IDENTIFICATION AND VERIFICATION OF SPEAKER USING MEL FREQUENCY CEPSTRAL COEFFICIENT

Viplav Gautam¹ (EC09125), Saurabh Sharma² (EC09096), Swapnil Gautam³ (2003PH10620), Gaurav Sharma⁴ (2011UMT1850)
ECE Department, Global Institute of technology, Jaipur¹,², Engineering Physics
Department, IIT, Delhi³, Metallurgy Department, MNIT, Jaipur⁴
viplovecgautam@gmail.com¹
sgnsharma123@gmail.com²
swapnilgautamiitd@gmail.com³
gauravsharma000005@yahoo.in⁴

ABSTRACT

Speech processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. Feature extraction is the most important step for speaker recognition. In this work, the Mel Frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. MFCC is based on the human peripheral auditory System. Generally, MFCC for feature extraction is used to improve the efficiency of speaker recognition.

KEYWORDS:
Feature extraction, Feature Matching, Mel frequency Cepstral coefficients (MFCC), Speaker recognition

INTRODUCTION

As human beings, we are able to recognize someone just by hearing him or her talk. Usually, a few seconds of speech are sufficient to identify a familiar voice. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of speaker recognition is to Extract, characterize and recognize the information about Speaker identity. At the primary level, speech conveys a message via words. But at other levels speech conveys information about the Language being spoken and the emotion, gender and, generally, the
identity of the speaker. To study the spectral properties of speech signal the concept of time varying Fourier representation is used.

Speaker recognition is basically divided into two-classification: speaker recognition and speaker identification and it is the method of automatically identify who is speaking on the basis of individual information integrated in speech waves. Speaker identification is the task of determining who is talking from a set of known voices or speakers and Speaker verification is the task of determining whether a person is who he/she claims to be. The main aim of this project is speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers. Below figure shows the fundamental formation of speaker identification and verification systems.

![Diagram of speaker identification and verification systems]

(a) Speaker identification

(b) Speaker verification

**Figure** Basic structures of speaker recognition systems
Speaker recognition can also divide into two methods, text-dependent and text independent methods. In text dependent method the speaker to say key words or sentences having the same text for both training and recognition trials. Whereas in the text independent does not rely on a specific texts being speak. Formerly text dependent methods were widely in application.

Like any other pattern recognition systems, speaker recognition systems also involve two phases namely, training and testing. Training is the process to upload the system with the voice characteristics of the speakers registering. Testing is the actual recognition task. The block diagram of training phase is shown in Figure below. In training phase the voice characteristics of the speaker are extracted from the training utterances and are used for building the reference models. During testing, similar feature vectors are extracted from the test utterance, and the degree of their match with the reference is obtained using some matching technique. The level of match is used to arrive at the decision.

![Figure](image_url)  
**Figure**: The block diagram of training phase.

![Figure](image_url)  
**Figure**: The block diagram of testing phase.
II. Speech Feature Extraction

A. Introduction

The purpose of this module is to convert the speech waveform, using digital signal processing (DSP) tools, to a set of features for further analysis. This is often referred as the signal-processing front end.

The speech signal is a slowly timed varying signal. An example of speech signal is shown in Figure below. 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 change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal.

B. Steps of MFCC

Step 1- Frame Blocking

In this step the continuous speech signal is blocked into frames of $N$ samples, with adjacent frames being separated by $M$ ($M < N$). The first frame consists of the first $N$ samples. The second frame begins $M$ samples after the first frame, and overlaps it by $N - M$ samples and.

Figure Example of 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 will be discussed in this paper, because MFCC is perhaps the best known and most popular and also, it shows high accuracy results for clean speech and also experiments show that the parameterization of the Mel frequency Cepstral coefficients is best for discriminating speakers and is different from the one usually used for speech recognition applications.
Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N - 2M samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 and M = 100.

**Step 2- 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. 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), 0 ≤ n ≤ N - 1, where N is the number of samples in each frame, then the result of windowing is the signal

\[ y_i(n) = x_i(n)w(n), \quad 0 \leq n \leq N - 1 \]

Typically the Hamming window is used, which has the form:

\[ w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N - 1 \]

**Step 3- Fast Fourier transform**

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 a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples \( \{x_n\} \), as follow:

\[ X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \quad k = 0,1,2,\ldots, N - 1 \]

**Step 4- Mel-frequency Wrapping**

In sound processing, MFCC’s are based on the known variation of the human ear’s critical bandwidths. It is derived from the Fourier Transform of the audio clip. In this technique the frequency bands are positioned logarithmically, whereas in the Fourier Transform the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. These coefficients allow better processing of data. Each tone with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the ‘Mel Scale’. The Mel frequency scale is linear frequency spacing below 1000 Hz and logarithmic spacing above 1 kHz. 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 formula to determine the Mels for a given frequency f in Hz.

\[ \text{Mel}(f) = 2595 \times \log_{10}(1 + f/700). \]

To obtain the subjective spectrum we use a filter bank which is spaced uniformly on the Mel scale is described on the figure below. 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.
Step 5-**Cepstrum**

Cepstrum name was derived from the spectrum by reversing the first four letters of spectrum. We can say Cepstrum is the Fourier Transformer of the log with unwrapped phase of the Fourier Transformer.

Mathematically we can say Cepstrum of signal = FT (log (FT (the signal)) +j6.28m)

Where m is the integer required to properly unwrap the angle or imaginary part of the complex log function. Algorithmically we can say – Signal - FT - log - phase unwrapping - FT - Cepstrum.

We can calculate the Cepstrum by many ways. Some of them need a phase-warping algorithm, others do not. Figure below shows the pipeline from signal to Cepstrum.

![Signal to Cepstrum Pipeline](image)

**Figure** Signal to Cepstrum Pipeline
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.

\[
\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos \left( n \left( k - \frac{1}{2} \right) \frac{\pi}{K} \right), \quad n = 0, 1, \ldots, K-1
\]

Where \( \tilde{S}_0 \), \( k = 0, 2, \ldots, K-1 \)

\( \tilde{c}_0 \), is excluded from DCT because it represents the mean value of the input signal, which carried little speaker specific information.

The complete figure which shows the calculation of the Mel frequency Cepstrum coefficient is shown below.

III. Speech Feature Matching

A. Vector Quantization (VQ)

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 codeword’s is called a codebook.

The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. Hence, Vector Quantization is also suitable for lossy data compression. It is a fixed-to-fixed length algorithm. VQ may be thought as an approximator. Figure shows an example of a 2-dimensional VQ.
Here, every pair of numbers falling in a particular region is approximated by a star associated with that region. In Figure 2, the stars are called codevectors and the regions defined by the borders Frame Blocking Windowing FFT Cepstrum Mel-frequency Wrapping 567 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.

B. 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. 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.
IV. CONCLUSION

The main idea of this paper was to discuss a speaker recognition system that could be applied to a speech of an unknown speaker. By determining the extracted features of the unknown speech and then comparing them to the stored extracted features for each different speaker in order to identify the unknown speaker.

The feature extraction was done by using MFCC (Mel Frequency Cepstral Coefficients). The figure below shows the result of Cepstral Coefficient Calculations for five users and first ten DCT coefficients are Cepstral coefficients. Each user having five vocalization of the word “hello”. Then it was averaged and represented in tabular form named “table”. Each column corresponds to a given speaker. The next column denoted as “ceps2” is Cepstral coefficient of 2nd speaker. We can clearly see its resemblance to 2nd column of “table”. The result obtained is shown in the next page.
The speaker was building up using Vector Quantization (VQ). By clustering the training feature of each speaker we produce the VQ codebook and then stored in the speaker database. In this method, the K-means algorithm was used for clustering purpose. In the recognition stage, a distortion measure which based on the minimizing the Euclidean distance was used when matching an unknown speaker with the speaker database. VQ based clustering approach is best as it provides us with the faster speaker identification process.

**Figure** Result of Cepstral Coefficient Calculation

**Figure** Result of Speaker Recognition
ACKNOWLEDGEMENT

To complete this work, we have got valuable suggestions and guidance from Dr. Rajesh Kumar (Dept. of electrical engineering, MNIT, Jaipur). I also thankful to all the researcher and authors of the manuscript from where I got valuable information for completing the work.

REFERENCES


