ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

Speech to Text Conversion Using Discrete Hidden Markov Model

S.Suganya[1], G.Premalatha[2]

Assistant Professor, Dept. of ECE, CK College of Engineering & Technology, Cuddalore, TamilNadu, India[1,2]

ABSTRACT:In recent years, Speech Recognition has the great development in the automation industry**.** This paper proposes an Automatic Speech Recognition (ASR) to facilitate an interaction between human and the electronic components. The main concern of this paper involves the suppression of various noises to achieve a robust speech recognition system. Discrete Hidden Markov Model is used to increase the speed of speech recognition. This paper explores the hardware realization of desired speech recognition system on the Field Programmable Gate array (FPGA). The accuracy has to be increased to get the clear and robust Speech Recognition. The speech features can be extracted through the cepstral coefficients by using warping filter banks. The cepstral coefficients are used to increase the robustness of Speech Recognition. To minimize the complexity of desired ASR system, the number of coefficients has to be minimized. The Speech- to-Text conversion is the main objective of this paper. This can be achieved by using Matlab software, Modelsim 6.4a for simulation and can be implemented in Altera DE2 board.

KEYWORDS: Discrete Hidden Markov Model, Feature Extraction, Speech-to-text, Warping filter banks.

I.INTRODUCTION

Speech Recognition has been the most dominant and convenient means of communication. Speech communication is not only a face-to-face interaction but also the individuals at any moment, via a wide variety of modern technological media. Speech Recognition plays an important role that a human to make an interaction with the electronic components. Actually, Speech Recognition is also a kind of Pattern Recognition technique. Various applications of Speech Recognition includes voice controlled devices, speech-to-text, etc. Automatic speech Recognition can be resolved into two phase viz., training phase and testing phase. In training phase, the speech feature vectors can be extracted and is trained in the codebook. In testing phase, the feature vectors can be obtained as in the testing phase and also comparing the testing features are matched with the codebook. If both the features are similar, this given speech can be recognized and that can be utilized for an authentication purpose.

Section II describes that the literature survey that have referred for proposed methodology. Section III describes the proposed methodology of this paper. Section IV explores the steps has been followed in the feature extraction. Section V explodes the theory of DiscreteHidden Markov Model. Section VI explains the hardware architecture which is designed to implement in FPGA. Section VII discussed the results which are obtained for the desired system. Section VIII explains the concepts which are concluded from this paper. Section IX explains the concepts from various papers that are referred.

II.RELATED WORK

Speech Recognition is an advance technique to be followed in the fields of Automation, Artificial Intelligence and so on. The improvement in the recognition accuracy, robustness may cause effects on the performance of ASR. Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are the various methods of feature extraction techniques available for ASR. MFCC and PLP are most widely used feature extraction techniques which are required to reconstruct the original signal [1]. The recent focus of researchers involved in the implementation of ASR into an embedded platform [2]. The speech-to-text conversion is a useful technique which is helpful for handicapped peoples [3].This paper focusing on the speech-to-text conversion and can be implemented in Altera DE2 board. The robustness of speech recognition can be improved by suppressing the noise from the speech signal. The conversion of speech-to-text conversion depends on the variation in the frequency. As the number of cepstral coefficients increases, the desired system gets complex to achieve the desired goal. So the number of coefficients here is 12.

DOI:10.15662/IJAREEIE.2016.0504131

The speech signal has to be in wave format, then only the signal has to be processed and feature extraction can be done. The speech signal is sampled at 16000 Hz and the number of bits per sample as 32. The reasonable modeling of speech signal can be done according to the assumption that such a small segment of speech is sufficiently stationary [4].

III.PROPOSED METHODOLOGY

The Proposed methodologyconsists of the feature extraction module and the codebook generation. The proposed architecture has been shown in Fig. 1, which explodes the steps has been followed in this paper. The original signal has to be split up into number of frames according to the frame length. The low frequency signals are selected by blocking the low frequency in every frame. After frame blocking of each and every frames, Hamming window is applied to reduce the discontinuity of the signal. Determine the DCT coefficients for the evaluation of cepstral coefficients by using filter bank spacing to each and every windowed frame. Apply logarithmic values to get the DCT values as a single value [5].

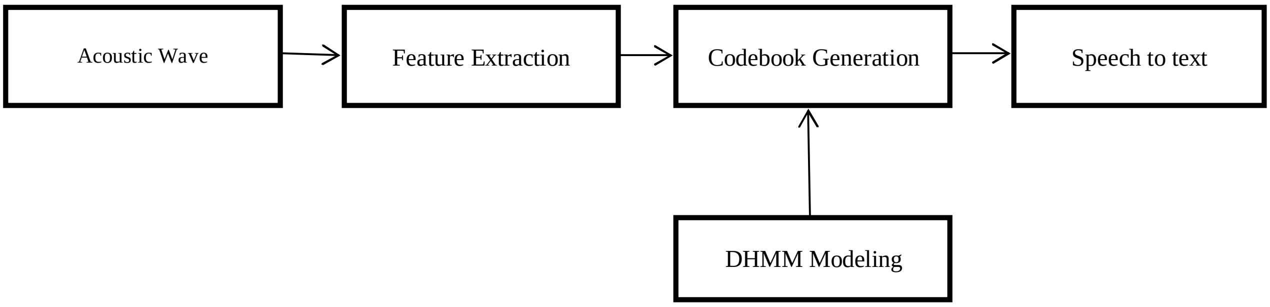


Fig. 1 Block Diagram of Proposed Methodology

IV.FEATURE EXTRACTION TECHNIQUES

Fig.2 describes the block diagram of feature extraction techniques. The steps which followed for the desired system has been explained below.

1. Frame blocking

With an appropriate time length for each frame, frame blocking is applied to divide the signals into matrix form. This speech signal is sampled at 16 kHz, the number of frames could be assumed as 320 samples within a frame. Overlapping of frames would have the factor of separation of samples due to the effect of frame blocking [5].

Hamming Windowing

To reduce the discontinuities ofsignal at the end of each frames, hamming windowing is applied to each and every frames after frame blocking. The equation (3.1) representing the discrete time representation of signal,

$$w(n) = 0.54 - 0.46Cos\left(\frac{2\pi n}{N-1}\right) \qquad (3.1)$$

By introducing hamming windowing to each frames, windowing generates the least distortion[5].

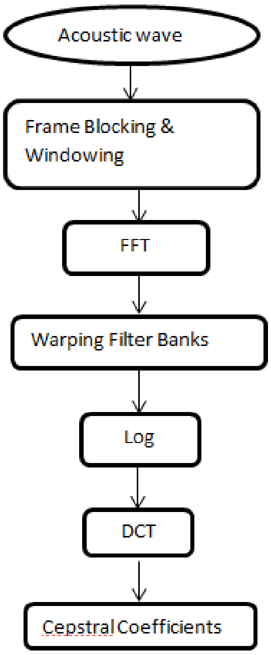


Fig. 2 Feature Extraction Techniques

2. Fast Fourier Transform

Time domain signal can beconverted into frequency domain by applying fast fourier transform to each and every windowed frames. The output of FFT can be complex numbers having both real and imaginary parts. Real time data has to be processed with the speech recognition system. The complex variables could be neglected by the FFT[5]. Equation (3.2) describes the spectral domain

$$|X(n)| = \sqrt{[Re(X(n))]^2 + [Im(X(n))]^2} \qquad (3.2)$$

3. Warping Filter Banks

The warping filter banks are used here to fetch the exact data without any loss. The comparison of two filter banks were done with the help of Mel Frequency filter banks and Bark frequency filter banks.

Mel Frequency Filter Banks

Based on the humanperception, the Mel frequency analysis is preferable. The human ear is very sensitive and it is proved that humans having high resolution to the low frequency rather than the higher frequency. Speech signal does not be linear. To make a linear scale conversion for thefrequency using Mel scale is used to warping a signal in frequency domain to the Mel scale. The conversion of speech signal from frequency domain to Mel scale can be done using the following equation (3.3).

$$Mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \qquad (3.3)$$

The Mel Filter bank spacing has to be applied to the FFT values to get the conversion for the frequency domain into the Mel scale. Triangular band pass filters are applied as a filter bank spacing which his non -uniformly spaced on the linear frequency axis and it is uniformly spaced on the linear frequency axis, with the larger number of filters in the low frequency region and lesser number of filters in the high frequency region and is shown in Fig. 3.

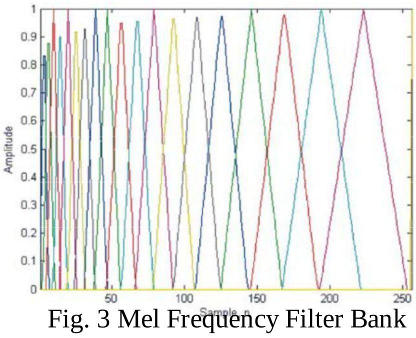


Fig. 3 Mel Frequency Filter Bank

Bark Frequency Filter Bank

The bark-scale is originally defined in Zwicker (1961). A distance of 1 on the bark scale is known as a critical band. The implementation provided in this function is described in Traunmuller (1990). An approximate expression for the Bark scale frequency warping, due to Schroedinger is used in these proposed method [6].

$$Bark\,(f) = 6\log\left[\left(\frac{f}{600}\right) + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\,\right] \qquad (3.4)$$

The Bark frequency filter bank spacing is applied to the FFT. It is uniform speed to collect more information with the input wave. It is shown in Fig.4

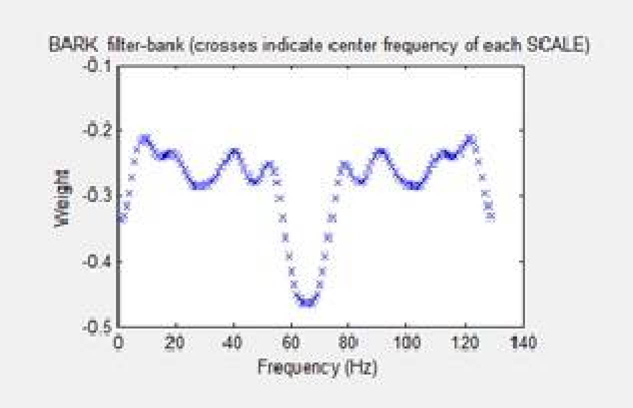


Fig. 4 Bark Frequency Filter Bank

4. Logarithm of Energies

To compute the log-energy, i.e.,the logarithm of the sum of filtered components for each filter.Equation(3.4) expresses the computing logarithm of weighted sum of spectral values in the filter-bank channel. At this stage, the number of rows equal to the number of columns equal to the number of filters in the filter bank.

$$S(m) = log_{10}\left[\sum_{n=0}^{N-1} |x(n)|^2 . Hm(n)\ \right], 0 \le m \le M \qquad (3.4)$$

5. Discrete Cosine Transform

The cepstral analysisincludes the conversion of spatial domain to frequency domain by applying DCT to the Mel Scale values. DCT expresses a finite set of data points in terms of a sum of cosine functions. The conversion of DCT is similar to the DFT in the conversion process, DCT is more preferable since the value obtained will provide us the accuracy for increasing the robustness of ASR. Equation (3.5) expresses the DCT,

$$C(n) = \sum_{m=0}^{M-1} S(m)\cos\,(\pi n(m+\frac{1}{2})/M\,],\, 0 < n < N \qquad (3.5)$$

The cepstral coefficients are obtained by applying DCT to the Mel scale values. The coefficients which are obtained after the evaluation of DCT called Mel Frequency Cepstral Coefficients (MFCC). When the number of coefficients increases the accuracy and also the increase in the complex of the designer complexity, there is a lag in design of ASR system. The number of coefficients taken here is 12 [5].

V.DISCRETE HIDDEN MARKOV MODEL

Discrete Hidden MarkovModel is used to accelerate the speed of Speech Recognition. A Codebook is to be first generated for the feature vectors. Feature vectors can be trained using DHMM in the codebook. From the training samples, the upper and lower bounds of each element has to be calculated to generate the codebook. The range of upper and lower bounds is divided into various sub-intervals from which the feature vectors are extracted. By randomizing the same number of vectors according to the number of classes, the initial codebook has to be formed. The codebook can be initialized with the values obtained from the feature vectors. DHMM is the only classifier based on probability. This paper utilizes this technique as a comparator based on the probability basis. Since DHMM is a time consuming process, it improves the efficiency of the desired system [7].

VI.HARDWARE ARCHITECTURE

The desired system can be implemented in the Altera DE2-70 Board. The desired system can be evaluated and can be implemented through the System On chip Architecture of FPGA. The SOC Architecture can be explained below as shown in Fig. 5. All the algorithms of the desired methodology can be implemented through the NIOS-II processor. The Altera DE2-70 development board in which the CYCLONE-II processor is included is used for this experiment. The Push button is used here for noise suppression. The toggle switch can be used as an input for the FPGA board. This can be used as an Authentication purpose. The microphone can be used as an output for checking the robustness of ASR. The Liquid Crystal Display will be used to display the words under the conversion of Speech-To-Text. An Audio controller is used to receive the speech signal. The I^2C protocol is used to control the register of the platform [7].

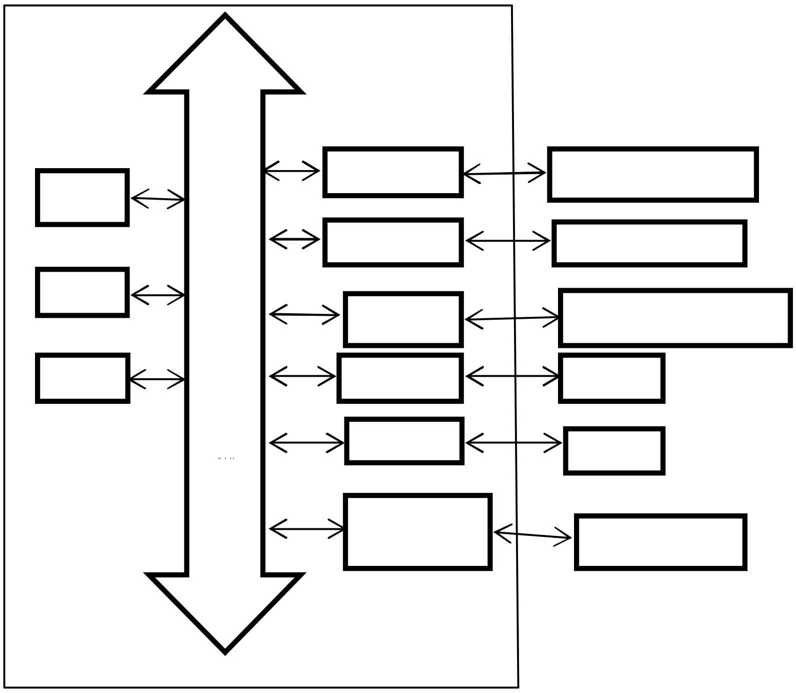
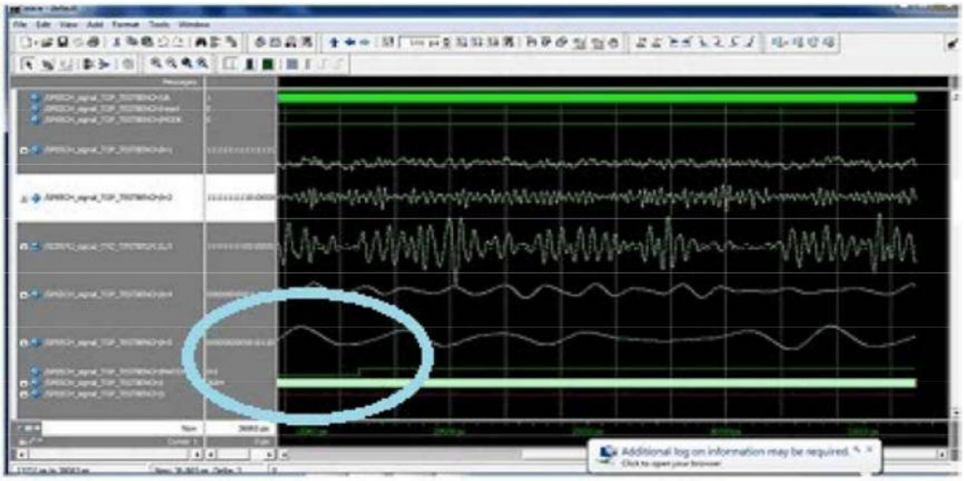


Fig. 5 SOC Architecture

VI.RESULTS & DISCUSSION

The codebook can be initialized with the help of feature vectors are obtained through the training phase. The speech vectors can be randomized and that can be evaluated through the suppression of environmental noises. The speech signal can be processed and that can be compared with the various hidden states using DHMM. When the feature vectors are similar and the pattern can be recognized using desired ASR. The simulation results can be obtained with the help of Modelsim 6.4a Software and this can be synthesized through the help of QUARTUS II software and which is helped to implement in the hardware realization. Fig 6.shows that the authorized speech, due to the match-out signal is high when the feature vectors are similar.



Additional log on information may be required.

Fig. 6 Authorized Speech

Fig. 7 shows that an Unauthorized speech, due to the match-out signal is low when the feature vectors are not similar as in the codebook values, i.e., the features extracted from the speech signal has been stored in the database as a codebook. The given speech signal is not similar with the trained speech feature vectors. So, the match out signal was low with the given input signal.

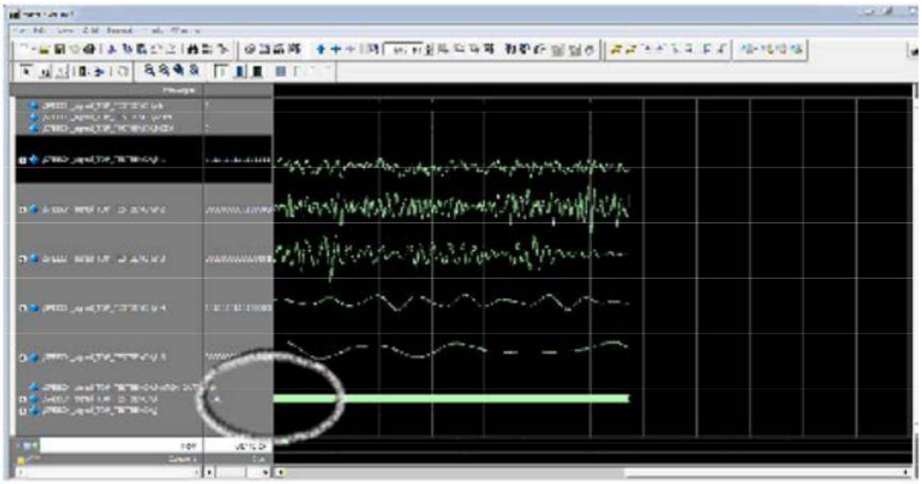
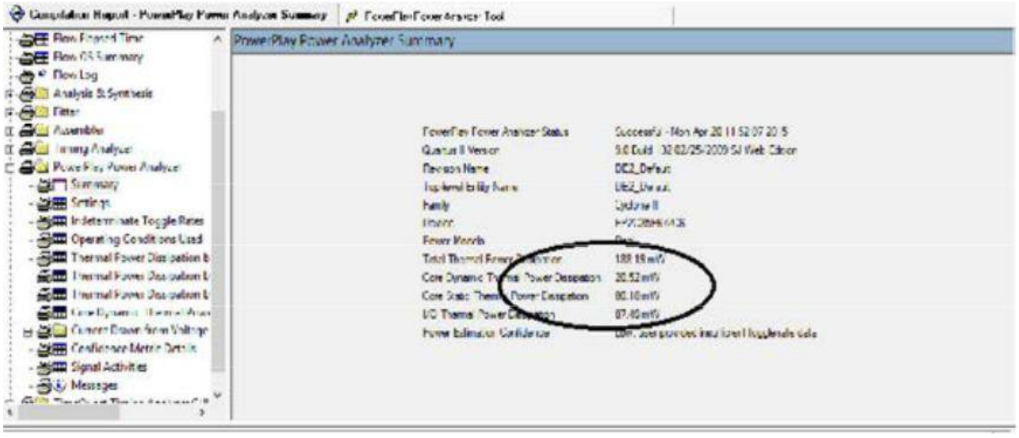


Fig. 7 Unauthorized Speech

Fig. 8 shows that the power dissipated from the desired system in the hardware. The number of logic gates requirement should be known before the hardware implementation, so the synthesis report would help us to get the data by using Modelsim 6.4a software.The power dissipated for the desired system is 188.18 mW. This analysed output explores the number of logic units required to design a robust speech recognition system.


Compilation Report - PowerPlay Power Analyzer Summary
PowerPlay Power Analyzer Tool
Flow Elapsed Time
Flow OS Summary
Flow Log
Analysis & Synthesis
Fitter
Assembler
Timing Analyzer
PowerPlay Power Analyzer
Summary
Settings
Indeterminate Toggle Rates
Operating Conditions Used
Thermal Power Dissipation b
Thermal Power Dissipation b
Thermal Power Dissipation b
Core Dynamic Thermal Pow
Current Drawn from Voltage
Confidence Metric Details
Signal Activities
Messages

PowerPlay Power Analyzer Summary

PowerPlay Power Analyzer Status	Successful - Mon Apr 20 11:52:07 2015
Quartus II Version	9.0 Build 132 02/25/2009 SJ Web Edition
Revision Name	DE2_Default
Top-level Entity Name	DE2_Default
Family	Cyclone II
Device	EP2C35F672C6
Power Models	Final
Total Thermal Power Dissipation	188.18 mW
Core Dynamic Thermal Power Dissipation	20.52 mW
Core Static Thermal Power Dissipation	80.10 mW
I/O Thermal Power Dissipation	87.56 mW
Power Estimation Confidence	Low: user provided insufficient toggle rate data

Fig. 8 Power Dissipation

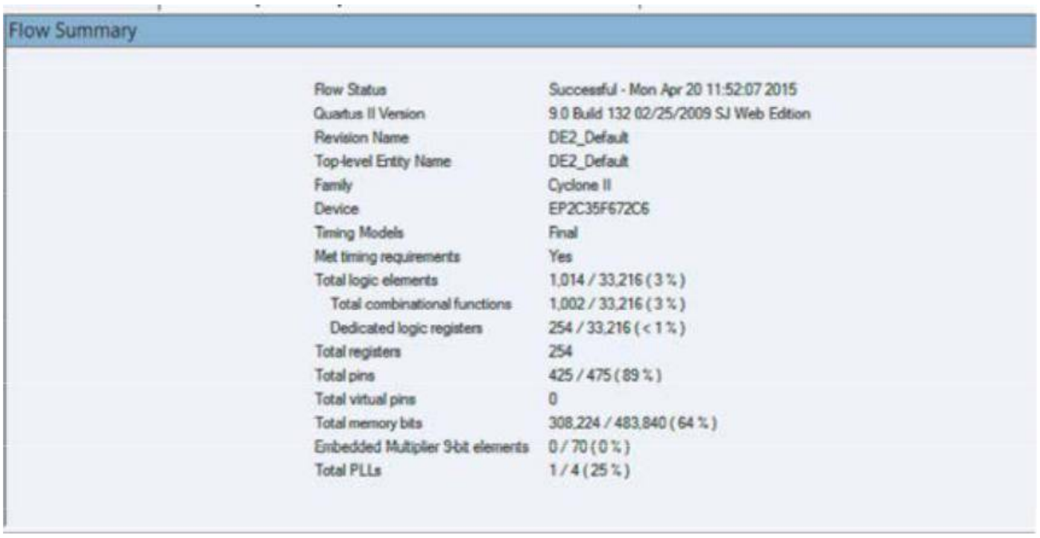
Fig.9 shows that the amount of area required occupying the logic elements in the desired proposed methodology. The synthesis report helps us to know the requirements of logic elements can be needed for the hardware implementation. Thetotal logic elements required to implement the desired system is 3% , the number of pins required is 429 and the number of registers used is 254 and the number of PLL required is 25%.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016



Flow Summary

Flow Status	Successful - Mon Apr 20 11:52:07 2015
Quartus II Version	9.0 Build 132 02/25/2009 SJ Web Edition
Revision Name	DE2_Default
Top-level Entity Name	DE2_Default
Family	Cyclone II
Device	EP2C35F672C6
Timing Models	Final
Met timing requirements	Yes
Total logic elements	1,014 / 33,216 (3 %)
Total combinational functions	1,002 / 33,216 (3 %)
Dedicated logic registers	254 / 33,216 (< 1 %)
Total registers	254
Total pins	425 / 475 (89 %)
Total virtual pins	0
Total memory bits	308,224 / 483,840 (64 %)
Embedded Multiplier 9-bit elements	0 / 70 (0 %)
Total PLLs	1 / 4 (25 %)

Fig. 9 Synthesis Report

VII.CONCLUSION

The speech signal can be processed and that can be trained and compared with the feature vectors that are obtained by processing the speech. DHMM technique is a slight time consuming process but it provides accuracy for robust speech recognition.The future work is to be implemented in ALTERA DE2 FPGA starter kit and this also can be used to convert the speech to text. The research work can be extended to activate the voice controlled devices for an authentication purpose.

REFERENCES

[1] Yuan Mang, “Speech Recognition on DSP : An Algorithm on Optimization & Performance Analysis”, The Chinese University of Hong Kong, pp.1-18, 2004.

[2] Huggins Daines. D, M.Kumar, A.Chan, A.Black, M.Ravishankar and A.Rudnicky, “PocketSphinx : A frees real-time continuous speech recognition system for hend-held devices”, in Proceedings of ICASSP, 2006.

[3] RumiaSulthana and Rajesh Palit, “A Survey on Bengali Speech-To-Text Recognition Techniques”, The 9th International Forum on strategic Technology, Cox’s Bazar, 2014.

[4] Muda Lindalsalwa, MumtajBegam and I.Elamvazhuthi, “Voice Recognition Algorithm using MFCC & DTW Techniques”, Journal of computing , ISSN 2151-9617, 2(3):138-143.

[5] Joshi, Siddhant C. and Dr. A.N.Cheeran, “MATLAB based Feature Extraction using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition”, IJESTR, 3(6), 2014.

[6] Ben J.Shannon, KuldipK.Paliwal, “A Comparitive study of Filter Bank Spacing for Speech Recognition”, 2003.

[7] Pan Shing-Tai and Xu-Yu Li, “An FPGA based Embedded Robust Speech Recognition system designed by combining Empirical Mode Decomposition and a Genetic Algorithm”, IEEE Transaction on Instrumentation and Measurement, 61(9),2012.

BIOGRAPHY

Ms. S. Suganya was born in Tamilnadu, India in 1991. She has received B.E., in Electronics & Instrumentation Engineering from Annamalai University (Autonomous University, Chidambaram) in 2012 and M.E., in Applied Electronics from Saveetha Engineering College, Thandalam(A constituent College of Anna University, Chennai) in 2015. She is interested in VLSI design, Image Processing, and Robotics. She is a member of ISTE. She is an Assistant Professor in Electronics & Communication Engineering in CK College of Engineering & Technology, Cuddalore (A Constituent College of Anna University, Chennai).

Ms. G.Premalathaborn in Tamilnadu, India in 1991. She pursued her Bachelors Degree in Adhiparasakthi Engineering College (A Constituent College of Anna University, Chennai) in 2013 and Masters Degree in CK College of Engineering & Technology, Cuddalore (A Constituent College of Anna University, Chennai) in 2015. She is interested in the fields of Image Processing, Wireless Sensor Networks and Robotics. She is an Assistant Professor in CK College of Engineering & Technology, Cuddalore (A Constituent College of Anna University, Chennai).