Proceedings of International Conference on Intelligent Knowledge Systems (IKS-2004), August 16-20,2004

VLSI IMPLEMENTATION OF SPEECH TO TEXT CONVERSION

Mrs. P. Nirmala Devi, ***Mr. R. Asokan,***
Email : nimmy_kec@yahoo.com, asokece@kongu.ac.in
Kongu Engineering College, Perundurai

ABSTRACT

In this paper an efficient method for conversion of speech into text is been proposed. The main aim of the project is to access the internet for the blind person. The project is designed using FPGA for the hardware realization. The commercial software available is mainly operating system dependent and less robust. This voice processor designed using FPGA so that it is operating system independent. We have taken up the complex speech recognition system and the word by word recognition is completed.

INTRODUCTION :

The main aim of the project is to help the blind people. We are designing a voice processor FPGA chip which helps a blind person to access the computer. The steps involved are training, database and pattern matching and text conversion. The complex part in this project is speech recognition. First the blind person's voice is trained and some list of frequent words pattern are stored in a database. If the blind person wants to access the internet, he has to speak the content. The words will be checked with the stored pattern, if the pattern is matched, then the matched patterns closest equivalent word is stored in target i.e., e-mail, word etc.

SPEECH RECOGNITION BASICS :

Communication helps us to convey our ideas. Speech is an important way of communication. For a blind person speech is the only way of communication.Talking to a computer instead of typing on a keyboard and being perfectly understood is an idea developed by Stanley kubrick in his famous film 2001, a Space Odyssey. We are now in 2004 and this dream is now becoming more and more a reality.

Automatic Speech Recognition is a powerful multimedia browsing tool: it allows us to easily search and index, recorded audio and video data. Speech recognition is also a useful form of input. It is especially useful when someone's eyes and hands are busy. It allows people working in active environment such as hospitals to use computers. It also allows handicapped people such as blind or palsy to use computers. Finally although knows how to talk, not as many people know how to type. With speech recognition, typing would no longer be a necessary skill for using a computer.

CHALLENGES :

In 1904, IBM was the first company to commercialize a dictation system based on speech recognition. Speech recognition has since been integrated in many applications such as

- Telephony applications
- Embedded systems
- Multimedia applications, like language learning tools

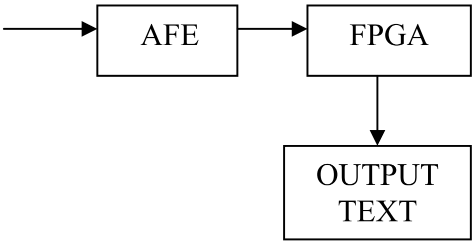
Many improvements have been realized since 50years but computers are still able to understand every single word pronounced by every one. Speech recognition is still a cumbersome problem. There are quite a lot of difficulties. The main one is that, two speakers uttering the same word, will say it very differently from each other. This problem is known as inter speaker variation(variation between speakers). In addition the same person does not pronounce the same word identically on different occasions. This is known as intra speaker variation. It means that even consecutive utterances of the same word by the same speaker will be different. A listner would not be confused by this, but a computer might. The waveforms of speech signals also depends on the environmental conditions (noise, reverberation). Noise and channel distortion are very difficult to handle, especially when there is no prior knowledge of the noise or distortion.

EXISTING SYSTEMS :

Today many software based voice processors are in the market. But all these are operating system dependent and also the reliability of the software is very low. Hence we are designing a dedicated voice processor hardware chip for speech to text conversion, which will be useful in home automation, on/off controls and particularly for the visually impaired to use the computer.

WORKING PRINCIPLE:

Recognition systems can be broken down into two main types. Pattern recognition systems compares the patterns to the known/trained patterns to determine a match. Acoustic Phonetic systems use knowledge of the human body (speech production and hearing) to compare speech features (phonetics such as vowels sounds). Most modern systems focus on pattern matching approach because it combines nicely with current computing techniques and tends to have higher accuracy.



The voice processor is designed for speaker dependent and word by word speech recognition system. The pattern recognition techniques is used for matching the words.

BLOCK DIAGRAM DESCRIPTION :

ADC_IF :

Converts 14 bit serial data from the analog front end board to parallel data.

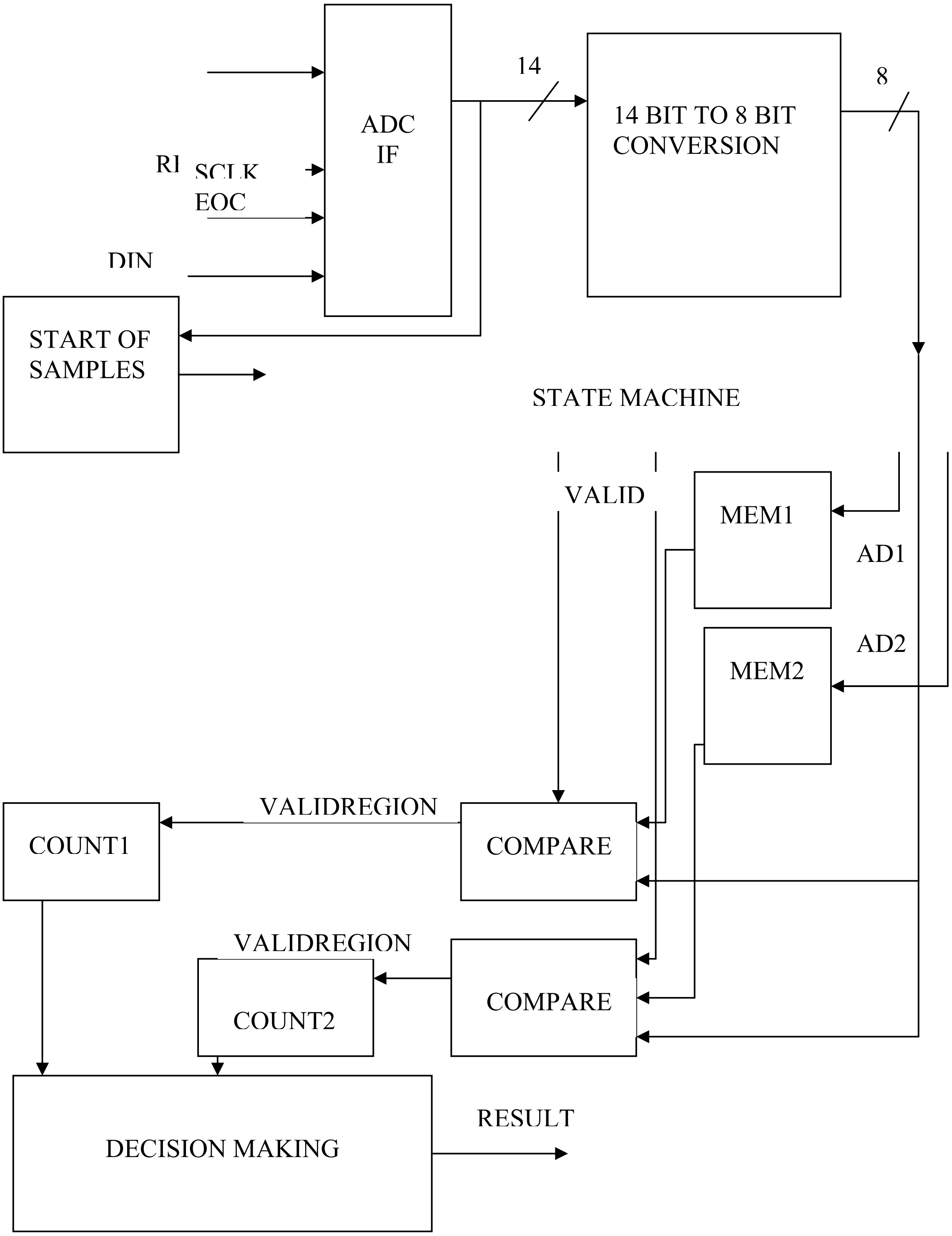
14 TO 8 BIT CONVERSION:

Digital representation of audio offers many advantages: high noise immunity, stability and reproducibility. Speech signals have a large dynamic range in the region of 60db, therefore requiring a large number of quantization levels. A typical channel would require a 12bit precision at a sampling rate of 8000 times/second. Hence there is a clear need for some form of speech compression techniques differ in the offered design complexities, audio quality, i.e. with minimal degradation and the amount of data compression performed. The ear is less sensitive to errors at high volume levels than at low volumes.i.e accuracy of lower amplitude parts of the signal is more important than the higher amplitude parts. Therefore logarithmic quantization can reduce this data rate to 8 bit with very little degradation. Standard techniques are the European A-Law PCM and the American U-Law PCM, both found in CCITT G.711 standard.

µ law audio compression :

The transformation is essentially logarithmic in nature and allows 8 bits per sample output to cover a dynamic range equivalent to 14 bits of linearly quantised values. Unlike the linear quantization, the logarithmic step spacing represent low-amplitude audio signals with more accuracy than the higher amplitude signals.

BLOCK DIAGRAM :



The μ law transformation is :

$c(x) / X_{MAX} = LN(1+\mu(X)/X_{MAX}) / \ln(1+\mu(X)) \quad 0 \leq (X)/X_{MAX} \leq 1$

START OF SAMPLES :

Output is obtained from the ADC irrespective of whether speech is there or not. i.e surrounding noise level also acts as input to the ADC. Therefore there is a need for a logic that would determine whether the ADC output has crossed the threshold. This block provides an output signal if and only if the ADC output has crossed the threshold of noise, thus enabling only valid compressed data to be written on to the RAM.

STATE MACHINE LOGIC :

This block produces all necessary control signals to other blocks. For the real time recognition of speech, the corresponding word is to be recognized from the memory. The address for the samples is issued by this unit. For comparison of data, it gives a VALID signal to the compare block. It has a a counter inside, which count the samples and correspondingly issues VALID signal to the COMPARE block, during that time the samples are compared. For the other periods there is no comparison.

MEMORY:

The block RAM in FPGA is accessed for both reading and writing of samples. I.e. it has to be selectively read/write enabled. There are eight block RAMs, of size 512x8 are there in the SPARTRAN FPGA chip used here. Out of that two memories(1024x8) are used for storing the samples of the word “HELLO” and one memory (512x8) is used for storing the word “GO”.

SPEECH COMPARE :

The word to be typed is compared with the compressed sample already stored. The comparison is done only for the period of VALID signal. In a similar manner all the samples are compared and the result is passed to the next block.

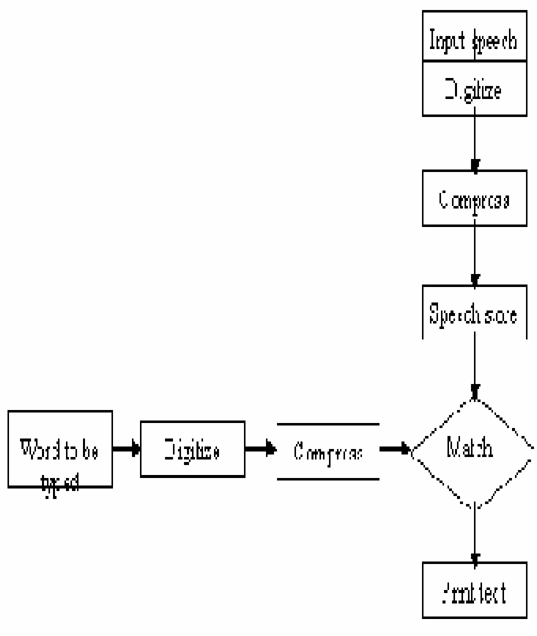
COUNT 1&2 :

The output from the compare block is counted here. If the input speech matches the stored samples then a count is increased. This is also done in the valid region, specified by the State machine block.

DECISION MAKING:

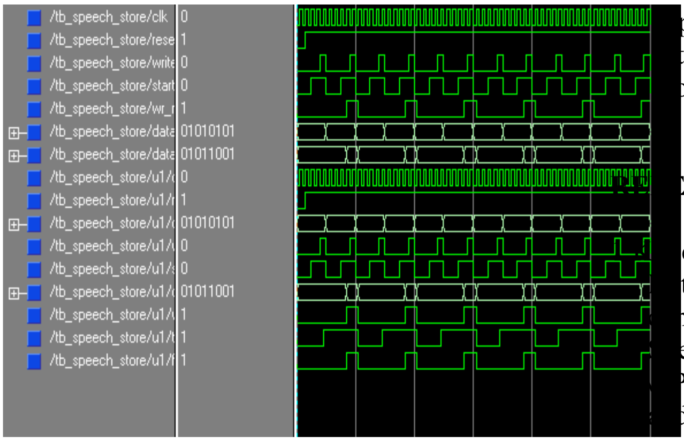
At the end of the comparison the decision making circuit analyses the count values and gives the result “HELLO” or “GO” or no word whichever has been detected.

PROPOSED ALGORITHM :

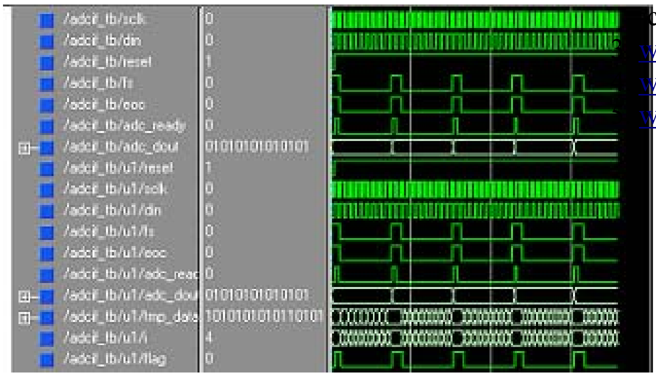


SIMULATION RESULTS :

The project is simulated using VHDL. Simulation for the encoder and speech store blocks are given below.



```
/tb_speech_store/clk   0
/tb_speech_store/rese  1
/tb_speech_store/write 0
/tb_speech_store/start 0
/tb_speech_store/wr_r  1
/tb_speech_store/data  01010101
/tb_speech_store/data  01011001
/tb_speech_store/u1/c  0
/tb_speech_store/u1/r  1
/tb_speech_store/u1/c  01010101
/tb_speech_store/u1/\  0
/tb_speech_store/u1/s  0
/tb_speech_store/u1/c  01011001
/tb_speech_store/u1/\  1
/tb_speech_store/u1/t  1
/tb_speech_store/u1/f  1
```



```
/adcif_tb/sclk          0
/adcif_tb/din           0
/adcif_tb/reset         1
/adcif_tb/fs            0
/adcif_tb/eoc           0
/adcif_tb/adc_ready     0
/adcif_tb/adc_dout      01010101010101
/adcif_tb/u1/reset      1
/adcif_tb/u1/sclk       0
/adcif_tb/u1/din        0
/adcif_tb/u1/fs         0
/adcif_tb/u1/eoc        0
/adcif_tb/u1/adc_reac   0
/adcif_tb/u1/adc_dou    01010101010101
/adcif_tb/u1/tmp_data   1010101010110101
/adcif_tb/u1/i          4
/adcif_tb/u1/flag       0
```

CONCLUSION :

The project is implemented in VHDL and downloaded in SPARTRAN FPGA chip. The voice processor chip proposed here thus produces recognition for two words “HELLO” and “GO”. At this stage this chip can be used for controlling portable devices. For example toys can be controlled by human voice. Thus the chip is designed for speaker dependent, word by word utterance. This will surely be a boom to the the blind people.

FUTURE WORK :

This project can be developed for connected words and continuous peech.

: can de developed to provide security) speech also.

RENCES :

e J. Palomaki, Guy Brown, "Missing ta Speech Recognition in Reverberant nditions". IEEE Transactions on eech Processing, Jan'2002.

. Cooke, P.D. Green, L. Josifovski l A. Vizinho "Robust automatic Speech Recognition with Missing and Unreliable acoustic data" Speech communication.

www.speechgroups.com

www.xilinx.com

www.comspeech.edu