Vol. 1, Issue VIII, Sep. 2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Development of Speech to Text Conversion System on OMAP 3530

Srivani Pokala[1], G. Radha Krishna[2], P.Sandeep[3]

[1]PG student, Department of Electronics and Communication Engineering,
VNR VignanaJyothi Institute of Engineering and Technology, Hyderabad, India.
srivani.pokala04@gmail.com

[2]Associate Professor, Department of Electronics and Communication Engineering,
VNR VignanaJyothi Institute of Engineering and Technology, Hyderabad, India
guntur_radhakrishna@yahoo.co.in

[3]Research Associate, Research and Consultancy Centre, VNR VignanaJyothi
Institute of Engineering and Technology, Hyderabad, India
sandeep_ec468@yahoo.com

Abstract: *In this paper, Development of Speech to text conversion system on OMAP 3530 is designed, in which the embedded chip and the programming techniques are adopted. The central monitor which adopts OMAP 3530 chip as a processor is the core of the whole system. Here we used speech recognition technology as the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program. Here the system was implemented using the HMM toolkit HTK by training HMMs of the words making the vocabulary on the training data. HMMs have found application in many areas interested in signal processing, and in particular speech processing. In this system speech is given through Microphone, that is interfaced to the beagle board using ALSA driver and Output is displayed in the form of text on Beagle board Touch screen. Its ultimate goal is to achieve natural language communication between man and machine.*

Keywords: *Speech to text conversion; Beagle Board; Continuous speech recognition; Hidden markov model*

1. INTRODUCTION

Speech-to-text conversion is the process of converting spoken words into written texts. This process is also often called speech recognition. All speech-to-text systems rely on at least two models: an acoustic model and a language model. In addition large vocabulary systems use a pronunciation model. Speech to text system has many potential applications including command and control, dictation, transcription of recorded speech, searching audio documents and interactive spoken dialogues.It is important to understand that there is no such thing as a universal speech recognizer. To get the best transcription Quality, all of these models can be Specialize for a given language, dialect, application domain, type of speech, and communication channel. Like any other pattern recognition technology, speech recognition cannot be error free. The speech transcript accuracy is highly dependent on the speaker, the style of the speech and environmental conditions. Humans are used to understanding speech, not to transcribing it and only Speech that is well formulated can be transcribed without ambiguity.

The core of all speech recognition systems [1] consists of a set of statistical models representing the various sounds of the language to be recognized. Since speech has temporal structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range, the hidden Markov model (HMM) provides a natural framework for constructing such models. A speech recognition system consists of the following: A microphone, for the person to speak into Speech recognition software, a computer to take and interpret the speech, a good quality soundcard for input and/or output.

2. HARDWARE DESIGN

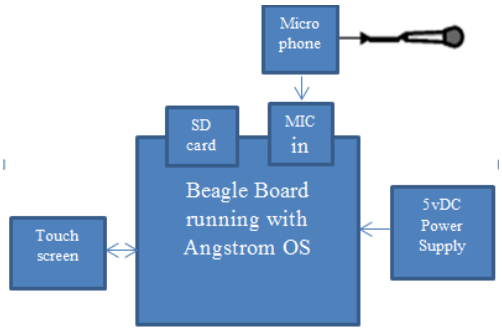

Figure 1: Block diagram of the hardware design

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

The hardware system consists of the Beagle board which is connected to the 5v dc supply. The maximum supply voltage limitation of the Beagle board is 5V.The output of the embedded environment can be viewed in Touch Screen monitor.
Secure Digital card (SD card) for mounting of uImage file and root file system (rfs) where the whole file system is present which is used for driving the audio. Micout and Micin are also a part of beagle board at which the audio jack is connected. Self-powered mic is connected to Micin for recording.

2.1. Power Supply System

There are two possible sources of the 5V required by the Beagle board. It can come from the USB OTG port connected to a PC, powered USB HUB, or a 5V DC supply. The USB Supply is sufficient to power the Beagle board. However, depending on the load needed by the expansion port on Beagle Board, additional power may be required.
It should also be noted that if an OTG configuration is used, for example tying two Beagle Boards together via a USB OTG cable, both of the Beagle boards must be powered by the DC supply. If the OTG port is used as a Host port, then the DC supply must also be used.

2.2. Audio input jack

External Audio input devices, such as a powered microphone or the audio output of a PC or MP3 player, can be connected to the via a 3.5mm jack. The audio cables are not provided with Beagle board, but can be obtained from just about any source. Figure 2 shows how the cable is connected to the stereo input jack.

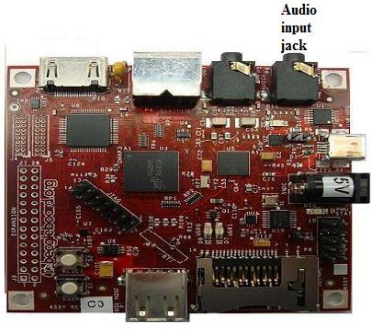

Fig.2: Audio Input Jack

2.3. System block diagram

The high level block diagram consist of OMAP3530[2] processor with SVideo, Touch Screen, Stereo In & Out, USB Host, SD MMC, JTAG, LCD, Expansion pins, Reset & User buttons. Beagle board high level diagram is shown in the figure 3.

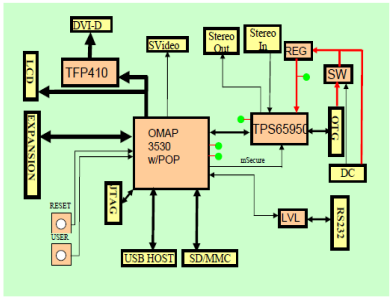

Figure 3: Beagle Board High Level Block Diagram

2.3.1 OMAP 3530 Processor

There are many features on this board which are useful for Open Embedded Developers. However, this project uses only few of the features. The Beagle Board is an OMAP 3530 platform. It has been equipped with a minimum set of features to allow the user to experience the power of the OMAP3530 [2]. By utilizing standard interfaces, the Beagle Board [3] is highly extensible to add many features and interfaces.

3. DESIGN OF THE SYSTEM

The prepared system if visualised as a block diagram will have the following components: Sound Recording and word detection component, feature extraction component, speech recognition component, acoustic and language model.

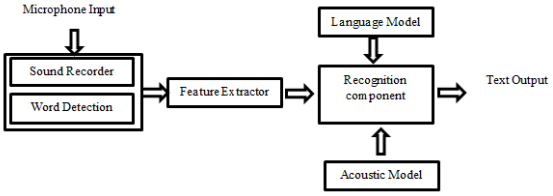

Figure 4: Block Diagram of Recognition System

3.1 Sound Recorder and Word Detection

The component is responsible for taking input from microphone and identifying presence of words. To record the speech samples AUDACITY software is used and it has a provision of saving the sound into WAV files. The recorder takes input from the microphone and saves it or forwards it depending on which function is invoked recorder supports changing of sampling rate, channel and size of the sample. Default sampling rate of the recorder is 44100 samples per second, at a size of 16 bits per sample and dual channel.
In speech recognition it is important to detect when a word is spoken. The system does detects the region of silence is considered as a spoken word by the system. The system uses energy pattern present in the sound signal and zero crossing rate to detect the silent region. Taking both of them is important as only energy tends to miss some parts of sounds which are important. This technique has been described in [9]

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

3.2 Feature Extractor

Humans have a capacity of identifying different types of sounds (phones). Phones put in a particular order constitute a word. If we want a machine to identify the spoken word, it will have to differentiate between different kinds of sound the way the human perceive it. The point to be noted in case of human is that although, one word spoken by a different people produces different sound waves humans are able to identify the sound wave as same. On the other hand two sounds which are different are perceived as different by humans. The reason being even when same phones or sounds are produced by different speakers they have common features. A good feature extractor should extract these features and use them for further analysis and processing. This extracted information is known as feature vector.MFCC and LPC are the few methods for implementing extracting feature factor.

3.3 Knowledge Models

For speech recognition, the system needs to know how the words sound. For this we need to train the system. During the training, using the data given by the user the system generates acoustic model and language model. These models are later used by the system to map a sound to a word or a phrase.

3.4 Acoustic Model

In speech recognition [6], basic unit of sound is phoneme. Phoneme is a minimal unit that serves to distinguish between meanings of words. For example 10 sequence of phoneme for “CAT” is K A and T. In English language there are nearly around 46 phonemes. We can construct any word from English dictionary using proper concatenation of this phoneme. In order to recognize a given word, we should extract phoneme from voice sample. Due of slowly timed varying nature of speech signal short-term spectral analysis is the most common ways to characterize speech signal. When examined over a sufficiently short period of time (between 10 and 25 m sec), its characteristics are fairly stationary. However, over long period of time the signal characteristic change to reflect the different speech sounds being spoken. Using this observation, we find that feature vector extracted over 10 to 25 m sec correspond to single phoneme.

The acoustic model is used to score the unknown voice sample. VQ-Code Book [7] and GMM models are the different types of acoustic model.

3.5 Language Model

Although there are words that have similar sounding phone, humans generally do not find it difficult to recognise the word. For example **car key** and **Khakee** has same phoneme sequence. In such case language structure comes in to picture. Language model uses context information to narrow down the recognized word to resemble the given grammar constructs. The language model specifies what are the valid words in the language and in what sequence they can occur.

4. ALGORITHM USED FOR SPEECH RECOGNITION SYSTEM

Both acoustic modeling and language modeling are important parts of modern statistically-based speech recognition algorithms The Hidden Markov Model (HMM) [5] is a powerful statistical tool for modeling generative se-quinces that can be characterized by an underlying process generating an observable sequence. HMMs have found application in many areas interested in signal processing, and in particular speech processing, but have also been applied with success to low level NLP tasks such as part-of-speech tagging, phrase chunking, and extracting target information from documents. Andrei Markov gave his name to the mathematical theory of Markov processes in the early twentieth century [4], but it was Baum and his colleagues that developed the theory of HMMs in the 1960s[8].

HMM [6] is very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of application.HMM model, when applied properly work well in practice for several important application.

4.1 Elements of hidden markov model

HMM is characterized by following,

1. Number of state N
2. Number of distinct observation symbol per state M, V = V1, V2, · · · , VM

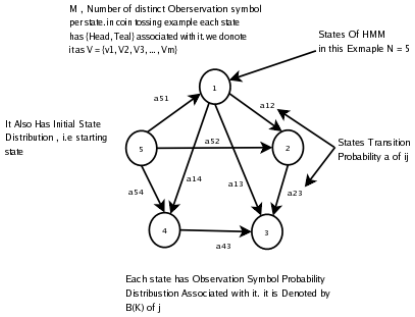

Figure 5: Elements of HMM

3. State transition probability, aij = P[qt+1 = Si|qt = Sj], 1 _ i, j _ N
4. Observation symbol probability distribution in state j,Bj(K) = P[Vk at t|qt = Sj]
5. The initial state distribution $\pi = \pi_i$ where π_i= P [q1 = Si] 1 ≤i ≤N[2]

Vol. 1, Issue VIII, Sep. 2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Given appropriate value of N, M, A, B andπ, HMM can be used as generator to give an observation sequence

$$O = O1\ O2\ O3 \cdots OT$$

5. IMPLEMENTATION

We developed the speech to text conversion system on Beagle Board by porting mobile operating system Angstrom With the programmable environment supporting component implementation in C++ assembly language, the entire code was compiled after several unsupported constructs in the code were removed.

5.1. Software

• Linux on the Beagle Board:
• Angstrom (mobile operating system
Which is Linux Distribution)

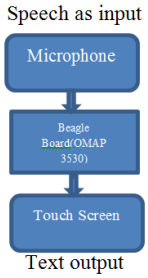

Figure 6: Speech to Text System

5.2. Porting Angstrom OS

Make two partitions on the SD/MMC card

- FAT partition(MLO, u-boot, uImage)
- Ext2 partition

5.3. The five (5) boot phases

- ROM loads x-load (MLO)
- X-load loads u-boot
- U-boot reads commands
- Commands load kernel(uImage)
- Kernel reads root file system.

6. RESULTS

The results have been depicted that the speech to text conversion system is capable of real time operation and is successfully developed on beagle board.

```
        ************************************************************
        * NOTICE: The first input may not be recognized, since      *
        *         no initial CMN parameter is available on startup. *
        * for MFCC01*
        ************************************************************

Stat: adin_oss: device name = /dev/dsp (application default)
Stat: adin_oss: sampling rate = 16000Hz
Stat: adin_oss: going to set latency to 50 msec
Stat: adin_oss: audio I/O Latency = 32 msec (fragment size = 512 samples)
STAT: AD-in thread created
<<< please speak >>>
```

```
        ************************************************************
        * NOTICE: The first input may not be recognized, since      *
        *         no initial CMN parameter is available on startup. *
        * for MFCC01*
        ************************************************************

Stat: adin_oss: device name = /dev/dsp (from AUDIODEV)
Stat: adin_oss: sampling rate = 16000Hz
Stat: adin_oss: going to set latency to 50 msec
Stat: adin_oss: audio I/O Latency = 32 msec (fragment size = 512 samples)
STAT: AD-in thread created
pass1_best: <s> SPEECH
WARNING: 00 _default: hypothesis stack exhausted, terminate search now
STAT: 00 _default: 0 sentences have been found
<search failed>
pass1_best: <s> SPEECH
WARNING: 00 _default: hypothesis stack exhausted, terminate search now
STAT: 00 _default: 0 sentences have been found
<search failed>
pass1_best: <s> SPEECH
sentence1: <s> SPEECH IS THE PRIMARY MEANS OF </s>
<<< please speak >>>
```

Figure 7: Speech to Text conversion Output

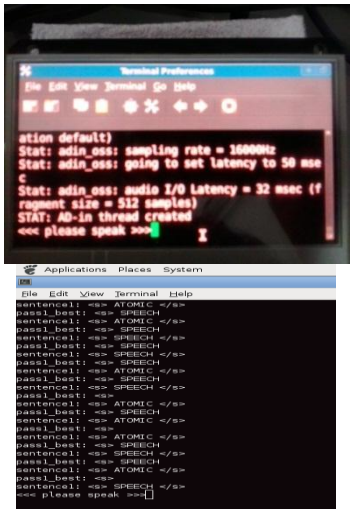

Figure 8: Terminal of Beagle Board displayed the text output

In the fig shown above is the terminal of the beagle board displayed the text as output when connected to Microphone and monitor .The input is the speech which has been saved in sample.wav format and the path has been given in the command. The output is displayed in text and clear English when the terminal of beagle board displays **<<<please speak>>>.**

Recognition was tried on three kinds of sounds

- Seen Sound: The sound files used to train the model.
- Unseen Sound seen user: Unused Sound file of the user whose other sound files were used for training.
- Unseen User: The user whose voice we not used for training.

The results of the experiment were as shown:

Type of Sound	No of Sounds	Recognition
Seen Sound	10	6
Unseen sound and Seen user	5	Search Failed
Unseen User	5	0

Table 1: Recognition Result

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

7. CONCLUSION AND FUTURE SCOPE

We have successfully developed speech to text conversion system on beagle board which is useful as PDA's for visually impaired, illiterate and can be used in many other applications. While in HMM, the recognition ability is good for unknown word. HMM is generic concept and is used in many area of research. In whole architecture of speech recognition, HMM is just one block which helps in creating search graph. It works in tandem with other blocks such as front-end, language model, lexicon to achieve desired goal. What we discussed in this paper is recognition of English sentence, we can extend this model to recognize multi-lingual sentence.

REFERENCES

[1] John Kirriemuir, "Speech Recognition Technologies," TSW 03-03, Issue 30th March 2003

[2] http://linux.OMAP.com/mailman/listinfo/ Linux- OMAP-open-source

[3] http://elinux.org/Beagleboardbeginners

[4] A. Markov. "An example of statistical Investigation in the text of Eugene onyegin, illustrating Couplingof tests in Chains". Proceedings of the Academy of Sciences of St. Petersburg, 1913.

[5] L. Rabiner. A tutorial on hidden markov models and selected applications Inspeech Recognition. Proceedings of IEEE, 1989.

[6] Nirav S. Uchat "Seminar report On Hidden Markov Model and Speech Recognition" Department of Computer Science and Engineering Indian Institute of Technology, Bombay Mumbai.

[7] Dan Jurafsky. CS 224S / LINGUIST 181 Speech Recognition and Synthesis. http://www.stanford.edu/class/cs224s/.

[8] L. Baum et. al. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. Annals of Mathematical Statistics, 41:164171, 1970.

[9] L. R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. Bell System Technical Journal, Vol. 54, pages 297-315, 1975,.