Characterization between Child and Adult voice using Machine Learning Algorithm

Gaurav Aggarwal
Department of Computer Science and Engineering,
ITM University, Gurgaon, India
mtech.gaurav@gmail.com

Dr. Latika Singh
Department of Computer Science and Engineering,
ITM University, Gurgaon, India
latikasingh@itmindia.edu

***Abstract* – Speech Feature Detection is a technique employed in speech processing in which different features of speech are used to distinguish between speech in different age groups. This paper implements a new approach for the extraction and classification of the speech features using the Mel-frequency cepstral coefficient, and Support Vector Machine. This paper presents the Mel-frequency cepstral coefficients (MFCC) for extracting the speech features of child and adult voices. Using the support vector machine, we classify the datasets in a child and an adult's speech.**

Keywords- Mel-frequency cepstral coefficient (MFCC); Support Vector Machine (SVM); speech feature extraction.

I. INTRODUCTION

Speech is the ability to express the thoughts and the feelings to others. It is a process through which humans are able to interpret and understand the speech sound of others. In this paper speech can be distinguished in two ways: 'Normal speech' and 'impaired speech'. When a person clearly communicates with another person, it is called normal speech. People who have normal speech can easily understand the views of others and also easily represent their thoughts to others.

Speech is the most important part of human life. Spoken language involves use of words; variations in pitch, loudness, tempo and rhythm in order to convey different meanings [1]. The word which we speak is created by the combination of vowels and consonants [2]. It can also be a very efficient method of communication to solve interactive problem.

MFCC are the popular coefficient for speech recognition [3]. It is used for the Speech Detection. MFCC has the ability to represent the speech amplitude spectrum in a compact form and has less complexity in implementation of feature extraction algorithm [4]. MFCC provides the information of varying speech coefficient at multiple band level [5].

Support Vector Machine is a supervised learning model with associated learning algorithms that can analyse data and recognize pattern. Support Vector Machine used a hyperplane [6] in a high dimension space that can be used for classification and regression.

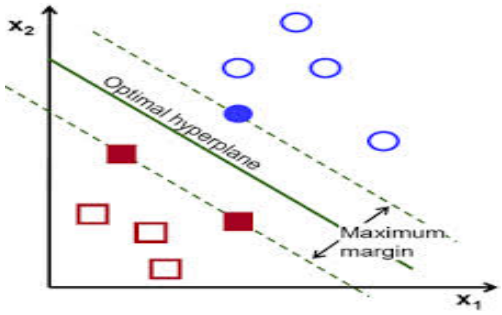

Fig.1. Maximum margin between training sets

Principally, SVM takes a set of input data and predicts for each given input which of two classes forms the possible output. It is used for dividing the categories [7] of two classes with clear and possible gap. There are lot of other applications of SVM- like it improves the speech emotion feature performance. Using SVM, classifications of images can easily take place.

SVM are used for performing a non- linear classification in efficient manner. The points are represented is space and these points are separately divided as well as maximum gap are possible.[8]

II. EXTRACTION OF SPEECH FEATURE

For speech features extraction, we use the MFCC (Mel-frequency cepstral coefficient).

Mel-frequency cepstral coefficient

MFCC is a way to represent short term power spectrum of speech based on linear cosine transform of a log power spectrum of speech signal on a non-linear Mel scale of frequency [9][10]. The first stage of speech processing is feature extraction. MFCC is used for feature extraction. It gives the information about extracted voiced sound rather than unvoiced sound [11]. MFCC is used to extract parameters of the sound wave. It is derived from the cepstral and represents the information of the audio clip. This frequency allow for the better representation of sound i.e. audio compression. In hyperplanes maximum distance are achieved in the neighbouring data points of any class. Larger the margins lower the generation of error takes place.

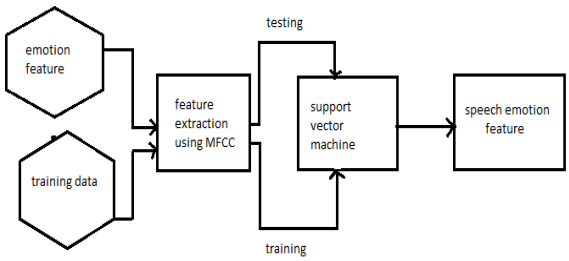

Fig.2. Block diagram for proposed speech feature

The speech feature and the training data set are extracted using the MFCC are shown in fig.2.

Basically, MFCC are used as a feature in speech recognition system, for example automatically recognize numbers spoken the numbers into a phone. In this research paper through MFCC we can extract the speech features of the different ages of people. When a child speaks something through the help of microphones, MFCC recognizes the child through his voice. Mainly MFCC are used for the speech recognition. MFCC are not very robust in the presence of additive noise and normalise their values in speech recognition system.

Fourier Transform

Fourier transform is employed to transform the signals between time domain and frequency domain. The Fourier transform is an extension of Fourier series that results when the period of the represented function is allowed to approach the infinity [12]. The voice is converted in the form of signal using the Matlab code. The formula for Fourier transform is:

$$X = \cos(2 * fs * t * pi) e^{-\pi t^2} \quad (1)$$

Where fs are the frequency of the signal, t the time interval and pi is the constant value.

Using this formula the transformation of signal is done for child voice signal and adult voice signal. The output of Fourier transform of the child voice signal is:

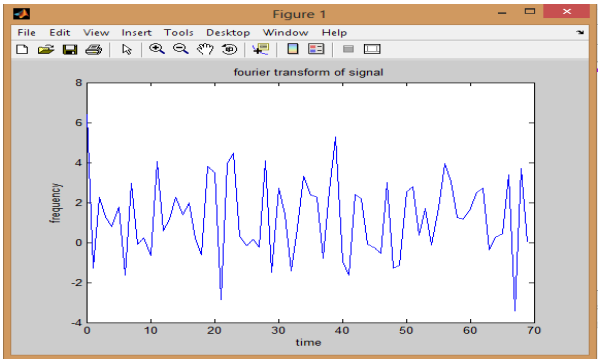

Fig.3. Fourier transform of the child voice signal

The output of Fourier transform for adult voice signal is:

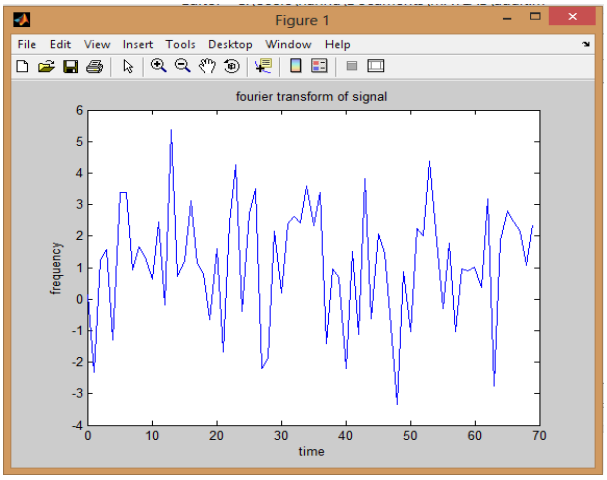

Fig.4. Fourier transform of the adult voice signal

Mel-scale

Mel-scale is a perceptual scale of pitches judged by listeners that have an equal distance from each other [13]. It consists of triangular filters triangular filters that are used to calculate the weighted sum of filter spectral components.

Map the powers of the spectrum obtained above onto recognize the children through their voice. Mainly MFCC Mel-scale, using triangular overlaps window. Take the logs of the powers at each of the Mel-frequencies.

$$mf = 2595 * log_{10}\left(1 + \frac{f_s}{700}\right) \quad (2)$$

fs is the frequency of the sample, m is the Mel-scale of the given frequency. According to this, the Mel-scale of given frequencies are:

TABLE I. MEL-SACLE VALUES FOR CHILD VOICE

Fs	Mf
0.04	0.06439782753
0.05	0.08049670947
0.07	0.11269378345
0.08	0.12879197551
0.16	0.25756923447
0.25	0.4024260641
0.42	0.67599373047
0.45	0.72426348487
0.7	1.12643108883
0.72	1.15859826583
0.76	1.22292992555

TABLE II. MEL-SACLE VALUES FOR ADULT VOICE

Fs	Mf
0.3	0.48289403604
0.7	1.12643105883
0.34	0.54726427762
0.56	0.90123489851
0.65	1.04600902429
0.7	1.12643105883
0.73	1.17468752503
0.8	1.28725791327
0.9	1.4480618203
0.92	1.4802198486
0.94	1.51237695931

Discrete Cosine Transform (DCT)

The last step of MFCC is DCT. DCT-II is used in signal processing mainly for lossy data compression [14]. DCT-II used has the property of energy compaction.

Take the discrete cosine transform of the list of Mel log powers, as if it were a signal:

$$X_K = \sum_{n=0}^{N-1} x_n \ Cos\left[\frac{\pi}{n}\left(n+\frac{1}{2}\right)k\right]$$

Where K=0, 1------, N-1. (3)

After finding the value of DCT-II, the original signals are obtained that shown in figure. The Fig 5 shows the DCT-II signal for child voice is:

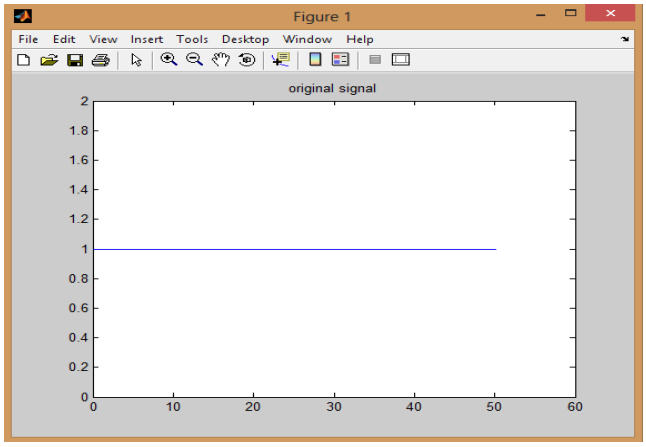

Fig.5: Original signal of DCT-II of child voice

The Fig 6 shows the DCT-II signal for adult voice is:

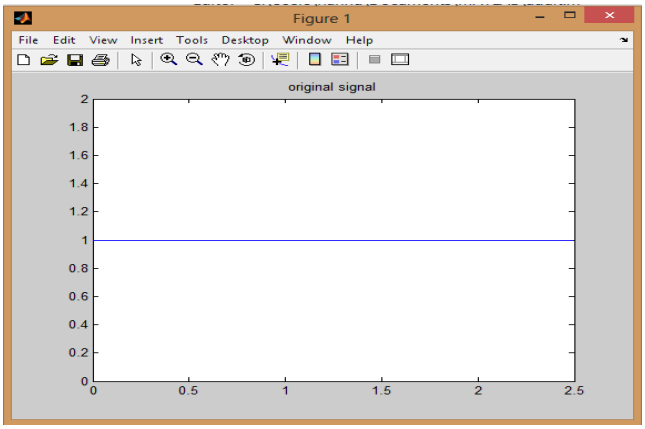

Fig.6. Original signal of DCT-II adult voice

III. CLASSITICATION OF SPEECH FEATURE

There are so many classification algorithms that we can use for the classification of the speech emotion feature. But we use Support Vector Machine for the classification of speech.

Support Vector Machine

SVM are used to classify the speech feature. Audio classification is performed using the SVM [15]. Read the file and pass the data to the SVM classification. Fig.4 shows the basic block diagram for the classification of speech [16].

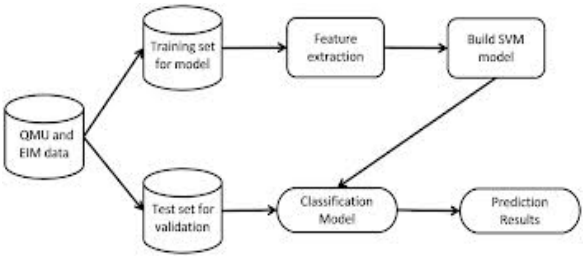

Fig.7. General block diagram of speech classification

The sample of given datasets are sorted. Each input speech has group of vector parameter [17] that describes the input speech. Afterwards, this paper can utilize above-mentioned features to provide the collected speech into classifier and commence training and launch the classification of the database. Given some training data D, a couple of n points from the form

$$D = \{(x_i, y_i) | x_i \in R^F, y_i \in \{-1, 1\}\}_{i=1}^{n} \quad (4)$$

Where the yi is either 1 or -1, indicating the class to which the xi point belongs. We want to find maximum margin hyperplane [7] as the set of point x satisfying,

$$w.x - b = 0$$

Where "." denotes the dot product and "w" the normal product of the hyperplane. This could select two hyperplanes in ways that they separate it datasets and there are no points with shod and non-shod [18] and attempt to maximize their distance as if the margin is a bit more compared to generation of errors will be the least.

$$w.x_i - b \geq 1$$ For xi of the first class

Or

$$w.x_i - b \leq -1$$ For xi of the second class

If there exists a hyperplane satisfies [19] the set is said to be linearly separable [20][21] and change w and b so that

$$y_i(w.x_i + b) \geq 1 \ , \ i = 1, 2 - - - - N.$$

This graph shows the classification of the speech feature using Support Vector Machine algorithm after doing the feature extraction using MFCC. These take the 22 samples of voice out of which 11 samples are of child voice and 11 samples are of adult voice and use it as an input for the Support Vector Machine algorithm. The resultant shown in Fig 8, the two sets of hyperplanes that classifies the child voice and adult voice, and also show the difference in child voice and adult voice in Fig.8

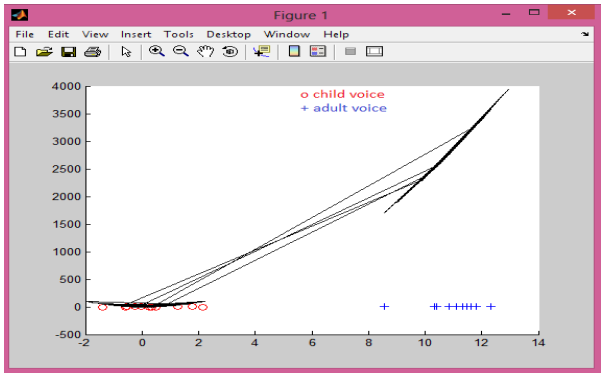

Fig.8. Showing the classification of child voice and adult voice between time(s) and frequency (Hz)

The Fig.9 shows the difference in child voice and adult voice. Take the MFCC output value as an input data for showing the differences in child voice and adult voice. Through the help of histogram, this paper shows the difference of both voices using the Matlab code.

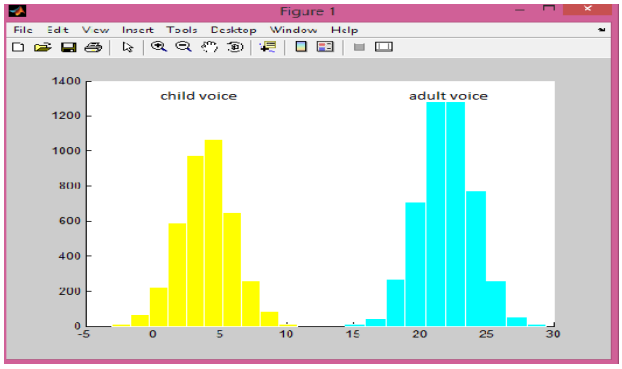

Fig.9. Showing the difference in voice of child voice and adult voice voice between time(s) and frequency (Hz)

IV. CONCLUSION

This paper implements the extraction and classifies the speech feature using different methods. For the extraction and classification of speech feature, we use MFCC and Support Vector Machine. The process of feature is done through MFCC. Feature exaction stage is most important in the entire process, since it is responsible for extracting relevant information from the feature parameters or vectors. Now, this paper classifies the speech feature of child voice and adult voice. This paper classifies the feature of child voice and adult voice and also finds out the difference in an adult's voice and a child's voice.

REFERENCES

[1] Amal alqahtani, Nouf Jaafar, Nourah Alfadda "Interactive Speech Based Games for Autistic Children with Asperser Syndrome" Information Technology Department King Saud University, Riyadh, KSA
[2] http://www.neurology.org
[3] Yi-lin lin, Gang wei "Speech emotion recognition based on hmm and svm" Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005
[4] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk "Speech Recognition using MFCC" International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)
[5] Ibrahim Patel1 Dr. Y. Srinivas Rao2 "Speech recognition using hmm with mfcc- an analysis using frequency specral decomposion technique Signal & Image Processing :" An International Journal(SIPIJ) Vol.1, No.2, December 2010
[6] http://www.biomedcentral.com/
[7] Olivier Chapelle, Patrick Haffner, Vladimir N. Vapnik "Support Vector Machines for Histogram-Based Image Classification" IEEE transactions on neural networks, vol. 10, no. 5, September 1999
[8] http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
[9] http://en.wikipedia.org/wiki/Mel-frequency_cepstrum
[10] Dalmiya C.P, Dr. Dharun, V.S, Rajesh K.P "An Efficient Method for Tamil Speech Recognition using MFCC and DTW for Mobile Applications" Proceedings IEEE Conference on Information and Communication Technologies, ICT 2013
[11] http://en.wikipedia.org/wiki/Mel-frequency_cepstrum
[12] http://en.wikipedia.org/wiki/Fourier_transform
[13] http://en.wikipedia.org/wiki/Mel_scale
[14] http://en.wikipedia.org/wiki/Discrete_cosine_transform

[15] Ruijie Zhang, Bicheng Li, Tianqiang Peng Zhengzhou "Audio Classification Based on SVM-UBM" Information Science and Technology Institute, Zhengzhou, China
[16] http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf
[17] http://www.neurology.org/content/79/4/358/F1.expansion.html
[18] Vipul Garg, Harsh Kumar , Rohit Sinha "Speech Based Emotion Recognition Based on Hierarchical Decision Tree with SVM, BLG and SVR Classifiers" Indian Institute of Technology Guwahati
[19] Meng-Chi Tu, Wei-Kai Laio, Yu-Hau Chin, Chang-Hong Lin, Wei-Jun, Liao, Szu-Hsien Lin, Jia-Ching Wang "Speech Based Boredom Verification Approach for Modem Education System" 2012 international symposium information technology in medicine and education
[20] Hai-yan yang, Xin-xing jing "Performance test of parameters for speaker recognition system based on svm-vq" Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012
[21] Lingli Yu1, Binglu Wu1 and Tao Gong2,3,4 "A hierarchical support vector machine based on feature-driven method for speech emotion recognition" Artificial Immune Systems - ICARIS