Speaker Recognition System Based On MFCC and VQ Algorithms

Nimesh V Bhimani
Electronics & Communication Department
Marwadi Education Foundation Group of Institutions, GTU
Gujarat, India

***Abstract*— The main aim of this paper is speaker recognition. This can be achieved by automatically identify who is speaking on the basis of individual information integrated in speech waves. Objective is comparing a speech signal from a unknown speaker to database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition needed two task, "feature extraction" and "feature classification". Feature classification further divided in two task, pattern matching and decision. For feature extraction, we are using mel frequency cepstral coefficient (MFCC) method. For feature classification, we are using vector quantization (VQ) method. In the feature matching stage "Euclidean distance" is applied.**

***Keywords*— MFCC, Vector quantization**

I. Introduction

Spoken language is the most natural way used by humans to communicate information. The speech signal conveys several types of information and speaker information. From the speech perception point of view, it also conveys information about the environment in which the speech was produced and transmitted [5].

In speaker identification, task is use the speech sample to select the identity of person that produced the input voice sample from the database. This technique makes it possible to use the speakers' voice to verify their identity and control access to services such as voice dialling, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers [2].

MFCC algorithm is used for extraction and vector quantization algorithm is used to reduce amount of achieved data in form of codebooks. These data are saved as acoustic vectors. In the matching stage, features of input command are compared with each codebook using Euclidean distance criterion [3].

It was Lawrence Kersta who made the first major step from speaker identification by humans towards speaker identification by computers when he developed spectrographic voice identification at Bell Labs in the early 1960s. His identification procedure was based on visual comparison of the spectrogram, which was generated by a complicated electro-mechanical device. Although the visual comparison method cannot cope with the physical and linguistic variation in speech, his work encouraged the introduction of automatic speaker recognition.

In the following four decades, speaker recognition research has advanced a lot. Some commercial systems have been applied in certain domains. Speaker Recognition technology makes it possible to use a person's voice to control the access to restricted services (automatic banking services), information (telephone access to financial transactions), or area (government or research facilities). It also allows detection of speakers, for example, voice-based information retrieval, recognition of perpetrator on a telephone tap,and detection of a speaker in a multiparty dialog [6].

II. Speaker Recognition principles

The goal of the speaker identification task is to determine which speaker out of a group of known speakers produces the input voice sample. There are two modes of operation that are related to the set of known voices. In the closed-set mode, the system assumes that the to-be-determined voice must come from the set of known voices. Otherwise, the system is in open-set mode. The closed-set speaker identification can be considered as a multiple-class classification problem. In open-set mode, the speakers that do not belong to the set of known voices are referred to as impostors. This task can be used for forensic applications, e.g., speech evidence can be used to recognize the perpetrator's identity among several known suspects.

According to the constraints placed on the speech used to train and test the system, Automatic speaker recognition can be further classified into text-dependent or text-independent tasks. In text-dependent recognition, the user must speak a given phrase known to the system, which can be fixed or prompted. The knowledge of a spoken phrase can provide better recognition results. In text-independent recognition, the system does not know the phrase spoken by the user. Although this adds flexibility to an application, it can have reduced accuracy for a fixed amount of speech.

III. Speech Feature Extraction

In this project the most important thing is to extract the feature from the speech signal. The speech feature extraction in a categorization problem is about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. As we know from the above fundamental formation of speaker identification systems, that number of training and test vector needed for the classification problem grows exponential with the dimension of the given input vector, so we need feature extraction.

The most commonly used acoustic vectors are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Prediction Cepstral (PLPC) Coefficients. All these features are based on the spectral information derived from a short time windowed segment of speech. They differ mainly in the detail of the power spectrum representation. MFCC features are derived directly from the FFT power spectrum, whereas the LPCC and PLPC use an all-pole model to represent the smoothed spectrum. The Mel-scale filter bank Centers and bandwidths are fixed to follow the Mel-frequency scale, giving more detail to the low frequencies. LPCC features can be considered as having adaptive detail in that the model poles move to fit the spectral peaks wherever they occur. The detail is limited mostly by the number of poles available. PLPC features are a hybrid between filter bank and all-pole model spectral representation.

The spectrum is first passed through a bark-space trapezoidal-shaped filter bank and then fit with an all-pole model. The detail of the PLP representation is determined by both the filter bank and the all-pole model order. The spectra representation is transformed to cepstral coefficients as a final step. This is done because of the (near) orthogonalizing property of the cepstral transformation. The filter bank representations are transformed directly by a Discrete Cosine Transform (DCT). The all-pole representations are transformed using the recursive formula between prediction coefficient and cepstral coefficients. In all cases, we have to discarding the zeroth cepstral coefficient results in energy normalization. PLPC and MFCC features are used in most state-of-the-art automatic speech recognition systems. MFCC features are used in more and more speaker recognition applications.

Following the trends in many state-of-the-art speaker recognition systems, MFCC coefficients are used as acoustic feature vectors in our project.

IV. MFCC COMPUTATION

A block diagram of an MFCC processor is given in the Figure 1. Speech input is recorded at a sampling rate of 22050 Hz, 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.

- In first step, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). Typical values for N and M are N= 256 and M= 100. We have to choose N value such a way that it is power of 2

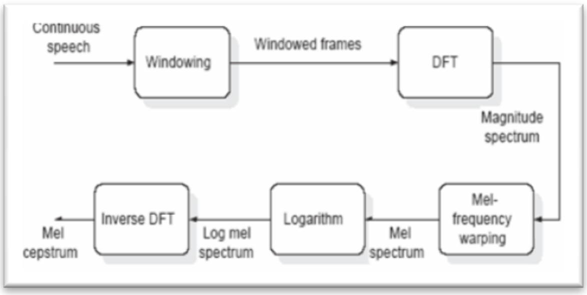

Figure 1 Block diagram of the MFCC process

- The next step is window each individual frame such that to minimize the signal discontinuities at the beginning and end of each frame. Usually Hamming window is used.
- Next step is the FFT, which converts each frame of N samples from the time domain into frequency domain.
- Next step is Mel Frequency Wrapping, which convert the frequency spectrum to the Mel spectrum.
- Next step is Cepstrum, which is the final step. In this step the log Mel spectrum is converted back to time. And the result is called MFCC.

V. VECTOR QUANTIZATION

Different techniques of feature matching are there such as Dynamic time wrapping (DTW), Gaussian mixture models (GMM), Hidden Markov model (HMM), Vector quantization (VQ). Vector quantization is most popular for text dependent speaker identification system. We have used VQ as Feature matching technique for our project.

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. In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ design algorithm based on a training sequence. The use of a training sequence bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to in the literature as an LBG-VQ.

It's a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution. The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features.

On a training sequence, *the use of a training sequence* bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to in the literature as an LBG-VQ.

Since storing every single vector that we generate from the training is impossible. By using Vector Quantization these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision.

VI. LBG-VQ DESIGN ALGORITHM

The LBG VQ design algorithm is an iterative algorithm which alternatively solves the following two optimality criteria.

NEAREST NEIGHBOUR CONDITION:

- This condition says that the encoding region should consists of all vectors that are closer to n-Code vector than any of the other code vectors. For those vectors lying on the boundary any tie-breaking procedure will do.

CENTROID CONDITION:

- This condition says that the n-code vector should be average of all those training vectors that are in encoding region. In implementation, one should ensure that at least one training vector belongs to each encoding region.

The algorithm requires an initial codebook C(0). 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.

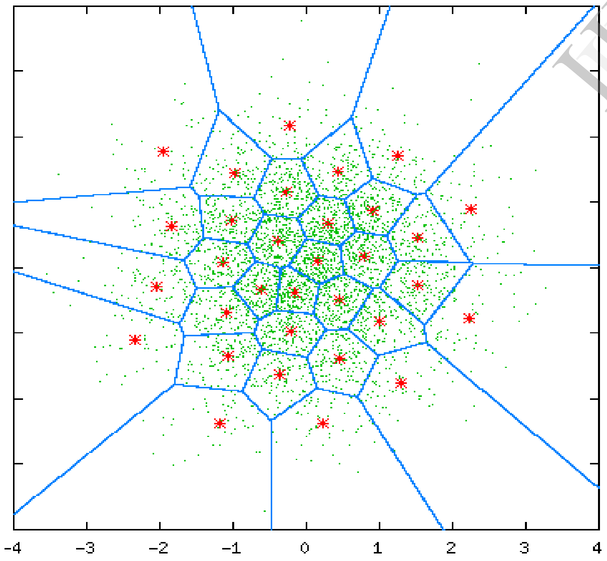

Figure 2 Example of a 2-Dimensional VQ

VII. 7. SCOPE OF FUTURE WORK

Our Future aspect is to design entire codebook and to use Euclidean distance technique for producing the desired result with minimum error probability. If accuracy will not be up to the mark we will try to improvise the algorithm.

VIII. CONCLUSION

There is trade- off between the time duration of speech sample and the probability of error. As the time duration of speech sample is increased, more amounts of data feature vectors can be extracted. So more amount of information for the particular speaker is obtained which helps to recognize the speaker so probability of error will be decreased. If time duration of speech sample is decreased, less amounts of data feature vectors will be extracted. So less amount of information for the particular speaker will be obtained, so probability of error will be increased.

REFERENCES

[1] Joseph P. Campbell, Jr. Department of Defence Fort Meade, MD j.campbell@ieee.org “SPEAKER RECOGNITION”.

[2] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman “SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS”.

[3] Mahdi Shaneh, and Azizollah Taheri “VOICE COMMAND RECOGNITION SYSTEM BASED ON MFCC AND VQ ALGORITHMS”.

[4] Vibha Tiwari International Journal on Emerging Technologies 1(1):19-22(2010). “MFCC AND ITS APPLICATIONS IN SPEAKER RECOGNITION”.

[5] Mr. Mohammed Imdad N, Prof. Shameem Akhtar N, Prof. Mohammad Imran Akhatar international Journal of advanced Research in computer science and electronic engg. Volume 1, Issue 4, June 2012.

[6] Qin jin CMU-CS-07-001 “ROBUST SPEAKER RECOGNITION” JANUARY 2007.

[7] WU Zunjing, CAO Zhigang TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 05/23 pp158-161 volume 10, Number 2, April 2005 **.”I**MPROVED MFCC-BASED FEATURE FOR ROBUST SPEAKER IDENTIFICATION”.

[8] Yoseph Linde, member IEEE, Andres Buzo, member IEEE, and Robert M. Gray, Senior member IEEE ”AN ALGORITHM FOR VECTOR QUANTIZER DESIGN”.

[9] Ahsanul Kabir, Sheikh Mohammad Masudul Hasan 6th WSEAS international conference CAIRO, EGYPT Dec 29-31, 2007.”VECTOR QUANTIZATION IN TEXT DEPENDENT AUTOMATIC SPEAKER RECOGNITION USING MEL- FREQUENCY CEPSTRUM COEFFICIENT”.

[10] Tomi Kinnunen,Haizhou Li,” AN OVERVIEW OF TEXT-INDEPENDENT SPEAKER RECOGNITION”, accepted 20 August 2009 elsevier.