International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

TEXT INDEPENDENT TECHNIQUE OF VOICE RECOGNITION ALGORITHMS USING MEL FREQUENCY CEPSTRAL COEFFICIENT (MFCC)

Shweta Tiwari , Bansal Institute Of Science and Technology Bhopal (RGPV),
Manish Saxena, Bansal Institute Of Science and Technology Bhopal (RGPV)

ABSTRACT

This paper we study of design a system to recognition voice commands. Speaker recognition is the task of establishing identity of an individual based on his/her voice. It has a significant potential as a convenient biometric method for telephony applications and does not require sophisticated or dedicated hardware. Many organizations like banks, institutions, industries etc are currently using this technology for providing greater security to their vast databases. Most of voice recognition systems contain two main modules as follow "feature extraction" and "feature matching". In this paper, MFCC algorithm is used to simulate feature extraction module. Using this algorithm, the cepstral coefficients are calculated on mel frequency scale. VQ (vector quantization) method will be used for reduction of amount of data to decrease computation time. In the feature matching stage Euclidean distance is applied as similarity criterion. Because of high accuracy of used algorithms, the accuracy of this voice command system is high. Using these algorithms, by at least 5 times repetition for each command, in a single training session, and then twice in each testing session zero error rate in recognition of commands is achieved. Digital processing of speech signal and voice recognition algorithm is very important for fast and accurate automatic voice or speaker recognition technology.

Keywords : Feature Extraction, Feature Matching, Mel Frequency Cepstral Coefficient (MFCC), dynamic Time Warping (DTW)

I. INTRODUCTION

There are a few types of similar system has been developed by other researcher. Each system has its own method as well as advantages and disadvantages. They also depend on how the mechanism works. Each system has the same purpose. The purpose is to provide medium security system based on Biometric Recognition, or in this paper it is known as Voice . It is a process of automatically recognizing who is speaking on the basis of individual information. The system uses physical characteristics and traits on human being which are unique for the recognition of the individuals.

Using the information, the system makes it possible to use speaker's voice to verify their identity and control access. Speaker recognition technology identifies people based on the differences in the voice resulting from physiological differences and learned speaking habits. When an individual is enrolled the system captures sample of the person's speech as the individual says certain scripted information in to a microphone multiple times. This information is known as extraction phase. The extraction phase is then converted to a digital format and distinctive characteristic (e.g. pitch, cadence, tone) are extracted to create a template for the speaker. Speaker recognition templates require the most data space of all biometric templates.

Speaker recognition technology requires minimal training for those involved. It is also fairly inexpensive and is very non-intrusive. The biggest disadvantage with the technology is that it can be unreliable and does not work well in noisy environments.

Speaker verification task is further classified in to open and closed set task. If the target speaker is assumed to be on of the registered speakers, there cognition task is a closed-set problem. If there is possibility that the target speaker is none of the registered speakers, the task is called as open-set problem. In general, the open set problem is much more challenging. In the closed-set task, the system makes a forced decision simply by choosing the best matching speaker from the speaker database no matter how poor this speaker matches. However, in the case of open-set identification, the system must have a predefined tolerance level so that the similarity degree between the unknown speaker and the best matching speaker is within the tolerance. In this way, the verification task can be seen as a special case of the open-set identification task with only one speaker in the database.

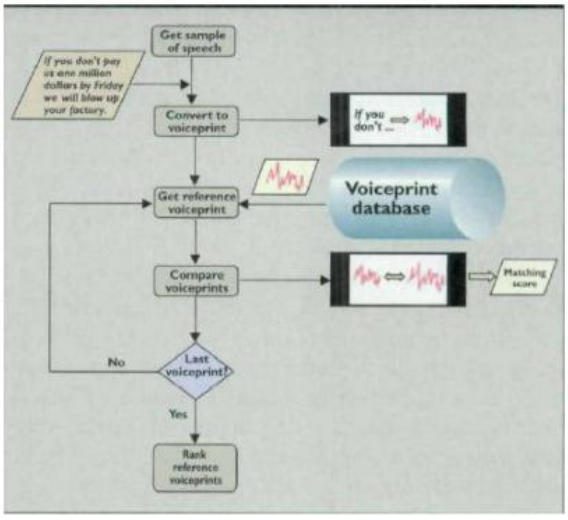


Fig1: Speaker Identification

II. SIGNAL PROCESSING

A speech signal is a form of wave motion carried by a medium (e.g. air particles), and it can be captured by a microphone, which converts the continuous air pressure changes in to continuous voltage changes. The analog signal sa(t) is then sampled to a digital form s[n] by an analog-to-digital converter (A/D converter). The A/D converter samples the analog signal uniformly with the sampling period T:

s[n] = sa (nT)

The inverse of T is the sampling frequency (or sampling rate) and marked here by Fs = 1= T. given that the original signal sa(t) contains frequencies only up to Fs=2, it can be fully reconstructed from the samples s[n]. The frequency Fs=2 is called the nyquist rate of the signal and it is the upper limit for frequencies present in the digital signal. For instance, if one wants to preserve frequencies up to 4 KHz, the sampling rate must be chosen Fs > 8 KHz. In addition to the sampling, the ADC quantizes the samples in to a finite precision. The number of bits used per sample determines the dynamic range of the signal. Adding one bit extends the dynamic range of the signal roughly +6 dB. Fourier analysis provides a way of analyzing the spectral properties of a given signal in the frequency domain. The Fourier analysis tools consider a signal as being composed of a superposition of sinusoidal basis functions of different frequencies, phases and amplitudes. An example is shown in Fig. 2 [12], which shown three sinusoids and their superposition (sum) Fourier analysis provides a tool for finding the parameters of the underlying sinusoids (Forward transform) or for synthesizing the original time-domain signal from the frequency domain presentation (inverse transform).

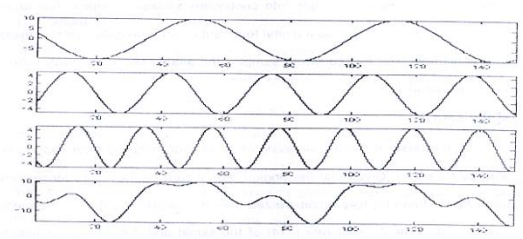


Fig 2 : Three sinusoids and their superposition (sum) Fourier analysis

DFT can be computed via a faster algorithm called fast Fourier transform of FFT. A requirement for FFT is that input signal (vector) has a length of 2M for some M 2 N+, i.e., a power of two. In practice, the input signal is first zero-padded to the next highest power of two and the zero-padded signal is given as an input for the FFT. For instance, if the length of signal is 230 samples, it is zero padded to length N = 256 for which the FFT can be computed. Zeros can be added to the beginning or end of the signal, and it does not affect the result of the DFT. Time complexity of the FFT is N log2 N. The savings in computation time is practices are on the order of hundred folds. For instance, for N = 1024, the ratio of DFT multiplications to FFT multiplications is about 200 and the ratio of additions about 100.

III. PHASES OF SPEAKER VERIFICATION.

For almost all the recognition systems, training is the first step. We call this step in SVS enrollment phase, and call the following step identification phase. Enrollment phase is to get the speaker models or voiceprints for speaker database. The first phase of verification systems is also enrollment in these phases we extract the most useful features from speech signal for speaker verification, and train models to get optimal systems parameters.

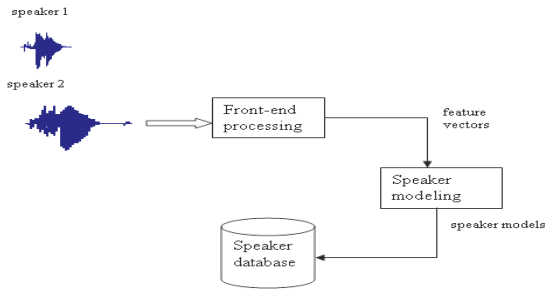


Fig 3. Enrollment phase for speaker verification

Enrollment phase is to get the speaker models or voiceprints to make a speaker database, which could be used later in the next phase, i.e. identification phase. The front-end processing and speaker modeling algorithms in both phases of SVS should be consistent respectively.

In verification phase, the same method for extracting features as in the first phase is used for the incoming speech signal, and then the speaker models getting from the enrollment phase are used to calculate the similarity between the new speech signal model and all the speaker models in the database. In closed-set case the new speaker will be assigned to the speaker ID which has the maximum similarity in the database. Even though the enrollment phase and verification phase are working separately, they are still closely related. The modeling algorithms used in the enrollment phase will also work on the identification algorithms.

Feature extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. This is often referred as the signal-processing front end.

The speech signal is a slowly timed varying signal (it is called quasi-stationary).An example of speech signal is shown in Fig 4.When examined over a sufficiently short period of time (between 5 and 100msec),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.

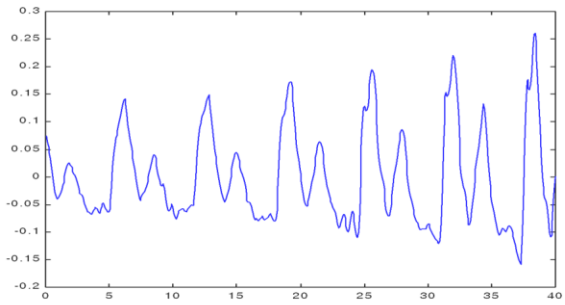


Fig 4: An example of speech signal

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 is perhaps the best known and most popular, and these will be used in this paper.

Feature matching

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern recognition. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching. Furthermore, if there exists some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. This is exactly our case since during the training session, we label each input speech with the ID of the speaker. These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm.

The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this paper, the VQ approach will be used, due to ease of implementation and high accuracy.

IV. MEL-FREQUENCY CEPSTRUM COEFFICIENTS PROCESS

The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.

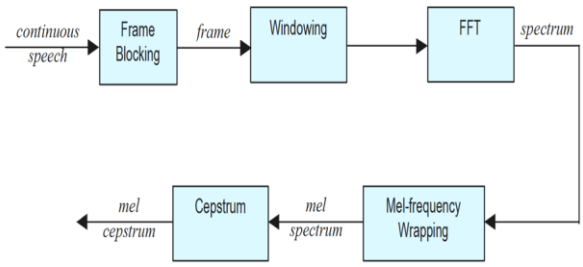


Fig 5: Block diagram of MFCC processor

Feature Matching Using VQ Technique

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 codeword. The collection of all code words is called a codebook.

Figure. 5 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 5 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. 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.

the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. As described above, the next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of L training vectors into a set of M codebook vectors.

V. SOFTWARE SETUP AND EXPERIMENT

MATLAB's built-in functions provide excellent tools for linear algebra computations, data analysis, signal processing, optimization and many other types of scientific computations. Most of these functions use state-of the art algorithms. There

are numerous functions for 2-D and 3-D graphics as well as for animation. Also, for those who cannot do without their FORTRAN or C codes, MATLAB even provides an external interface to run those programs from within MATLAB. The user, however, is not limited to the built-in functions the person can write his own functions in the MATLAB language. Once written, these functions behave just like the built-in functions. MATLAB's language is very easy to learn and to use.

There are also several optional Toolboxes available from the developers of MATLAB. These Toolboxes are collections of functions written for special applications such as Symbolic Computation, Image Processing, Statistics, Control System Design, and Neutral Networks. The basic building block of MATLAB is the matrix. The fundamental data type is the array. Vectors, scalars, real matrices are all automatically handled as special cases of the basic data-type. The thing is the user need not have to declare the dimensions of a matrix. The built-in functions are optimized for vector operations. Consequently, vectorized commands or run much faster.

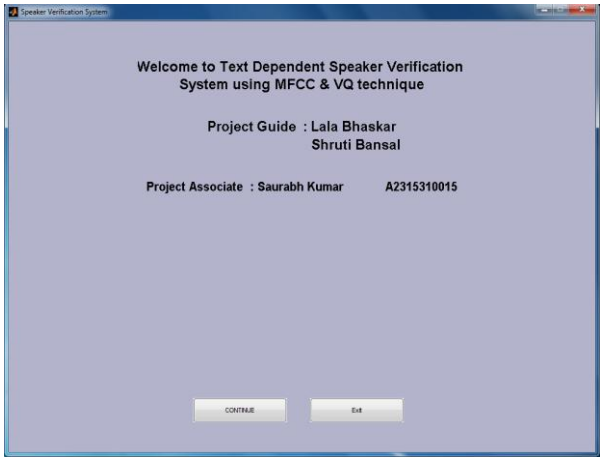


Fig 6. Set up screen

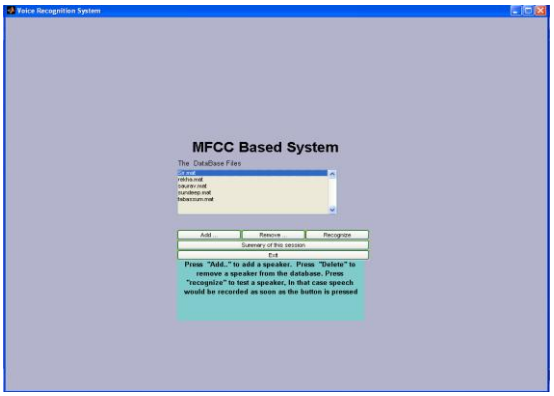


Fig 7. Database maintenance interface

Fig 8. Add user interface

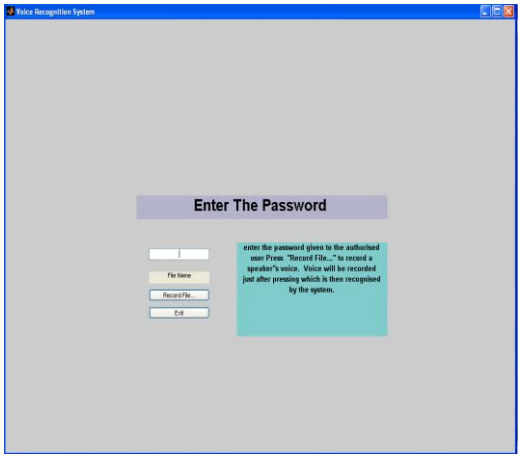


Fig 9. Password window

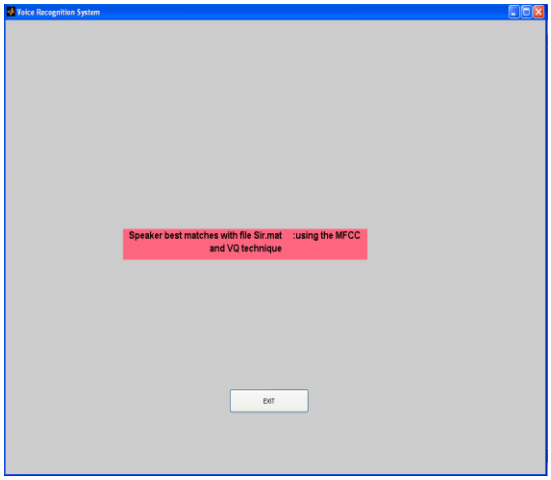
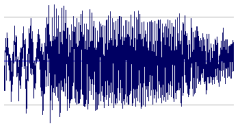
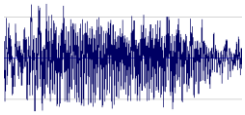


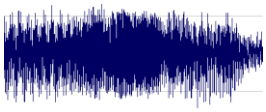
Fig 10. Final result window



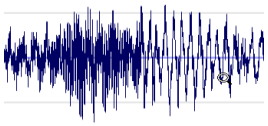
Sir



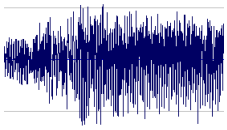
Rekha



Saurabh



Sundeep



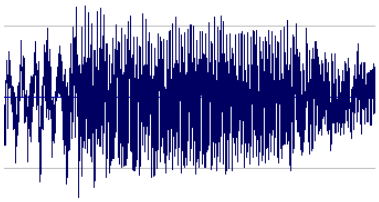
Tabassum

Fig 11 Different Speech Sample

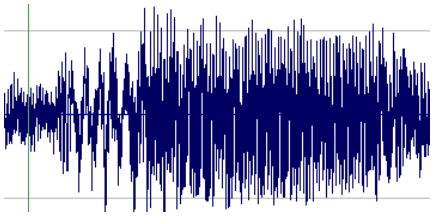
VI. RESULT

The tested speech is best matches with sir

Final result is displayed in this window.



Speech waveform of sir



waveform of the person to be tested

CONCLUSION

By applying the procedure described, each speech frame of around 30msec with overlap, a set of Mel-frequency Cepstrum coefficients is computed. These are results of a cosine transform of the logarithm of the short-term power spectrum on a Mel-frequency scale and the voice recognition carried out with the help of data base maintained in MATLAB.

REFERENCES

1. Campbell, J.P., Jr.; "Speaker recognition: a tutorial" Proceedings of the IEEE Volume 85, Issue 9, Sept. 1997 Page(s):1437 – 1462.
2. Seddik, H.; Rahmouni, A.; Sayadi, 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 IEEE 2004 Page(s):631 – 634.
3. Childers, D.G.; Skinner, D.P.; Kemerait, R.C.; "The cepstrum: A guide to processing" Proceedings of the IEEE Volume 65, Issue 10, Oct. 1977 Page(s):1428 – 1443.
4. Roucos, S. Berouti, M. Bolt, Beranek and Newman, Inc., Cambridge, MA; "The application of probability density estimation to text-independent speaker identification" IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '82. Volume: 7, On page(s): 1649- 1652. Publication Date: May 1982.
5. Castellano, P.J.; Slomka, S.; Sridharan, S.; "Telephone based speaker recognition using multiple binary classifier and Gaussian mixture models" IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 Volume 2, Page(s) :1075 – 1078 April 1997.
6. Zilovic, M.S.; Ramachandran, R.P.; Mammone, R.J "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions".; IEEE Transactions on Speech and Audio Processing, Volume 6, May 1998 Page(s):260 - 267
7. Davis, S.; Mermelstein, P, "Comparison of parametric representations for monosyllabic word

recognition in continuously spoken sentences" , IEEE Transactions on Acoustics, Speech, and Signal Processing Volume 28, Issue 4, Aug 1980 Page(s):357 – 366
8. Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, issue 1, Jan 1980 pp.84-95.
9. S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol.34, issue 1, Feb 1986, pp. 52-59.
10. Fu Zhonghua; Zhao Rongchun; "An overview of modeling technology of speaker recognition", IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Page(s):887 – 891, Dec. 2003.
11. Moureaux, J.M., Gauthier P, Barlaud, M and Bellemain P."Vector quantization of raw SAR data", IEEE International Conference on Acoustics, Speech, and Signal Processing Volume 5, Page(s):189 -192, April 1994.
12. Nakai, M.; Shimodaira, H.; Kimura, M.; "A fast VQ codebook design algorithm for a large number of data", IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, Page(s):109 – 112, March 1992.
13. Xiaolin Wu; Lian Guan; "Acceleration of the LBG algorithm" IEEE Transactions on Communications, Volume 42, Issue 234, Part 3Page(s):1518 - 1523, February/March/April 1994.
14. B. P. Bogert, M. J. R. Healy, and J. W. Tukey: "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking".Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15, 209-243. New York: Wiley, 1963.
15. Martin, A. and Przybocki, M., "The NIST Speaker Recognition Evaluations: 1996-2000", Proc. OdysseyWorkshop, Crete, June 2001
16. Martin, A. and Przybocki, M., "The NIST 1999 Speaker Recognition Evaluation—An Overview", Digital Signal Processing, Vol. 10, Num. 1-3. January/April/July 2000
17. Claudio Becchetti and Lucio Prina Ricotti, "Speech Recognition", Chichester: John Wiley & Sons, 2004.
18. John G. Proakis and Dimitris G. Manolakis, "Digital Signal Processing", New Delhi: Prentice Hall of India. 2002.
19. Rudra Pratap. Getting Started with MATLAB 7. New Delhi: Oxford University Press, 2006
20. R. Chassaing, DSP Applications Using C and the TMS320C6x DSK, Wiley, New York, 2002.