IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013)

Speech Recognition Using Euclidean Distance

Akanksha Singh Thakur[1], Namrata Sahayam[2]

[1]ME IV Sem, [2]Asst Professor

***Abstract*-Digital processing of speech signal and voice recognition algorithm is very important for fast and accurate automatic voice recognition technology. The voice is a signal of infinite information. A direct analysis and synthesizing the complex voice signal is due to too much information contained in the signal. Therefore the digital signal processes such as Feature Extraction and Feature Matching are introduced to represent the voice signal. This paper describes an approach of speech recognition by using the Mel-Scale Frequency Cepstral Coefficients (MFCC) extracted from speech signal of spoken words. Verification is carried out using a weighted Euclidean distance. For speech recognition we implement the MFCC approach using software platform MatlabR2010b.**

***Keyword*— Feature Extraction, Feature Matching, Mel Frequency Cepstral Coefficient (MFCC), Euclidean distance, Vector Quantization.**

I. INTRODUCTION

Speech is the most natural way to communicate for humans. While this has been true since the dawn of civilization, the invention and widespread use of the telephone, audio-phonic storage media, radio, and television has given even further importance to speech communication and speech processing. The advances in digital signal processing technology has led the use of speech processing in many different application areas like speech compression, enhancement, synthesis, and recognition. In this paper, the issue of speech recognition is studied and a speech recognition system is developed forward using MFCC algorithm [1].

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 recognition is widely applicable in use of speaker's voice to verify their identity and control access to services such as banking by telephone, database access services, voice dialing telephone shopping,

Information services, voice mail and security control for secret information areas. Speaker recognition technology is the most potential technology to create new services that will make our everyday lives more secured. Another important application of speaker recognition technology is for forensic purposes.

Speaker recognition has been seen an appealing research field for the last decades which still yields a number of unsolved problems. The main aim of this paper is speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker [2, 5].

II. SPEECH RECOGNITION

A Recognition Algorithms

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and recognition (Matching) of the spoken word.

B Feature Extraction (MFCC)

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech [1, 3]. The overall process of the MFCC is shown in Figure I.

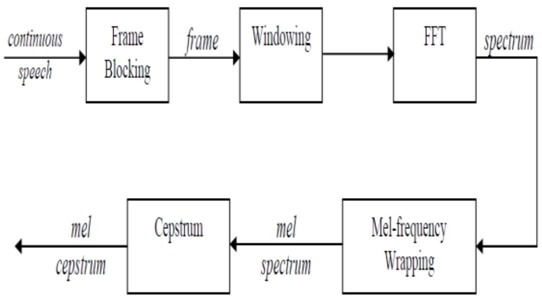


Figure I Block diagram of MFCC[2]

As shown in Figure I MFCC consists of six computational steps. Each step has its function and mathematical approaches as discussed briefly in the following:

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256

Step 3: Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

$$Y(n) = X(n) \times W(n) \qquad Y(\omega) = 0.54 - 0.46cos[{}^{2\pi n}/_{N-1}]0 \leq n \leq N-1$$

If the window is defined as W (n), 0 ≤ n ≤ N-1 where

N = number of samples in each frame
Y[n] = Output signal
X (n) = input signal
W (n) = Hamming window

Step 4: Fast Fourier Transform

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$Y(\omega) = FFT[h(t) * X(t)] = H(\omega)X(\omega)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale [1]. The bank of filters according to Mel scale as shown in. Fig. II.

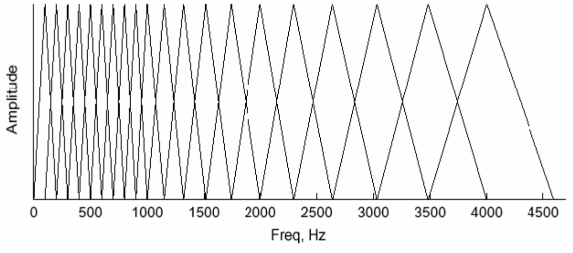


Fig II Mel scale filter bank[1]

This figure II shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale [3]. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

$$m = 1127.01048log_e(1 + {}^{f}/_{700})$$

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

C Vector Quantization

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 centroid [2,4] The collection of all code words is called a codebook.

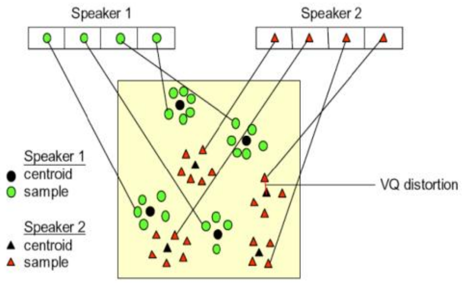


Figure III [3]

IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013)

One speaker can be discriminated from another based of the location of centroid Fig III shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result code words (centroids) are shown in Figure III by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion [4,7]. 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 smallest total distortion is identified.

D Distance measure:

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vector {x1, x2 ….xi), and then it is compared with the codebooks from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance[6].The formula used to calculate the Euclidean distance can be defined as following:

The Euclidean distance between two points P = (p_1, $p_2 \ldots p_n$) and Q = (q_1, $q_2 \ldots q_n$),

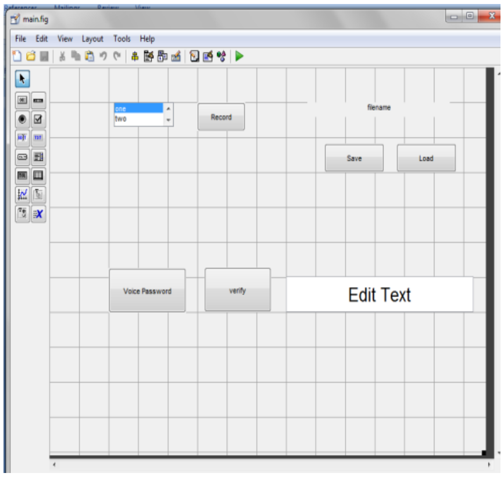
$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

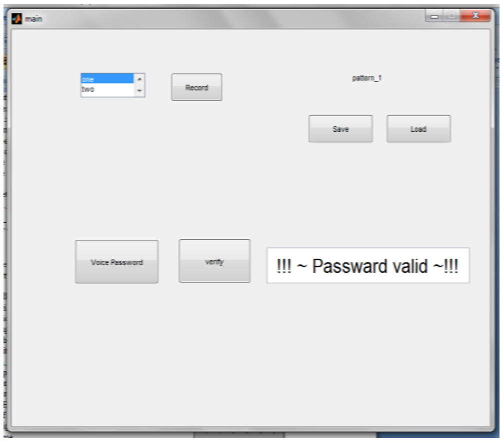
III. Simulation And Results

The simulation was carried out in MATLAB. The input speech signal considered in my work is word "HELLO". We are using GUI for recording input signal, we save it in codebook then we use it as voice password and verify with previously stored data (shown in figure) if this voice password is match with stored data we get password valid in edit text this means we get correct password and have minimum distortion distance, this is shown in figure IV and V.



Edit Text

Figure IV



!!! ~ Passward valid ~!!!

Figure V

IV. Conclusion

These papers discussed speaker recognition system, and apply it to a speech of an unknown speaker. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the unknown speaker. The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients). The speaker was modeled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. In Matlab simulation we get minimum Euclidean distance.

REFERENCES

[1] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi "voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques" Issue 3, March 2010, ISSN 2151-9617.

[2] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, **"**Speaker identification using mel frequency cepstral coefficients" ICECE *2004, 28-30* December 2004, Dhaka, Bangladesh.

[3] Ahmed Mezghani, Douglas O'Shaughnessy, "Speaker Verification Using a New Representation Based on a Combination of MFCC and Formants".

[4] Sheeraz Memon, Margaret Lech and Ling He "Using Information theoretic vector quantization for inverted MFCC based speaker verification" IEEE CCECE/CCGEI, Saskatoon, May 2005.

[5] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun "An efficient MFCC extraction method in speech recognition" Department of Electronic Engineering" The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006.

[6] Jon Gudnason and Mike Brookes "Voice source cepstrum coefficient for speaker identification" 2008 IEEE.

[7] "1999 IEEE.Donghoon Hyun and Chulhee Lee "Optimization of mel-cepstrum for speech recognition