International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

Speech Recognition For Hindi Language

Anushree Srivastava
Computer Science and Engineering
Institute of Technology and Management
Gorakhpur, India

Nivedita Singh
Computer Science and Engineering
Institute of Technology and Management
Gorakhpur, India

Shivangi Vaish
Computer Science and Engineering
Institute of Technology and Management
Gorakhpur, India

Abstract

Speech interface to computer is the next big step that computer science needs to take for general users. Speech recognition will play an important role in taking technology to them. Speech Synthesis and Speech Recognition together form a speech interface. A speech synthesizer converts text into speech. Thus it can read out the textual contents from the screen. Speech recognizer had the ability to understand the spoken words and convert it into text. Our goal is to create speech recognition software that can recognize Hindi words.

Keywords- *Acoustic, Corpus, Hidden Markov Model, Hindi, Speech Recognition, Sapi*

1. Introduction

Keyboard and mouse, although are popular medium but not very convenient as it requires a certain amount of skill for effective usage. Current computer interfaces also assume a certain level of literacy from the user. It also expects the user to have certain level of proficiency in English. Speech interface can help us tackle these problems. In this paper, we discuss about the survey done in Hindi language for building large vocabulary speech recognition systems.

Speech recognition refers to the ability to listen (input in audio format) spoken words and identify various sounds present in it, and recognize them as words of some known language.

Speech recognition in computer system domain may then be defined as the ability of computer systems to accept spoken words in audio format - such as the steps required to make computers perform speech recognition are: Voice recording, word boundary detection, feature extraction, and recognition with the help of knowledge models. Word boundary detection is the process of identifying the start and the end of a spoken word in the given sound signal. This can be attributed to various accents people have, like the duration of the pause they give between words while speaking. Knowledge models refer to models such as phone acoustic model, language models, etc. which help the recognition system. To generate a knowledge model one needs to train the system.

2. Speech Recognition (SR)

Speech recognition is the process of mapping an acoustic waveform into a text (or the set of words) which should be equivalent to the information being conveyed by the spoken words.

2.1 Modules of Speech Recognition

A speech recognition system comprises of modules as shown in the Fig 1[1].

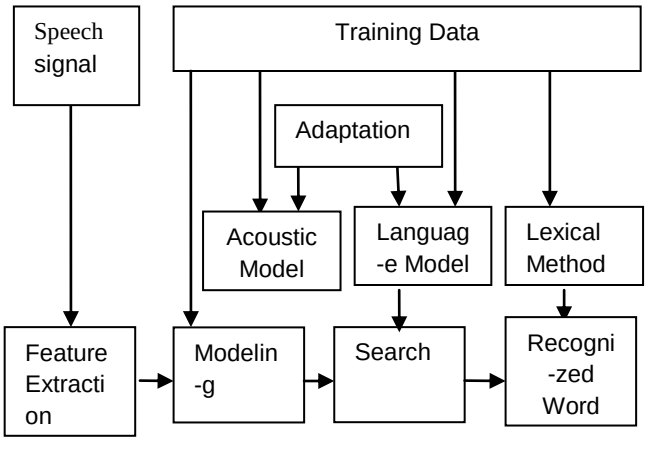

Figure 1: Block Diagram of a SR System

2.1.1 Speech Signal acquisition

At this stage, Analog speech signal is acquired through a high quality, noiseless, unidirectional microphone in .wav format and converted to digital speech signal.

2.1.2 Feature Extraction

Feature extraction is a very important phase of SR system development during which a parsimonious sequence of feature vectors is computed so as to provide a compact representation of the given input signal. Speech analysis of the speech signal acts as first stage of Feature extraction process where raw features describing the envelope of power spectrum are generated.

2.1.3 Acoustic Modeling

Acoustic models are developed to link the observed features of the speech signals with the expected phonetics of the hypothesis word/sentence.

2.1.4 Language & Lexical Modeling

Word ambiguity is an aspect which has to be handled carefully and acoustic model alone can't handle it. Lexical model provides the pronunciation of the words in the specified language and contains the mapping between words and phones. Generally a canonical pronunciation available in ordinary dictionaries is used. The purpose of performing adaptation is to minimize the system's performance dependence on speaker's voice, microphones, transmission channel and acoustic environment so that the generalization capability of the system can be enhanced.

2.1.5 Recognition

Recognition is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed. Two approaches are being used to match the patterns.

3. Approaches to Speech Recognition System

Speech Recognition has been an active research field since the invention of the vocoder by Homer Dudley in the late 1930s [2]. Different approaches have been developed to cope with the challenges while meeting the constraints of reasonable recognition rates and affordable computational requirements. In the following we present an overview of three common approaches to automatic speech recognition:

3.1 Hidden Markov Models (HMMs)

Hidden Markov Models (HMMs) is currently the most common approach to speech recognition. A Hidden Markov Model is a Markov chain in which the actual state of the chain is hidden from the observer. A Markov chain is a chain in which each state depends only on the previous state and does not depend in any way on any state other than the previous state. The different states of the HMM represent different distributions. The speech signal is modeled as a piecewise stationary stochastic process and in many applications time intervals are held constant at 10 ms. .A feature vector is computed for each time interval. Typically, a feature vector has 13elements which are the cepstral coefficients of the sampled speech signal in the current time interval. The features are then used to determine the state which represents the distribution associated with the specified time interval. Finally the Viterbi algorithms are used to perform Maximum A-Posteriori (MAP) analysis of the data and produce the sequence with the highest likelihood of occurrence. There has been a significant amount of work in the field of HMM-based automatic speech recognition systems and many theoretical and application specific algorithms exist [3].

3.2 Neural Networks

Neural Networks (NN) based systems were popular in the late 80s, however due to the relative success of HMM they have been somewhat neglected. Rabiner et al. [4] demonstrate the importance of spectral parameterization of a speech signal that serves as input to a NN system. Since linguistic isomorphism does not imply acoustic isomorphism, we can expect different spectral representations to similar words/phonemes. Two methods of parameterization that are commonly used are a bank of filters and an all-pole linear prediction model. A bank of filters is a set of overlapping filters that are spaced in frequency according to either a uniform or non-uniform law. The Linear prediction analysis technique models speech as an all-pole filter and looks at the distance from the coefficients of an actual known utterance as an optimization criteria measure.

3.3 Hybrid Systems

Hybrid systems as their name implies combine different strategies with the objective of improving recognition rates. Common hybrid systems are Neural Network Hidden Markov Model as described in [5]. Makhoul et al. suggest an N-best paradigm that uses multiple hypotheses instead of a single one. A segmental neural net is constructed to model the different phonemic connections. Such modeling is not possible to perform with a HMM since the basic assumption of the HMM restricts the dependency of

the current state only on the previous one. By using multiple connections a consistent improvement of performance is obtained.

4. Problem Formulation

In India, where people prefer Hindi over English for daily use, if information technology has to reach the grass root level, these constraints have to be thought over. Speech interface can help us tackle these problems. Most of the information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so that digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages. These technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages. While Hindi written in Devnagari script, is the official language, the other languages recognized by the constitution of India are: Assamese, Tamil, Malayalam etc. Seamless integration of speech recognition, machine translation and speech synthesis systems could facilitate the exchange of information between two people speaking two different languages.

5. Methodology

As an emerging technology, not all developers are familiar with speech recognition technology. While the basic functions of both speech synthesis and speech recognition takes only few minutes to understand (after all, most people learn to speak and listen by age two), there are subtle and powerful capabilities provided by computerized speech that developers will want to understand and utilize. Most importantly, speech technology does not always meet the high expectations of users familiar with natural human-to-human speech communication. Understanding the limitations, as well as the strengths is important for effective use of speech input and output in a user interface and for understanding some of the advanced features of the Speech API. An understanding of the capabilities and limitations of speech technology is also important for developers in making decisions about whether a particular application will benefit from the use of speech input and output. In our project we are using an open-source speech development kit of Microsoft, Speech API sometimes abbreviated as SAPI.

5.1 About SAPI

The Speech Application Programming Interface or SAPI is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. Broadly the Speech API can be viewed as an interface or piece of middleware which sits between applications and speech engines (recognition and synthesis). We are using SAPI 5.1, a member of version 5.0 of Microsoft's speech development kit.

5.1.1 SAPI 5.1

This version was launched in late 2001 as part of the Speech SDK version 5.1. Automation-compliant interfaces were added to the API to allow use from Visual Basic, scripting languages such as JScript, and managed code [10]. This version of the API and TTS engines was shipped in Windows XP. Windows XP Tablet PC Edition and Office 2003 also include this version, but with substantially improved version 6 recognition SAPI 5.1 supports OLE automation. The languages themselves need to support OLE automation.

5.2 Fundamentals to Speech Recognition

Speech recognition is basically the science of talking with the computer, and having it correctly recognized [9]. To elaborate it we have to understand the following terms [8], [11].

5.2.1 Utterances

When user says some things, then this is an utterance [11] in other words speaking a word or a combination of words that means something to the computer is called an utterance. Utterances are then sent to speech engine to be processed.

5.2.2 Pronunciation

A speech recognition engine uses a process word is its pronunciation, that represents what the speech engine thinks a word should sounds like [8]. Words can have the multiple pronunciations associated with them.

5.2.3 Grammar

Grammar uses particular set of rules in order to define the words and phrases that are going to be recognized by speech engine, more concisely grammar define the domain with which the speech engine works [8]. Grammar can be simple as list of words or flexible enough to support the various degrees of variations.

5.2.4 Accuracy

The performance of the speech recognition system is measurable [8]; the ability of recognizer can be measured by calculating its accuracy. It is useful to identify an utterance.

5.2.5 Vocabularies

Vocabularies are the list of words that can be recognized by the speech recognition engine [8]. Generally the smaller vocabularies are easier to identify by a speech recognition engine, while a large listing of words are difficult task to be identified by engine.

5.2.6 Training

Training can be used by the users who have difficulty of speaking or pronouncing certain words, speech recognition systems with training should be able to adapt.

6. Results

While providing voice input to the software it recognized the spoken words in few attempts, this is due to noisy environment, variation in the voice and multiple user factors.

This work can be taken into more detail and more work can be done on the project in order to bring modifications and additional features. The current software doesn't support a large vocabulary; the work will be done in order to accumulate more number of samples and increases the efficiency of the software. The current version of the software supports only few areas of the notepad but more areas can be covered and effort will be made in this regard.

7. Advantages

- Able to write the text through both keyboard and voice input.
- Voice recognition of different notepad commands such as open save and clear.
- Open different windows software's, based on voice input.
- Requires less consumption of time in writing text.
- Provide significant help for the people with disabilities.
- Lower operational costs.

8. Disadvantages

- Low Accuracy.
- Not good for noisy environment.

9. Conclusion

Lot of research in the field of Speech recognition is being carried out for Oriya, Malayalam, Bengali, Assamese, Marathi, Urdu and Sinhala languages of Indo-Aryan languages family. But there is a long way to go so as to enhance the performance standards set for other languages. Results of some SRE reported here show encouraging results. However, one should not forget that most of these experimental systems end in the lab; very few experimental systems are converted to real systems or products. Hindi, being a widely spoken Indo-Aryan language, is still trailing in the research and development for the field of automatic speech recognition. So far the work done for Hindi language is isolated word speech recognition using Acoustic template matching technique on MATLAB. In this paper, almost all the efforts made by various researchers for the research and development of SR systems have been analyzed.

10. Acknowledgement

The elation and gratification of this project will be incomplete without mentioning all the people who helped us to make it possible. Firstly, we would like to thank GOD, the almighty; we express our sincere gratitude to Mr. Rajeev Ranjan Kumar Tripathi, Head of Department, Computer Science and Engineering and Mr. Abhishek Kumar Srivastava, Professor of Department for his support and guidance.

11. References

[1] Wiqas Ghai and Navdeep Singh, "Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi A Case Study", Vol-2, Issue-1, March 2012.

[2] Dudley H., the Vocoder, Bell Labs Record, Vol. 17, pp. 122-126, 1939.

[3] Renals.et.al. "Connectionist Probability Estimators in HMM Speech Recognition" IEEETran.On Speech and Audio Processing, Vol. 2, No. 1, Part 11, pp. 161-174, Jan. 1994.

[4]Juang, B.H. and Rabiner L.R.;"Spectral representations for speech recognition by neuralnetworks-a tutorial",Neural Networks for Signal Processing [1992] II., Proceedings of the 1992IEEE-SP Workshop, pp. 214–222, Sep. 1992.

[5] Makhoul J.et al. "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition" IEEE Trans. on Speech and Audio Processing. pp. 151-160, Vol. 2, No. 1, Part 2, Jan. 1994.

[6] M.A. Anusuya and S.K. Katti "Speech Recognition by a Machine: A Review", Vol. 6, No. 3, 2009.

[7] T.Sakai and S.Doshita, The phonetic type writer information processing 1962, Proc.IFIP Congress, 1962.

[8] "Fundamentals of Speech Recognition". L. Rabiner & B. Juang. 1993. ISBN: 0130151572.

[9] "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and

Speech Recognition". D. Jurafsky, J. Martin. 2000. ISBN: 0130950696.

[10] http://www.nextup.com/sapi5doc/Getting_Started.html

[11] Stephen Cook “”Speech Recognition HOWTO” Revision v2.0 April 19, 2002. Source: http://www.scribd.com/doc/2586608/speechrecognitionhowto