International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

A Review on Basic of Voice Recognition

Gaganpreet Kaur
Research Scholar, Dept. of CSE,
P.T.U, (Punjab) India

Dr.Dheerendra Singh
Prof. & Head, Dept. of CSE
SUSCET, Tangori, (Punjab) India

Ramandeep Kaur
Student, Dept. of CSE
S.G.G.S.W.U., FGS. (Punjab) India

Abstract—

This paper presents the main paradigms for speaker recognition. Biometric systems are mostly used as person identification or verification systems. Speaker recognition is a process where a person is recognized on the basis of his/ her voice samples instead of images. Security is big issue of human life. There is need of reliable and efficient systems which can recognize and authenticate humans by using biometric features like voice face etc. Voice recognition is one of the top security measures in banking systems that are used to provide protection to human's belongings by his/her voice characteristic . In this article, voice sample is observed with MFCC for extracting features and then used to train Artificial Neural for classification. Speaker recognition is used an automatic recognition, text-dependent and Speaker verification in security systems. It will recognize the speaker by comparing the current voice from the pre-stored database against the password stored in the Database system. It is implemented in Mat lab 7.0 version and showing results as correct acceptance and correct rejections with the error rate in Speaker verification.

Keywords— Extracting Features, Mel Frequency Cepstral Coefficients, Speaker Verification, Artificial Neural Networks (ANNs).

I. INTRODUCTION

Biometric recognition systems are used for the recognition of Humans. Instead of remembering passwords and PINs (which can be stolen or forgotten) or written signatures (which can be forged), biometric used such as fingerprints, voice and face are specific to an individual (and hence cannot easily be stolen or forged) and characterizes that individual (and hence cannot be forgotten)[1].

A computer system that automatically identifies and verifies the person by capturing the voice sample from a source like microphone is known as voice recognition. Voice recognition is the terms of biometric technology. It uses to provide any authentication to any system on the basis of his/her acoustic features of voice sample instead of images. The behavioral aspect of human voice is used for identification and verification systems. Voice is used for identification by converting spoken phrase/words from analog to digital format, and extracting unique vocal characteristics, such as pitch, frequency, and tone to establish a speaker model or voice sample. In voice recognition, is consists of an enrollment phase and verification phase. Enrollment process describes the registration of speaker by training his/her voice features [2] [3]. And verification process contains to verify the speaker by comparing his/her current voice features to pre stored features of voice sample. In real time, the verification process divides into two mechanisms. It first compares the unknown speaker to the pre stored database of known speakers on the basis of 1:N. and then it make decision of speaker to the exact match of 1:1. Where the one voice sample finally matched to only 1 template stored in the database [4]. Speaker verification systems can be classified into two broad categories, text-dependent and text-independent systems. In a text-dependent system, verification is based on a specific text or password, which is the same for both, the enrollment and verification phases. In the text-independent systems, the system requires a user to utter different text for each authentication attempt. [3] [5].

Speaker verification is an automatic process that uses human voice characteristics, obtained from a recorded voice signal, as the biometric measurements to verify claimed identity of a speaker [6], Speaker verification is used in a variety of application: such as shopping by telephone, voice mail, security for the e-commerce transactions, phone banking and trading, password resetting, accessing customer care services, credit card activation, transactions and payments, etc.[6].

II. BASIC ALGORITHMS

A vast range of methods and techniques have been developed. The current commonly used methods under speaker recognition for feature extraction; such as Mel Cepstral Frequency Coefficients (MFCCs). Linear Prediction Cepstrum Coefficient (LPPC), and for classification algorithms such as Vector Quantization(VQ), Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Artificial Neural Networks (ANNs).

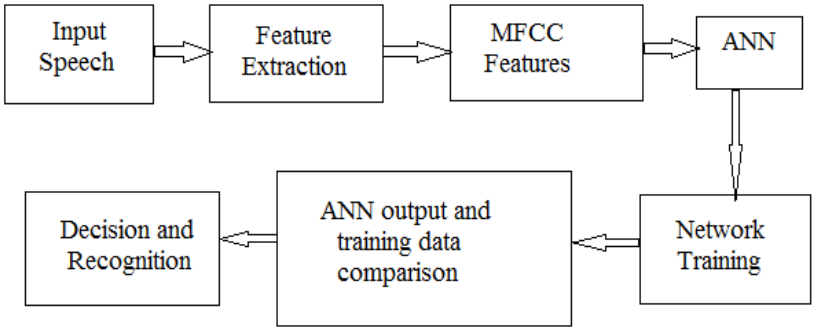

Fig. 1 Block diagram of speaker recognition phase.

A.) FEATURE EXTRACTION

1) Mel Cepstral Frequency Coefficients:-MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency Cepstral these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as: [7-8]

$$m=2595 \log_{10}(1+f/700) \quad \ldots. (1)$$

i.) *Frame Blocking:* - Framing is the first applied to the speech signal of the speaker. The signal is partitioned or blocked into N frames.

ii.) *Windowing:*-The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame.

iii.) *Fast Fourier Transform:*-Fast Fourier Transform which converts each frame of N samples in time domain to frequency domain.

iv.) *Mel-Frequency Wrapping:*-The spectrum obtained from the FFT step is Mel Frequency Wrapped, this process is to convert the frequency spectrum to Mel spectrum.

v.) *Cepstrum:*-In this final step, we convert the log Mel spectrum back to time. The result is called the Mel frequency Cepstrum coefficients (MFCC).

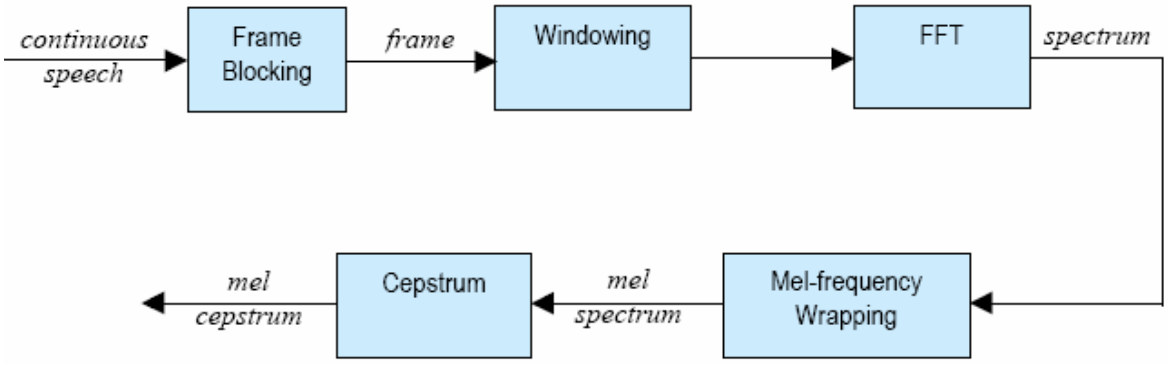

Fig. 2 Block diagram of the MFCC processing [7]

2) *Linear Predictive Coding:*-Another popular feature extraction technique is Linear Predictive Coding (LPC). The LPC algorithm produces a vector of coefficients that represents a smooth spectral envelope of the DFT magnitude of a temporal input signal. But use of Linear Prediction coefficients alone for speech recognition process is not efficient because all assumption of the vocal cord transfer function is not accurate and this method is not efficient enough to separate the convolution of the excitation from the glottis and the pole transfer function [9].

B.) PATTERN MACHING

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern matching. The goal of pattern matching is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns. Pattern matching methods include Dynamic Time Warping (DTW), the Hidden Markov Model (HMM), Artificial Neural Networks (ANNs), and Gaussian Mixture Models (GMM).

1) *Dynamic Time Warping:*- DTW technique compares words with reference words. Every reference word has a set of the spectra; but there is no distinction between separate sounds in the word. Because a word can be pronounced at different speeds, pitch etc.. This is used specifically to deal with variance in speaking rate and variable length of input vectors because this algorithm calculates the similarity between two sequences, which may vary in time or speed. To normalize the timing differences between test utterance and the reference template, time warping is done non-linearly in time dimension. After time normalization, a time normalized distance is calculated between the patterns. The speaker with minimum time normalized distance is identified as authentication speaker.[10]

2) *Gaussian Mixture Model:*-GMM is the most widely used modelling technique for text independent data. The Gaussian components (mixtures) of a GMM represent speaker dependent spectral shapes. For each speaker a GMM is built using the MFCC vectors obtained. A mixture in a model is characterised by its mean, covariance matrix and prior probability, which are estimated using the Expectation Maximization (EM) algorithm. The GMM is a density estimator. The distribution of the feature vector x is modelled clearly using a mixture of M Gaussians. Expectation maximization (EM) algorithm is used to estimate mean, covariance parameters. During recognition, sequences of features are extracted from the input signal. Then the distance of the given sequence from the model is obtained by computing the log likelihood of given sequence. The GMM model that provides the highest likelihood score is verified as the identity of the speaker. [10]

3) *Support Vector Machines:*-SVM is a supervised learning algorithm. It needs training of the tool before classification procedure gets started. This is best tool for binary classification of the data. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes form the input. The hyper plane is constructed defined by set of weights W, data points X and a bias or offset b, such that:

$$W.X + b = 0$$

Where W.X denotes the dot product the data and the normal vector to the hyper plane. The parameter b determines the offset of the hyper plane from the origin along the normal vector.[11]

4) *Artificial Neural Network:*- It is a machine learning process . ANN aims to work the way the human brain work. It consists of neurons connected by weights. The network learns the knowledge by adjusting its weights. ANN has many classifications depending on configurations and methods of working. The main classification is to classify ANN into supervised and unsupervised. In supervised ANN the network is trained by giving it the input and the corresponding output. In the unsupervised ANN the network is only given the input samples and it will adjust its weights so that it will have similar response for similar inputs. The Artificial Neural Network that will use is Back Propagation which is the most used form of supervised Neural Network. Feed Forward Neural Network is used form of supervised Neural Network and unsupervised Neural Network .It consists of an input layer, one or more hidden layers, and output layer.

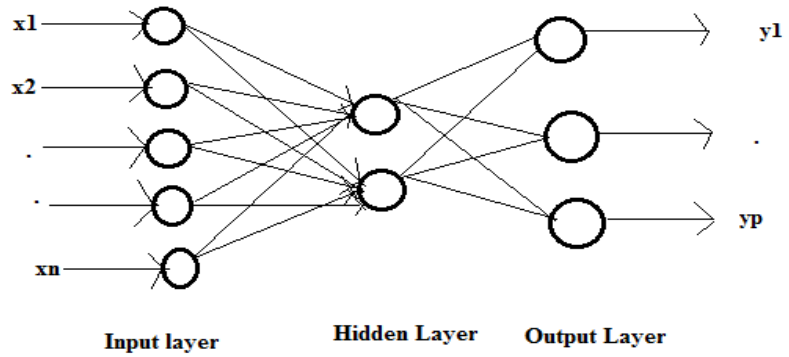

Fig. 3 Architecture of Feed Forward Neural Network

5) *Hidden Markov Modelling:*-HMM is defined as a finite state machine with fix number of states. It is statistical process to characterize the spectral properties of voice signal. It has two types of probabilities. There should be a set of observation or states and there should be a certain state transitions, which will define that model at the given state in a certain time. [12] In hidden markov model the states are not visible directly they are hidden but the output is visible which is dependent on the states. Output is generated by probability distribution over the states. It gives the information about the sequence of states but the parameters of states are still hidden. HMM can be characterized by following when its observations are discrete:

- N is number of states in given model, these states are hidden in model.
- M is the number of distinct observation symbols correspond to the physical output of the certain model.

III. REVIEW OF THE EXISTING APPROACHES

Kirandeep Kaur, Neelu Jain *et al.* [11] Automatic speaker recognition (ASR) has used many applications in the industries like banking, security, forensics etc. for its advantages such as easy implementation, more secure, more user friendly. To have a good recognition rate is a pre-requisite for any ASR system, which can be achieved by making an optimal choice among the available techniques for ASR. In this paper, different techniques for feature extraction such as MFCC, LPCC, LPC, Wavelet and VQ, GMM, SVM, DTW, HMM for feature classification. All these techniques are also compared with each other to find out best suitable candidate among them. On the basis of the comparison done, MFCC has upper edge over other techniques for feature extraction as it is more consistent with human hearing. GMM comes out to be the best among classification models due to its good classification accuracy and less memory usage.

Shi-Huang Chen and Yu-Ren Luo *et al.* [13] This presents in the MFCC as to extract features and trained and recognized using SVM. They defined the MFCC as the unique and reliable feature extraction technique. In recognition phase SVM

(super Vector Machine) technique based on two class classifiers by defining the decision in binary form was introduced. It discriminate claimed speaker and imposter by +1 and -1 by maximizing the margins or minimizing the structural risks. It shows results averaged to 95.1% with ERR of 0.0%. That was considered as the best results under second order of MFCC.

Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi *el at.* [14] It also extracted the voice features using MFCC that was trained and recognized using DTW. DTW (dynamic time Wrapping) a non linear sequence alignment is another technique that is used for recognition process. They find it best for time sequence between two speeches. Here, the optimal wrapping path is achieved by wrapping the time distance between two signals.

Ibrahim Patel and Y. Srinivas Rao el at. [15] They represent the voice recognition with improvement of MFCC with frequency decomposition technique. They introduces sub band coding in their research. The integration of MFCC with sub band coding increases its efficiency. These two features of MFCC and integrated sub band decomposition with MFCC are used in HMM to train and recognize the speaker.

Abushariah *et al.* [16**]** recognize digits from 0 to 9 in English using the Hidden Markov Model [8]. They map the wave forms obtained from the training audio to the ones that need to be recognized. Their experiments have been carried out using sound produced in a professional visual-audio studio and other speech samples were collected in noisy environments (class rooms and student's rooms). The resulting certainty rate varies from 79% to 99.5% in this controlled environment. However, subsequent experiments in noisy environments have achieved certainty rates between 76.67% and 82.5%. Abushariah et al. claim that his audio database is unbiased, since it stores sound from 34 varied persons (men, women, teenagers, adults, Asians, and Arabs).

Abushariah *et al.* [17] It identifies voice by means of two algorithms: Mel Frequency Cepstral Coefficients for extracting features and Vector Quantization for classifying and training. They have created a database containing the voices of 100 persons, 50 women and 50 men, between 18 and 37 years old, from 25 different countries. For the training phase, each person had to record 30 files. Their best results show that the recognition rate of female voices (94.2%) is higher than the one of masculine voices (91%), since the female voices are clearer and have better diction.

IV. CONCLUSIONS

In this paper, a text-dependent speaker verification system is discussed that uses an artificial neural network as the primary classification method and MFCC coefficients as the input features extracted from store database. MFCC coefficients and their corresponding first order derivatives as the training feature vectors are taken which gives acoustic features of each individual depending upon password spoken. The results, that the optimized ANN-based speaker verification system could be successfully applied to clarify passwords, PINs and in general be employed as a part of a security system. A wide variety of systems requires more reliable person recognition schemes to either confirm or determine the identity of an individual features. MFCC features and neural network can be a reliable combination for authenticating users. Here all algorithms for feature extraction and pattern matching are discussed.

REFERENCES

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Systems Video Technol*., Vol. 14,Issue No. 1, pp. 4–20, 2004.

[2] Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov A. and Minkyu Choi, "Biometric Authentication: A Review", International *Journal of u- and e- Service, Science and Technology* ,Vol. 2, Issue No. 3, September, 2009 .

[3] Judith A. Markowitz, "Voice Biometrics", *Communications of the ACM*, Vol. 43, Issue No. 9, September 2000.

[4] http://en.wikipedia.org/wiki/Speaker_recognition.

[5] http://www.globalsecurity.org/security/systems/biometrics-voice.htm.

[6] I. D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Commun.*, Vol. 17, Issue No. 1-2, pp. 91-108. , 1995.

[7] Anjali Bala, Abhijeet Kumar and Nidhika Birla, "Voice Command Recognition System Based On MFCC and DTW," *Anjali Bala et al / International Journal of Engineering Science and Technology* ,Vol. 2 (12), 7335-7342, 2010.

[8] http://en.wikipedia.org/wiki/Mel_scale.

[9] W. B. Mikhael, P. Premakanthan, "Speaker verification /recognition and the importance of selective feature extraction: review," *44th IEEE Proceedings on Midwest Symposium on Circuits and Systems, Ohio*, Vol. 1, pp. 57-61, 2001.

[10] R. P. Ramachandran, K.R. Farrell, R. Ramachandran, R. J. Mammone, "Speaker Recognition—General Classifier Approaches and Data Fusion Methods," *Pattern Recognition in Information Systems*, Vol. 35, Issue No.12, pp. 2801-2821, December 2002.

[11] Kirandeep Kaur, Neelu Jain, "Feature Extraction and Classification for Automatic Speaker Recognition System-A Review,"*International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 5, Issue No.1, January 2015.

[12] Lawrence R. Rabiner, Fellow, "A Tutorial On Hidden Markov Model And Selected Applications In Speech Recognition,"*Proceedings Of The IEEE*, Vol. 77, Issue No. 2, February 1989.

[13] Shi-Huang Chen and Yu-Ren Luo, " Speaker Verification Using MFCC and Support Vector Machine", *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1, IMECS 2009, March 2009 .

[14] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal Of Computing*, Vol. 2, Issue No. 3, March 2010 .

[15] Ibrahim Patel and Dr. Y. Srinivas Rao, "Speech Recognition Using Hmm With Mfcc- An Analysis Using Frequency Specral Decomposion Technique," *an International Journal (SIPIJ)* ,Vol. 1, Issue No. 2, December 2010.

[16] A.A.M. Abushariah, T.S. Gunawan, O.O. Khalifa, and M.A.M. Abushariah. " English digits speech recognition system based on hidden markov models," *In Computer and Communication Engineering (ICCCE), 2010 International Conference,* pp1–5, may 2010.

[17] A.A.M. Abushariah, T.S. Gunawan, J. Chebil, and M.A.M. Abushariah. " Voice based automatic person identification system using vector quantization," *In Computer and Communication Engineering (ICCCE), 2012 International Conference*, pp 549–554, july 2012.

