Vol III Issue VI July 2013
Impact Factor : 0.2105
ISSN No : 2230-7850

Monthly Multidisciplinary
Research Journal

Indian Streams Research Journal

IMPACT FACTOR : 0.2105

Welcome to ISRJ

RNI MAHMUL/2011/38595 **ISSN No.2230-7850**

Indian Streams Research Journal is a multidisciplinary research journal, published monthly in English, Hindi & Marathi Language. All research papers submitted to the journal will be double - blind peer reviewed referred by members of the editorial Board readers will include investigator in universities, research institutes government and industry with research interest in the general subjects.

International Advisory Board

Editorial Board

Address:-Ashok Yakkaldevi 258/34, Raviwar Peth, Solapur - 413 005 Maharashtra, India
Cell : 9595 359 435, Ph No: 02172372010 Email: ayisrj@yahoo.in Website: www.isrj.net

Indian Streams Research Journal
ISSN 2230-7850
Volume-3, Issue-6, July-2013

A Review of Voice Analysis and Recognition Techniques

Aziz Ur Rahaman Makandar , SairaAltaf Shaikh
Associate Professor, Department of Computer Science, KSW University, Bijapur.
Department of Computer Science, KSW University, Bijapur.

Abstract: The Speech is the most prominent and primary mode of Communication among human being.Speech recognition and text-to-speech synthesis technologies continue to be adopted successfully by government agencies, industries and research areas. These organizations have typically deployed large enterprise-grade proprietary platforms into their call centers and realized significant business benefits despite the high costs of deploying such technology.This paper addresses about the interaction between the system and the user through voice response.The user can retrieve information using voice effectively whereas the system too gives output using systems voice. The system in turn will notify the user about the current processing that are been carried out by them. It uses Speech recognition to detect the voice from the user and uses the speech control to deliver the voice output.People with disabilities can benefit from speech recognition programs.

Keyword: Analysis, Speech processing, Speech Application Platform, Training of the system, Recognition.

1.INTRODUCTION

As this is the age of speed everything happens in the speed of supersonic. The data can be transferred at the speed of light in the digital medium; hence there is a need of information inflow in the same speed. Here is one such need ofinformation fast enough. We have experienced in waiting atterminals to get information about the transport facility. We encounter so many times therewill be no person for providing these information which in factually wastes the time just toknow whether there is any facility or not. Here is one solution for such a problem which lessensthe human intervention in providing such information at the terminals.
Voice Automated System is asystem which operatesbased on the voice input given by the user. Voice/speech recognition is a field of computer science that deals with designing computer systems that recognize spoken words. It is a technology that allows acomputer to identify the words that a person speaks into a microphone or telephone.

"Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words" [11].Automatic speechrecognition (ASR) is one of the fastest developing fields in the framework of speech science and engineering. As the new generation of computing technology, it comes as the next major innovation in man-machine interaction, after functionality of text-to-speech.

SpeechRecognitionTechniques

Template based approaches:Matching unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match [2]. This has the advantage of using perfectly accurate word models. But it also has the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. Dynamic time warping is such a typical approach, the templates usually consists of representative sequences of features vectors for corresponding words [3]. The basic idea here is to align the utterance to each of the template words and then select the word or word sequence that contains the best.

For each utterance, the distance between the template and the observed feature vectors are computed using some distance measure and these local distances are accumulated along each possible alignment path. The lowest scoring path then identifies the optimal alignment for a word and the word template obtaining the lowest overall score depicts the recognized word or sequence of words.

Knowledgebasedapproaches:An expert knowledge about variations in speech is hand coded into a system. This has the advantage of explicit modeling variations in speech but unfortunately such expert knowledge is difficult to obtain and use successfully. Thus this approach was judged to be impractical and automatic learning procedure was sought instead.

Statistical based approaches:In which variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models, or HMM. The approach represents the current state of the art. The main disadvantage of statistical models is that they must take priori modeling assumptions which are liable to be inaccurate, handicapping the system performance. In recent years, a new approach to the challenging problem of conversational speech recognition has emerged, holding a promise to overcome some fundamental limitations of the conventional Hidden Markov Model (HMM) approach [4], [5].

VOICE AUTOMATED SYSTEM

Voice Automated System is developed for providing the information for the enquiry in public associated terminals. It uses Speech recognizers to detect the

ISSN 2230-7850
Volume-3, Issue-6, July-2013

voice from the user and uses the speech control to deliver the voice output. This also displays the results on the screen for further verification.

WORKING OF ASR

The goal of an ASR system is to accurately and efficiently converts a speech signal into a text message transcription ofthe spoken wordsindependent ofthe speaker, environment orthe device used to record the speech.

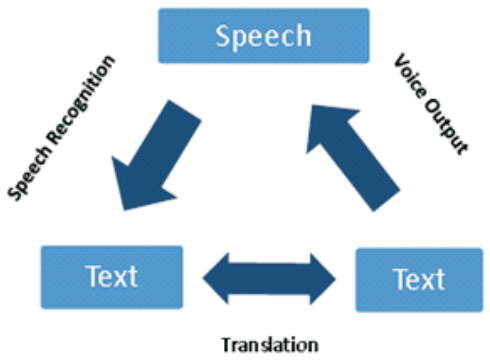

Figure 1: Working of Automatic Speech Recognition

This process begins when a speaker decides whatto say and actually speaks a sentence (This is a sequence of words possibly with pauses, u's, and um's.) The software then produces a speech wave form, which embodies the words ofthe sentence as well asthe extraneoussounds and pausesin the spoken input. Next, the software attemptsto decode the speech into the best estimate ofthe sentence. Firstit convertsthe speech signal into a sequence of vectors which aremeasured throughout the duration ofthe speech signal. Then, using a syntactic decoderit generates a valid sequence of representations [1].

VOICE RECOGNITION MODEL

The voice recognition model firstly recognizes voice which is converted into text. The speech engine trains the system through voice inputs. The voice recognizer then recognizes the voice commands later is converted into sound waves, which is interpreted using the training wizard which results in voice output.

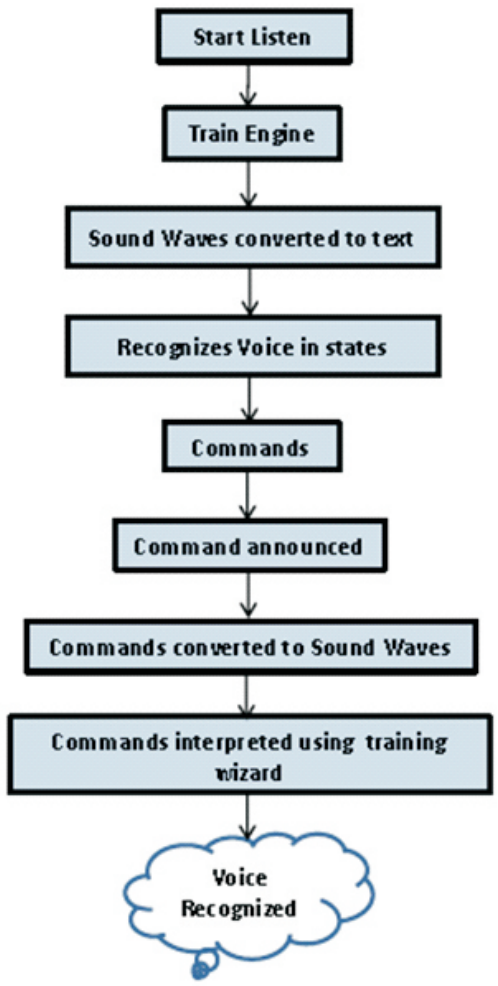

Figure 2: Voice Recognition Model

SPEECH ANALYSIS

Voice problems that require voice analysis most commonly originate from the vocal folds or the laryngeal musculature that controls them, since the folds are subject to collision forces with each vibratory cycle and to drying from the air being forced through the small gap between them, and the laryngeal musculature is intensely active during speech or singing and is subject to tiring. However, dynamic analysis of the vocal folds and their movement is physically difficult. The location of the vocal folds effectively prohibits direct, invasive measurement of movement. Less invasive imaging methods such as x-rays or ultrasounds do not work because the vocal cords are surrounded by cartilage which distorts image quality. Movements in the vocal cords are rapid, fundamental frequencies are usually between 80 and 300 Hz, thus preventing usage of ordinary video. Stroboscopicand high speed videos provide an option but in order to see the vocal folds, a fiberoptic probe leading to the camera has to be positioned in the throat, which makes speaking difficult. In addition, placing objects in the pharynx usually triggers a gag reflex that stops voicing and closes the

larynx. In addition, stroboscopic imaging is only useful when the vocal fold vibratory pattern is closely periodic.

The most important indirect methods are currently inverse filtering of either microphone or oral airflow recordings and electroglottography (EGG). In inverse filtering, the speech sound (the radiated acoustic pressure waveform, as obtained from a microphone) or the oral airflow waveform from a circumferentially vented (CV) mask is recorded outside the mouth and then filtered by a mathematical method to remove the effects of the vocal tract. This method produces an estimate of the waveform of the glottal airflow pulses, which in turn reflect the movements of the vocal folds. The other kind of noninvasive indirect indication of vocal fold motion is the electroglottography, in which electrodes placed on either side of the subject's throat at the level of the vocal folds record the changes in the conductivity of the throat according to how large a portion of the vocal folds are touching each other. It thus yields one-dimensional information of the contact area. Neither inverse filtering nor EGG are sufficient to completely describe the complex 3-dimensional pattern of vocal fold movement, but can provide useful indirect evidence of that movement [10].

SPEECH APPLICATION PLATFORM

The Speech Application Platform provides development and deployment environments for speech-enabled Web applications. The Speech Platform consists of two major components:

Speech Application SDK (SASDK): The SASDK is the component of the Speech Platform that enables developers to create and debug multimodal and voice-only applications.

Speech Server (SS): SS is the server-based infrastructure that deploys and runs distributed speech-enabled Web applications. SS provides scalable, secure, and manageable speech-processing services. MSS also enables deployment of both telephony applications and multimodal applications.

Process

When we use a computer system to perform a certain task, the computer system acts both as a tool and as a partner in communication.The information that is being exchanged between the user and the system during task performance can be represented in different forms, or modalities, using a variety of different input/output devices.

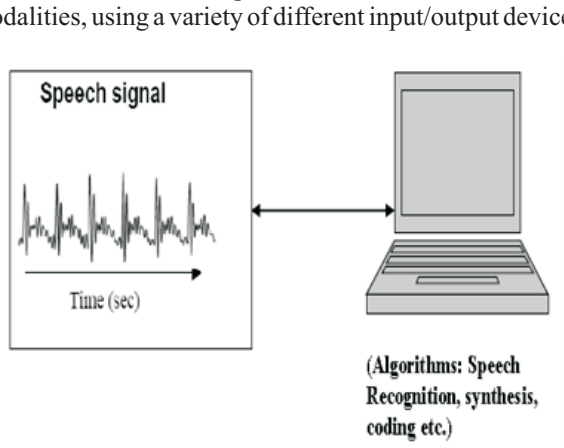

Figure 3: Speech recognition process by machine

A user interface is that component of any product, service, or application that interacts directly with the human user. The job of the interface is twofold. First, it must present information to the user - information about the task at hand as well as information about the interface itself. Second, it must accept input from the user - input in the form of commands oroperations that allow the user to control the application [Ballentine99].

ARCHITECTURAL DESIGN

This document will give you a technical overview of speech limitations occurring in the technology. Speech recognition fundamentally functions as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech.The figure below shows the design architecture related to this project.

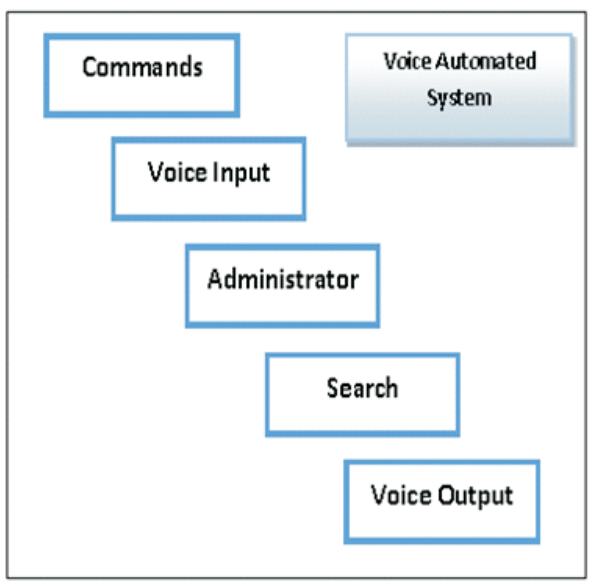

Figure 4:Architectural Design of Voice Automated System

The elements of the pipeline are:

1. Transform the PCM digital audio into a better acoustic representation
2. Apply a grammar so the speech recognizer knows what phonemes to expect. A grammar could be anything from a context-free grammar to full-blown English.
3. Figure out which phonemes are spoken.
4. Convert the phonemes into words.

Transform the PCM Digital Audio

The first element of the pipeline converts digital audio coming from the sound card into a format that is more representative of what a person hears. The digital audio is a stream of amplitudes, sampled at about 16,000 times per second. If you visualize the incoming data, it looks just like the output of an oscilloscope. It's a wavy line that periodically repeats while the user is speaking. While in this

ISSN 2230-7850
Volume-3, Issue-6, July-2013

form, the data is not useful to speech recognition because it is too difficult to identify any patterns that correlate to what was actually said.

To make pattern recognition easier, the PCM digital audio is transformed into the "frequency domain." Transformations are done using a windowed fast-Fourier transform. The output is similar to what a spectrograph produces. In frequency domain, you can identify the frequency components of a sound. From the frequency components, it is possible to approximate how the human ear perceives the sound.

The Fast Fourier transform analyzes every 1/100th of a second and converts the audio data into the frequency domain. Each 1/100thof second resultsare a graph of the amplitudes of frequency components, describing the sound heard for that 1/100thof a second. The speech recognizer has a database of several thousand such graphs (called a codebook) that identify different types of sounds the human voice can make. The sound is "identified" by matching it to its closest entry in the codebook, producing a number that describes the sound. This number is called the "feature number." (Actually, there are several feature numbers generated for every 1/100thof a second but the process is easier to explain assuming only one.)

The input to the speech recognizer began as a stream of 16,000 PCM values per second. By using Fast Fourier transforms and the codebook, it is boiled down into essential information, producing 100 feature numbers per second.

This does not work because of a number of reasons:

Every time a user speaks a word it sounds different. Users do not produce exactly the same sound for the same phoneme. The background noise from the microphone and users office sometimes causes the recognizer to hear a different vector than it would have if the user was in a quiet room with a high quality microphone. The sound of a phoneme changes depending on what phonemes surround it. The "t" in "talk" sounds different than the "t" in "attack" and "mist". The sound produced by a phoneme changes from the beginning to the end of the phoneme, and is not constant. The beginning of a "t" will produce different feature numbers than the end of a "t".

TEXT-TO-SPEECH

The technique of text-to-speech is used in various products, and maybe even incorporated it into your own application, but you still do not know how it works. This document will give you an overall technical overview of text-to-speech so one can understand how it works, and better understand some of the capabilities and limitations of the technology.

Text-to-speech fundamentally functions as a pipeline that converts text into PCM digital audio.
The elements of the pipeline are:

1. Text normalization
2. Homograph disambiguation
3. Word pronunciation
4. Prosody
5. Concatenate wave segments

1.Understanding of Voice

The Components which makes the current system is shown above. It has Four Components which are listed below.

Commands: This is one of the major components of the current system which recognizes the commands given by the user. This component is responsible for recognizing the commands and interpreting the command and sending appropriate request to the Search component.

Voice Input:It takes input form the Search component.

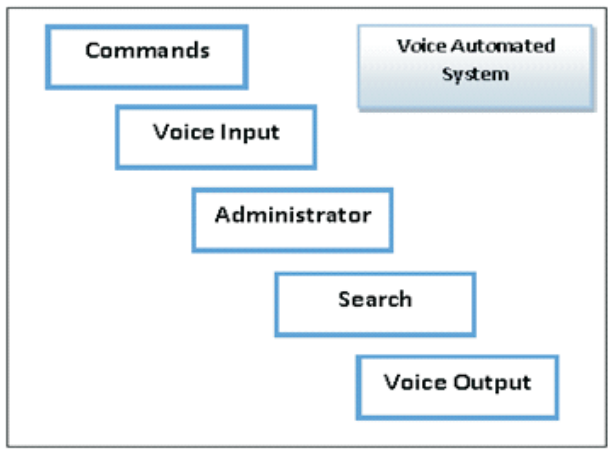

Figure 5: Voice Automated System

Administrator: Through this component the maintenance personnel can update the information and also the commands to the system.

Search: Search components take the input as the request from the Command component and retrieve the appropriate result from the database. It gives back to the display component and the speech component.

Speech: This component is used

Flow Diagram to deliver the result in the form of the voice using speech control.

The flow diagram for the following system is:

ISSN 2230-7850
Volume-3, Issue-6, July-2013

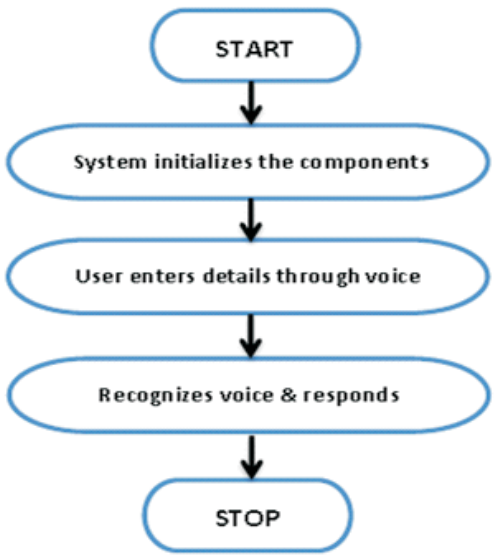

Figure 6: Flow Diagram of Voice Automated Systems

CONCLUSIONS

The objective of this work is mainly to build a speech recognizer for recognizing voice and notifying the user. In order to meet this objective a limited word grammar was constructed, a dictionary is created and data from different speakers was recorded and trained thereafter.

The system was tested using testing engine and live data. This implies that the objective of creating a system that can recognize spoken language was achieved. As much as it has created a basis for research, this work can be expanded to cater for more extensive language models and larger vocabularies.

REFERENCES:-

I]Rabiner, Lawrence R. and Juang, B.H. (2004).Statistical Methods for the Recognition and Understanding of Speech.Rutgers University and the University of California, Santa Barbara; Georgia Institute of Technology, Atlanta.

II]Rabiner L.R., S.E.L.evinson: (1981) "Isolated and connected word recognition – Theory and selected applications", IEEE Trans. COM-29, pp.621-629

III] Tolba, H., and O'Shaughnessy, D., (2001). Speech Recognition by Intelligent Machines, IEEE Canadian Review (38).

IV] Bridle, J., Deng, L., Picone, J., Richards, H., Ma, J., Kamm, T., Schuster, M., Pike, S., Reagan,R.,1998. An investigation of segmental hidden dynamic models of speech Co-articulation for automatic speech recognition.Final Report for the 1998 Workshop on Language Engineering, Centerfor Language and Speech Processing at John Hopkins University, pp. 161.

V] Ma, J., Deng, L., 2004. Target-directed mixture linear dynamic models for spontaneous speech recognition.IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL.12, NO.1, JANUARY 2004.

VI]Stolcke A., Shriberg E., Ferrer L., Kajarekar S., Sonmez K., Tur G.(2007)," Speech Recognition As Feature Extraction For Speaker Recognition" SAFE, Washington D.C., USA pp 11-13.

VII] Kesarkar M. (2003), "Feature Extraction For Speech Recogniton" M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dep , IIT Bombay.

VIII] Becchetti, C. and Ricotti, L. (2004), "Speech Recognition Theory and C++ Implementation", John Wiley & Sons, Wiley Student Edition, Singapore, pp. 121-188.

IX] [Ballentine99] Ballentine, B., Morgan, D. How to Build a Speech Recognition Application. A Style Guide for Telephony Dialogues, CA: Enterprise Integration Group, Inc., San Ramon. 1999

X] Voice analysis
(http://en.wikipedia.org/wiki/Voice_analysis)
(Accessed on date: June 11, 2013)

XI] Techniques for Feature Extraction in Speech Recognition System: A Comparative Study.
http://arxiv.org/abs/1305.1145.
(Accessed on date: June 11, 2013)

XII]http://www.en.wikipedia.org(Accessed on date: June 12, 2013)