Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 78 (2015) 29733-29737

ISSN: 2229-712X

Finite vocabulary text dependent speaker identification and speech recognition

Suma Swamy[1], Radha.K[2], Shwetha Rani G[2], Smitha Crystal D'Almeda[2], Sunil M[2] and K.V Ramakrishnan[1]
[1]Department of Electronics and Communications Engineering, Research Scholar, Anna University of Technology, Coimbatore
[2]Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore.

ARTICLE INFO

Article history:
Received: 18 August 2011;
Received in revised form:
21 December 2014;
Accepted: 9 January 2015;

Keywords
MFCC,
VQ,
DSP,
FFT,
DCT.

ABSTRACT

The major aim of bringing forth this project is to design a system for Text Dependent Speech Recognition. The speech recognition system contains two main modules, namely "feature extraction" and "feature matching". In this project, the MFCC (Mel-Frequency Ceptral Co-efficient) is used to simulate the feature extraction module. In MFCC algorithms the Ceptral co-efficients are calculated on the Mel frequency scale where the frequency bands are equally spaced on the Mel scale. For the reduction of amount of data in order to reduce the computation time the VQ (Vector Quantization) is used. VQ is the data compression method based on the principle of block coding. Because of the accuracy of the used algorithms the accuracy of this speech recognition system is high. Using these algorithms we achieve the text dependent speaker identification and speech recognition and thus providing improved efficiency

Introduction

Speech Processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation, so speech processing is regarded as a special digital signal processing. Digital signal processing (DSP) is concerned with the representation of discrete time signals by a sequence of numbers or symbols and the processing of these signals. DSP includes subfields like audio and speech signal processing.

In sound processing the MFC (Mel-Frequency Ceptrum) is used to represent the short-term power spectrum of sound. It is based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency. Mel-Frequency Ceptral Co-efficients (MFCC) are the co-efficients that collectively make up an MFC. The MFCC values are usually normalized in speech recognition systems to lessen the influence of noise as these values are sensitive to the additive noise.

VQ (Vector Quantization) is a data compression method based on the principle of "Block Coding". The design of VQ earlier was considered to be a challenging task due to the need for multi-dimensional integration. In 1980 Linde, Buzo and Gray (LBG) proposed a design of VQ algorithm that is based on a training sequence. Thus the use of training sequence bypasses the need for multi-dimensional integration. Thus the VQ designed this way is also called as LBG VQ.

In this paper section II deals with Text Dependent Speech Recognition System, section III presents the Experimental Results, section IV is the Conclusion.

Text dependent speech recognition system

Speech signal processing

Speech signal processing refers to the acquisition, manipulation, storage, transfer and output of vocal utterances by a computer. The main applications are the recognition, synthesis and compression of human speech. The various applications are explained in detail as follows:

Speaker Recognition: Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's 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.

Unknown Speaker requesting access
Speakers known to the recognition system
Which of the speakers in the recognition system is the unknown speaker ?

Fig.1 depicts the known and unknown speaker to the system [7]

It can be classified into two phases namely, identification and verification.

Speaker identification:

It is the process of determining which registered speaker provides a given utterance. It also implies, identification is the task of determining an unknown speaker's identity**.**

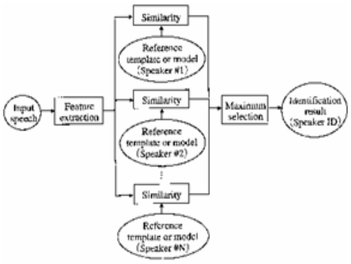
Input speech
Feature extraction
Similarity
Reference template or model (Speaker #1)
Similarity
Reference template or model (Speaker #2)
Similarity
Reference template or model (Speaker #N)
Maximum selection
Identification result (Speaker ID)

Fig.2 depicts the process of speaker identification[8]

For identification systems, the utterance is compared against multiple voice prints in order to determine the best match. It is a 1:N match where a voice is compared against "N" templates

Tele:
E-mail addresses: suma_swamy@yahoo.com, ramradhain@rediffmail.com

Speaker verification: It is the process of accepting or rejecting the identity claim of a speaker. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication.

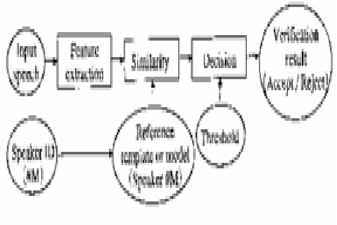
Input speech
Feature extraction
Similarity
Decision
Verification result (Accept / Reject)
Speaker ID (#M)
Reference template or model (Speaker #M)
Threshold

Fig.3 depicts the process involved in speaker verification[8]

In the verification phase, a speech sample or "utterance" is compared against a previously created voice print. Verification systems compare an utterance against a single voice print. Speaker verification is the 1:1 match if one speaker's voice is matched to one template. It is usually employed as a "gatekeeper" in order to provide access to a secure system like telephone banking.

In most of the applications a voice is used as the key to confirm the identity of a speaker that is classified under speaker verification.

There is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said).

Speech Recognition Method is of two types, Text-dependent and Text-independent methods. This paper deals with the Text-dependent method.

Open set identification

Open set identification is one of the cases where in it depicts how the designed system reacts to an unknown speaker. In this situation, an additional decision alternative, the "unknown" does not match any of the models. In both verification and identification processes, an additional threshold test can be used to determine if the match is close enough to accept the decision or if more speech data are needed.

Text-dependent speech recognition

It requires the speaker to say key words or sentences having the same text for both training and recognition trails. This system requires speaker pronounce in accordance with the contents of the text. Each person's individual sound profile model is established accurately. People can be identified by the contents of the text during recognition to achieve better effect. This system has a drawback because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. In order to overcome this problem, we use methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used.

Software algorithms

MFCC: MFCC are co-efficients that represent audio. They are derived from a type of cepstral representation of the audio clip. The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are positioned logarithmically which approximates the human auditory systems response more closely than the linearly spaced frequency bands obtained directly from the FFT(Fast Fourier Transform) or DCT(Discrete Cosine Transform). This can allow for better processing of data like in audio compression.

Block diagram of MFCC (overall process) is as shown below:

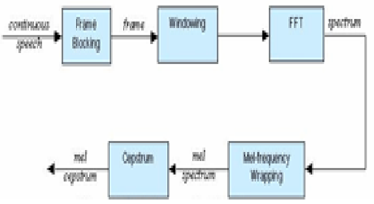

Fig.4 Block diagram of overall MFCC process[10]

MFCC is currently the most popular feature coefficient used in the speech recognition, and it can obtain the more accurate results of speech recognition under a non-noise condition. In the MFCC algorithm, we first use the FFT to calculate the signal frequency spectrum, then we use DCT to further reduce the speech signals redundant information, and reach the aim of regulating the speech signal into feature coefficients with small dimensions. The FFT and DCT algorithm can be used for any speech segment whose time-frequency resolution is fixed.

MFCCs are commonly derived as follows:

1. Take the Fourier transform of the signal.
2. Map the log amplitudes of the spectrum obtained above on to the Mel scale using triangular overlapping windows.
3. Take the DCT of the list of Mel log-amplitudes, as if it were a signal.
4. The MFCCs are the amplitudes of the resulting spectrum.

Feature Extraction: MFCC feature coefficient extraction flow chart is as shown:

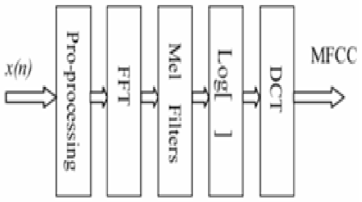

Fig.5 represents the flowchart for the MFCC feature extraction.[11]

MFCC consists of seven computational steps; each step has its own function and mathematical approaches, briefly explained as follows:

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.

$$Y[n] = X[n] - 0.95\,X[n-1] \qquad (1)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

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:

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

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below

Y(n)= X(n) x W(n) (2)

w(n)= $0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right]$ (3)

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(w) =$\mathrm{FFT}[h(t) * X(t)] = H(w) * X(w)$ (4)

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. The bank of filters according to Mel scale as shown in figure 4 is then performed.

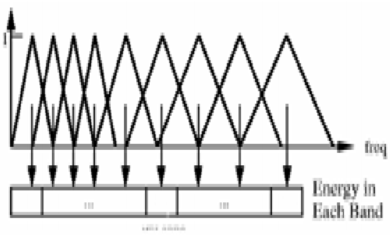

Fig.6 represents the Mel scale filter bank[7]

This figure 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. 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 [7, 8]. 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:

F(Mel)=$[2595 * \log 10[1+f]\ 700]$ (5)

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.

Step 7: Delta energy and delta spectrum

The voice signal and the frames changes, such as the slope of a format at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time.13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented at the equation below:

Energy= $\sum X^2\,[t]$ (6)

Each of the 13 delta features represents the change between frames in the equation 8 corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

d(t) = $\frac{c(t+1)-c(t-1)}{2}$ (7)

Vector quantization

VQ is a classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. It is used for data compression which works by dividing a large set of vectors into groups having approximately the same number of points closest to them. Each group represented by its centroid points, as in k-means and some other clustering algorithms.A VQ is an "approximator" i.e., it is similar to "rounding off" to the nearest integer.

An example of 1-D VQ is shown below:

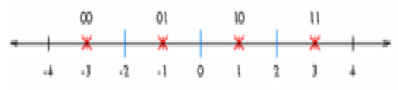

Fig.7 represents the 1-dimensional VQ.[14]

An example of 2D VQ:

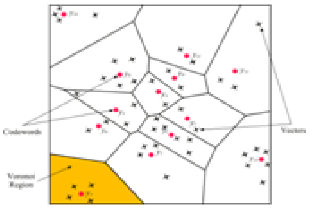

Fig.8 represents the 2-Dimensional VQ.[14]

An example of 3D VQ

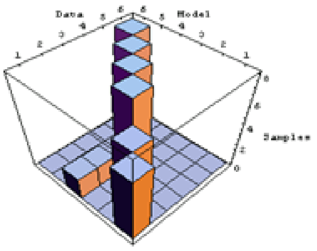

Fig.9 represents the 3-Dimensional VQ[14]

Feature matching:

The goal of pattern recognition is to classify objects of interest into one of a number of classes. The objects of interest are called acoustic vectors that are extracted from an input speech. The classes here refer to individual speakers. Since the classification procedure is applied on extracted features it can also be referred to as feature matching. The various feature matching techniques used in speaker recognition includes DTW (Dynamic Time Wrapping), HMM (Hidden Markov Modeling) and VQ. In this project VQ approach is used, due to the ease of implementation and high accuracy. 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 conceptual diagram illustrating VQ codebook formation is as shown:

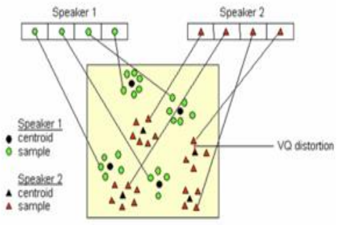

Fig.10 represents the conceptual diagram illustrating VQ codebook formation.[15]

Designing a codebook that best represents the set of input vectors in NP-hard. Hence it requires an exhaustive search for the best possible codeword's in space, and the search increases exponentially as the number of codeword increases. We therefore resort to suboptimal codebook design schemes which are named as LBG (Linde-Buzo-Gray) algorithm. The algorithm

1. Determine the number of codeword's, N, or the size of the codebook.
2. Select N codeword's at random, and let that be the initial codebook. The initial codeword's can be randomly chosen from the set of input vectors.
3. Using the Euclidean distance measure clusterize the vectors around each codeword. This is done by taking each input vector and finding the Euclidean distance between it and each codeword. The input vector belongs to the cluster of the codeword that yields the minimum distance.
4. Compute the new set of codeword's. This is done by obtaining the average of each cluster. Add the component of each vector and divide by the number of vectors in the cluster.

$$y_i=\frac{1}{m}\sum_{j=1}^{m} x_{ij}$$

Where *i* is the component of each vector (x, y, z directions), *m* is the number of vectors in the cluster.

5. Repeat steps 2 and 3 until the either the codeword's don't change or the change in the codeword's is small.

The performance of the VQ can be measured using "Good old Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR). MSE is defined as follows:

$$\text{MSE}=\frac{1}{m}\sum_{i=1}^{m}(\hat{x}_i - x_i)^2$$

Where M is the number of elements in the signal or image.

The PSNR is defined as follows:

$$\text{PSNR}=10\,log_{10}\left(\frac{(2^n-1)^2}{MSE}\right)$$

Where n is the number of bits per symbol.

Block diagram of LBG algorithm is shown below:

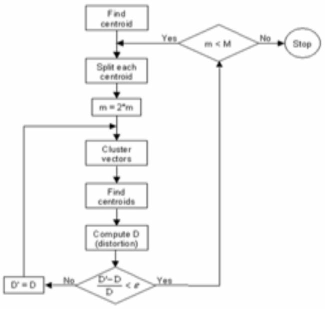

Fig.11 depicts the flowchart for the LBG algorithm[15]

Result analysis

To implement the proposed system, a combination of five words is considered namely: welcome, thank you, mat lab, window, book. These words are being tested under three speakers.

The below snapshot gives the final menu of the project using which different phases of project is accessed. It acts as a guide for proceeding to use the software.

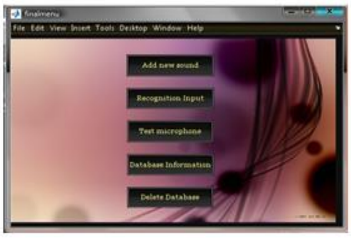

Fig. 12 depicts final menu

The following snapshot shows the training phase of the speech recognition. It has the indication for start and end of recording the voice., The voice which is recorded is represented in a signal form using cepstral coefficients.

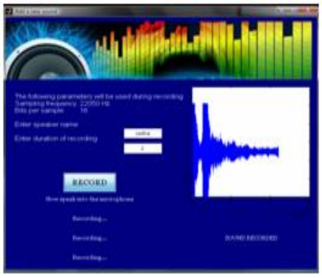

Fig. 13 depicts testing phase

The below snapshot shows the pictorial form which shows start and end for recording a trained voice. after comparing the two voices the result is put up on the same screen

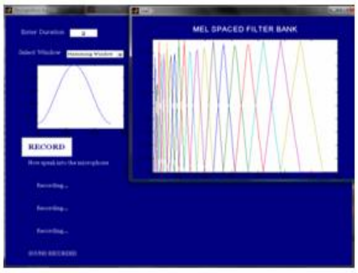

Fig. 14 depicts recording phase

This is implemented using two phases namely the training phase and then the testing phase.

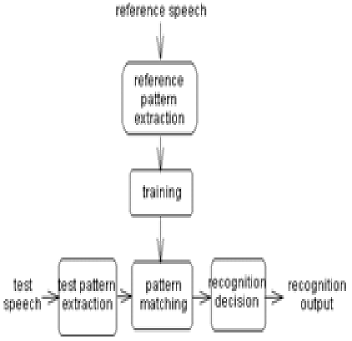

Fig.15 represents the phases in result analysis.[15]

In training phase, each speaker was to repeat the respective word once and the corresponding database was created for a particular word by all the speakers. In the testing phase, five trials were conducted for each speaker for respective round of the word and this is done on a random basis for the turn of the speaker in each round of the respective word. The overall efficiency of the speakers is as tabulated:

The overall efficiency of speakers

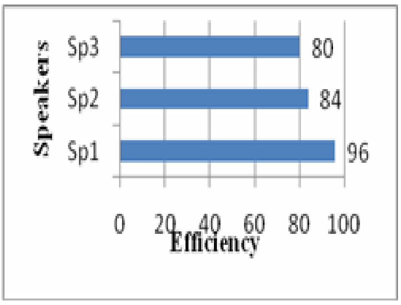

Fig 16 Represents the overall efficiency of the speaker

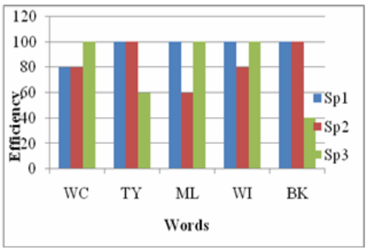

Fig 17. Represents the efficiency

The graph below represents the words overall efficiency:

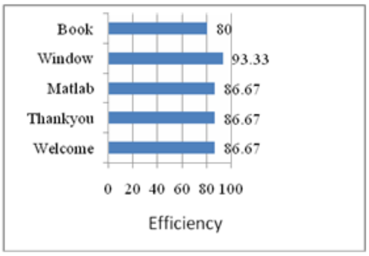

Fig.18 Bar-graph to depict the overall efficiency of the words

Conclusion

This project depicts the speech recognition system using the well known algorithms namely, MFCC and VQ by which the feature extraction and feature matching is implemented. The technique authenticates the speaker based on the speech recorded during the training phase. Therefore the results thus implement the improved efficiency of the system.

Acknowledgement:

The author acknowledges Anna University of Technology, Coimbatore for the encouragement and the permission to publish this paper.

The author would like to thank the Principal of Sir MVIT, Dr. M.S. Indira for her constant encouragement and Prof. Dilip K. Sen, Head of the Department, CSE/ISE for his invaluable guidance and suggestions at all stages.

References:

[1]. Matlab Programming for Engineers, Thrid edition, Stephen J. Chapman

[2]. Discrete Time-Speech Signal processing, by Thomas F

[3]. Fundamentals of Speech Recognition, by Lawrence Rabiner, Biing-Hwang Juang.

[4]. Digital Processing of Speech Signals,by L.R Rabiner,R.W Schafer.

[5]. Speech and Audio Signal Processing, by Ben Gold, Nelson Morgan.

[6]. Fundamentals of speech Synthesis, J.L Flanagan (Basics).

[7].http://www.cslu.ogi.edu/HLTsurvey/ch1node9.html (FEATURES)

[8]. Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. “Learning Internal Representations by Error Propagation”. David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986.

[10].http//..ncbi.nlm.nih.gov/pmc/articles/PMC2676177/(sample equations)

[11].http://www.journal.au.edu/ijcem/jan98/article5.html

[14]. A.Gersho and R. M. Gray, Vector Quantization and Signal Compression.

[15]. Junqua, J.-C.; Haton, J.-P. (1995). Robustness in Automatic Speech Recognition: Fundamentals and Applications. Kluwer Academic Publishers.

Table1: represents the overall efficiency of the speaker

Speakers	Overall Efficiency
Speaker1	96
Speaker2	84
Speaker3	80

Table 2: Represents the efficiency

Words	Speaker1 (%)	Speaker2 (%)	Speaker3 (%)
Welcome	80	80	100
Thank you	100	100	60
Mat lab	100	60	100
Window	100	80	100
Book	100	100	40

Table.3 represents overall efficiency of words

Words	Efficiency (%)
Welcome	86.67
Thankyou	86.67
Mat lab	86.67
Window	93.33
Book	80.00