



**“A Study On Speech Recognition Technology And Speaker
Identification Using Data Mining”**

Prof. Yogeshkumar J. Patel

Prof. Rasikkumar D. Patel

Prof. Piyush A. Patel

Assistant Professor



ABSTRACT

The purpose of this paper is to develop a Speaker Identification System which can recognize speakers by their acoustic characteristics of speech. The proposed system would be a text independent system means the user is free to speak any word or sentence. Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information include in speech waves. This technique makes it possible to use the speech's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, Forensic speaker recognition ,security control for confidential information areas, and remote access to computers. Wavelet Transform particularly Discrete Wavelet Transform (DWT) is used in order to extract the vocal characteristics of the speakers in speech signal whereas KNearest Neighbor (KNN) algorithm is used for feature matching, which shows a very much improvement in the identification rate. The feature extraction is done by six levels wavelet decomposition and these features are extracted from wavelet coefficients by mean, standard deviation and ratios between them.

Keywords –MFCC, KNN, DCP, DWT, Cepstrum, Feature Extraction, LPC, Feature Matching, HMM, SVM.

I. INTRODUCTION:

In speaker identification, we match a given (unknown) speaker to the set of known speakers in a database. The database is constructed from the speech samples of each known speech. Features vectors are extracted from the samples by short term spectral analysis and processed further by vector Quantization for locating the clusters in the feature space. For speaker identification we are extracting many features of speech like real Cepstral coefficients, MFCC (Mel-Frequency Cepstral Coefficients), Linear predictive Cepstral Coefficients to get 100 % result or finding particular speaker from database. For finding who is speaker from large amount of database Data Mining concept is use. Study the role of vector Quantization (VQ) in the speaker identification. The vocabulary of digit is use very often in testing speaker



recognition because of its applicability to many security applications by checking the voice characteristics of the input utterance (sound), using an automatic speaker recognition system similar to the one we will develop,

Acoustic communication is one of the fundamental prerequisites for the existence of human society. Textual language has become extremely important in modern life, but speech has dimensions of richness that text cannot approximate. From speech alone fairly accurate guesses can be made as to whether the speaker is male or female, adult or child. In addition, expert can extract information from speech regarding the speaker's state of mind. As computer power increased and knowledge about speech signals improved, research of speech processing became aimed at automated systems for many purposes. Speaker recognition is the complement of speech recognition. Both techniques use similar methods of speech signal processing. In automatic speech recognition the speech processing approach tries to extract linguistic information from the speech signal to the exclusion of personal information. Speaker identification focused on the characteristics unique to the individual, disregarding the current word spoken. The uniqueness of an individuals' voice is a consequence of both the physical feature of the person vocal tract and the person mental ability to control the muscles in the vocal tract.

Speaker recognition can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance (sound). Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim (state) of a speaker.

Speaker recognition method can also be divided into Text – independent method and Text – dependent method. In text – independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In text – dependent system, on the other hand, the recognition of the speaker identity is based on his or her speaking one or more specific phrases, like password, card number, PIN codes, etc.

A human speech is different from person to person by pitch and ferments. In order to recognize a speaker those speech characteristics are to be extracted which are varying with persons' dialects and sex.

Speech Analysis

In order to understand the speech analysis it is important to understand how the speech produces. Speech production can be divided in two parts.

- i. Basic sound production source – vocal cords
- ii. A filter through the vocal cords which creates acoustic disturbances.

Human speech is produced by flowing air from lungs to vocal cords: basic sound production source produces a pulse known as pitch which is basic frequency of speech. Many parts plays role to produce sound such as nasal cavity, teeth, lips, tongue, vocal cords and lungs. Speech may categorize in two types: voiced and unvoiced speech. Voiced speech is produced when the vocal folds vibrates during air flow from lungs to vocal cords and unvoiced speech is produced when these vocal folds does not vibrates. Pitch can be changed by modified vocal cord tension governed by a control input to musculature.

Challenges in Speaker Identification

- Speaker identification is a complex task due to complexity of the speech signal itself. There are the following challenges associated with Speaker Identification system.
- Identification of speech with noisy background environment.
- Extraction of relevant acoustic parameter such as pitch, formants, slopes, zero crossing distribution etc.
- Effect of situational parameter such as cold, emotion, loudness, pitch, whispering, distortion due to talker's acoustical environment and distortions by communication

systems (telephone, transmitter-receiver, public address, face mask) and non std. environments.

Factors affecting Speaker Identification

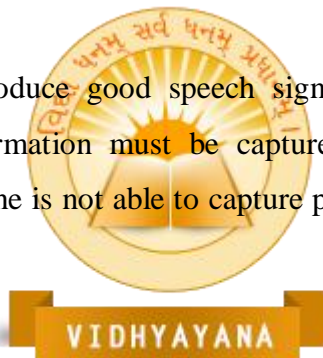
There are the following factors affect the Speaker Identification System.

Background of the Environment

Background of the environment affects greatly to the accuracy of the system. Naturally if the speech samples contain much noise, the features would be noisy and to identify a speaker, features must be robust.

Quality of Microphone

A good quality microphone produce good speech signal containing maximum acoustic information. This acoustic information must be captured properly since it represents a speaker. A low quality microphone is not able to capture proper acoustic information present in speech of individual.



Emotional state of the speaker

Here it is assumed that the speaker speaks normally. The loudness of the speech can be varied but the emotional speech is not allowed to speak. The speaker must speak in normal mode.

Computing Power

The computing power of PC should be enough in order to process the speech. Low configuration PC takes much time to process and even possibility of hang-up.

Silent part of speech

Silent part of speech not so matter because it is removed by preprocessing of the speech signal.



Signal Energy

The signal energy must be normal in order to normalize the loudness of speech.

II. SPEECH FEATURE EXTRACTION

The purpose of this phase is to convert the speech waveform, using Digital Signal Processing (DSP) tools, to a set of features (at a considerably lower information rate) for further analysis. This is often referred as the signal – processing front end. The speech signal is a slowly timed varying signal (it is called quasistationary).

When examined over a sufficiently short period of time (between 5 and 100 msec.), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic changes to reflect the different speech sound being spoken. Therefore, short time spectral analysis is the most common way to characterize the speech signal.

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel – Frequency Cepstrum Coefficients (MFCC) and other. MFCC is perhaps the best known and most popular.

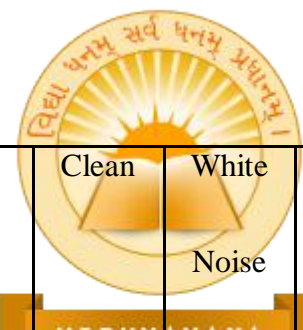
Linear Predictive Coding (LPC) of speech has proved to be a valid way to compress the spectral envelope in an all-pole model (valid for all non-nasal sounds, and still a good approximation for nasal sounds) with just 10 to 16 coefficients, which means that the spectral information in a frame can be represented in about 50 bytes, which is 10% of the original bit rate. Instead of LPC coefficients, highly correlated among them (covariance matrix far from diagonal), pseudo orthogonal cepstral coefficients are usually used, either directly derived as in LPCC (LPC-derived Cepstral vectors) from LPC coefficients, or directly obtained from a perceptually-based Mel-filter spectral analysis as in MFCC (Mel-Frequency based Cepstral Coefficients). Some other related forms are described in the literature, as PLP (Perceptually based Linear Prediction), LSF (Line Spectral Frequencies) and many others, not detailed here

for simplicity. By far, one of the main factors of speech variability comes from the use of different transmission channels (e.g. testing telephone speech with microphone-recorded speaker models).

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale, which is linear frequency spacing above 1000 Hz. The process of computing MFCC is described in more detail next.

2.1 MEL – FREQUENCY CEPSTRUM COEFFICIENTS (MFCC) PROCESSOR.

Figure 7: Block diagram of the MFCC processor.



	Clean	White Noise	Lung Sound	Speech
Mel Frequency Cepstral Coefficient	97.25 %	42.55 %	45.75 %	52.00 %
Mel – Scale Wavelet Transform	94.00 %	85.59 %	93.26 %	79.00 %

Table 1: Result of Mel Frequency Cepstral Coefficient

III. FEATURE MATCHING

Feature matching is important stage in speaker identification system and several techniques exist that can be used to model speakers based on the features extracted from speech samples. In this thesis two types of standard classifier used, the first one is K Nearest Neighbor (KNN) which is simple and very efficient with large dataset and the other one is Artificial Neural Network (ANN) having discriminate-training power used for many years in various fields such as speech and image processing.

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called Pattern Recognition. The goal of pattern recognition is to classify object of interest into one of a number of categories or classes. The object of interest is generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speaker. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching.

Furthermore, if there exist some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one evaluate the performance of the algorithm.

The state-of-the-art feature matching techniques used in speaker recognition include, NET talk, Time-Delay Neural Network (TDNN), Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). A new approach, Support Vector Machine is used here, due to high accuracy. In this project, the VQ approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. The speaker identification system, used in the experiments outlined below, uses a vector Quantization classifier to build the feature space and to perform speaker classification. The LPC-Cepstrum is used as features with the

Euclidean distance between test utterances and the trained speaker models as the distance measure.. Each vector y_i is called a codeword and the set of all the code words is called a codebook.

Speaker Identification using HMM

Mangesh S. Deshpande and Raghunath S. Holambe proposed a close-set, text independent speaker identification approach using Continuous Density Hidden Markov Model (CDHMM). Baum-Welch algorithm is used to train HMM for each enrolled speaker. A non-stationary speech signal is represented as a sequence of states in HMM. Actually HMM is a classical approach both used in speech and speaker recognition successfully since last many years. The system used TIMIT database having 630 speakers with 70% male and 30% female to evaluate the system performance. For 2 stage single mixtures CDHMM the speaker identification result is 96.75%.

Speaker Identification using SVM

Rabbani N., [13] proposed an approach of speaker identification using support vector machines. An extra training set is applied to train a discrete density hidden markov model to improve the performance of the identification. Multiclass SVM classifier is used for each feature vector during testing phase. For decision making HMM model is used with the class sequence. The paper claimed the existing approach reduces the identification error rates up to 57.14%.

K-NN Algorithm

K-NN is a supervised learning algorithm[7] also known by K-Nearest Neighbor used in many applications such as pattern classification, image processing, speech analysis, data mining and many others. This is very simple algorithm classified the new sample based on minimum distance from the query sample to the training sample. The classification is done based on majority of k neighbor (where k is an integer number) and training. It is important to find out



VIDHYAYANA

the k number of objects nearest to query sample. Here nearest is taken as the smallest distance in n-dimensional feature space.

The major distances used by K-NN algorithm are-

1. Euclidean distance

Euclidean distance can be given by the following equation

2. Cityblock distance – This is Sum of absolute differences. Also known as Manhattan distance can be given by equation

3. Cosine distance – This is $1 - \text{Cosine}$ (the included angle between points (treated as vectors)?)

4. Correlation distance – $1 - \text{sample correlation}$ between points (treated as sequences of values)

5. Hamming – % of bits that differ (suitable only for binary data)



VIDHYAYANA

IV RESULT AND DISCUSSIONS

Speech Corpus

For the evaluation of the speaker identification methods, two speech corpora were used, namely VoxForge Speech corpus and my own in house speech corpus.

In-House Dataset

In-House dataset, compiled by myself contains the speech samples of my institute's students. This contains 10 speech samples of 20 speakers. Matlab7.6 was used to record speaker's voice. The Speech recording parameters are shown in table

Number Of Speaker	20
Sampling Rate	450
Bit Depth	16
Duration Of Recording	7 sec. and 13 sec
Channel	mono

Table 2. Speech Recording Parameters.

Ten different sentences used to record a speaker's voice. Out of ten, five sentences used in training and five for testing the system. The sentences were rich in vowel so that the signal contains maximum information which is helpful to extract robust features in order to identify a speaker.

VoxForge Speech Corpus

A VoxForge speech corpus contains 250+ speakers' speech samples of both male and female. This is available in two categories, the first one is available with sampling rate 8KHz and 16 bit and other one is 16 KHz with 16 bit. The speech samples recorded in different time and different environment conditions.

Performance of Discrete Wavelet Coefficients on dataset

The table 3 shows the accuracy with number of coefficients for VoxForge speech corpus. The experiment was performed on a set of 100 speakers. The coefficients decreased as per order cD1 and cA1 then cD2 and cA2 and so on.

No. of Coefficients	Accuracy Rate (%) (On 100 Speaker)
2	72.6
4	74.2
6	79.6
8	78.8
10	82.8
12	85.2

Table 3 . Performance of Discrete Coefficients of VoxForge Corpus

V. CONCLUSION

In this work, we studied and analyzed different technique for speaker identification. In the first part, we started from the identification background, which is best on the digital signal theory and modeling of the speaker vocal tract. Then discuss various techniques for reducing amount of time and feature extraction.

Speaker Recognition system is a complex process for identification of particular person form particular speech. In this project, particular person is identified using some clustering algorithms. Some features are extracted form speech signal. Comparing this feature with sampled speech signal, then using clustering algorithms particular actor is identified.

In this paper Wavelet Transform based approach has been used for speaker identificationsystem. Discrete Wavelet Transform successfully used to extract feature in

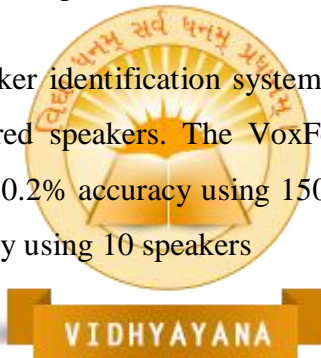


order to build robust speaker identification system. Pre processing of signal being used to improve the speech signal before feature extraction method application. Pre-processing techniques involved silence removal; DC offset removal, pre-emphasis etc.

Feature extraction was the main process of speaker identification system. This involved speaker specific features extraction from speech signal and discrete wavelet transform being used here for this purpose. The system's accuracy depends upon the feature vector which was used to create speaker model for classification purpose.

The design and development approach of speaker identification system has been presented in this thesis. The speech dataset used for experiment contains ten samples of two hundred voice profiles. The experiment results indicate that discrete wavelet transform produce best performance as compare with wavelet packet transform.

The proposed design of the speaker identification system used ten feature vector containing forty-four features of two hundred speakers. The VoxForge speech corpus achieved 98% accuracy using 10 speakers and 80.2% accuracy using 150 speakers while Alternative corpus (In-House) achieved 80% accuracy using 10 speakers



References

[1] D.A.Reynolds and R.C.Rose, "Robust text-independent Speaker identification using Gaussian mixture speaker

models," IEEE Transactions on Speech and Audio Processing, vol.3, no.1, pp.72–83, 1995.

[2] S.Melnik, S.F.Quigley, and M.Russell, "Speech recognition On an FPGA using discrete and continuous hidden Markov models, " in Proceeding of the International Workshop on Field- Programmable Logic, pp.202–211, 2002.



[3] S. Melniko?, S. F. Quigley, and M. Russell, "Implementing a Simple continuous speech recognition system on an FPGA," in Proceedings of IEEE Symposium on Field Programmable Custom Computing Machines, pp. 275–276, Los Alamitos, Calif, USA, 2002.

[4] K. Miura, H. Noguchi, H. Kawaguchi, and M. Yoshimoto, "A low memory bandwidth Gaussian mixture model (GMM) Processor for 20,000-word real-time speech recognition FPGA system," in Proceedings of the International Conference on Field-Programmable Technology (ICFPT'08), pp. 341–344, December 2008.

[5] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete system," IEEE Transactions on Circuits and Systems I, vol. 53, no. 1, pp. 70–77, 2006.

[6] David Michael Graeme Watts, "Speaker Identification – Prototype Development and Performance" Research Project, University of Southern Queensland, Faculty of Engineering & Surveying, 2006.



[7] D.A. Reynolds, "An overview of Automatic Speaker Recognition Technology", international conference on Acoustic Speech and Signal processing, Signal Processing Society IEEE 2002.

[8] Tridibesh Dutta, "Dynamic Time Warping Based Approach to Text Dependent Speaker Identification Using Spectrograms" in Proceedings of Congress on Image and Signal Processing, Vol. 2, 2008.

[9] Rabbani N., "Novel approach in speaker identification using support vector machines", 9th International Symposium on Signal Processing and Its Applications, ISSPA 2007, Sharjah, UAE, 2007

[10] Daqroug K., "Speaker Identification Wavelet Transform based method", IEEE, 5th International Multi-Conference on Systems, Signals and Devices, Amman, Jordan, SSD-2008.



VIDHYAYANA

ISSN 2454-8596
www.MyVedant.com

An International Multidisciplinary Research E-Journal

[11] VoxForge Speech Corpus <http://en.wikipedia.org/wiki/VoxForge>

[12] Michael Negnevitsky, "Artificial Intelligence: a guide to Intelligent Systems", Addison Wesley, 2002

[13] R. Polikar, The Wavelet Tutorial,

<http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html>

[14] Rabinar, L. and R.W. Schafer, "Fundamentals of Speech Recognition", PrenticeHall, 1993.



VIDHYAYANA