RESEARCH ARTICLE OPEN ACCESS

Performance Evaluation of Human Voice Recognition System based on MFCC feature and HMM classifier

Mr. Anand Mantri*, Mr. Mukesh Tiwari**, Mr. Jaikaran Singh***
*PG Student (Department of Electronics and Communication, SSSIST, Sehore, India)
** (Department of Electronics and Communication, SSSIST, Sehore, India)
*** (Department of Electronics and Communication, SSSIST, Sehore, India)

ABSTRACT
The speech processing is very exciting field in research area. The human voice recognition is one of the applications of this field. There are various techniques which are evaluated by scientist for voice recognition and available in daily use. In this paper we will design and develop MFCC (Mel-Frequency Cepstral Coefficients) feature and HMM classifier based human recognition system, which will identify the human based on their voice input. The performance of recognition system is evaluated based on recognition rate, which is given in this paper for different data set of human voice.
Keywords **-** Voice Recognition, MFCC (Mel-Frequency Cepstral Coefficients), HMM (Hidden Markov Model), VAD (Voice Activity Detection), LPC.

I. INTRODUCTION

Voice is the sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box. When a human speaks, his or her voice changes in power and tone during the utterance. A graph of the voice signal shows these changes in the height of the wave during very small changes in time.

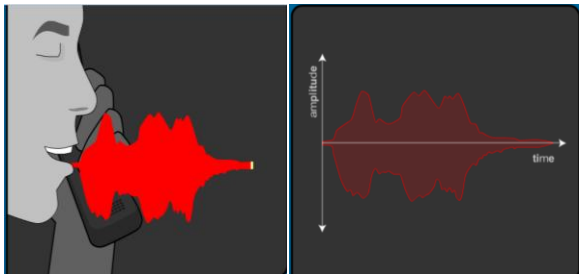


Fig 1: Voice Signal

Automatic recognition of speech by machine has been a goal of research for more than four decades [1, 2]. Generally speaking, there are three approaches to voice recognition: the acoustic phonetic approach, the pattern recognition approach and the artificial intelligence approach. One well-known and widely used pattern-recognition approach to voice recognition is Hidden Markov Model (HMM) approach, which is a statistical method of characterizing the spectral properties of the frames of a pattern. This provides a natural and highly reliable way of recognizing speech for a wide range of applications.

The earliest attempts to devise systems for automatic voice recognition by machine were made in the 1950's, when various researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, at Bell Laboratories, Davis, built a system for isolated digit recognition for a single speaker [12]. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In 1959 another attempt was made, at MIT Lincoln Laboratories. Ten vowels embedded in a/b/-vowel-/t/ format were recognized in a speaker independent manner [10]. Mainly voice recognition began as early as the 1960's with exploration into voiceprint analysis, where characteristics of an individual's voice were thought to be able to characterize the uniqueness of an individual much like a fingerprint. The early systems had many flaws and research ensued to derive a more reliable method of predicting the correlation between two sets of voice utterances. Speaker identification research continues today under the realm of the field of digital signal processing where many advances have taken place in recent years. The performance of the voice recognition systems is given in terms of a word error rate (%) as measured for a specified technology, for a given task, with specified task syntax, in a specified mode, and for a specified word vocabulary.
In the 1970's voice recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on the fundamental studies in Russia [9], Japan [10] and United States [5]. The Russian studies helped advance the use of pattern recognition ideas in speech recognition; the Japanese research showed

how dynamic programming methods could be successfully applied; and United States research showed how the ideas of linear predicting coding(LPC). At AT&T Bell Labs, began a series of experiments aimed at making speech recognition systems that were truly speaker independent [03]. They used a wide range of sophisticated clustering algorithms to determine the number of distinct patterns required to represent all variations of different words across a wide user population. In the 1980's a shift in technology from template-based approaches to statistical modeling methods, especially the hidden Markov model (HMM) approach [1].

The main goal of this work is to create a device that could recognize one's voice as a unique biometric signal and compare it against a database to choose the person's identity or deny an unregistered person while being as standalone as possible. A human can easily recognize a familiar voice however; getting a computer to distinguish a particular voice among others is a more difficult task. Immediately, several problems arise when trying to write a voice recognition algorithm. The majority of these difficulties are due to the fact that it is almost impossible to say a word exactly the same way on two different occasions. Some factors that continuously change in human speech are how fast the word is spoken, emphasizing different parts of the word, etc… In order to analyze two sound files in time domain, the recordings would have to be aligned just right so that both recordings would begin at precisely the same moment.

II. VOICE RECOGNITION

Voice recognition is "the technology by which sounds, words or phrases spoken by humans are converted into electrical signals, and these signals are transformed into coding patterns to which meaning has been assigned". We focus here on the human voice because we most often and most naturally use our voices to communicate our ideas to others in our immediate surroundings, so voice recognition is aimed toward identifying the person who is speaking .voice recognition works by analyzing the features of speech that differ between individuals. Everyone has a unique pattern of voice stemming from their anatomy (the size and shape of the mouth and throat) and behavioral patterns (their voice's pitch, their speaking style, accent, and so on).The applications of voice recognition are markedly different from those of voice recognition. Most commonly, voice recognition technology is used to verify a human's identity or determine an unknown person identity. Human verification and human identification are both common types of voice recognition. The most common approaches to voice recognition can be divided into two classes: "template matching" and "feature analysis". Template matching is the simplest technique and has the highest accuracy when used properly, but it also suffers from the most limitations. As with any approach to voice recognition, the first step is for the user to speak a word or phrase into a microphone. The electrical signal from the microphone is digitized by an "analog-to-digital (A/D) converter", and is stored in memory. To determine the "meaning" of this voice input, the computer attempts to match the input with a digitized voice sample, or template that has a known meaning. The program contains the input template, and attempts to match this template with the actual input using a simple conditional statement. Since each person's voice is different, the program cannot possibly contain a template for each potential user, so the program must first be "trained" with a new user's voice input before that user's voice can be recognized by the program. A more general form of voice recognition is available through feature analysis (MFCC) and this technique usually leads to "speaker-independent" voice recognition. Recognition accuracy for speaker-independent systems is usually between 90 and 95 percent.

One more method is used for voice reorganization, which is based on HMM (Hidden Markov Model) algorithm.

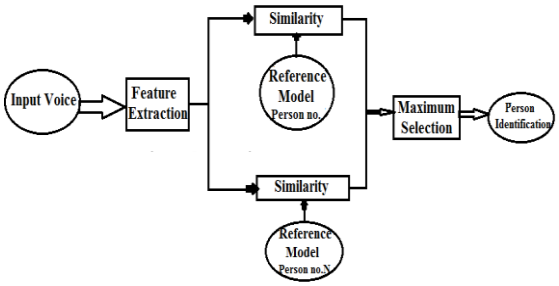


Fig 2: Voice recognition process using feature extraction.

Voice recognition technologies based on Hidden Markov Model (HMM) have developed considerably and can provide high recognition accuracy. HMM is a statistical modeling approach and is defined by three sets of probabilities: the initial state probability, the state transition probability matrix, and the output probability matrix. The computation cost of a typical HMM-based speech recognition algorithm is very high, which depends on the number of states for each word and all words, the number of Gaussian mixtures, the number of speech frames, the number of features for each speech frame and the size of the vocabulary.

In this paper, HMM classifier which will accept the MFCC feature of voice as an input and pass through different model to generate the output. The description of the method is subsequently given in this paper.

III. SYSTEM DESCRIPTION

Hidden Markov Model filter receive two inputs. One is sample voice database and other is real time input voice. To create data base as HMM model first human voice sample is taken, and then Voice Activity Detection (VAD) separate actual date from the samples. MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the
Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation. Out of this extracted MFCC, HMM is generated which is also training phase for filter which models the given problem as a "doubly stochastic process" in which the observed data are thought to be the result of having passed the "true" (hidden) process and that is how database model is created . Same process of MFCC extraction for real time input voice is performed. HMM filter compare this two hmm values and short out best match between input voice and database. And thus human voice is recognized.

.

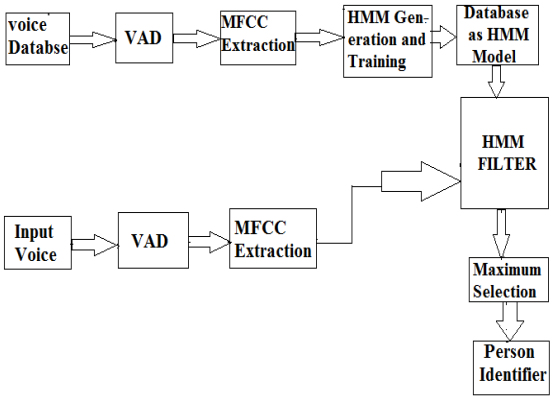


Fig 3: Training and Testing of HMM

3.1 VOICE ACTIVITY DETECTION

Voice Activity Detection (VAD) determines which parts of a voice signal are actual data and which are silence. The VAD algorithm used here utilizes the short-time energy, and zero crossing rates to decide if there is voice activity. Mel-Frequency Cepstral Coefficients (MFCC) was used to extract characteristic information from the speech vectors.

3.2 MEL-FREQUENCY CEPSTRAL COEFFICIENTS

The MFCC is the most evident example of a feature set that is extensively used in voice recognition. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation [8] [9]. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by Mel (f) = 2595*log10 (1+f/700). The following figure shows the steps involved in MFCC feature extraction.

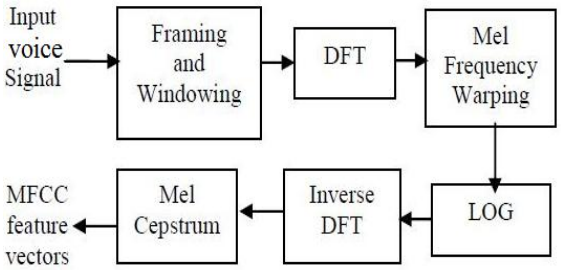


Figure 4: Steps involved in MFCC feature Extraction

3.3 HIDDEN MARKOV MODEL (HMM)

The HMM is a stochastic approach which models the given problem as a "doubly stochastic process" in which the observed data are thought to be the result of having passed the "true" (hidden) process through a second process. Both processes are to be characterized using only the one that could be observed. The problem with this approach is that one do not know anything about the Markov chains that generate the speech. The number of states in the model is unknown, there probabilistic functions are unknown and one cannot tell from which state an observation was produced. These properties are hidden, and thereby the name hidden Markov model. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

IV. IMPLEMENTATION & RESULTS

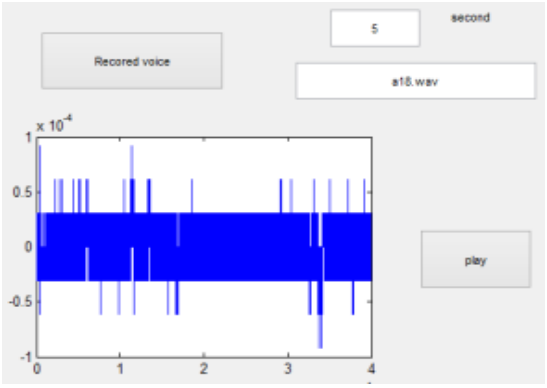
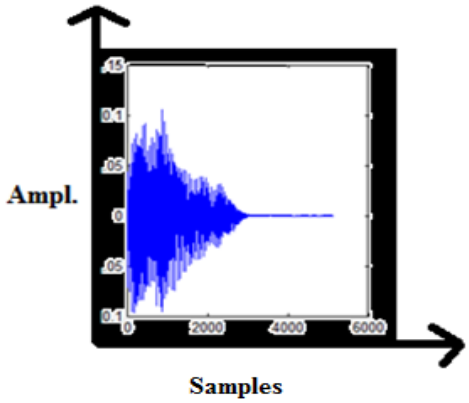
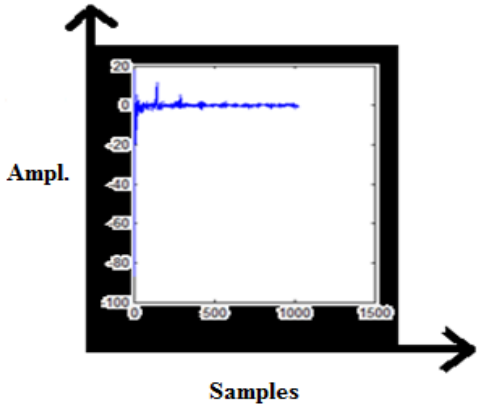


Fig 5: Recoded Voice



(a)



(b)

Fig 6: (a) Input Voice samples, (b) MFCC samples

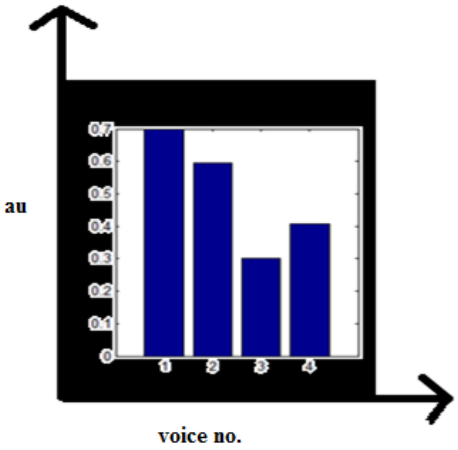
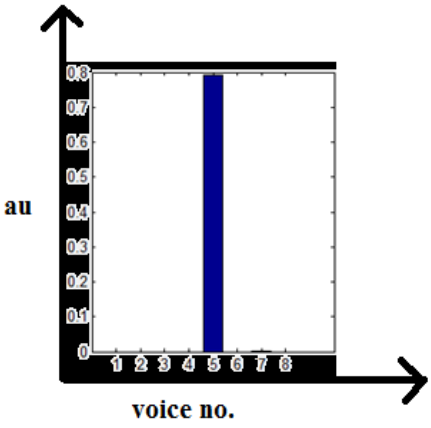
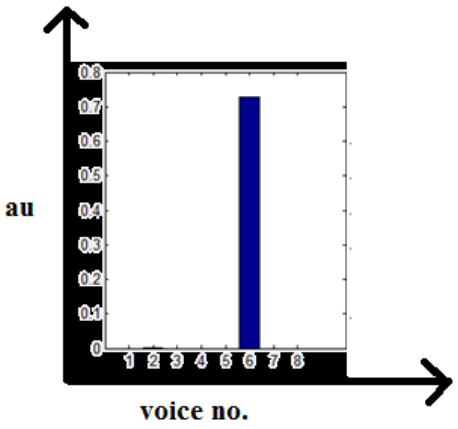


Fig 7: HMM coefficients after training



(a)



(b)

Fig 8: (a) and (b) Final Output with Voice no

Table1. Recognition rate with different input voice samples

S. No.	No. of training voice	Recognition rate
1	8	0.875
2	16	0.937
3	24	0.958

V. CONCLUSION

Voice Recognition technology is relate to taking the human voice and converting it into words, commands, or a variety of interactive applications. In addition, voice recognition takes this application one step further by using it to verify, identity, and understand basic commands. These technologies will play a greater role in the future and even threaten to make the keyboard obsolete. Initially looking at the experiment, the plan was to have a text-dependent system, or a speaker verification system, or something that could actually determine what word was being spoken to the system by the user. It has become painfully clear that that would be a very difficult task to accomplish, and would require much more time, effort. Our system is, relatively successful –it identified speakers at a rate of almost 80-95% - a very good recognition rate for a basic system.

REFERENCES

[1] M. Yuan, T. Lee, P. C. Ching, and Y. Zhu, "Speech recognition on DSP: Issues on computational efficiency and performance analysis," in *Proc. IEEE ICCCAS*, 2005.

[2] B. Burchard, R. Roemer, and O. Fox, "A single chip phoneme based. HMM speech recognition system for consumer applications," *IEEE Trans. Consumer Electron.*, vol. 46, no. 3, Aug. 2000.

[3] U. C. Pazhayaveetil,"Hardware implementation of a low power speech recognition system," Ph.D. dissertation, Dept. Elect. Eng., North Car-olina State Univ., Raleigh, NC, 2007.

[4] Chadawan I., Siwat S. and Thaweesak Y.," Speech Recognition using MFCC".International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)July 28-29, 2012 Pattaya (Thailand)

[5] W. Han, K. Hon, Ch. Chan, T. Lee, Ch. Choy, K. Pun, and P. C. Ching, "An HMM-based speech recognition IC," in *Proc. IEEE ISCAS*, 2003,

[6] P. Li and H. Tang, "Design a co-processor for output probability Calculation in speech recognition," in *Proc. IEEE ISCAS*, 2009,

[7] Jia-Ching Wang, Jhing-Fa Wang*, Yu-Sheng Weng "Chipdesign of MFCC extraction for speech recognition" INTEGRATION, the VLSI journal 32 (2002)

[8] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia

[9] DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003,

[10] Mahdi Shaneh, and Azizollah Taheri "Voice Command Recognition System Based on MFCC and VQ Algorithms"World Academy of Science, Engineering and Technology

[11] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi"Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617

[12] Kashyap Patel, R.K.Prasad ."Speech Recognition and Verification Using MFCC &VQ" International Journal of Emerging Science and Engineering(IJESE) ISSN: 2319–6378,Volume-1, Issue-7, May 2013