KANNADA SPEAKER RECOGNITION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) AND ARTIFICIAL NEURAL NETWORK (ANN)

Kailashnath J K [1], Priyanka Burhan [2]
[1]Assistant Professor, Masters of Engineering
[1]Department of Electronics Instrumentation Engineering,
[2]Department of Biomedical Engineering,
[1]KMIT, Hyderabad, India
[2]BIGCE, Solapur, India

Abstract: Speaker Recognition is a most prominent emerging technology in today's society Speaker recognition is a process of authentication of the claimed of a person from voice characteristics. The main goal of speaker recognition is to extract, classify and identify the information about speaker identity. In this paper shows the usage of Mel Frequency Cepstral Coefficients (MFCC) and Artificial Neural Network for Speaker Recognition System of Kannada Languages. The standard multilingual database is not available experiments are carried out on our own created database of ten speakers. The code developed in MATLAB Environment.

Index Terms - MFCC, ANN, Feature Extraction, Feature Matching, Windowing

I. INTRODUCTION

Several useful biometric signals exist today including Retina, Face, Finger print and voice .Different biometrics has different strengths and weakness. A Primary Strength of applying a voice signal to biometrics is derived from fact that it is a natural modality for Communication and user are generally Comfortable with allowing entities –both human and Machine –hear and analyze their voice . Analysis and Synthesizing the speech signal is more complex due to too large information contained in the signal. Therefore the digital signal processes such as Feature Extraction and Feature Matching are introduced to represent the voice signal .There are many algorithms and techniques such as Linear Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and etc.

Initially human voice is converted into digital signal form to produce digital data representing each level of signal at every discrete time step. The digitized speech samples are then processed using Mel frequency Cepstral Coefficients (MFCC) to produce voice features. After that, the coefficient of voice features can go through ANN to select the pattern that matches the database and input frame in order to minimize the resulting error between them. This paper present the speaker recognition system for Kannada language with comparison of Different implementation of Mel Frequency Cepstral Coefficients (MFCC) during Feature Extraction and Artificial Neural Networks for feature matching for designing an accurate/Robust Speaker recognition for Languages i.e. Kannada, English ,Hindi.

II. SPEAKER RECOGNITION SYSTEM

Speaker recognition system can be divided into two categories.

2.1 Text Dependent

If the text must be the same for enrollment and verification, the system and process is said to be text dependent.

2.2 Text Independent

In this procedure text–independent technology does not compare what was said at enrollment and verification.

III. SPEAKER RECOGNITION STAGES

Analysis of the input voice is done after taking a speech sample through microphone from a user. The different operations are performed on the input Signal such as Pre- processing, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the Spoken speech samples. The Speaker recognition system consists of two important Stages. The first one is training stage, whilst, the second one is referred to as testing stage as described below [1].

3.1 Training Stage

Voice samples of individual speaker has to provide so that the reference template model can be build.

3.2 Testing Stage

Matching the voice with stored reference template to certify and recognition decision is made.

IV. FEATURE EXTRACTION BY MFCC

Mel Frequency Cepstral Coefficient is the most popular module implemented in speaker Recognition System for feature extraction .Firstly this module is used to convert the speech waveforms to some type of the parametric representation for Analysis and Processing in next phase's .MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz [2]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz [4].The Process flow block diagram of the MFCC is shown in figure 1[7].

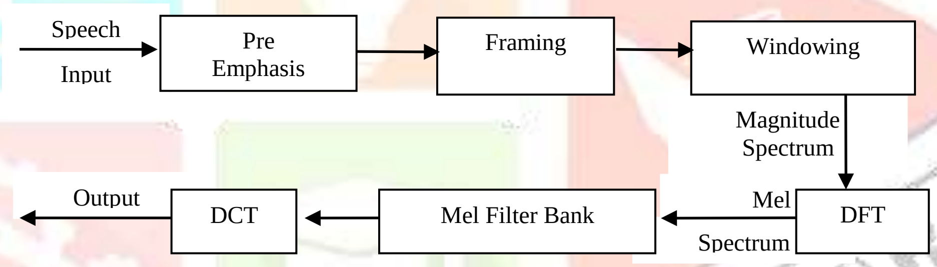

Fig. 1: Block diagram of MFCC

Mel Frequency Cepstral coefficients (MFCC) module consists several computational Phases. Each phase has its function and mathematical approaches as discussed briefly in the following:

Phase 1: Pre–emphasis

This phase processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95\ X[n-1] \tag{1}$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Phase 2: Framing

In this phase the speech samples acquired from analog to digital conversion is segmented into a small frame with the length within the range of 20 to 40 msec. The speech signals are divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Phase 3: Windowing

To avoid the discontinuities in the speech segment and distortion in the underlying spectrum windowing is performed .To prevent an abrupt change at the end points ,it gradually attenuates the amplitudes at both ends and also produces Convolution for the Fourier Transforms of the window function and the speech spectrum [2].

Phase 4: Fast Fourier Transform

The conversion each frame of N samples from time domain into frequency domain. To obtain the magnitude frequency response of each frame the FFT is performed .This statement supports the equation below:

$$Y\ (w) = FFT\ [\ h\ (t)\ {*}X\ (t)\ ] = H\ (w) * X\ (w) \qquad (2)$$

If X (w), H (W) and Y (W) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Phase 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 2 is then performed

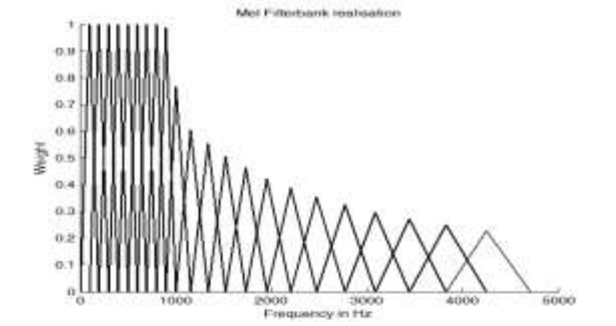

Fig. 2: Mel Filter Bank

In figure 2 shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ [11].

$$F\ (Mel) = [2595 * \log 10\ [1 + f\ \ ]/700] \qquad (3)$$

Phase 6: Discrete Cosine Transform

To get the Mel frequency Cepstral Coefficients, log Mel spectrum is to be converted into time domain using Discrete Cosine Transform (DCT). The set of coefficient is called acoustic vectors.

5.1 Effect of Number of Filters

Results of the speaker recognition performance by varying the number of filters of MFCC to 12, 22, 32, and 42 are given [8].The recognizer reaches the maximal performance at the filter *nu*mber $K = 32$. Too few or two many filters do not result in better accuracy. Hereafter, if not specifically stated, the number of filters is chosen to be $K = 32$.

V. COMPARISION OF DIFFERENT IMPLEMENTATION OF MFCC

The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by (1) Number of Filters, (2) Type of window. In this paper, several comparison experiments are done to find a best implementation.

5.1 Effect of Number of Filters

Results of the speaker recognition performance by varying the number of filters of MFCC to 12, 22, 32, and 42 are given [8].The recognizer reaches the maximal performance at the filter *nu*mber *K* = 32. Too few or two many filters do not result in better accuracy. Hereafter, if not specifically stated, the number of filters is chosen to be *K* = 32.

Table 5.1: MFCC with 12 filters

Speakers	No of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	1
S3	4	0	2
S4	4	0	0
S5	4	0	2
Total	20	0	5

Threshold Value of Distance =130, Efficiency=75%

Table 5.2: MFCC with 22 Filters

Speakers	No of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	1
S3	4	0	2
S4	4	0	0
S5	4	0	2
Total	20	0	5

Threshold Value of Distance =150, Efficiency=85%

Table 5.3: MFCC with 32 Filters

Speakers	No of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	1
S3	4	0	2
S4	4	0	0
S5	4	0	2
Total	20	0	5

Threshold Value of Distance =150, Efficiency=85%

Table 5.4: MFCC with 42 Filters

Speakers	No of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	1
S3	4	0	2
S4	4	0	0
S5	4	0	2
Total	20	0	5

Threshold Value of Distance =85, Efficiency=80%

5.2 Effect of Variation in Type of Window Using 32 Filters

Considering 32 filters as a standard number of filters we have changed the window type [8]. In this experiment we have used only the Hamming Window. Results show that efficiency is 75% while using the hamming window.

Table 5.5: Hamming Window

Speakers	No of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	2
S3	4	0	0
S4	4	0	0
S5	4	0	3
Total	20	0	5

Threshold Value of Distance =150, Efficiency=75%

VI. FEATURE MATCHING BY ANN

Artificial Neural Network is information processing devices with the capability of performing computations similar to human brain or biological neural network. Feature matching involves assigning speech signals of each speaker a different class based on its feature. Features are taken from known samples and then unknown samples are compared with those known samples. In this Paper, we have opted for Artificial Neural Networks. The main advantage of using Neural networks is that it is unaffected by the differing shape and style of testing samples as the network is already trained with large variations. The architecture of a generalized neural network back propagation is shown in figure.4 [6].

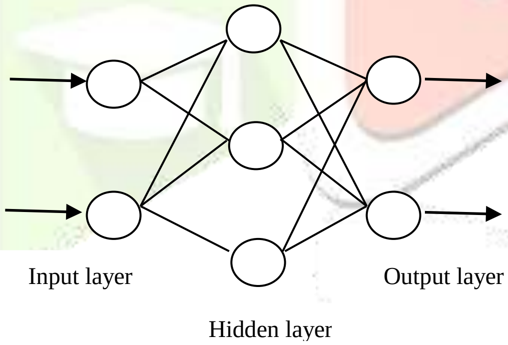

Fig.3. Architecture of Neural Network back propagation

Table 6: Features Sets of NN Training from MFCC Calculation of 5 Speakers

Speakers Values	SP1	SP2	SP3	SP4	SP5
1	2.737	2.946	3.043	2.825	3.897
2	1.937	2.131	2.420	2.350	1.949
3	3.622	3.317	3.358	3.326	3.733
4	2.350	1.949	2.420	2.131	1.937
5	4.593	3.831	4.356	3.890	4.092

6	2.272	2.576	2.228	2.722	3.622
7	2.050	2.276	1.705	2.595	2.272
8	1.957	1.960	2.218	2.360	2.218
9	5.942	4.779	7.302	5.422	7.165
10	2.213	2.312	2.266	2.577	2.376

VII. DATABASE

Data base required for this experiment is created for three Languages, i.e. Kannada, English and Hindi. As the Standard Data base is not available for multilingual. We have created our own database with 10 Speakers. Two of them are male and three of them are females with age group of 22-25 years. The voice recording was done in the engineering college laboratory. The training and testing data were recorded in different sessions with a minimum gap of two days. The approximate training and testing data length is two minutes. Recording was done using free downloadable Wave Pad sound Editor Masters Edition Version 5.68 software and Panasonic Head phone. The speech files are stored in .wav format. The detail requirements used for collecting the database are shown in Table 7.

Table 7: Requirements for Creating Data Base

Item	Descriptions
Number of Speaker	10
Sampling Rate	11Khz
Sessions	Training and Testing
Language Covered	Kannada ,English ,Hindi
Head phone	Panasonic Headphone with Noise Cancellation
Recording Software	Wave Pad sound Editor Masters Edition Version 5.68
Environment	Department Laboratory
Speakers	Male and Female
Speech Style	Random Kannada Words

VIII. RESULTS

The Performance of the Speaker recognition System for 32 Filters in MFCC for three different languages is shown in detail in table 8.

Table 8: Performance of Speaker Recognition System

Languages	Kannada	Hindi	English

Efficiency (%)	56.66%	64.69%	66.66%

IX. CONCLUSION

MFCC is well known techniques used in speaker recognition to describe the signal characteristics, relative to the speaker discriminative vocal tract properties. The goal of this project was to create a speaker recognition system and apply it to a speech of an unknown speaker of 10 speech samples recorded for each speaker, 5 samples are taken for feature extraction and training for kannada Language and Compare the efficiency with other languages .The speaker Recognition system does not yield satisfactory performance with Kannada as training and/or testing language. The presence of ottakshara, arka and anukaranavyayagalu leads to long pause and hence the less number of energy frames (features) in Kannada words.

REFERENCES

[1] Lindasalwa Muda, Mumtaj Begaum and I.Elamvazuthi “Voice Recognition Algorithms Using Mel Frequency Cepstral (MFCC) and Dynamic Time Wrapping (DTW) Technique” ,Universiti Teknologi Petronas ,Tronoh, Perak.Vol 2 ,Issue 3 ,2010,138-143

[2] Anand Vardhan Bhalla, Shailesh Kharparkar, Mudit Ratna Bhalla, “Performance Improvement of Speaker Recognition System” Gyan Ganga College of Technology, Jabalpur (M.P).India, Vol 2, Issue 3 2012.

[3] Bansood, N.S Seema Kawathekar and Dabhade S.B, “Review of Different Techniques for Speaker Recognition System”, Dept of CS & IT, Dr Babashaheb Ambedkar Marathwada University, Aurangabad, MH, India,Vol 4 ,Issue1, 2012,57-60.

[4] Jamal Price, Sophomore Student, “Design an Automatic Speech Recognition System Using Malta”, University of Maryland Eastern Shore Princess Anne, 2006.

[5] Douglas A. Reynolds, Member, IEEE, and Richard C. Rose, Member, IEEE, “Robust Text- Independent Speaker Identification Using Gaussian Mixture Speaker Models”, Transactions On Speech And Audio Processing, 1995.

[6] Hui Kong, Xuchun Li, Lei Wang, Earn Khwang Teoh, Jian-Gang Wang, Venkateswarlu.R “Generalized 2D Principal Component Analysis”,Proc. 2005 IEEE International Joint on Volume 1, Aug. 2005.

[7] Zaidi Razak, Noor Jamilah Ibrahim, Emran Mohd Tamil, Mohr Yamani Dina Iris, Mohd yaakob Yusoff, “Quranic Verse Recitation Feature Extraction Using Mel Frequency Cepstral Coefficient (MFCC)”,Universiti Malaysia, Vol 8,Issue 8,2008.

[8] Vibha Tiwari, “MFCC and Its Applications in Speaker Recognition‖” International Journal on Emerging Technologies 1(1): 2009 ,19-22

[9] Nagaraj B G ,”Kannada Language parameters for Speaker Identification with Constraints of Llimited Data,” Department of Information Science and Engineering, SIT, Tumkur.2013,9,14-20

[10] Kshamamayee Dash, Debananda Padhi, Bhoomika Panda ,Prof.Sanghamitra Mohanty “Speaker Identification using Mel frequency Cepstral Coefficient and BPNN ,Dept Of CS ,Utkal University ,Vani Vihar ,Bhuvaneshwar,Oddisha,Vol 2 ,Issue 4 2012.

[11] Anjali Bala, Abhijeet Kumar, Nidhika Birla – ―Voice Command Recgnition using system based on MFCC and DTW”, International journal of engineering science and technology vol.2 (12), 2010