Development of FPGA based Human Voice Recognition System with MFCC feature

Mr. Anand Mantri[#1], Mr. Mukesh Tiwari[*2], Mr. Jaikaran Singh[#3]

[#]*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 voice recognition development on the hardware is also a challenging field due to minimal hardware resource utilization with higher accuracy demand. This paper described the FPGA based human voice recognition system. This system includes voice activity detection, MFCC feature extraction, HMM filter generation and classification of voice. The system is train for different person with the repeated voice for achieving the higher accuracy. The HMM filter coefficient for different person is only stored for reducing the memory utilization. The standard FFT and DCT blocks are used to reduce the hardware utilization. The development results are given in this paper for illustrating the effectiveness of system.**

***Keywords*— FPGA, Voice Recognition, MFCC (Mel-Frequency Cepstral Coefficients), HMM (Hidden Markov Model), VAD (Voice Activity Detection), LPC.**

I. Introduction

There are billions of human beings around the world speaking different languages and yet we are able to recognize someone by listening to someone's conversation or speech as long as we can understand the language. We can also usually recall someone's voice even we have not seen that person for years. In the movies, we have all seen how robots and ultimately computers can understand human voice commands and even speeches and in some cases they can even speak our language and have an intelligent and interactive conversation with us. The entire subjects of voice & speech recognition are often confused and the terms are often misused. Knowing what's being spoken is very different than knowing who is speaking. Voice recognition focuses on who is speaking it and the speech recognition concentrates on what is being spoken. Base on above then there are the difference between speaker verifications and identifications. Today, we have good success with voice & speech recognitions in controlled environment. However, the application becomes very limited for a technology that can only work in controlled environment. Voice recognition technology utilizes the distinctive aspects of the voice to verify the identity of individuals. Voice recognition is occasionally confused with speech recognition, a technology which translates what a user is saying (a process unrelated to authentication).

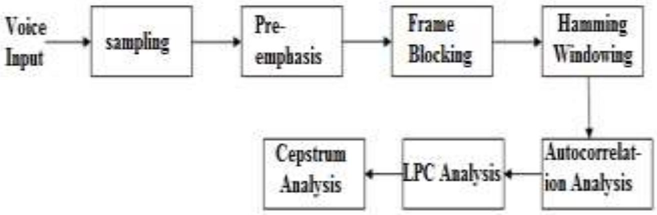


Fig1:- Basic process of voice recognition

Voice recognition technology, by contrast, verifies the identity of the individual who is speaking. The two technologies are often bundled – speech recognition is used to translate the spoken word into an account number, and voice recognition verifies the vocal characteristics against those associated with this account. Most of the speech recognition systems for embedded consumer electronics applications are implemented by general purpose processors like the micro-controller (MCU), digital signal processor (DSP), or a combination of both [1-13]. Custom-designed processors [9] implemented in application-specific integrated circuit (ASIC) or field-programmable gate array (FPGA) for the entire CHMM-based voice recognition algorithm are usually considered as an effective solution for low-power high-performance embedded voice recognition systems compared to general purpose processor-based implementations. One of the most commonly used acoustic features are Mel-frequency cepstral coefficients (MFCCs) [2] which give us a smooth PSD estimation of the speech frames. During the last 30 years several different approaches are presented for speech recognition to cope with its complexities and difficulties [14-18]. One of the widely adopted classification techniques in many patent recognition systems is the *k*-NN classification [8]. Voice recognition is increasingly popular in embedded applications. For fast time-to-market, a general-purpose DSP, MCU (Micro-controller), or a combination of both [1]. Since these implementations are not dedicated designs for speech recognition systems, they are not able to process efficiently the vast amount of vector operations required by complex speech recognition algorithms [4,19] are main implementation approaches for embedded speech recognition systems. Over the last decade, Hidden Markov Model (HMM) based speech recognition has become increasingly popular and many of today's state-of-the-art software systems rely on the use of HMMs to calculate the probability that a particular audio sample matches a specific acoustic characteristic of a given word [2, 13]. Hardware-based voice recognition systems meet these requirements. Previous research on custom hardware described the implementation of the HMM algorithm using application-specific integrated circuits (ASICs) [1,10] and field-programmable gate arrays (FPGAs) [3, 6]. Word speech recognition employs a word HMM or a phoneme HMM in acoustic models. In particular, the word HMM is adopted in [2, 7], and the phoneme HMM is adopted in [1, 21].

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. THEORETICAL BACKGROUND OF VOICE RECOGNITION

Voice recognition is a technique in computing technology by which specialized software and systems are created to identify, distinguish and authenticate the voice of an individual speaker. Voice recognition evaluates the voice biometrics of an individual, such as the frequency and flow of their voice and their natural accent. Voice recognition is also known as speaker recognition.

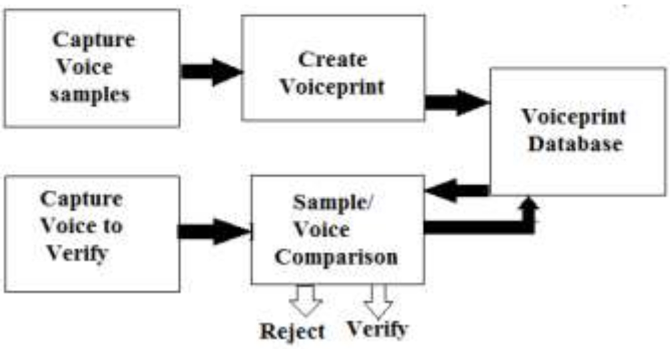


Fig 2:- Voice Enrollment and verification process

Voice recognition is the ability of a device or program to receive and interpret dictation, or to identify and execute spoken commands. There are a number of voice recognition systems available on the market and the most powerful can identify thousands of words. Voice recognition is where the computer understands commands that are spoken into a microphone by the user. This technology still isn't quite reliable, but it can be used to dictate documents if properly trained. Voice Recognition is software and hardware that is able to decipher the human voice to either perform commands or to transcribe speech. You find it on cell phones with voice calling and can buy software for your PC that will transcribe your words as you speak. Voice recognition powered systems are primarily designed to recognize the voice of the person speaking. Before being able to recognize the voice of the speaker, voice recognition techniques require some training in which the underlying system will learn the voice, accent and tone of the speaker. This is generally accomplished through a series of textual words and statements that the person has to speak through the built-in or external microphone. Voice recognition systems are related to speech recognition systems but the former only identifies the speaker whereas the latter can understand and evaluate what has been said.

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.

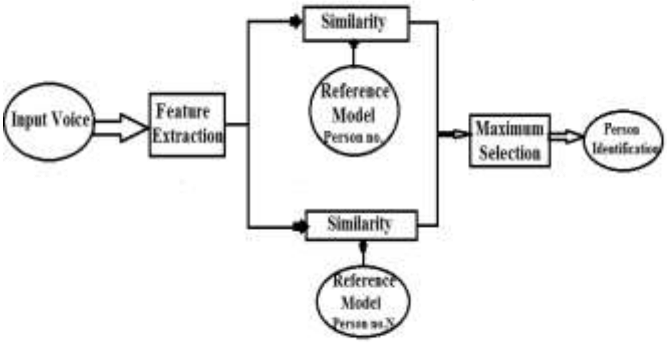
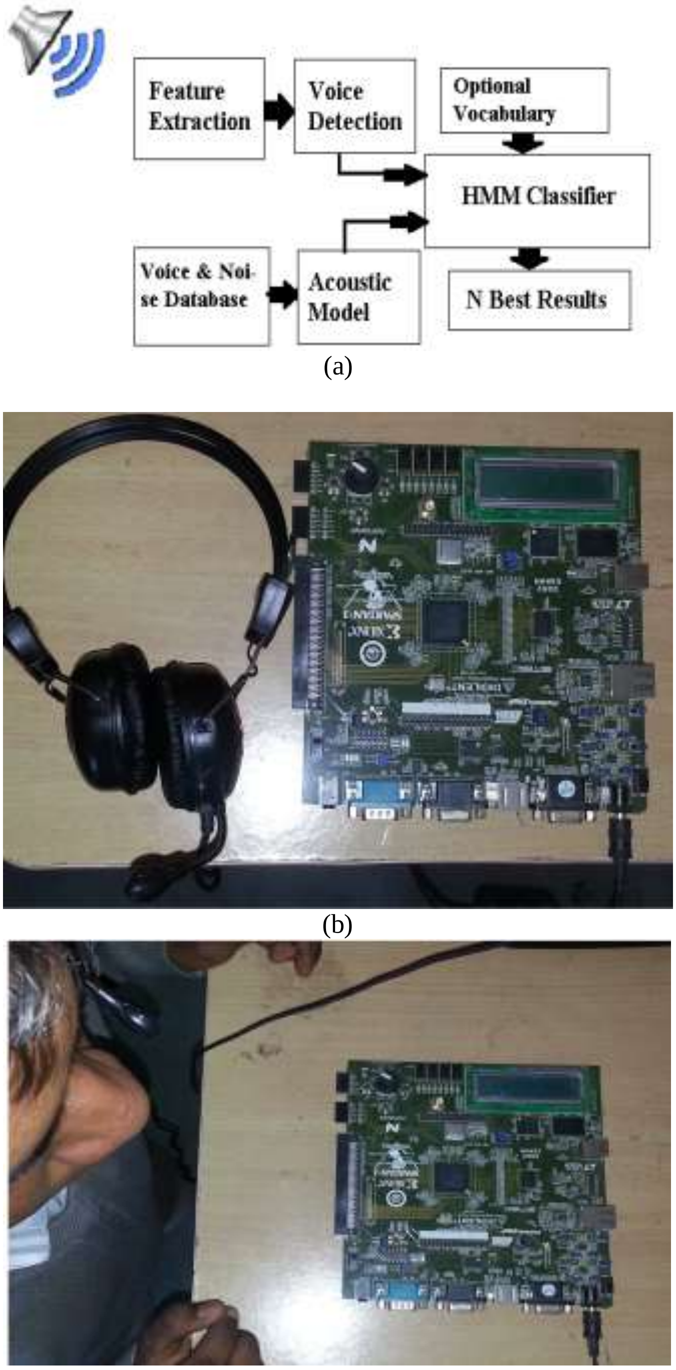


Fig 3: Voice recognition process using feature extraction.

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.

III. SYSTEM DESCRIPTION

The FPGA based voice recognition system is shown in figure given below.



(a)

(b)

(c)

Fig 4: - (a) Block diagram of system, (b,c) FPGA Board with microphone

This system is based on audio input of sparten-3 based FPGA board which is digitized and interfaced with sparten-3 FPGA processor. This processor is used to process the digitized audio input and extract the voice activity signal . This signal is processed to generate the MFCC coefficient which is pass through HMM filter for getting the feature database and classified output. 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.

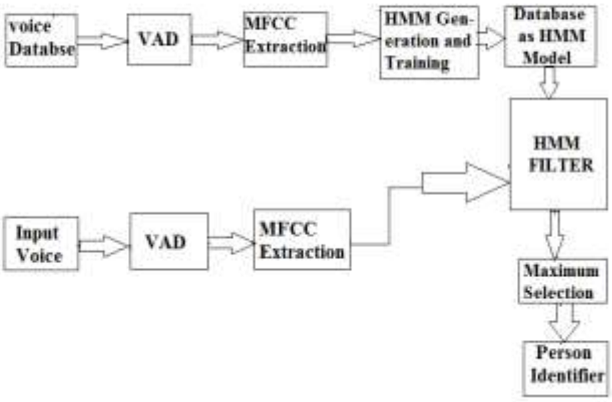


Fig 5: Training and Testing of HMM

A. *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.

B. *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 centre 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.

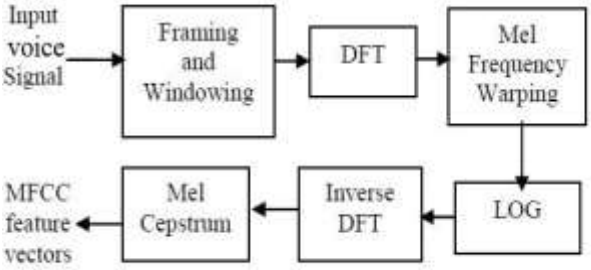


Figure 6: Steps involved in MFCC feature Extraction

C. 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.

IV. Implementation and Results

The various results of different develop subsystem are given as follows

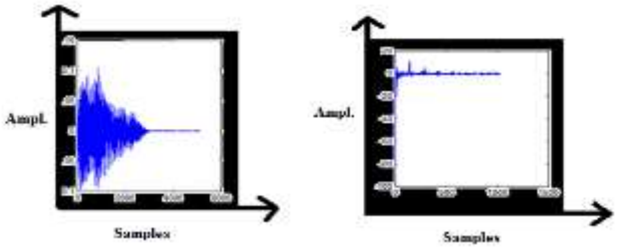


Fig 7: Input Voice samples and MFCC samples

Table1. Recognition rate with different input voice samples

S. No.	No. of training voice	Recognition rate
1	8	0.789
2	16	0.855
3	24	0.885

V. Conclusions

The implementation of voice recognition system has been completed. The optimized hardware utilization and low complexity has been achieved for the system development. 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. Our system is, relatively successful –it identified speakers at a rate of almost 70-90% - a very good recognition rate for a basic system.

References

[1] X. Zhu, Y. Chen, J. Liu, R. Liu, ”A Novel Efficient Decoding Algorithm.For CDHMM-based Speech Recognizer on Chip.” Proc. of IEEE ICA-SSP, 2003.

[2] M. Yuan, T. Lee, P. C. Ching, Y. Zhu, ”Speech Recognition on DSP:Issues on Computational Efficiency and Performance Analysis ”Proc. Of IEEE ICCCAS, 2005.

[3] Zh. Yang, J. Liu, E. Chan, L. Guan, Ch. Ching, ”DSP-based Systemon-Chip Moves Speech Recognition from the Lab to Portable Devices”,www.embedded.com, Jan 2007.

[4] S. Yoshizawa, N.Wada, N. Hayasaka, Y. Miyanaga, ”Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete system”. IEEE Trans. on Circuits and Systems I, Vol. 53, No. 1, Jan 2006, pp. 70-77.

[5] 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, pp. 852–856.

[6] S. Dobler, “Speech recognition technology for mobile phones,” *Ericsson Rev.*, vol. 77, no. 3, pp. 148–155, 2000.

[7] [Online]. Available: http://www.sensoryinc.com

[8] [Online]. Available: http://www.voicecontrol.com

[9] [Online]. Available: http://www.voicegate.com/voiceics.htm

[10] P. Placeway, *et al*, “The 1996 HUB-4 Sphinx-3 System”, *Proc. DARPA Speech Recognition Workshop*, Feb. 1997.

[11] K.K. Agaram, S.W. Keckler, D. Burger, “Characterizing the SPHINX Speech Recognition System”, University of Texas at Austin, Department of Computer Sciences, *Technical Report TR2001-18*, January 2001.

[12] J. Pihl, T. Svendsen, and M. H. Johnsen, “A VLSI implementation of pdf computations in HMM based speech recognition,” in *Proc. IEEE TENCON’96*, 1996, pp. 241–246

[13] W. Han, K. Hon, and C. Chan, “An HMM-based speech recognition IC,” in *Proc. IEEE ISCAS’03*, vol. 2, 2003, pp. 744–747.

[14] S. J. Melnikoff, S. Quigley, and M. J. Russell, “Implementing a simple continuous speech recognition system on an FPGA,” in *Proc. IEEE Symp. FPGAs for Custom Computing Machines (FCCM’02)*, 2002, pp. 275–276.

[15] F. Vargas, R. Fagundes, and D. Barros, “A FPGA-based Viterbi algorithm implementation for speech recognition systems,” in *Proc. IEEE ICASSP’01*, vol. 2, May s2001, pp. 1217–1220.

[16] L. R. Rabiner, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1211–1234, 1985.

[17] M. Karnjanadecha and S. A. Zahorian, "Signal modeling for isolated word recognition," in *Proc. IEEE ICASSP'99*, vol. 1, Mar. 1999, pp. 293–296.

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

[19] S. Nedevschi, R. K. Patra, and E. A. Brewer, "Hardware speech recognition

[20] for user interfaces in lowcost, lowpower devices," in *Proc. IEEE DAC*, 2005, pp. 684–689.

[21] 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, pp. 744–747.

[22] R. Krishna, S. Mahlke, and T. Austin, "Architectural optimizations for low-power, real-time speech recognition," in *Proc. ACMCASES*, 2003, pp. 220–231.