Asian Journal of Information Technology 6 (9): 938-942, 2007
ISSN: 1682-3915

Automatic Speaker Identification Using Vector Quantization

Poonam Bansal, Amita Dev and Shail Bala Jain
Amity School of Engineering and Technology, 580, Delhi Palam Vihar Road,
Bijwasan, New Delhi 110061, India

Abstract: An automatic speaker identification scheme is purposed and developed, to identify or verify a person, by identifying his/her voice, using a novel method. All speaker identification system contains two main phases, training phase and the testing phase. In the training phase the features of the words spoken by different speakers are extracted and during the testing phase feature matching takes place. Feature extractor transforms the raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal. The feature or the template thus extracted is stored in the database. During the recognition phase the extracted features are compared with the template in the database. In the purposed Speaker Identifier (SI) the features extracted are LPCC, Mel-Frequency Cepstrum coefficients (MFCC), Delta MFCC (DMFCC) and Delta-Delta MFCC (DDMFCC). Vector Quantization (VQ) is used for speaker modeling process. The final recognition decision is made based on the matching score: Speaker model with the smallest matching score is selected as a speaker of the test speech sample. Speaker identification rate was observed to be 96.59% in text independent case and increases by 3.5% in reference to text dependent, as we increase the feature vector size to 36 by including 12 DMFCC and 12 DDMFCC recognition rate gets increased by 0.4%. Better performances could be seen when applying this approach itself or mixed with Hidden Markov Model (HMM) in isolated-word speech recognition.

Key words: Speaker identification, mel frequency cepstrum coefficient, delta and delta delta MFCC, vector quantization

INTRODUCTION

The problem of speaker identification is one that is rooted in the study of the speech signal. A very interesting problem is the analysis of the speech signal and therein what characteristics make it unique among other signals and what makes one speech signal different from another. When an individual recognizes the voice of someone familiar, he/she is able to match the speaker's name to his/her voice. This process is called speaker identification. Speaker identification exists in the realm of speaker recognition, which encompasses both identification and verification of speakers. Speaker verification is the subject of validating whether or not a user is who he/she claims to be. There is an increasing need for person authentication in the world of information, applications ranging from credit card payments to border control and forensics (Prabhakar *et al*., 2003).

In general, a person can be authenticated in three different ways:

- Something the person has, e.g. a key or a credit card, signature.
- Something the person knows, e.g. a PIN number or a password.
- Something the person is, e.g., fingerprints, voice, facial features.

The first two are traditional authentication methods that have been used for several centuries. However, they have the shortcoming that the key or credit card can be stolen or lost and the PIN number or password can be easily misused or forgotten. Such shortcomings will not be there in the last class of authentication methods, known as biometric person authentication (Prabhakar *et al*., 2003). Each person has unique anatomy, physiology and learned habits such that familiar persons use in everyday life to recognize the person. Increased computing power and decreased microchip size has given impetus for implementing realistic biometric authentication methods. The interest in biometric authentication has been increasing rapidly in the past few years. Speaker

Corresponding Author: Poonam Bansal, Amity School of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India

recognition refers to task of recognizing peoples by their voices. The goal of this research is to build a simple, yet complete and representative automatic speaker identification system.

PRINCIPLES OF SPEAKER IDENTIFICATION

Speaker identification further divides into two subcategories, which are text-dependent and text-independent speaker identification. Text-dependent speaker identification differs from text-independent because in the aforementioned the identification is performed on a voiced instance of a specific word, whereas in the latter the speaker can say anything.

At the highest level, all speaker identification systems contain two main modules: Feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. We will discuss feature extraction and feature matching in detail in this study. Automatic speaker identification work is based on the premise that a person's speech exhibits characteristics that are unique to the speaker (Fig. 1). However, this task has been challenged by the highly variant nature of input speech signals. The principle source of variance is the speaker himself. Speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time, health conditions (e.g. the speaker has a cold), speaking rates, etc. There are also other factors, beyond speaker variability, that present a challenge to speaker recognition technology. Examples of these are acoustical noise and variations in recording environments (e.g. speaker uses different telephone handsets).

FEATURE EXTRACTION

The purpose of this step is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing, which is referred as the signal-processing front end (Rabiner and Schafer, 1979; Rabiner and Juang, 1978; Furui, 2001). The speech signal is a slowly time-varying signal (called quasi-stationary).

When examined over a sufficiently short period of time (5 ~ 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, the 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 speech recognition task, such as Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and others (Gold and Morgen, 2000). MFCC is perhaps the best known and most popular and it will be used in this study (Shannon and Paliwal, 2004). MFCC is based on the known variation of the human ear's critical bandwidths with frequencies, 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, linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Mel-Frequency Cepstral Coefficients (MFCC), introduced by Davis and Mermel stein constitute a parametric sound representation widely used in automatic speech recognition systems. MFCC provide a substantial data reduction, because a few coefficients are sufficient to represent the cepstrum of the acoustic signal.

The block diagram of MFCC processor is shown in Fig. 2.

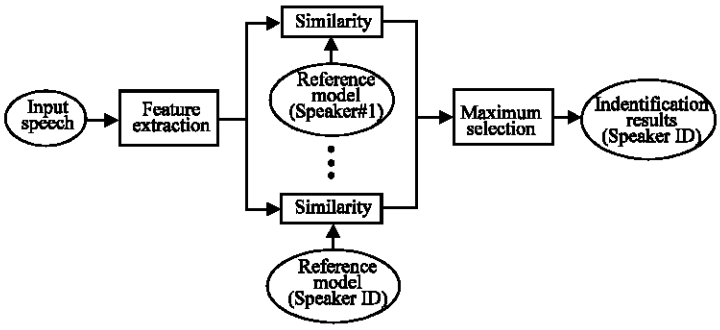

Fig. 1: Speaker identification

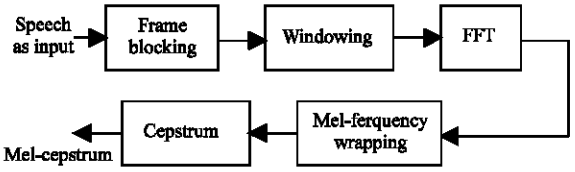

Fig. 2: Block diagram of MFCC processor

Frame blocking: The continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). There is a overlapping of (N-M) samples (Gold and Morgen, 2000). This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 128.

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, window to taper the signal to zero at the beginning and end of each frame. A hamming window function is used.

$$W(n) = 0.54 - 0.46\text{Cos}\left(\frac{2\pi n}{N-1}\right) \qquad 0 \le n \le N-1$$

FFT: The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is defined on a set of N samples X_n as:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \qquad n = 0,1,2,.....,N-1$$

Mel-frequency wrapping: Human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale Haykin (2002). The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz:

$$\text{mel}(f) = 2595 * \log 10(1 + f/700)$$

Our's approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular bandpass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The mel scale filter bank is a series of L triangular bandpass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale.

Cepstrum: In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The Discrete Cosine Transform is done for transforming the mel coefficients back to time domain.

$$C_n = \sum_{k=1}^{K} (\log \tilde{S}_k)\text{Cos}\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1,2 \ ..... \ K$$

Where as $\tilde{S}_k$, k = 1, 2,.....K are the outputs of the last step.

FEATURE MATCHING

The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM) and Vector Quantization (VQ). 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. Each region is called a cluster and can be represented by its center called a codeword (Linde *et al*., 1980). The collection of all such codewords is called a codebook. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all codewords is called a codebook for a known word. Vector Quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm.

VQ design: The VQ design can be stated as follows. Given a vector source with its statistical properties known, given a distortion measure and given the number of code vectors, we can find a codebook and a partition which result in the smallest average distortion.

We assume that there is a training sequence consisting of M source vectors:

$$T = \{x_1, x_2, x_3, \ldots\ldots, x_M\}$$

This training sequence can be obtained from some large database. M is assumed to be sufficiently large so that all the statistical properties of the source are captured by the

$$X_m = \{x_{m,1}, x_{m,2}, x_{m,3}, \ldots\ldots, x_{m,k}\}, \quad m = 1,2,3.....M$$

training sequence. We assume that the source vectors are K-dimensional, e.g., Let N be the number of code vectors and let

$$C = \{c_1, c_2, c_3, \ldots\ldots, c_N\}$$

represents the codebook. Each code vector is K-dimensional, e.g.,

$$c_n = \{c_{n,1}, c_{n,2}, c_{n,3}, \ldots\ldots, c_{n,k}\}, \quad n = 1,2,3.....N$$

Let S_n be the encoding region associated with code vector C_n and let Denote the partition of the space. If the source vector $X_m$is in the encoding region S_n, then its approximation (denoted by $Q(X_m)$) is C_n:

$$P = \{S_1, S_2, S_3, \ldots\ldots, S_N\}$$

$$Q(x_m) = c_n, \quad \text{if } x_m \in S_n$$

Assuming a squared-error distortion measure, the average distortion is given by:

$$D_{ave} = \frac{1}{Mk}\sum_{m=1}^{M} \| x_m - Q(x_m) \|^2$$

Where,

$$\|e\|^2 = e_1^2 + e_2^2 + \ldots + e_k^2$$

The design problem can be succinctly stated as follows: Given T and N find C and P such that D_{ave} is minimized.

Data set:

Language	:	Standard Hindi
Vocabulary size	:	A set of 1000 most frequently occurring hindi words
No. of Speakers	:	50 (30 Male and 20 Female)
Average duration of training and testing utterances	:	500-800 msec.
Audio recording	:	S/N > 40 db
Sampling and quantization	:	16Khz, 16-bit

RESULTS

The performance of the SI was evaluated in terms of Speaker identification rate. We have used the following identification measure for computing the identification rate.

$$\text{Speaker Identification rate (\%)} = S_c / S_T * 100$$

Where, S_c is the No. of times the correct speaker has been identified and S_T is the Total No. of speakers used in the testing session. Experimental analysis was done in reference to Text dependent speaker identification and Text Independent speaker identification.

The performance of speaker identification rate (%) got improved with the increase in codebook size, in case the speaker is identified with the same utterances by which he or she has been trained (Text Dependent). It reaches to 99% with codebook size of 64. If the speaker is tested with some other random utterances (Text Independent) the recognition rate decrease by about 3.5% (Fig. 3).

Further analysis was done with improved feature vector set. By incorporating the new parameters (DMFCC and DDMFCC) in the feature vector set the identification rate got improved by 0.4%.

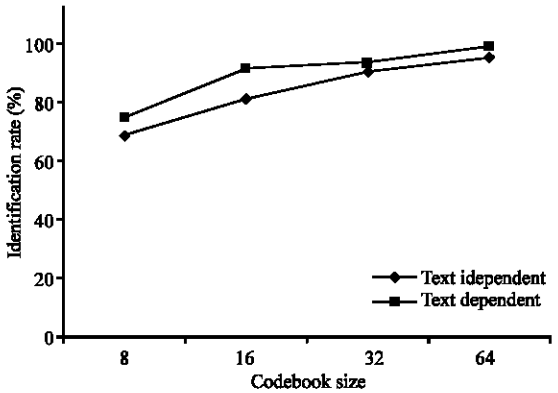

Fig. 3: Performance of SI with codebook size

CONCLUSION

The human voice is variable through temporal variations of the voice, caused by a cold, hoarseness, stress, emotional different states or puberty vocal change. MFCC and LPCC are well known techniques used in speaker identification to describe signal characteristics, relative to the speaker discriminative vocal tract properties. All-pole model used in the LPC provides a good model for the voiced regions of speech and quite bad for unvoiced and transient regions. The main drawback of LPCC is that unlike MFCC it does not resolve the vocal tract characteristics from the glottal dynamics, which vary from person to person and might be useful in speaker identification. By enhancing the feature vector set with DMFCC and DDMFCC performance of the SI gets improved in both the cases of text dependent and text independent speaker identification. By incorporating more features like pitch and formant values (Which are speaker dependent) the speaker identification rate can be further enhanced.

REFERENCES

Furui, S., 2001. Digital Speech Processing, Synthesis and Recognition. (2nd Edn.), Marcel Dekker, Inc., New York.

Gold and Morgen, 2000. Speech and audio signal processing. John Wiley and Sons Publication, (1st Edn.).

Haykin S., 2002. Adaptive filter theory. Pearson Education Publication, (4th Edn.).

Linde Y., A. Buzo and R. Gray, 1980. An algorithm for vector quantizer design. IEEE. Trans. Commun., 28: 84-95.

Prabhakar, S., S. Pankanti and A. Jain, 2003. Biometric recognition, security and privacy concerns. IEEE. Security and Privacy Mag., 1: 33-42.

Rabiner L.R. and B.H. Juang, 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, N.J.

Rabiner, L.R. and R.W. Schafer, 1978. Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, N.J.

Shannon and Paliwal, 2004. B.J. Shannon and K.K. Paliwal, MFCC Computation from Magnitude Spectrum of higher lag autocorrelation coefficients for robust speech recognition. Proc. ICSLP., pp: 129-132.