Vector Quantization In Text Dependent Automatic Speaker Recognition Using Mel-frequency Cepstrum Coefficient

AHSANUL KABIR, SHEIKH MOHAMMAD MASUDUL AHSAN
Department of Computer Science and Engineering
Khulna University of Engineering and Technology
Fulbarigate, Khulna 920300
BANGLADESH

Abstract: - Automatic speaker recognition is a field of study attributed in identifying a person from a spoken phrase. The technique makes it possible to use the speaker's voice to verify their identity and control access to the services such as biometric security system, voice dialing, telephone banking, telephone shopping, database access services, information services, voice mail, and security control for confidential information areas and remote access to the computers. This thesis represents a development of a Matlab based text dependent speaker recognition system. Mel Frequency Cepstrum Coefficient (MFCC) Method is used to extract a speaker's discriminative features from the mathematical representation of the speech signal. After that Vector Quantization with VQ-LBG Algorithm is used to match the feature.

Key-Words: - Speaker Recognition, Human Speech Signal Processing, Vector Quantization

1 Introduction

The speech is a natural communication mechanism between two persons but it is very complex from automatic analysis viewpoint. The human speech is, technically speaking, air pressure variations caused by movements of muscles and other tissue within the speaker. The speech organ system is a complicated mechanism but in short one can say that the lungs pump air through the windpipe to the surrounding environment via mouth and nose and speech sounds are formed during this process. The brain drives the muscle system that controls the lungs and vocal cords, the shape and volume of the windpipe, size and shape of the oral cavity, nasal passage controlling airflow through the nose, and finally, the lips. The signal can be though of arising from the speaker articulating the message that he wants to express. The audio signal is sampled and quantized. This digital speech signal is the input of speaker profile management and automatic speaker recognition systems.

2 Principles

Speaker recognition can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. All speaker recognition systems have to serve two distinguishes phases. The first one is referred to the enrollment sessions or training phase while the second one is referred to as the operation sessions or testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In case of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training samples. During the testing (Operational) phase, the input speech is matched with stored reference models and recognition decision is made. Speaker recognition methods can also be divided into text-independent and text-dependent methods. In a text-independent system, speaker models capture the characteristics of somebody's speech which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc. At the highest level, all speaker recognition systems contain two main modules: Feature Extraction and Feature Matching[1][5].

3 Feature Extraction

The main objectives of feature extraction are to extract characteristics from the speech signal that are unique to each individual which will be used to differentiate speakers. The purpose of this module is to convert the speech waveform to some type of parametric representation for further analysis and processing. This is often referred as the signal-processing front end. The speech signal is a slowly timed varying signal (It is called quasi-stationary). When examined over a sufficiently short period of time (Between 5 and 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, 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 others. MFCC is perhaps the best known and most popular, and these will be used here[4][5].

3.1 Preprocessing

The first step of speech signal processing involves the conversion of analog speech signal into digital speech signal. This is a crucial step in order to enable further processing. Here the continuous time signal (Speech) is sampled at a discrete time points to form a sample data signal representing the continuous time signal. The method of obtaining a discrete time representation of a continuous time signal through periodic sampling, where a sequence of samples, x[n] is obtained from a continuous signal s (t). It is apparent that more signal data will be obtained if the samples are taken closer together[4].

3.2 Frame Blocking

Framing is the process of segmenting the speech samples obtained from the analog to digital conversion into small frames with time length in the range of 20ms to 40 ms. In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The human speech production is known to exhibit quasi-stationary behavior over a short period of time (20ms to 40 ms)[1][2].

3.3 Windowing

It is necessary to work with short term or frames of the signal. This is to select a portion of the signal that can reasonably be assumed stationary. Windowing is performed on each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as w (n), then the result of windowing is given by-

$$y_1 = x_1(n)w(n), \qquad 0 \le n \le N-1$$

If Hamming window is used, then it has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),$$
$$0 \le n \le N-1$$

Where,

n = 0, 1, 2……………….N-1
N = Number of samples in each frame
$y_1(n)$ = Resultant signal after windowing

Typically Hamming window is used[1][3].

3.4 Mel-frequency Wrapping

Psychophysical studies have shown that 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. The mel-frequency scale has 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[1][6]. Therefore the following approximate formula is used to compute the mels for a given frequency f in Hz:

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

3.5 Mel Filterbank

One approach to simulating the subjective spectrum is to use a filter bank. That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The modified spectrum of S (w) thus consists of the output power of these filters when S (w) is the input. The number of mel spectrum coefficients K, is typically chosen as 20.

3.6 Cepstrum

In this final step, the conversion of the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). Discrete Cosine Transform (DCT) is used to convert them back to the time domain[1]. Therefore it leads to the definition of those mel power spectrum coefficients, the result of the last step, such as-

$$\tilde{S}_k, k = 1,2,......,K$$

Now MFCC's can be calculated as-

$$\tilde{c}_n = \sum_{k=1}^{K} \left(\log \tilde{S}_k\right) \cos\left[n(k-1/2)\pi / K\right]$$

$$n = 1,2,......,K$$

4 Feature Matching

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 objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in this case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. 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). The Vector Quantization approach will be used here due to ease of implementation and high accuracy.

4.1 Vector Quantization

Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. In the earlier days, the design of a vector quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. The results of the feature extraction are a series of vectors characteristic of the time-varying spectral properties of the speech signal. These vectors are 24 dimensional and are continuous. We can map them to discrete vectors by quantizing them. However, as we are quantizing vectors this is Vector Quantization. VQ is potentially an extremely efficient representation of spectral information in the speech signal[7][8].

4.2 The LBG-VQ Algorithm

In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ design algorithm based on a training sequence. The LBG-VQ design algorithm is an iterative algorithm which alternatively solves the optimality criteria. The algorithm requires an initial codebook. This initial codebook is obtained by the splitting method. In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are splitted into four and the process is repeated until the desired number of code vectors is obtained[3].

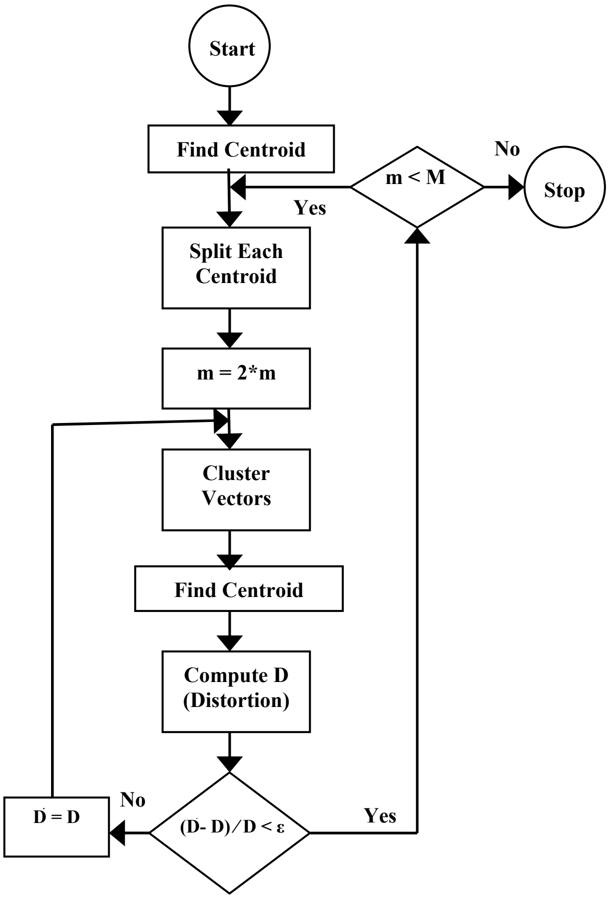

Figure 1 Flowchart of VQ-LBG Algorithm

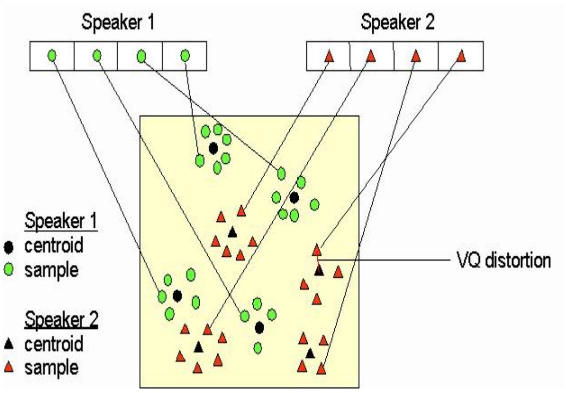

Figure 2 Conceptual Diagram to illustrate the VQ

5 Experimental Results

In this thesis, we attempted to identify speakers by comparing a sample of their voice to a database of speakers. There were two phases to out method. In the first phase, we created a "codebook" of the speakers to characterize their vocal characteristics using training sentences. In the second phase, we compared a sample of a speaker's voice to the codebook to determine the identity of the speaker.

We have taken 32 samples of different speakers. Then Triangular, Hanning and Hamming window is used respectively to test the recognition rate in linear scale of the system varying the number of centroids up to sixteen. Our system is at its best 100% accurate in identifying the correct speaker when Hamming window is used and number of centroids is eight or more.

Table 1: Recognition Rate (%) **[Linear Scale]**

No. of Centroids	Triang	Hann	Hamm
1	50.0	50.0	62.5
2	62.5	62.5	75.0
4	62.5	75.0	87.5
8	87.5	75.0	100
16	100	87.5	100

Similar tests are performed using the mel scale. Here, we included rectangular window. Because it's not possible to calculate the MFCC in linear scale using rectangular window as zero of log is zero. Our system also show 100% accurate result in identifying the correct speaker when Hamming window is used and number of centroids are four or more. Triangular window can have 100% accurate result if the number of centroids is sixteen or more.

Table 2: Recognition Rate (%) **[Mel Scale]**

No. of Centroids	Triang	Rect	Hann	Hamm
1	37.5	37.5	50.0	62.5
2	62.5	37.5	62.5	75.0
4	75.0	50.0	62.5	100
8	87.5	50.0	75.0	100
16	100	62.5	75.0	100

6 Conclusion

The unique characteristics of human speech, Mel Frequency Cepstrum Coefficients (MFCC) are used for feature extraction and Vector Quantization is used for feature matching technique. Clustering algorithm such as VQ-LBG algorithm is taken to implement the vector quantization for this purpose. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises. From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. Furthermore various other analyses could be carried out on Voiceprints, Age and Emotion Identification, Combination of Features, Spoken Language, Natural Language etc. from the voice pattern of human speech.

References:

[1] Minh N. Do, "*An Automatic Speaker Recognition System*" Swiss Federal Institute of Technology, Lausanne, Switzerland.

[2] Etienne Perron, Martin Klauser, and Minh N.Do, "*An Automatic Speaker Recognition System*", Digital Signal Processing Mini Project, January 2001.

[3] Y. Linde, A. Buzo & R. Gray, "*An algorithm for vector quantizer design*", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

[4] Othman O. Khalifa, S. Khan, Md. Rafiqul Islam, M. Faizal, and D. Dol "*Text Independent Automatic Speaker Recognition*", ICECE 2004, 28-30 December 2004 Dhaka, Bangladesh.

[5] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani, and Md. Saifur Rahman, "*Speaker Identification Using Mel-Frequency Cepstral Coefficients*", ICECE 2004, 28-30 December 2004 Dhaka, Bangladesh.

[6] Md. Zulfiquar Ali Bhotto and Md. Ruhul Amin, "*Bengali Text Dependent Speaker Identification Using Mel-Frequency Cepstrum Coefficient and Vector Quantization*", ICECE 2004, 28-30 December 2004 Dhaka, Bangladesh.

[7] comp.speech Frequently Asked Questions, http://svr-www.eng.cam.ac.uk/comp.speech/

[8] "Vector Quantization", http://www.data-compression.com/vq.html