International Journal of Electrical and Electronics Engineering (IJEEE)
ISSN(P): 2278-9944; ISSN(E): 2278-9952
Vol. 5, Issue 2, Feb - Mar 2016, 9-18

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

VOICE CONTROLLED MUSIC PLAYER USING SPEECH RECOGNITION SYSTEM

M. NAGAMANI & SHAIK FATHIMA

South Campus, University of Hyderabad, Gachibowli, Hyderabad, Telangana, India

ABSTRACT

Speech enabled devices are gaining a strong footholds, in the man's day today life. Several speech recognition applications have implemented, and are available for public or commercial use. Very few of these, however, are available for the Indian Market. We present this paper with the intent of developing a speech controlled music player viable for the Indian customers this paper aims at introducing the concept of handling the existent music player with speech commands. Here we investigated from the speech recognition problem point of view. This system is based on the open source CMU Sphinx-3, from the Carnegie Mellon University.

KEYWORDS: Speech Recognition, Speech Commands, Speech Controlled Music Player

INTRODUCTION

Voice driven systems are becoming increasingly important in both personal and enterprise world. Progress in Speech enabled devices has progressed to the point where it is now feasible to implement simple, but useful, task oriented systems. It is the purpose of this paper to describe one such system, a voice controlled music player, which has *Speech recognition* to play an important role in its implementation.

Speech enabled devices are gaining a strong footholds, in the man's day today life. Several speech recognition applications have implemented, and are available for public or commercial use. Very few of these, however, are available for the Indian Market. We present this paper with the intent of developing a speech controlled music player viable for the Indian customers. This paper aims at introducing the concept of handling the existent music player with speech commands. It describes the hardware-software co-design for speech activated music player. A general framework of the system is presented and special attention is paid to the speech recognition part used within the framework. A partial reference to implementation has been written for making the framework itself robust, feasible and extensible. The framework is based on the small vocabulary isolated word recognition coupled with command and control strategy.

The process described in this paper involves conversion of audio signal of human speech command into text[1], extraction of the command word from the recognized text, compare the recognized text with the predefined player functions, and finally trig the resulting event. Development of this system requires the technologies, such as digital signal processing, automatic speech recognition and command control respectively [2]. Each of these modules takes input and passes output to another module in sequential order.

Sphinx-3 tool-kit is used to build the speech recognition system. Sphinx -3 [3] is the well-known tool-kit that is freely available. These are used to build any kind of speech recognition system for any language. Recognizable work is being done in India in the field of in spoken word recognition [4], modeling out-of-vocabulary Words for robust speech recognition [5], and multi-domain speech understanding [6]. The system that we have developed is a speaker Independent

Isolated word recognition system. The recognition accuracy of this system is around 85%. Here in this paper we present a brief overview of developing speech controlled music player in English language. In order to build this system we collected the data, which covered adequately the wide range of potential mp3 player consumers (users) with different dialect zones in different environmental conditions.

BASIC ARCHITECTURE

Since the area of our research is speech recognition we are not very good at coming up with circuit designs on my own. So we employed the built in mp3 player (XMMS player) in the system. The approach we take here is just using laptop/computer with a microphone. **Figure 1.1** describes the basic design implemented to control the music player.

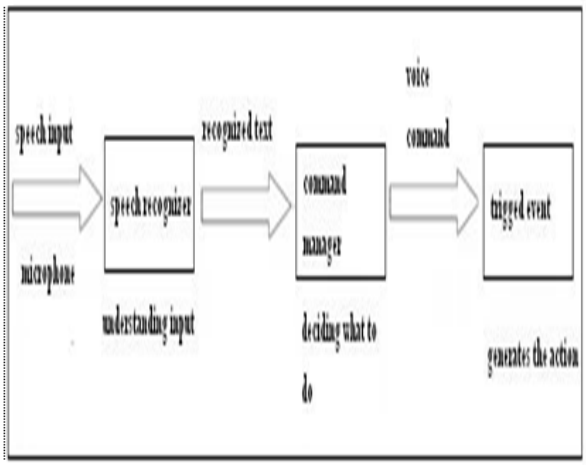

Figure 1.1: Basic Architecture

Components of the System

Two major components are employed in this system.

- Automatic speech recognition
- Command manager

Automatic Speech Recognition

Automatic speech recognition (ASR) can provide a rapid means of controlling electronic assistive technology. Speech recognition emphasizes recognizing simple discrete speech utterances that match a fixed vocabulary of commands. In this system the input from the user is translated into text that the computer can parse by automatic speech recognition (ASR). The output from the recognizer is sent to a component that interprets the semantic meaning of the input. The interpretation is then used by the command manager to determine what to do. Sphinx –3[3] is used for the Automatic Speech recognition

The crucial components needed for building an ASR.

Front End Digital Signal Processing

It parameterizes an input signal into a sequence of output features. It performs Digital Signal Processing (DSP) on the incoming data [1]. Table 1.2 describes the parameters and their values

Table 1: Parameters for Front-End Signal Processing

Parameter	Values
Sampling rate(Hz)	16000.0 Hz
Frame Rate	100 frames/sec
Window length	0.025 seconds
Filterbank type	Mel
Number of ceptra	13
Number of Mel filters	40
DFT size	512 Lower
Filter Frequency	133.3333 hz
Upper filter Frequency	6855.4976Hz
Pre-Emphasis	0.97 α

Below figure 1.2 shows the front end DSP in ASR.

Since the speech commands we intended to use are probably not even grammatical sentences (e.g. “play”, “stop”), and are generally fairly short. A trigram model probably captures most of the information needed. Since the Sphinx uses an HMM trigram language model as opposed to a structural language model. This is the best option for me.

Sphinx-3[3] tool-kit is used to deploy the speech recognition part. The figure 1.3 describes the internal flow of the ASR

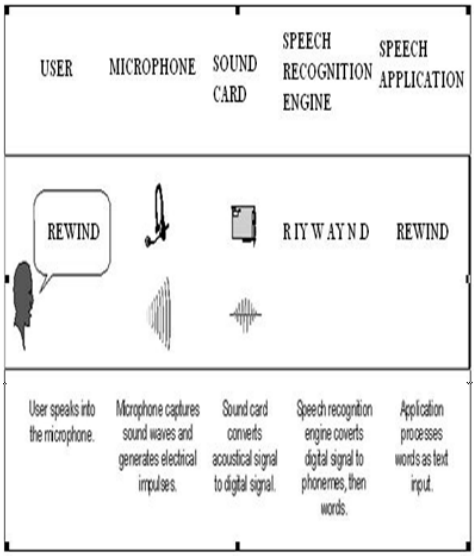

Figure 1.2: Front End Signal Processing

Linguist

Or knowledge base, it provides the information the decoder needs to do its job.

Acoustic Model

Contains are presentation (often statistical) of a sound, created by training using many acoustic data

The Phonetic Lexicon Dictionary

It is responsible for determining how a word is pronounced.

THE Language Model

It contains a representation (often statistical) of the probability of occurrence of words.

Decoder

It is the main bloc of thesystem, which performs the bulk of the work. It reads features from the front end, couples this with data from the knowledge base and feedback from the application, and performs a search to determine the most likely sequences of words that could be represented by a series of features.

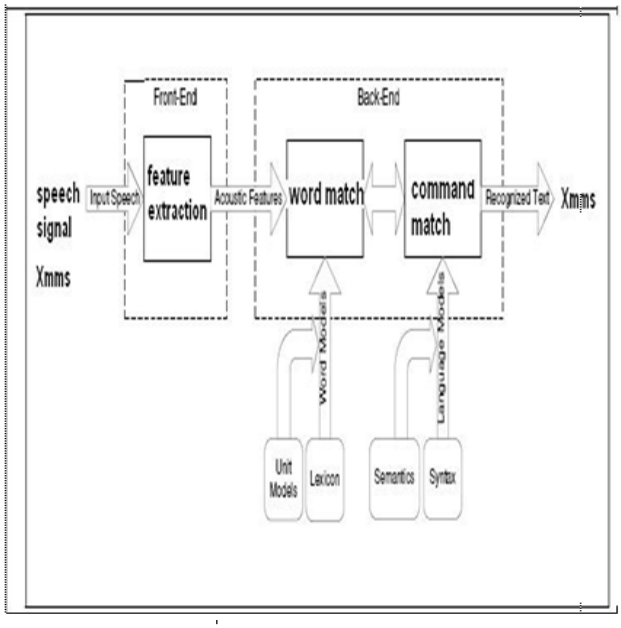

Figure 1.3 Internal Workflow of the Automatic Speech Recognition Part

The voice input to the microphone goes to the sound card. The output from the sound card—digital audio is processed using HMMs. First, the users—give a voice command over the microphone, which is passed to the sound card in your system. This analog signal is sampled 16,000 times a second and converted into digital form using a technique called Pulse Code Modulation (PCM) [7]

Command Manager

The role of the command manager unit is to extract the command word front the recognized output of the automatic speech recognition part and compare with the predefined functions of the music player and generate the resulting action. If the extracted word is a match with the predefined functions then it will trig the resultant action otherwise it gives the error message. Figure 1.4 describes the internal workflow the CM

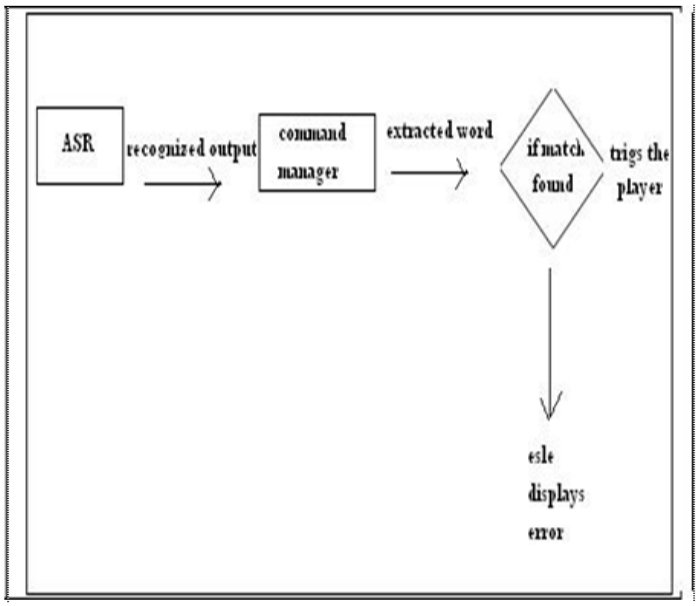

Figure 1.4: Command Manager Workflow

DESIGN CONSIDERATIONS AND REQUIREMENTS OF THE VOICE DRIVEN MUSIC PLAYER

Since the area of our research is speech recognition we are not very good at coming up with circuit designs on my own. So we employed the built in mp3 player (XMMS player) in the system. The approach we take here is just using laptop/computer with a microphone. The basic motivation behind this project is to make some common functions of the existent music player (play, pause, rewind, stop, forward, etc) enabled though speech commands.

Parameters for this Speech Driven System

Isolated Word Recognition

Isolated word recognizers require single utterance at a time [1].

Command and Control Strategy

In this strategy the user speaks a single command and the player upon correctly recognizing the command acts appropriately [1].

Speaker Dependent to Speaker Independent

In Speaker dependent system, speaker will initially train the system before using it. On the other hand, speaker independent system can identify any speaker's voice [1].

Vocabulary for Recognition

Vocabulary is the lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. For this implementation we decided to hold small number of words in the speech recognition vocabulary. Since there are fewer input candidates for this system to examine, a small vocabulary may improve recognition speed and accuracy. The basic functions of the player are deployed as the basic vocabulary. Below is the list of the 11-word vocabulary used.

Table 2: Vocabulary for the Player

Xmms	Forward
Play	Queue
Pause	Enqueue
Rewind	Stop
Repeat	Shuffle
Volume	

Voice Commands

Voice commands are a quick and fairly natural way to interact. The difficulties of voice activated devices come from the serial nature of speech [1]. Only one instruction can be offered at a time, compared to a graphical user interface (GUI) with several active areas and buttons. The command manager of the system contains the complete voice commands list that is supported in the system. Creating new grammar rules for the speech engine can expand voice commands.

By limiting the vocabulary, though, one can increase the single word recognition, and by keeping the commands as short "stock" phrases I can increase the total utterance recognition accuracy. We decided to use the 11 words as the voice commands as well.

Hard Ware Requirements

Sound Cards

We used the Sound cards with the 'cleanest' A/D (analog to digital) conversion.

Microphone

We equipped a highly directional, noise-canceling microphone. A good microphone is an essential component of a SR system.

Laptop/ PC

ASR applications can be heavily dependent on processing speed. That is why the system with P4 (2.4 GHZ), processor, RAM of 128 MB is used in this process.

Software Requirements

OS Platform: Red Hat Linux 9.0

Speech recognition engine: CMU Sphinx-3 [8]

Command manager: Perl scripts

Speech Data Base Collection

A prerequisite for a successful development of speech controlled application is a comprehensive definition of the speech data to be collected [9]. Here we collected the data, which covered adequately the wide range of potential mp3 player consumers (users) dialectal zones, which includes the prominent region in our country and state. Our database text corpus consists of 11command words spoken by different genre of speakers (adults, males, females, and children).

Recording

The speech was recorded in a computer laboratory and under different environmental conditions. Attention has been devoted to research into the environment of the recordings, which are, like the typical surroundings of PLAYER applications, home, office, school, public places and moving vehicles. The properties of audio files are set as, sampling rate is l60OOHz, and bit rate is 16bits/sec. and the number of channels 'is mono. While recording we maintained a distance of 5-7cm between the microphone and mouth. The audio was recorded, with gnome-sound-recorder using a high quality microphone. Each utterance was recorded and stored as separate file in wave format. The data is recorded from 60 males, 60 females, 60 children, from dialect zones of India. The data is divided into training data and testing data.

Training

The data in the training data set is trained and templates are created which learn the parameters of the sound units. The training is done using the CMU Sphinx -3 trainer. The language can be obtained fro the language model tool kit of the CMU Sphinx [10]

IMPLEMENTATION

This section presents the implementation of the above-described system. Whenever the user speaks the command words, the microphone captures the sound waves and generates electrical impulses. The sound card converts this analog signal to digital signal and the same is saved in the wav format. This wav file is then converted to raw file and subsequent MFCC's are calculated. After this calculation of ceptral features the word is decoded in the text format. The command word is extracted from the recognized output of the automatic speech recognition part and compared with the predefined functions of the music player and generates the resulting action. If the extracted word is a match with the predefined functions then it will trig the resultant action otherwise it gives the error message

EXPERIMENTS AND RESULTS

Different persons with different quality of spoken English, dialect zones and from different age groups are chosen as test persons. The experiments are done in a room with different environmental conditions. The performance of the system is registered.

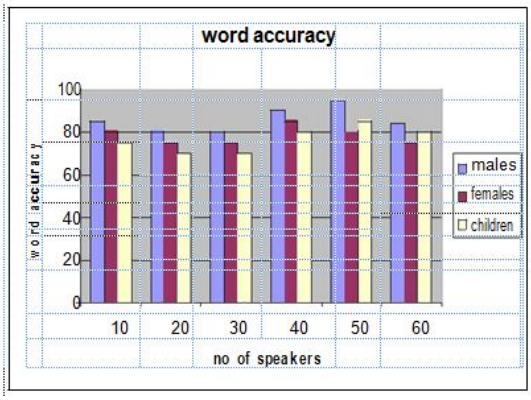

Figure 1.5: Word Accuracy for Different Speakers

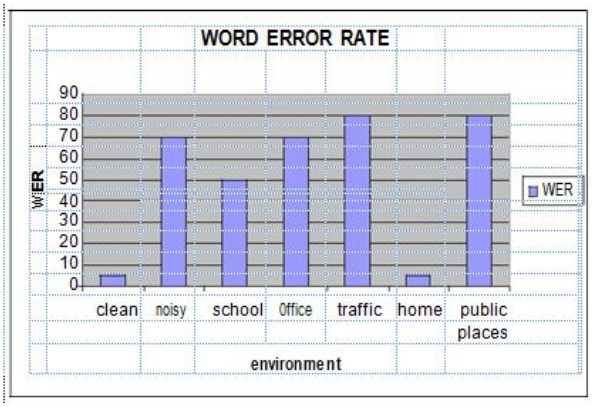

Figure 1.6: Word Error Rate

For these trials, ASR word accuracy, over the whole trial, of 85%.

EVALUATION AND ANALYSIS

The evaluation of the system is done on the Basis on word err roe rate WERs for children's speech are higher than for Adult speech. It can also be analyzed that there are Higher WERs for females and Lower WERs for higher age.

The command manager part of the system works perfectly. This is not surprising, because the Perl script program and grammar file has a reasonable size. If a command is not recognized properly, the command can be spoken again until correctly recognized. Then the commands were spoken. In case of false recognition, the commands were spoken again until the right commands were recognized.

There are many factors that influence the error rate. And they are speaker variability, environment variability, usage of low quality microphones, and high background noise. Automatic recognition of certain speech (e.g., tele phone numbers) is very feasible today with very good ac curacy, even when using telephone lines and serving a large population. However, even such simple recognition tasks suffer decreased.

CONCLUSIONS

The goal of this paper was to present the overview of the speech controlled music player application. CMU Sphinx-3 was used for the ASR part. The goals were fulfilled with quite good results.

The bottleneck of the performance lied on the speech recognition part of the system. For users with a strong English accent the performance is very good and perfect. For people who do not know English even for rural native speakers, the performance is very poor. The overall performance is satisfactory the system would not be accurate on every word spoken, but the overall there was about an 80% success rate with a false alarm rate below 10%.

Analysis and modeling of speaker variability, such as gender, accent, age, speaking rate, and phone realizations, are important issues in speech recognition.

REFERENCES

1. "Fundamentals of Speech Recognition". L. Rabiner & B. Juang. 1993. ISBN: 0130151572.
2. http: //www.ece.msstate.edu/ research/isip/publications/courses/ ece_8643/lectures/current/

3. http://www.speech.cs.cmu.edu
4. Spoken Word Recognition: Lexical Vs Sublexical, Amit Gupta and Sandeep M. Workshop on Spoken language Processing, TIFR, Mumbai, January 9-1 I, 2003.
5. Bazzi, Modeling Outsf-Vocabulary Words from Robust Speech Recognition MIT Department of Electrical Engineering and Computer Science, June 2002.
6. *G.* Chung, Towards Multi-DomainSpeech Understanding' with Flexible and Dynamic Vocabulary. MIT Department of Electrical Engineering and Computer Science, June 2001.
7. http: //www.lawrencenajjar.com/ papers /User_interface_design _guidelines for _speech.html
8. http://www2.cs.cmu.edu/~robust Tutorial
9. SPEECH DAT CAR. A Large Speech Database for Automotive Environments. Asunción Moreno (1), Borge Lindberg (2), Christoph Draxler(3), Gael Richard (4), Khalid Choukry(5), Stephan Euler (6), Jeff Allen (5) (1) Universidad Politécnica de Cataluña, Spain; (2) CPK, Denmark(3) IPSK of the University of Munich.(4)Lernout & Hauspie, France (5), ELRA, France, (6), Bosch Germany,
10. http://www.speech.cs.cmu.edu/SLM/toolkit.html