts Institute of Technology (TIMIT) Database.

Among these the Berlin Emotional Database is mostly used because of its easy availability. It consists of a set of speech waveforms file which has been made by the 10 professional actors (5 male, 5 Females) uttering 10sentences in seven different emotions (Anger, Neutral, Happiness, Disgust, Boredom, Sadness, and Fear).

The basic features which are required to determine from a speech signal which has been taken from the above mentioned databases are Pitch, Formants, Speech Rate, and Cepstral Coefficients.

The features can be extracted as:

1. Pitch Analysis: Pitch corresponds to the fundamental frequency of the voiced speech. They are analyzed by using various methods such as Autocorrelation method, Cepstral analysis, SIFT (Simple Inverse Filtered Tracking), etc.[4] From these methods the various statistical values can be acquired are Pitch mean, variance, standard deviation, pitch peak, pitch range. Pitch peak in case of neutral is having the least value [10]. The pitch contours for different emotional speech are different [14]. In [14] a complete comparison has been done between various emotional speeches on the basis of pitch analysis.
2. Formant Analysis: Formant is defined as the fundamental frequency of the vocal tract. They can be determined by using LPC (Linear Prediction Coefficients) or by using cepstral analysis. The various values which are required to be extracted are formant contour and power. The main criteria of getting a formant frequency is :

$$r_k = \sqrt{(\mathrm{Re}(c_k)\,\hat{}\,2 + \mathrm{Im}(c_k)\,\hat{}\,2)}$$

Where *ck* is the roots of the equations derived by using LPC. [10]

3. Intensity Analysis: Intensity gives the information about the energy content within every frame. It gives the information about the amplitude variations of a speech signal. The various values which are needed to be extracted are Mean, Variance, and Standard deviation of the intensity. It can be determined by using the following flowchart [2]:

Fig 1: Block Diagram of energy estimation of a Speech signal [2]

It can be used to distinguish between the anger speeches from the rest of the emotions. [12]

4. Speech Rate Analysis: Speech rate is defined as the number of words uttered per unit time. The speech duration can also be extracted as given in [10]. Speech rate in case of fear has been found to be lower as compared to the rest of the emotional speech as given in [17].
5. Cepstral analysis: The cepstral analysis can be determined by using three methods:

i. *Linear Prediction Cepstral Coefficients* (LPCC): It gives information about the channel characteristics of a person utterance. [6]

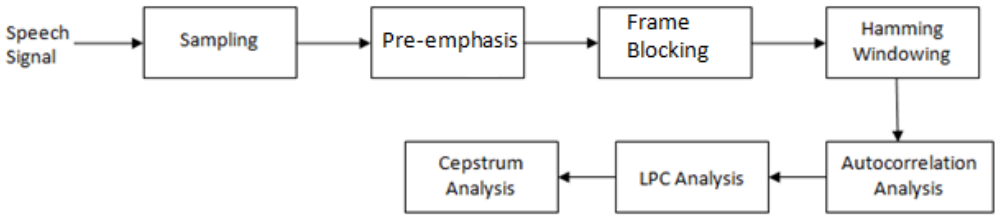

Fig 2: Block Diagram of Linear Prediction Cepstral Coefficients.

ii. *Mel Frequency Cepstral Coefficients* (MFCC): It is the most efficiently used method to determine the cepstral coefficients because of its several advantages such as robust to noise, better frequency resolution in case of low frequency as compared to the high frequency region.

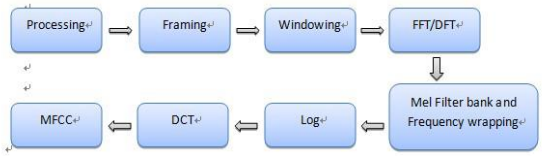

Fig 3: Block Diagram of MFCC [6]

iii. *Mel Energy Dynamic Coefficients* (MEDC): The procedure of extracting the cepstral coefficients by using MEDC is similar to MFCC but the only difference between the two is, in case of MEDC the logarithm of the frequency warped values are estimated.

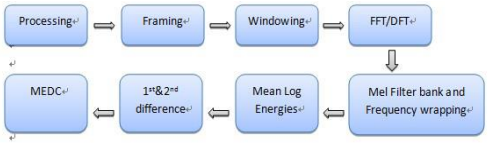

Fig 4: Block Diagram of MEDC [6]

The values which can be determined by using these methods are skewness, Kurtosis, Jitter and shimmer. The kurtosis (K) and Skewness (S) is mathematically expressed as:

$$S = \frac{\left(E(x-\mu)^3\right)}{\sigma^3} \qquad K = \frac{\left(E(x-\mu)^4\right)}{\sigma^4}$$

By using the cepstral coefficients, it is possible to distinguish between the different emotions. [9]

3. FEATURE RECOGNITION SYSTEM

An automatic emotion recognition system is an important part which is being used by the robot to recognize the expressions from the features which has been extracted from the speech signal recorded in real time.

There are around seventy (70) different features which can be extracted, but among them only the important features are to be selected. Thus, it requires a feature selection system.

After selecting the feature set, these feature set are applied to the decision making system which is known as Classifier. There are various types classifiers used are Hidden Markov Model, Bayes Classifiers, Support Vector Machine, Gaussian Mixture Model, K-Nearest Neighbor, Ada-boost classifiers etc.[11]

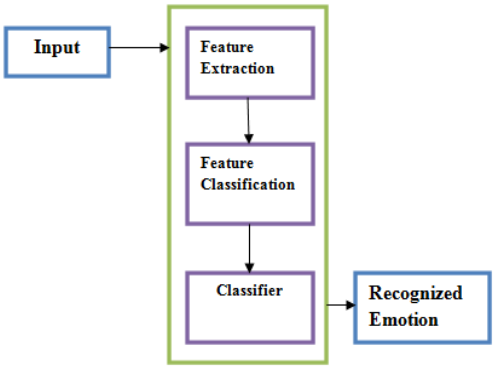

Fig5: Automatic Emotion Recognition System

i. *Hidden Markov Model*: It is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved states.

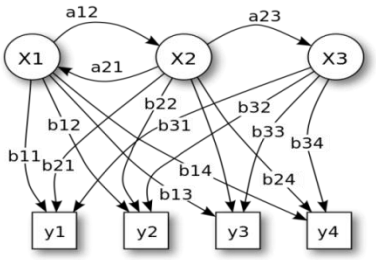

Fig6: Hidden Markov Model

Where X1, X2, X3 are the transition states which remains hidden from the observer, a12, a21, a23 are the transition probabilities, y1, y2, y3, y4 are the outcomes and b11, b12, b13…. are emission probability values.

ii. *Neural Networks*: The neural networks which are man made are known as Artificial Neural Network. These have emerged as an important and attractive acoustic modeling approach. It is implemented by having a number of inputs which are applied to each of the hidden states in which the neurons processes the input signal and after completion of its whole iteration the hidden state activates the output neurons.

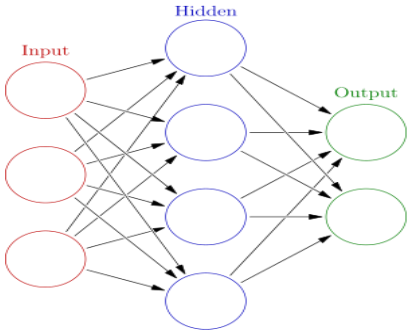

Fig. 7: Artificial Neural Network

iii. *K-NN Classifiers*: This type classifier is a non parametric method used for classification. It is mainly dependent on the minimum distance algorithm. An object is classified by a majority vote of its neighbors. The vote is related to the distance between the feature and the neighbors which have been collectively plotted in the Euclidian space. In [15] it has been mentioned that the recognition rate has been found to be as 50%. In [13] it was found that by having a proper feature selection the recognition rate can reach up to 74.39%.

iv. *Gaussian Mixture model*: Gaussian Mixture Model is a probabilistic model which can be realized as a mixture of Gaussian distribution function with unknown parameters [11]. In mainly works on the principle of expectation-maximization algorithm or Maximum A Posteriori (MAP) Parameter Estimation.

v. *Support Vector Machine*: SVM is a type of classifier which works on the principle of boundary making algorithm. In this case the feature values are plotted in the Euclidean space and then a boundary is being made which is known as hyper planes. The Support vector Machine can be considered as non-probabilistic Binary Linear Classifier and if a non linear quantity is required to be classified it can be performed by using Kernel Tricks.

vi. *Adaboost Classifier*: Adaboost classifier is short form of 'Adaptive Boosting'. In this type of classifiers initially all the data which are applied as the inputs are weighted with same value but after every iteration these weights are changed according to the error values produced at the output.

4. EXPERIMENTS, RESULTS AND APPLICATIONS

In this paper a small MATLAB program has been implemented for extracting the features such as:

i. *Pitch Peak* (By using Cepstral method)
ii. Intensity Mean
iii. *Cepstral Coefficients* (using LPCC)
iv. *Formant Power* (using LPC method)

The whole signal was framed into small sections such that each frame is of 15ms duration with 7.5ms overlapping.

Emotions	Formant Power	Intensity	Pitch Peak (Hz)
Anger	0.4803	0.3505	402.8
Neutral	0.3029	0.2250	147.8
Happiness	0.6825	0.3456	397.6
Boredom	0.3050	0.2375	221.8
Disgust	0.1547	0.2458	203.2
Sadness	0.2096	0.3035	272.3
Fear	0.1291	0.3298	338.4

Table 1: Formant power Intensity and Pitch peak

Emotions	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Anger	0.244	-0.76	0.393	-0.008	0.093	0.096
Neutral	0.004	-0.09	0.207	-0.280	0.036	-0.279
Joy	0.097	-0.78	0.307	-.0005	0.190	-0.022
Joy	0.045	-0.22	0.288	-0.390	-0.027	-0.168
Boredom	0.014	-0.06	0.106	-0.214	0.136	0.0098
Disgust	0.082	0.01	0.395	-0.359	0.019	-0.105
Fear	0.019	0.12	0.200	-0.220	0.050	-1.016

Table 2: Mean Cepstral coefficients using LPCC

Table 1&2 shows the results came out from the MATLAB programs.

The emotion based speech analysis system can be used in many applications. Some of them are mentioned below: [5]

i. *Tutoring*: By using such type of system, robots can be deployed as a tutor in various institutions or any industrial firms as a professional trainer.
ii. *Automatic Connectors*: The robots which are equipped with emotion recognition system can be used as receptionists, personal assistant, etc. They are able to automatically hand over the call to a more equipped person or robot.
iii. *Personal Use*: They can be used as a personal friend so that the robot can have an efficient interaction with the human beings.
iv. *Alerting*: They can be used in alerting the systems when an emotion of panic or stress is recognized in any emergency condition.

5. CONCLUSIONS

From the above experiments following can be inferred:

i. *Happiness*: Highest formant power, high Intensity (less than Anger), high pitch peak (Less than angry).
ii. *Disgust*: Low formant power (Less than sadness), low intensity (less than fear), Low pitch peak (less than Boredom).
iii. *Fear*: Least formant power, high intensity (less than happiness), High pitch peak (Less than happiness).
iv. *Boredom*: High formant power(less than anger), low intensity(less than disgust), Low pitch peak (less than sadness).
v. *Anger*: High formant power(less than happiness), highest intensity, highest pitch peak.
vi. *Sadness*: Low formant power (Less than Neutral), low intensity(less than fear), Low pitch peak (less than Fear).
vii. *Neutral*: Low formant power (Less than boredom), least intensity, least pitch peak.

The above conclusions have been taken by considering only the prosodic features whereas, by using only spectral features the emotions can be recognized significantly because of the occurrence of different values of the cepstral coefficients.

6. REFERENCES

1. http://www-mobile.ecs.soton.ac.uk/speech_codecs/speech_properties.html
2. http://spl.telhai.ac.il/speech/project_summary/project_book/Speech_Morphing_2_pbook.doc
3. http://www.sfu.ca/sonic-studio/handbook/Formant.html
4. Rabiner, Lawrence, et al. "A comparative performance study of several pitch detection algorithms." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 24.5 (1976): 399-418.
5. Cowie, Roddy, et al. "Emotion recognition in human-computer interaction." Signal Processing Magazine, IEEE 18.1 (2001): 32-80.
6. Pan, Yixiong, Peipei Shen, and Liping Shen. "Feature Extraction and selection in speech emotion recognition." IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2005), Como, Italy. 2005.
7. Hyun, Kyung Hak, Eun Ho Kim, and Yoon Keun Kwak. "Emotional feature extraction based on phoneme information for speech emotion recognition." Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on. IEEE, 2007.
8. Grimm, Michael, et al. "Primitives-based evaluation and estimation of emotions in speech." Speech Communication 49.10 (2007): 787-800.
9. Přibil, J., and A. Přibilová. "Statistical analysis of spectral properties and prosodic parameters of emotional speech." Measurement Science Review 9.4 (2009): 95-104.
10. Khulage, A. A. "Extraction of pitch, duration and formant frequencies for emotion recognition system." (2012): 7-9.
11. Joshi, Aastha, and Rajneet Kaur. "A Study of speech emotion recognition methods." Int. J. Comput. Sci. Mob. Comput.(IJCSMC) 2.4 (2013): 28-31.
12. Mohamed, Masnani, Chee Chuan Lee, and Ida Laila Ahmad. "Feature extraction of speech signal and heartbeat detection in angry emotion identification." International Journal of Computer Science and Electronics Engineering (IJCSEE) 1.1 (2013).
13. Fulmare, Nilima Salankar, Prasun Chakrabarti, and Divakar Yadav. "Understanding and estimation of emotional expression using acoustic analysis of natural speech." International Journal on Natural Language Computing (IJNLC) 2.4 (2013).
14. Rabiei, Mohammad, and Alessandro Gasparetto. "A system for feature classification of emotions based on Speech Analysis; Applications to Human-Robot Interaction." Robotics and Mechatronics (ICRoM), 2014 Second RSI/ISM International Conference on. IEEE, 2014.
15. Demircan, S., and H. Kahramanlı. "Feature Extraction from Speech Data for Emotion Recognition." Journal of Advances in Computer Networks 2.1 (2014).
16. Devi, J. Sirisha, Y. Srinivas, and Siva Prasad Nandyala. "Automatic Speech Emotion and Speaker Recognition based on Hybrid GMM and FFBNN." International Journal on Computational Sciences & Applications (IJCSA) 4.1 (2014): 35-42.
17. Sudhakar, Rode Snehal, and Manjare Chandraprabha Anil. "Analysis of Speech Features for Emotion Detection: A Review." Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on. IEEE, 2015.

