Speaker Identification Using Mel Frequency Cepstral Coefficients

Mustafa Jamil
Tanta University
2 PUBLICATIONS **192** CITATIONS
SEE PROFILE

Golam Rabbani
University of Washington Seattle
13 PUBLICATIONS **258** CITATIONS
SEE PROFILE

Md. Saifur Rahman
Bangladesh University of Engineering and Technology
90 PUBLICATIONS **423** CITATIONS
SEE PROFILE

3rd International Conference on Electrical & Computer Engineering
ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh

SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS

Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman
Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology,
Dhaka-1000
E-mail: saif672@yahoo.com

ABSTRACT

This paper presents a security system based on speaker identification. Mel frequency Cepstral Co-efficients{MFCCs} have been used for feature extraction and vector quantization technique is used to minimize the amount of data to be handled .

1. INTRODUCTION

Speech is one of the natural forms of communication. Recent development has made it possible to use this in the security system. In speaker identification, the task is to use a speech sample to select the identity of the person that produced the speech from among a population of speakers. In speaker verification, the task is to use a speech sample to test whether a person who claims to have produced the speech has in fact done so[1]. This technique makes it possible to use the speakers' 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, security control for confidential information areas, and remote access to computers.

2. PRINCIPLES OF SPEAKER RECOGNITION

Speaker recognition methods can be divided into *text-independent* and *text-dependent* methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up *irrespective of what one is saying*. [1] 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. Every technology of speaker recognition, identification and verification, whether text-independent and text-dependent, each has its own advantages and disadvantages and may require different treatments and techniques. The choice of which technology to use is application-specific. At the highest level, all speaker recognition systems contain two main modules *feature extraction* and *feature matching* [2,3].

3. SPEECH FEATURE EXTRACTION

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate). The speech signal is a slowly time 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 0.2s or more) the signal characteristics 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 this feature has been used in this paper. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [1,4].

3.1 The MFCC processor

A block diagram of the structure of an MFCC processor is given in Figure 1. The speech input is recorded at a sampling rate of 22050Hz. This sampling frequency is chosen to minimize the effects of *aliasing* in the analog-to-digital conversion process. Figure 1. shows the block diagram of an MFCC processor .

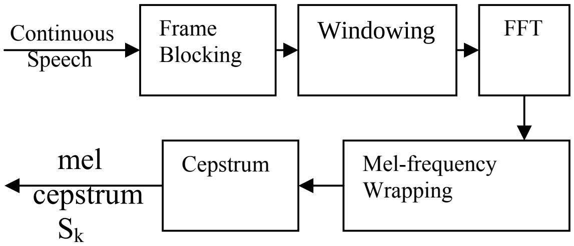

Figure 1 Block diagram of the MFCC processor

3.2 Mel-frequency wrapping

The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following formula to compute the mels for a given frequency *f* in Hz[5]:

mel(f)= 2595*log10(1+f/700) ……….. (1)

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

3.3 CEPSTRUM

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs). 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 are real numbers(and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT). The MFCCs may be calculated using this equation [3,5]:

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

where n=1,2,….K

The number of mel cepstrum coefficients, K, is typically chosen as 20. The first component, $\tilde{c}_0$, is excluded from the DCT since it represents the mean value of the input signal which carries little speaker specific information. By applying the procedure described above, for each speech frame of about 30 ms with overlap, a set of mel-frequency cepstrum coefficients is computed. This set of coefficients is called an *acoustic vecto*r. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker [4]. Therefore each input utterance is transformed into a sequence of acoustic vectors. The next section describes how these acoustic vectors can be used to represent and recognize the voice characteristic of a speaker.

4. FEATURE MATCHING

The state-of-the-art feature matching techniques used in speaker recognition include, Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). The VQ approach has been used here for its ease of implementation and high accuracy.

4.1 Vector quantization

Vector quantization (VQ) is a lossy data compression method based on principle of blockcoding [6]. It is a fixed-to-fixed length algorithm. VQ may be thought as an aproximator. Figure 2 shows an example of a 2-dimensional VQ.

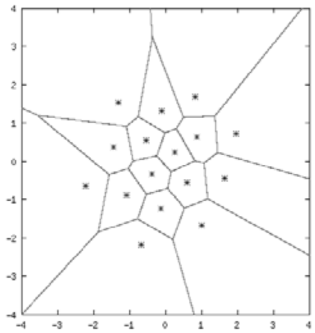

Figure 2 An example of a 2-dimensional VQ

Here, every pair of numbers falling in a particular region are approximated by a star associated with that region. In Figure 2, the stars are called *codevectors* and the regions defined by the borders are called *encoding regions*. The set of all codevectors is called the *codebook* and the set of all encoding regions is called the *partition* of the space [6].

4.2 LBG design algorithm

The LBG VQ design algorithm is an iterative algorithm (as proposed by Y. Linde, A. Buzo & R. Gray) which alternatively solves optimality criteria [7]. The algorithm requires an initial codebook. The initial codebook is obtained by the *splitting* method. In this method, an initial codevector is set as the average of the entire training sequence. This codevector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two codevectors are split into four and the process is repeated until the desired number of codevectors is obtained. The algorithm is summarized in the flowchart of Figure 3.

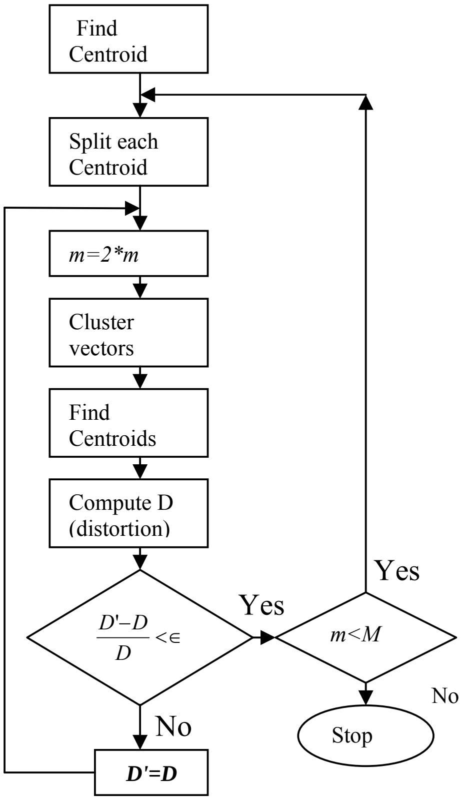

Figure 3 Flowchart of VQ-LBG algorithm

In figure 4 , only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from speaker 1 while the triangles are from speaker 2.

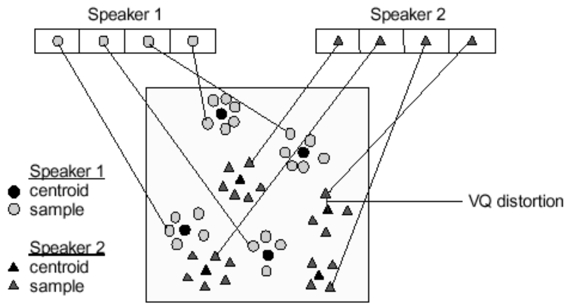

Figure 4 Conceptual diagram to illustrate the VQ Process.

In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The resultant codewords (centroids) are shown in Figure 4 by circles and triangles at the centers of the corresponding blocks for speaker1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with the smallest total distortion is identified. Figure 5 shows the use of different number of centroids for the same data field.

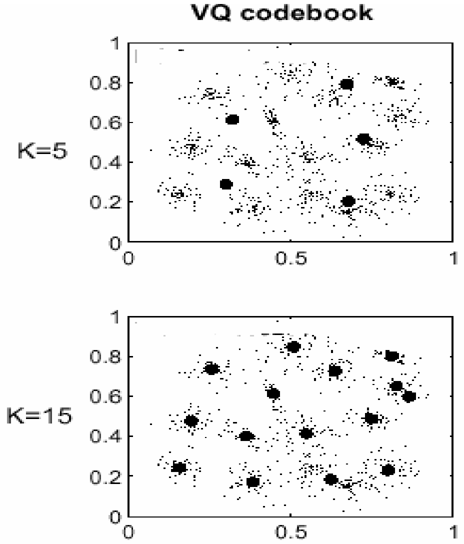

Figure 5 Pictorial view of codebook with 5 and 15 centroids respectively.

5. RESULTS

The system has been implemented in Matlab6.1 on windowsXP platform.The result of the study has been presented in Table 1 and Table 2. The speech database consists of 21 speakers, which includes 13 male and 8 female speakers. Here, identification rate is defined as the ratio of the number of speakers identified to the total number of speakers tested.

Table 1: Identification rate (in %) for different windows [using **Linear scale**]

Code book size	Triangular	Rectangular	Hamming
1	66.67	38.95	57.14
2	85.7	42.85	85.7
4	90.47	57.14	90.47
8	95.24	57.14	95.24
16	100	80.95	100
32	100	80.95	100
64	100	85.7	100

Table 2: Identification rate (in %) for different windows [using **Mel scale**]

Code book size	Triangular	Rectangular	Hamming
1	57.14	57.14	57.14
2	85.7	66.67	85.7
4	90.47	76.19	100
8	95.24	80.95	100
16	100	85.7	100
32	100	90.47	100
64	100	95.24	100

Table 1 shows identification rate when triangular, or rectangular, or hamming window is used for framing in a linear frequency scale. The table clearly shows that as codebook size increases, the identification rate for each of the three cases increases, and when codebook size is 16, identification rate is 100% for both the triangular and hamming windows.However, in case of Table2 the same windows are used along with a Mel scale instead of aLinear scale. Here, too, identification rate increases with increase in the size of the codebook. In this case, 100% identification rate is obtained with a codebook size of 4 when hamming window is used.

6. CONCLUSION

The MFCC technique has been applied for speaker identification. VQ is used to minimize the data of the extracted feature. The study reveals that as number of centroids increases, identification rate of the system increases. It has been found that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has to be increased as the number of speakers increases. The study shows that the linear scale can also have a reasonable identification rate if a comparatively higher number of centroids is used. However, the recognition rate using a linear scale would be much lower if the number of speakers increases. Mel scale is also less vulnerable to the changes of speaker's vocal cord in course of time.

The present study is still ongoing, which may include following further works. HMM may be used to improve the efficiency and precision of the segmentation to deal with crosstalk, laughter and uncharacteristic speech sounds. A more effective normalization algorithm can be adopted on extracted parametric representations of the acoustic signal, which would improve the identification rate further. Finally, a combination of features (MFCC, LPC, LPCC, Formant etc) may be used to implement a robust parametric representation for speaker identification.

REFERENCES

[1] Lawrence Rabiner and Biing-Hwang Juang, Fundamental of Speech Recognition”, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[2] Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. (1999). “Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification” in *IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1*, January 1999. IEEE, New York, NY, U.S.A.

[3] F. Soong, E. Rosenberg, B. Juang, and L. Rabiner, "A Vector Quantization Approach to Speaker Recognition", AT&T Technical Journal, vol. 66, March/April 1987, pp. 14-26

[4] Comp.speech Frequently Asked Questions WWW site, http://svr-www.eng.cam.ac.uk/comp.speech/

[5] Jr., J. D., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.

[6] R. M. Gray, ``Vector Quantization," IEEE ASSP Magazine, pp. 4--29, April 1984.

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