Control System with Speech Recognition Using MFCC and Euclidian Distance Algorithm

Hiren Parmar
Electronics & Communication Department, Dr. Subhash Technical Campus, Gujarat, India

Bhagwan Sharma
Electronics & Communication Department RKDF Institute of science & Tech., MP, India

Abstract

In this paper we describe the implementation of control system with speech recognition. To implement this, we used the MFCC and Euclidian distance algorithm. Using COLEA tool we give the input acoustic wave as a speech signal. In this paper, the simulation of simple digital hearing aid was developed using MATLAB programming language. Speaker recognition systems contain two main modules: Speaker Identification and Speaker Verification. With the help of MFCC we extract the information from the recognized speech signal. MFCC, the main advantage is that it uses Mel frequency scaling which is very approximate to the human auditory system. We also used VQLBG algorithm (as proposed by Y. Linde, A. Buzo & R. Gray) to generate the codebook and after that using the Euclidian distance algorithm we compare the codebook with stored data base. The primary objective of this paper is to compare and summarize some of the well known methods used for speech recognition.

Keywords: Speech recognition, MFCC, Feature Extraction, VQLBG, *Automatic Speech Recognition (ASR)*

1. Introduction

Speech is the most natural way of communication. Speech is the most basic, common and efficient form of communication method for people to interact with each other. Automatic Speech Recognition (ASR) system which allows a computer to identify the words that a person speaks into a microphone or telephone and convert it into written text. Mel frequency Cepstral Coefficients (MFCC) are one of the most popular spectral features in ASR.

1.1. Speech Recognition Component

Fig1.1 shows the basic component for speech recognition algorithm.

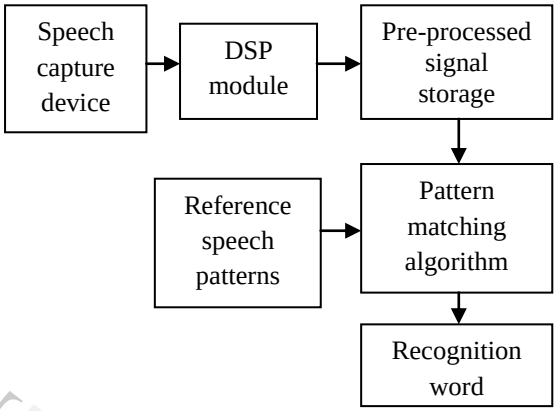

Fig 1.1 Component of Speech Recognition

All speaker recognition systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. It is also subdivided into two parts:

(i) Speaker verification

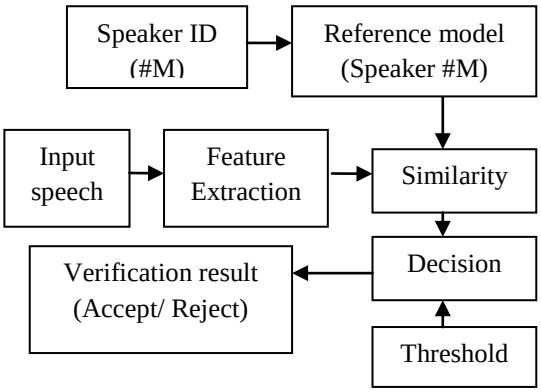

Fig 1.2 Speaker Verification

(ii) Speaker identification

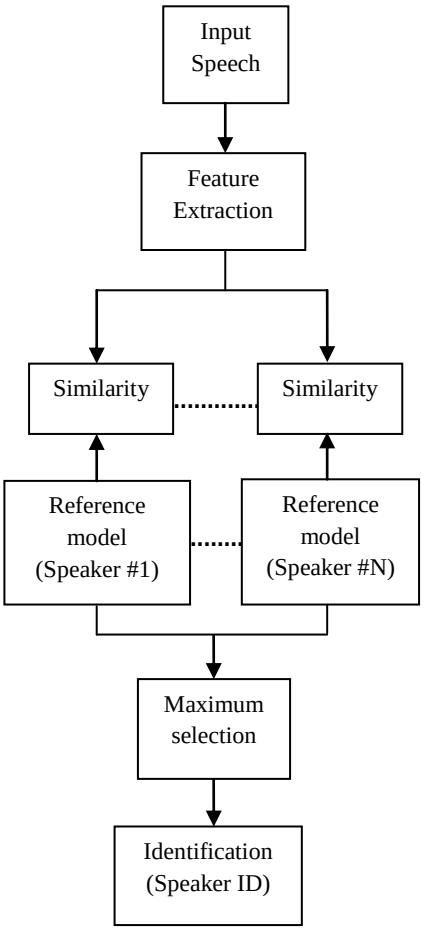

Fig 1.3 Speaker Identification

1.2. Feature Selection and Measures

To apply mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. The selection of appropriate features along with methods to estimate (extract or measure) them is known as feature selection and feature extraction.

Pattern-recognition models are divided into three components: feature extraction and selection, pattern matching, and classification. In speaker verification, the goal is to design a system that minimizes the probability of verification errors. Thus, the objective is to discriminate between the given speaker and all others.

2. Techniques of Feature Extraction

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Some of the audio features that have been successfully used for audio classification include Mel-frequency cepstral coefficients (MFCC), linear predictive coding (LPC), and Local discriminate bases (LDB). Few techniques generate a pattern from the features and use it for classification by the degree of correlation. Few other techniques use the numerical values of the features coupled to statistical classification method.

2.1. LPC

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding .The coefficients of the difference equation (the prediction coefficients) characterize the formants.

2.2. MFCC

MFCC is based on the human peripheral auditory system. The 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 *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 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

2.3. LDB

LDB is an audio feature extraction and a multi group classification scheme that focuses on identifying discriminatory time-frequency subspaces. Two dissimilarity measures are used in the process of selecting the LDB nodes and extracting features from them. The extracted features are then fed to a linear discriminate analysis based classifier for a multi-level hierarchical classification of audio signals.

3. Mel Frequency Cepstral Coefficients

The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly affects the recognition

performance. A compact representation would be provided by a set of Mel-frequency Cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a Mel-frequency scale. The MFCCs are proved more efficient. The calculation of the MFCC includes the following steps.

3.1. Mel-Frequency Wrapping

Human perception of frequency contents of sounds for speech signal 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 is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz .As a reference point ,the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 Mels. Therefore we can use the following approximate formula to compute the Mels for a given frequency *f* in Hz.

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

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired Mel-frequency component. 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 Mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a Mel frequency scale.

3.2. Cepstrum

In this final step, we convert the log Mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).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 (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). 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.

$$Cn = \sum_{K=1}^{K} (\log Sk) \cos\{n\left(k - \frac{1}{2}\right) * \frac{\pi}{k}\},$$

Whereas $n = 1, 2, \ldots k$
$Sk, K = 1, 2, \ldots K$ are the outputs of last step.
Complete process for the calculation of MFCC is shown in fig 1.4

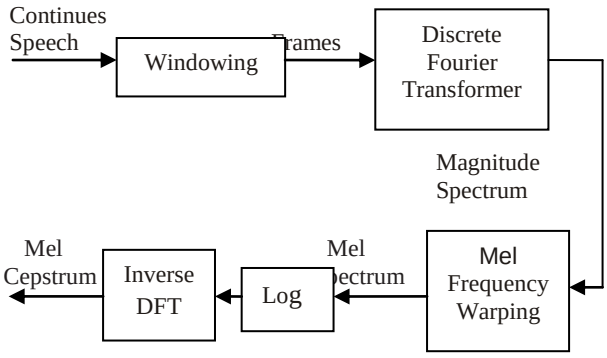

Fig 1.4 Complete pipeline for MFCC

4. Euclidean Distance

In vector quantization function, depending on the size determined for the codebook, training patterns are chosen to form the code vectors that make up a codebook; Euclidean distance is calculated between Centroids and Cepstrum. DIST is the minimum of Euclidean distance. This returns the pair wise Euclidean distance between columns of two matrixes. This distance will be returned and stored as results 1 for a comparison between identity claimed by the speaker and any one of previously recorded file.

The Euclidean distance measure is the" standard" distance measure between two vectors in feature space (with dimension DIM):

$$d^2 \text{Euclid}(x, p) = \sum_{i=0}^{\text{DIM}-1} (xi - pi)^2$$

To calculate the Euclidean distance measure, you have to compute the sum of the squares of the differences between the individual components of x and p. This can also be written as the following scalar product:

$$d^2 \text{Euclid}(x, p) = (xi - pi).(xi - pi)'$$

5. Flow of Software Implementation

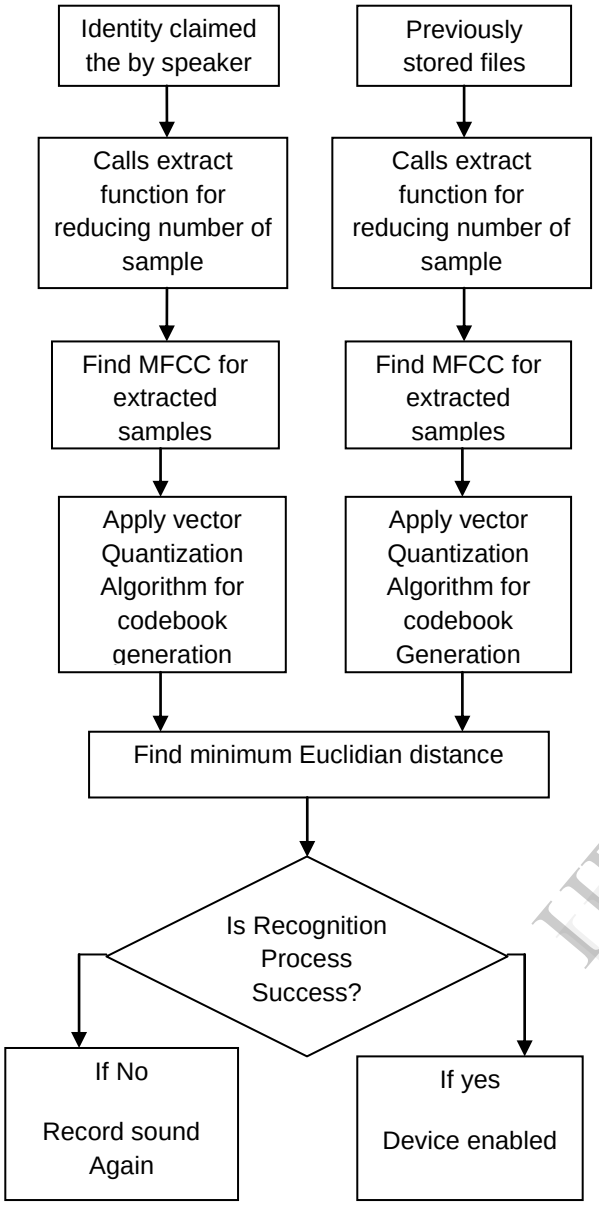

Fig1.5 Flow of Software Implementation

6. Simulation Result

Figure 1.6 shows the acoustic wave as L.WAV input signal. The input signal is screenshot of the displayed wave in the colea tool. The input signal is represent the original wave which is produced from the speech of the human. Over here the speech input signal is represented between X-axis and Y-axis, where X-axis shows Time and Y-axis shows Amplitude.

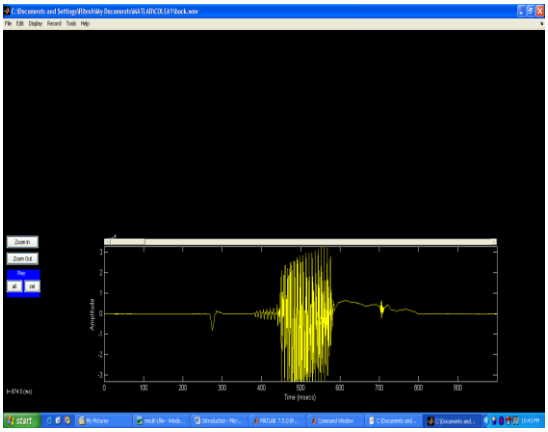

Fig 1.6 Original wave for L.WAV

The code written in MATLAB, loads the input wave signal, takes the sampling frequency and the number of bits of that signal. Then, Adaptive White Gaussian Noise (AWGN) and random noise are added to the signal before they are processed by various MATLAB function to get an output which is audible to the hearing impaired person.

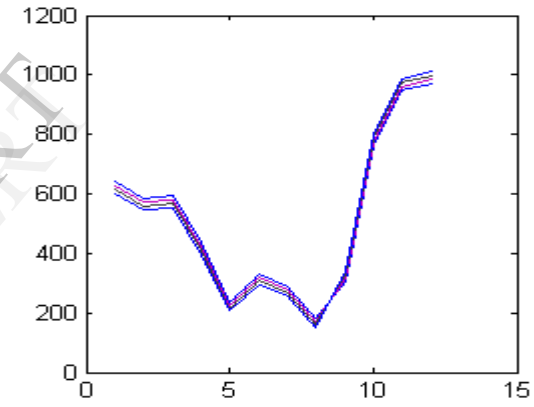

Fig 1.7 Euclidean distance of L.WAV signal

The Euclidean distance measure is the "standard" distance measure between two vectors in feature space. In this step we find the minimum Euclidean distance between the two codebook one which is already generated as stored database and the other one which is generated from the speaker input signal.

7. Conclusion

The goal of this paper was to implement a speaker-independent, device control using speech recognition. The feature extraction is done using Mel Frequency Cepstral Coefficients (MFCC). The purpose of the speech is communication. There is several way of characterizing the communication potential of speech. The speech can be efficiently processed with the help of the digital signal processing techniques. The speech signal can be modelled using vector quantization technique. Also

using the extraction feature, a codebook for each digit is built by clustering the feature vector. Codebook for all digits is collected in database. Euclidean Distance is used to compare the test speech with the speech stored in the database. Based on the results, data was send to Parallel Printer Port to the computer & using different command devices are controlled.

8. Future Work

In this paper, we have just considered few input signal as an example and with this we controlled only few devices. This can be extended up to more number of devices. This project can be downloaded on hardware by using embedded chip. This will avoid the use of computer. Size of hardware will be reduced to a great extent. We have done this project using VQLBG. It can be done by various other ways such as HMM (Hidden Markov Model) as well as DTW (Dynamic Time Warping).This project can also be minimized by making it wireless by using transmitter for Speech Signal and receiving it at another nearer place.

9. References

[1] L.R. Rabiner and R.W. Schafer, "Digital Processing of speech signals", Printice Hall Signal processing series, second edition.
[2] Wai C. Chu, " Speech coding algorithms, foundation and Evaluation of standard of Standardized coders", A john wiley & sons, inc. publication, 2003
[3] Sheikh Hussain sheikh sallesh, Ahmed zuri, Zulkarnian Yusoff, syed Rahman Lim, soon chieh, "Implemation of speaker identification system by means of personal computer".
[4] Thomas F. Quatieri, "Discrete- Time speech signal processing principle and practice", Pearson education signal processing series, first Indian reprint.
[6] Luc Vincent, "Exact Euclidean Distance Function by Chain Propagations" Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 520-525, Maui, Hawaii June 1991.
[7] Allen Gersho and Robert M. Gray, "Vector quantization and signal compression", Kluwer Academic publishers, 4th printing, 1995, Netharland.
[8] John G. Proakis, Dimitirs G. Manolakis, " Digital signal processing principles, Algorithm, and Application", Pearson education, fourth edition.
[9] Kinnunen T., Karpov E., Franti P. Audio, "Real-time speaker identification and verification", Speech and Language Processing, IEEE, Jan. 2006, Volume: 14 Issue:1,On page(s): 277 – 288 Page 59
[10] Sadaoki Furui, Speaker Characterization in Speech Technology, "Speaker-dependent-feature extraction, recognition and processing techniques" Volume 10, Issues 5–6, December 1991, Pages 505–520
[11] Shanks. J.E., Wilson. R.H.Larson, Williams. D., "Speech recognition performance of patients with sensor neural hearing loss under un aided and aided conditions using linear and compression hearing aids". 2002 Aug :23(4), page No. 280-90.
[12] Moore. B.C., Stainsby . T.H., Alcantara. J.I., Kuhnel. V., "The effect of speech intelligibility of varying compression time constants in a digital hearing aid". International Journal on Audio logy. 2004 Jul-Aug: 43(7), Page No. 399-409.
[13] Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. "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.
[14] 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 Page 60
[15] R. M. Gray, ``Vector Quantization", IEEE ASSP Magazine, pp. 4--29, April 1984.
[16] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Process. Mag.*,vol. 13, no. 5, pp. 58–71, Sep. 1996.
[17] M. Rahim,Y. Bengio, and Y. Lecun, "Discriminative feature and model design for automatic speech recognition," in *Proc. Eurospeech'97*, Rhodes, Greece, 1997, pp. 75–78.