Speaker Recognition: Feature Extraction Using MFCC and Recognition using Modified Vector Quantization and Retina Scanning

Dipangi Nayak[1], Suresh S. Gawande[2]
[1]M-Tech B.E.R.I. Bhopal, Madhya Pradesh, India
[2]Prof. and Head (Electronics & communication)

Abstract* - Speaker recognition and processing has received a lot of attention during the last few decades. Many applications of speaker recognition had been employed in different fields like in telephony, medical or command and control area. For decades humans have practiced to develop machines that can produce speech and understand as humans. Obviously such an interface would yield more benefits. Practices have been made to process vocally interactive computers to realize voice/speech recognition. The project work starts with recording the speech samples spoken by different speakers. The recorded samples are such that two or more words combined form a meaningful sentence. Background noise is also running. Then samples of speech are passed through Mel-filter banks to generate Mel Frequency Cepstral Co-efficient (MFCC).In [1], it was demonstrated that MFCC performs better than other feature extraction techniques. In the next step speech classification is carried out by using modified vector quantization. The system is successful in recognition rate of approx 85% for speaker in noisy environment. But this accuracy may increase by using upgrade version of MATLAB. Because in upgraded version there is no need to convert your original speech signal to wav file. Retinal-scan technology is interesting and developing to the biometric field and offers notable promise. One of the continuously promises and challenges for the biometric industry is to identified and define the environment in which the technology provides the better and strongest benefit to institutions.

Keywords: MFCC, Vector Quantization, Feature Extraction.

I. INTRODUCTION

The ability to communicate by speech is fundamental to our nature. Speaker recognition is one of the most oldest and favourite research of work. It is the process to identify/verify a person/speaker on the basis of his individual voice/speech’s information [2]. It has many applications and growing daily. The advantages of an accurate voice interface between humans and machines are that approaches the capabilities of human speech perception, would be enormous. This technique used for speaker's voice to verification and their identity and provides control access to services such as database access services; voice dialing, voice mail, information services, and security control for secrete information areas, remote access to computers and several other fields where security is the main area of concern. Now a day this technology becomes a part of our life and more comfortable as well. It has a history of about 40 year ago [3]. Since then it used several of applications. Now research is mainly focused on ASR (Automatic Speaker Recognition) [4]. Military applications of speaker recognition systems, although potentially widespread, focus on today's advanced aircraft environment. A recent study accident of MH-370 indicates that speaker recognition systems would have prevented many of these accidents.

In speaker identification human speech an individual verification is used to identify who that individual is. There are two different operational phases. In training it is also called Enrolment. In speaker verification human speech from an individual is used for verification and claimed identity of that individual.

The last biometric technology to identify is retinal scanning. Retina-scan technology creates use of the retina, which is the area surface on the behind of the eye ball that processes light entering through the pupil. Retinal Scan technology is dependent on the blood vessel idea in the retina of the eye. The principle for the technology is that the blood vessels at the retina provide a unique and natural idea and pattern, which may be used as a tamper-proof personal identifier and developer.

II. SYSTEM MODEL

Speech signal and features - Speech is a dynamic acoustic signal with numbers of sources of variation. The speech signal is mostly presented and showed in electrical pulses which are amplitude modulated in form of the envelope of the signal. The speech signal is divided into a many bands and the envelope of each band is extracted and used to modulate the pulse trains.

Speaker Modes- Here is two types of speaker modes in speech recognition.

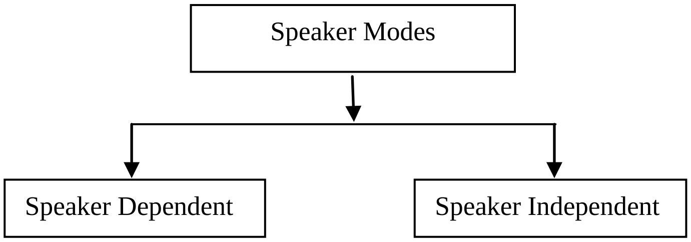


Fig1: Speaker's modes in Speech Recognition

Feature Extraction- Important step in speech recognition process is to find the suitable front-end features. The goal of feature extraction is for representing speech signal by a determinable number of measures of the signal. This is just because of information in the acoustic signal is too much to process, and not all of the information is relevant for specific tasks [26]. In pattern recognition and in speech processing, feature extraction is a very special task of dimensionality reduction. When input data to an algorithm is very huge to be processed and it is guessed to be notoriously redundant then the input data has to be transformed into a less representation bunch of features also named features vector.

Generating MFCC - In MFCC, the main advantage is that it uses Mel frequency scaling which is very approximate to the human auditory system [12].

Analysis of the enhanced retinal blood vessel image then takes place to find characteristic patterns. Retina-scan devices are used for physical applications and are usually used in environments that require high and very ups degrees of security such as high-level government military needs. Infrared energy is assimilated faster by blood vessels in the retina than by surrounding tissue which is used to illuminate the eye retina.

Processing of Calculation MFCC is give below-

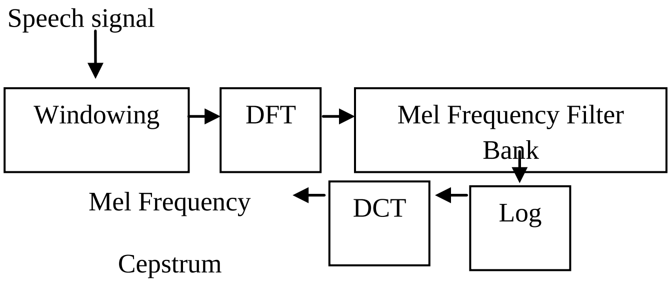


Fig 2: Computation of Mel Frequency Cepstral Coefficient

III. PREVIOUS WORK

Feature Extraction Techniques- Speaker recognition process includes different phases such as preprocessing, feature extraction and matching the feature or recognition classification. Numbers of research papers have been referred to understand the basics of speech recognition process. Numbers of algorithms have been studied to carry out project work related to feature extraction and feature matching or recognition classification techniques. A brief explanation and important results about these are discussed below.

Xinhui Zhou, Daniel Garcia-Romero, Ramani Durais wami, Carol Espy-Wilson, Shihab Shamma in, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition" [5] compares MFCC and Linear Prediction by their performances. It stated that a linear scale in frequency might provide some benefits in speaker recognition over the Mel scale.

Namrata Dave in,"Feature Extraction Methods like LPC, PLP and MFCC in Speech. Recognition" [6] state that Speech features are extracted from recorded speech by phone of a female or male speaker and compared with templates available in database. "Speaker Features and Recognition Techniques: A Review" [7] tells an overview of various techniques and Methods are used for modelling in speech recognition system and feature extraction. "A Novel Approach for MFCC Feature Extraction" [8] states that The MFCC (Mel-Frequency Cepstral Coefficients) feature extraction method is a leading close for speech signal feature extraction and current researcher aims is to increase performance enhancements.

A Novel Approach for Speech Feature Extraction in MFCC" by Cubic-Log Compression [9] state that Pre-processing for speech signal is measured as major step in enhance in the results of feature vector extraction for efficient Automatic Speaker Recognition system.

Recognition- Speaker Recognition Based on Neural Networks" [15], proposed a method to reduce information signal redundancy and transfer the sampled for human speech signal from time domain to frequency domain we will use one dimensional discrete cosine transform (DCT) for feature extraction. Speech Recognition Using MATLAB [16] enlightens upon the invention and technological advancement in the field of speaker recognition and they also focuses upon different steps involved for speech identification using MATLAB Programming.

IV. PROPOSED METHODOLOGY

Mel Frequency Filter Bank- Psychophysical studies have revealed that human perception of frequency content of sounds for speech signals do not follow a linear scale. For each sound with an actual frequency (f), a subjective pitch is studied and measure on a scale parameter called the “Mel” scale. The Mel-Frequency scale provides linear frequency spacing below 1 KHz and a logarithmic spacing above 1 KHz [28] [29].

The Mel Frequency Scale is given by:-

$$=2595*\log 1+ \frac{\quad}{700}$$

LOG- John Napier in 1614 publicly propounded the method of logarithms, in a book titled *Mirifici Logarithmorum .Canonis Descriptio* (*Description of the Wonderful Rule of Logarithms*). The logarithm is the reverse process of exponentiation, while the complex logarithm is reverse function of the exponentiation applied to complex numbers.

Logarithmic pattern scales reduce wide and broad-ranging quantities to smaller scopes [31]. For example, the decibel is a logarithmic unit measuring sound pressure and signal power ratios.

DCT (Discrete Cosine Transform) - Pre-processing is the process of the signal reduces the computational complexity while operating on the speech signal. After conditioning the speech signal i.e. after pre-processing the next step is to extract the features of the training signal. The method is calculating the Cepstral Coefficients of the signal using DCT (Discrete Cosine Transform). The Cepstral Coefficients are calculated using the following formula:-

ceps= dct(log(abd(DFT(Ywindowed))))

Speaker Recognition- The problem of speaker recognition has always been a much wider topic in engineering field so called pattern recognition. The aim of pattern recognition lies in classifying objects of interest into a number of classes or categories. These objects of interest are called patterns and in case are sequences of feature vectors that are extracted using the techniques described in the previous portion of this same chapter from an input speech. Since here we are only dealing with classification procedure based upon extracted features, it can also be abbreviated as feature matching. Feature matching problem has been sorted out with many class-of-art efficient algorithms like MVQ, stochastic models such as GMM & HMM. In our study project we have put our focus on MVQ algorithm.

VECTOR QUANTIZATION- Vector quantization (VQ in short) involves the process of taking a huge bunch of feature vectors of a particular user and producing a smaller bunch of feature vectors which is represent the centred of the distribution, such that points spaced so as to lessen the average distance to every other point. Vector quantization is used since it would be highly impractical to represent every single feature vector in feature space that we generate from the training utterance of the corresponding speaker. A vector quantize maps k-dimensional vectors in the vector space *Rk* into a finite set of vectors Y = {y_i : i = 1, 2... N}. Here k-dimension refers to the no of feature coefficients in each feature vector. Each vector y_i is called a code vector or a codeword and the set of all the codeword’s is called a codebook. Hence for a given number of users, code books are generated for each speaker during the training phase using VQ method. For each codeword y_i , there is a “nearest neighbour” region associated with it called Voronoi region [32], and is defined by

$$\boldsymbol{V_i = \{x \in R^k : \left| x - y_i \right| \leq \left| x - y_j \right| \}, \textbf{ for all } j \neq i}$$

The set of voronoi regions partition the entire feature space of a given user such that-

$$\overset{N}{\underset{i=1}{}} V_i = R^k$$

$$\overset{N}{\underset{i=1}{}} V_i = \Phi$$

For all $i \neq j$

The Voronoi region associated with the given centred is cluster region for the vector. The Euclidean distance is defined by given formula:-

$$d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2}$$

Where x_j is the j^{th} component of the input vector, and y_{ij} is the j^{th} component of the codeword y_i

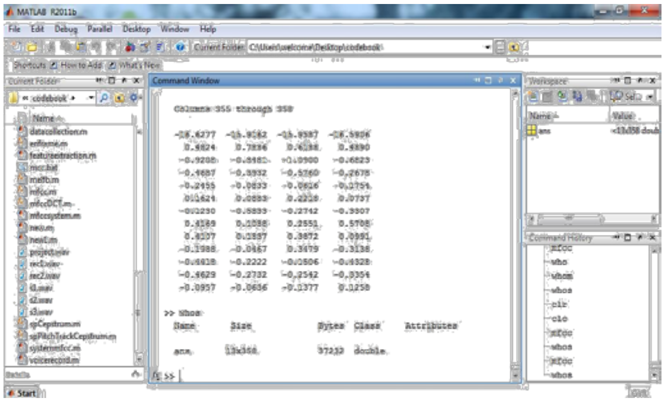
Order for a retinal image and compression to be required, the operator must gaze directly into the lens and remain still constant; movement losses the acquisition process requiring another attempt. A low intensity light source is used in order to scan the vascular quality at the retina. This involves a 360 degree circular scan of the area taking over 400 readings in order to establish the blood vessel pattern. This is then reduced to 192 reference points before being distilled into a digitized 96 byte template and stored in memory for subsequent verification purposes. Normally it takes 3 to 5 acceptable images to ensure enrolment. Because of this, enrolments process can be lengthy. Enrolments can take over 1 minute with some users not being able to be enrolled at all. It seems the more that a user is acclimated to the process, the faster the enrolment process works. After image acquisition, software is used to compile unique features of the retinal blood vessels into a template. Retina-scan technology has its benefits and losses. Among its benefits are its resistance to wrong matching or wrong positives and the fact that the pupil, like the fingerprint remains a stable physiological trait throughout one's life.

V. SIMULATION/EXPERIMENTAL RESULTS

This paper is based on identifying an unknown speaker given a set of registered speakers. Here we have assumed the unknown speaker to be one of the known speakers and tried to develop a model to which it can best fit into. CEPSTRAL COEFFICIENTS- When we take sample a spoken syllable, we'll be having many samples. Then we try to extract features from these sampled values. Cepstral coefficients calculation is one of such methods. Here we initially derive Short Term Fourier Transform of sampled values, then take their absolute value (they can be complex) and calculate log of these absolute values.

There after we go for converting back them to time domain using Discrete Cosine Transform (DCT). We have done it for five users and first ten DC T coefficients are

Cepstral coefficients. The result obtained is shown below:-

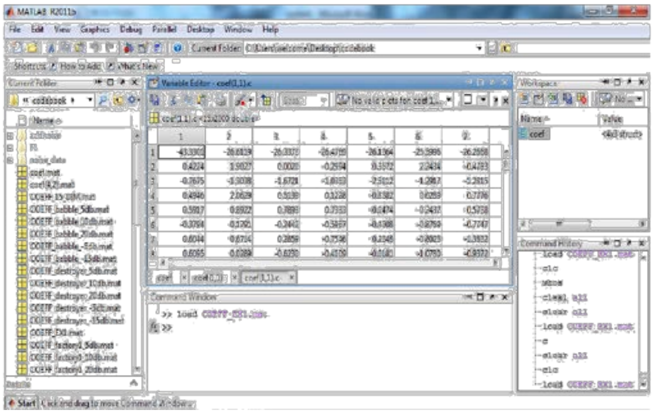


```
  Columns 355 through 358

  -16.6277  -15.9162  -15.9587  -16.5906
    0.4824    0.7836    0.6138    0.4890
   -0.9208   -0.8482   -0.0900   -0.6823
   -0.4687   -0.8932   -0.5760   -0.2678
   -0.2455   -0.0833   -0.0616   -0.2751
    0.1624    0.0883    0.2218    0.0737
   -0.2230   -0.5833   -0.2742   -0.3807
    0.4189    0.1088    0.2551    0.5708
    0.4107    0.1837    0.3872    0.0991
   -0.1988   -0.0467    0.3479   -0.3138
   -0.4818   -0.2222   -0.0506   -0.4328
   -0.4629   -0.2732   -0.2542   -0.3354
   -0.0957   -0.0636   -0.1377    0.1258

>> whos
  Name      Size            Bytes  Class     Attributes

  ans       13x358          37232  double

>>
```

Fig 3: Result of Cepstral Coefficient Calculation

Above figure was obtained after calculating the Cepstral coefficients in MATLAB for a user having utterances of the word "hello".

MFCC- we have only shown few feature vectors. Each column refers to a feature vector. The elements of each column are the corresponding MFCCs.



coef{1,1}.c <13x2000 double>

	1	2	3	4	5	6	7
1	-43.3300	-26.6819	-26.3373	-26.4719	-26.1364	-25.3996	-26.2558
2	0.4224	1.9027	0.0020	-0.2594	0.3572	2.2434	-0.4783
3	-0.7675	-1.9308	-1.6728	-1.8053	-2.5012	-1.2987	-0.2815
4	0.4946	2.0629	0.5138	0.1228	-0.8382	0.6289	0.7776
5	0.5917	0.8922	0.7898	0.7353	-0.0474	-0.2437	0.5758
6	-0.3764	-0.1795	-0.2442	-0.5857	-0.4388	-0.8759	-0.7747
7	0.6044	-0.6714	0.2859	-0.7536	-0.2348	-0.8823	-1.3932
8	0.6065	0.0289	-0.6290	-0.4109	-0.0180	-1.0783	-0.9372

```
>> load COEFF_DX1.mat
>>
```

Fig4: elements of MFCC

MFCC plot for a signal in capstral coefficient are given below:-

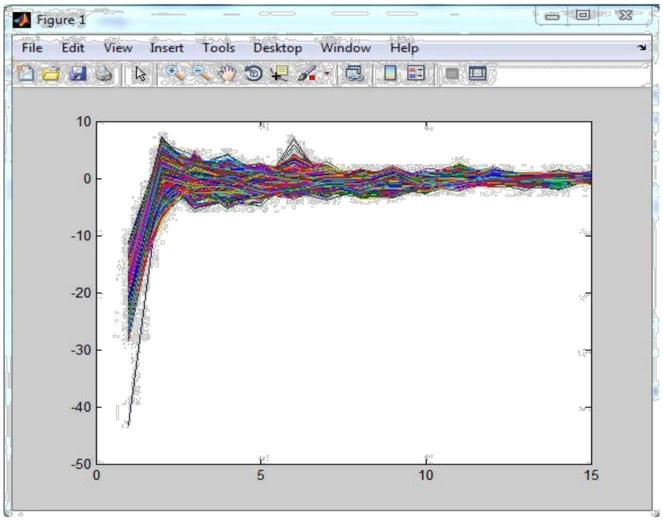


Fig5: MFCC plot for a signal in capstral coefficient

VI. CONCLUSION

Speaker Recognition technology has been gaining its importance day-by-day. Various attempts have been made in developing voice operated system and making them available globally. Attempts have also been made to create a human-computer interface thereby enabling humans to have a real time communication with computers. Various nations around the globe have incorporated voice operated systems that accomplish large number of tasks.

Systems like "Interactive Voice Response System", "JAWS" etc. are some of the best technologies available in this domain. Speaker recognition system contains two main important tasks (i) feature matching and (ii) feature extraction. Feature extraction is process which extracts a few amounts of data that can later be used for presenting each and every speech sample from the voice signal. Important step in speaker recognition process is to find t he suitable front-end features.

Mel-Frequency Cepstrum Coefficients (MFCC) technique gives superior performance attributed to the fact that they better perform the perceptually relevant aspects of the short-term speech spectrum. Regarding to previous study, there are many approaches in speaker recognition system. In this paper a very efficient speaker recognition system is designed. MFCC technique is used for feature extraction. Total 12 speakers Cepstral coefficients are extracted. Feature matching is carried out using centroid based modified vector quantization approach. It is found that continuous word recognition of speaker gives best results on an average 82%. Also it is identified that when testing is made in speaker in soundly environment, accuracy is less i.e. 75% compare to when testing is made in speaker less noisy environment for which accuracy is 85%.the system is designed to be speaker self dependent. Hence, it provides a higher flexibility in terms of advantage by any human and also eliminates the time requirement for training. A generalized coding facilitates the user to get access of the system without change in basic coding. In order to make technology more closely to the operator its access should be made easier and operative.

Retinal scanning confrontation to wrong matching is due to the fact that retinal scans produce tasks that have highly defining and distinctive characteristics, abundant to enable identification. Well-trained users find retina scan capable of reliable identification. Like fingerprints, retina traits remains stable throughout life.

Disadvantages include the fact that the technology is difficult to use, users complain not comfortable with eye-related terminology and technology in general and the fact that retina scan technology has limited uses.

VII. FUTURE SCOPES

This speaker recognition system has greater scope for development in its functionalities. Further development can be done by increasing the reference database size, by using advance version of MATLAB software (because here we convert our original signal in to wav files. This loses some information), by using fast processor. Implementing recognition by recognizing speech sounds can increase the accuracy greatly .One can use an alternative feature extraction technique to represent the features of the speech signal. Retinal scan uses are its resistance to wrong matching or wrong positives and the fact that the pupil.

REFERENCES

[1] M. Ramya Devi, DR. T. Ravichandran,"A Novel Approach for Speech Feature Extraction by Cubic- Log Compression in MFCC", Proceedings of the International Conference on Patterns Recognition, Informatics and Mobile Engineering, Feb 21-22, 2013.

[2] Juraj Kacur,Mario varga,Gregor Rozinaj,"Speaker Identification in a multimodel Interface"55th international Symposium ELMAR,25-27,2013.

[3] Diana Van Lancker, Jody Kreiman and Karen Emmorey,"Familier Voice Recognition Patterns and parameter part 1:Recognition of backward voice",journal of phonetics,pp 19-38,1985.

[4] Vimla. C, Dr. V. Radha, "A Review on Speech Recognition Challenges and Approaches" world of computer science and information technology journal (WCSIT) vol.2, No. 1, 1-7, 2012.

[5] R L K Venkateswarlu,V.Kumari,A.K.V. Nagayya, "Efficient Speech Recognition By using Modular Neural Network", Int. J. Comp. Tech. Appl., vol. 2, no.3, pp.463-470.

[6] Sunil Kumar Kopparapu and M Laxminarayana, "Choice of Mel Filter Bank In Computing MFCC of A Re-sampled Speech" 978-1-4244-7157-6/10/$26.00©2010.

[7] Namrata Dave1, "Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition" International journal for advance research in engineering and technology Volume 1, Issue VI, July 2013.

[8] Dr. Mahesh S. Chavan and Mrs. Sharada V. Chougule, "Speaker Features And Recognition Techniques: A Review", International Journal Of Computational Engineering Research / ISSN: 2250–3005.

[9] Md. Afzal Hossan, Sheeraz Memon, Mark A Gregory, "A Novel Approach for MFCC Feature Extraction", 978-1-4244-7907-8/10/$26.00 ©2010 IEEE.

[10] M. Ramya Devi, Dr. T. Ravichandran "A Novel Approach for Speech Feature Extraction by Cubic-Log Compression in MFCC", Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22,2013.

[11] Sajid A. Marhon, Duaa N. Ubaid Al-Aghar, in" Speaker Recognition Based On Neural Networks",IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics, Vol. 14, No. 1, January 2004.

[12] Aseem Saxena, Amit Kumar Sinha, Shashank Chakrawarti, Surabhi Charu,"Speech Recognition Using MATLAB",

International Journal of Advances In Computer Science and Cloud Computing, ISSN: 2321-4058 Volume- 1, Issue- 2, Nov-2013.

[13] Campbell, J.P., Jr.;“Speaker recognition: a tutorial”,Proceedings of the IEEE Volume85, Issue 9, Page(s):1437 – 1462, Sept. 1997.

[14] Seddik, H.; Rahmouni, A.; Samadhi, M.; “Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier ”, First International Symposium on Control, Communications and Signal Processing, Proceedings of Page(s):631 – 634,IEEE 2004.

[15] Shirali, Shailesh, “A Primer on Logarithms”, Hyderabad Universities Press, ISBN 978-81-7371-414-6, esp. section 2,2002.