A Review on Improvising Robustness of Speaker Recognition System

Kailashnath J K
Master of Technology
Bio Medical Signal Processing and Instrumentation
Sri Jayachamarejendra College of Engineering,
Mysore -570006

Rathnakara. S
Assistant Professor
Department of Instrumentation Technology
Sri Jayachamarejendra College of Engineering,
Mysore -570006

ABSTRACT

Speaker Recognition is a process by which a machine authenticates the claimed of a person from voice characteristics. A Major application includes biometric identification and security. Speaker recognition consists of the process to convert a speech waveform into features that are useful for further processing. A direct analysis and Synthesizing the complex voice signal is due to too much 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. Firstly, 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 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 with modification in the Computation Phases of Mel Frequency Cepstral Coefficients (MFCC) during Feature Extraction and Artificial Neural Networks for Feature matching for designing an accurate/Robust Speaker recognition.

General Terms

Artificial Neural Network, Mel Frequency Cepstral Coefficients, Speaker Recognition System, Feature extraction, Feature Matching.

Keywords

ANN, MFCC, Speaker recognition system, windowing.

1. INTRODUCTION

Speaker Recognition is the form of bio metric personal identification by using voice/speech characteristics of a person .Bio metric identification is generally considered to be more reliable than artifact identification because it is based on intrinsic characteristics of the individual which are difficult, if not impossible, to, mimic. While many other biometric systems like finger print recognition, retinal scans, face recognition etc .are more reliable means of identification or recognition and are used in various security and access control systems. Digital processing of speech signal and voice recognition algorithm is very important for fast and accurate speaker recognition technology. The voice is a signal of infinite information. Therefore the signal processes such as feature extraction and feature matching are introduced to represent the voice signal. Several methods such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN), Dynamic Time Wrapping (DTW) and etc are evaluated with a view to identify a straight forward and effective method for voice signal. The extraction and matching process is implemented right after the Pre Processing or filtering signal is performed. The non-parametric methods for modeling the human auditory perception system, Mel Frequency Cepstral Coefficients (MFCCs) are utilized as extraction techniques [1]. Artificial Neural Network (ANN) is used as feature matching technique and show better performance for speech and need less training data than other methods. Previously a Mel Frequency Cepstral Coefficient (MFCC) with computation phases using windowing as Hamming and Discrete Fourier Transform (DFT) utilized during extraction Technique. Modification in these computational phases to increase the robustness and accuracy of speaker recognition system can be improved [2] .This paper presents the viability of finding the Speaker recognition study by MFCC with Changes in Computational Phases and using ANN as Feature or pattern matching technique .The MFCC and ANN techniques can be implemented using MATLAB.

2. 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.

2.3 Speaker recognition Applications

There are two major application of speaker recognition

2.3.1 Verification

If the speaker claims to be certain identity and the voice is used to verify this claim, the process is called speaker verification

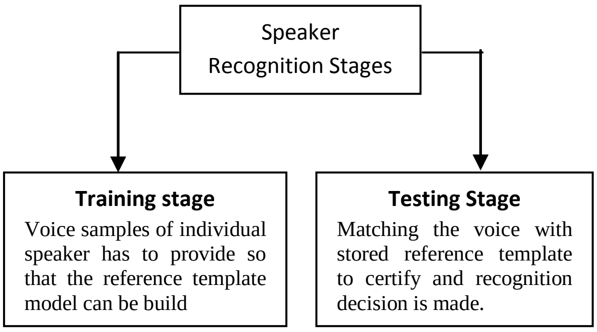
2.3.2 Identification

It is the task of determining an unknown person’s identity.

3. 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 in figures 1[1].



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

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

Fig 1: Speaker Recognition Stages

4. 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]. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The Process flow block diagram of the MFCC is shown in figure 2[10].

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 20to 40 msec.The speech signals is 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].

Hamming window is the most commonly used windowing technique in speaker recognition systems.

The Hamming window equation is given as:

$$Y(n) = X(n)\ xW(n) \tag{2}$$

$$W(n) = 0.54 - 0.46\ \cos\frac{2\pi n}{N-1} \quad 0 \le n \le N-1 \tag{3}$$

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:

$$Y(w)=FFT[h(t)*X(t)]=H(w)*X(w) \tag{4}$$

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 3 is then performed

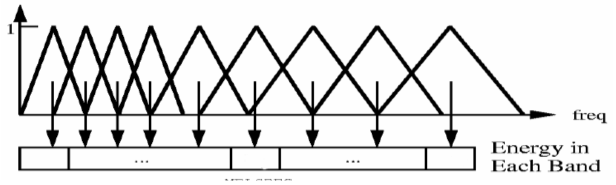


Fig 3: Mel scale filter bank

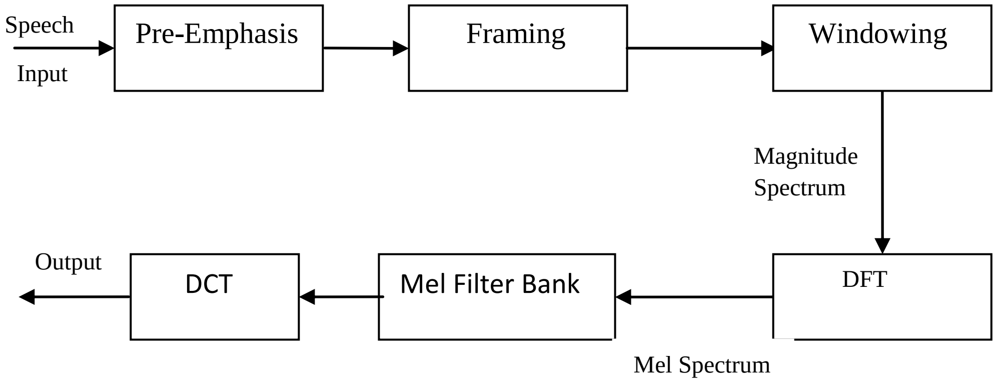


Fig 2: Block Diagram of Computational Phases of MFCC

In figure 3 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.

$$F(Mel) = [2595 * log10[1 + f\ ]/700] \qquad (5)$$

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. Therefore, each input utterance is transformed into a sequence of acoustic vector

5. ALTERATION IN COMPUTATIONAL PHASES OF MFCC[2]

5.1 Windowing

In MFCC technique Hamming window is used. In place of which a more efficient Kaiser Window that is based on the concept of minimizing the mean square error rather than maximum error is used. The Kaiser window has an adjustable parameter α, which controls how quickly it approaches zero at the edges [2].

$$Kaiser(x,r,\alpha) = \begin{cases} \dfrac{I_0(\alpha\sqrt{1-(x/r)^2})}{I_0(\alpha)} & |x\ | \le r \\ 0 & else \end{cases}$$

Where $I_0(x)$ is the zeroth order modified Bessel Function .the higher the α narrower gets the window

5.2 Mel Filter Bank

Alteration in the number of triangular filters banks increases robustness and accuracy of the Speaker Recognition /Identification system as suggested by “Bansod N S, Seema kawathekar and Dabhade S.B”, MFCC with Mel filter bank of 32 filters has maximum Performance for the language Marathi and Hindi [3].

5.2 Absolute of FT

Before applying to the Mel filter banks only the absolute of the FT of the frame is taken. This not only reduces the cost of computing but also is an attempt of making the algorithm more robust [2].

6. FEATURE MATCHING BY ANN

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. Different techniques such as Neural Networks, Minimum distance classifier, Bayesian classifier, Quadratic classifier, Correlation are used for this purpose. In this Paper, we have opted for Artificial Neural Networks [6].

6.1 Neural Network Backpropogation

Neural network is approach is implemented when we have large number of speech samples of each speaker with variations among them which are used to train the network and correspondingly weights are updated. Finally, the weights are applied to the testing samples to get the correct output. 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 backpropogation is shown in figure.4 [8].

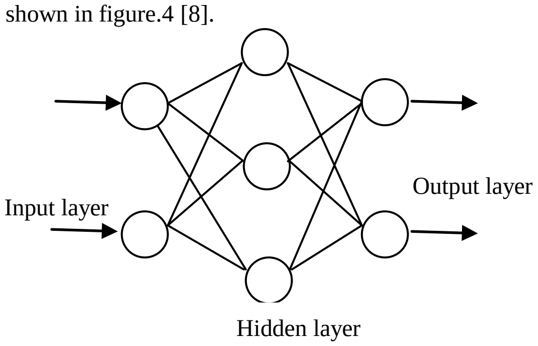


Fig.4.The Architecture Neural Network Backpropagation

Each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value. The weighted inputs are summed, and passed through a limiting function which scales the output to a fixed range of values. The output of the limiter is then broadcast to all of the neurons in the next layer. So, to use the network to solve a problem, we apply the input values to the inputs of the first layer, allow the signals to propagate through the network, and read the output values [10].

Table .1 Approaches used for Speaker Recognition

Reference	Feature Extraction Technique	Feature Matching Technique	Performance
[Eko Riyanto ‘13]	MFCC without modification in computational phases	Artificial Neural Network (ANN)	81.67%
[Anand Vardhan bhalla‘12]	MFCC with Modifications but without alterations in filter bank	Vector Quantization(V Q)	Efficient Speaker Recognition System
[Lindasalwa ‘10]	MFCC without Modification in Computational Phases	Dynamic Time Wrapping (DTW)	Techniques could used effectively for voice recognition

[Adjoudj Reda , ‘05]	MFCC without Modifications	Artificial Neural Network (ANN)LBG Algorithm	ANN= 96%
			LBG=92%

7. CONCLUSION

As considering the work of references mentioned in the above table, Speaker Recognition system using the Characteristic extraction method Mel Frequency Cepstral Coefficient (MFCC) and by alterations in the computational phases of Mel Frequency Cepstral Coefficient (MFCC) based technique used for feature extraction and using Artificial Neural

Network (ANN) matching technique for feature matching .The speaker recognition system is made more robust and efficient and hence the overall performance of Speaker recognition system is Enhanced

8. REFERENCES

[1] Lindasalwa Muda, Mumtaj Begaum and I.Elamvazuthi Voice Recognition Algorithms using Mel Frequency Cepstral (MFCC) and Dynamic Time Wrapping(DTW) Technique ,university Teknologi PETRONAS,Tronoh, Perak

[2] Anand Vardhan Bhalla, Shailesh Kharparkar, Mudit Ratna Bhalla , Performance Improvement of Speaker Recognition system,http://www.ijarcsse.com/ docs/ papers/March2012/volume_2_Issue_3/V2I30050..

[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, 2012.

[4] Jamal Price, sophomore student, Design an automatic speech recognition system *Using Malta,* University of Maryland Eastern Shore Princess Anne.

[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] Sujit kumar Behera, Jetendra, Speaker verification using Mel frequency cepstral coefficient and artificial neural ,network NIT ,Rourkela. http://ethesis.nitrkl.ac.in /3745/1/final_yr_project__thesis.pdf

[7] Speaker Recognition System, minhdo, teaching/speaker recognition, DSP mini Project.

[8] 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.

[9] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed et.al. “Deep Neural Networks For Acoustic Modeling In Speech Recognition”, IEEE Signal Processing Magazine, November 2012.

[10] Zaidi Razak,Noor Jamilah Ibrahim, Emran mohd tamil,mohd Yamani Idna Idris, Mohd yaakob Yusoff,Quranic verse recitation feature extraction using Mel frequency costrel coefficient (MFCC),Universiti Malaya.

[11] Eko Riyanto ,Suryono ,Informatics Engineering STMIK HIMSYA, Semarang, Indonesia

[12] Adjoudj Reda ,Boukelif Aoued ,Evolutionary Engineering and Distributed Information System Laboratory, EEDIS, Computer Science Department, University of sidi Bel- Abbes, Algeria