Speech Recognizer for Tamil Language

Mr. R. Arun Thilak* & Mrs. R. Madharaci**
*Final Year B.Tech <arun_thilak@yahoo.com>;
**Senior Lecturer <madharaci@yahoo.com>,
Sri Venkateswara College of Engineering,
Sriperumbudur, Tamilnadu, India

Abstract

Tamil being the most popular language in South East Asia, developing a Speech Recognizer for Tamil Language will be a base for many applications. This paper deals with development of a Speech Recognition Engine for Tamil. The Speech Recognition Engines for English are not accurate enough to understand what the user speaks perfectly. These engines are based on syllables, which correspond to more than single character. There is certain amount of ambiguity in the recognition. The use of Tamil in Speech Recognition Engine can improve its accuracy. This is because each character in Tamil has distinct pronunciation and hence each word has a distinct pronunciation. The use of literary Tamil can also enhance the accuracy further as this form of Tamil doesn't vary from region to region or from person to person by a great extent. The proposed implementation for this Speech Recognition Engine does a step-by-step process for the recognition. Firstly, we build a lexicon that contains the words and the corresponding pronunciation. Next, we try to predict the words by parsing the phonemes based on the lexicon. Next, we arrive at the best-matched word. Finally, we apply grammar rules to fit the proper word in the context of the sentence.

1 Introduction

Building a Speech Recognition Engine with more accurate output needs an in depth study about the features of Tamil and the technical details, which will affect the accuracy of the Speech Recognizer. This paper gives an insight into Speech Recognition in Tamil, all the technical, practical feasibilities and the way in which it is done. We also give an idea about how effectively this Speech Recognition in Tamil can work against those already available for English Language. By better recognition we can have diverse applications from Speech to Text Converter, Text to Speech Converter, Voice Keyboards, Voice Websites, etc.

Since we already have Speech Recognition Engines in English, why haven't they replaced the keyboards? The reason is simple - they are not accurate enough. The implementation of Speech Recognition Engine in Tamil will expose its unique features and simply highlight the accuracy that can be achieved by using Tamil as language for the Speech Recognition Engine.

2 Basics of Speech Recognition

2.1 Characteristics of Sound

The various characteristics of a sound wave are frequency, amplitude and time-period.

These characteristics are essential in studying further about human voice. The various *Tamil Internet 2004, Singapore* 118 mappings of frequency, amplitude and time-period in terms of voice are pitch, loudness and duration.

Spectrum

A spectrum is a representation of the different frequency components of a wave. It can be computed by a Fourier Transform, a mathematical procedure which separates out each of the frequency components of a wave. Rather than using the Fourier Transform spectrum directly, most speech applications use a smoothened version of the spectrum called the *LPC spectrum* [Atal and Hannauer, 1971; Itakura, 1975]. LPC (Linear Predictive Coding) is a way of coding the spectrum that makes it easier to see where the spectral peaks are. Analysis of the different sounds uttered is much easier with the LPC spectrum of the sounds.

2.2 Speech Recognition Process

By speech recognition, we have in mind the computational technique that processes spoken language. Without a doubt, the ability of human mind's speech recognition capabilities cannot be 100% perfect when it is simulated in the computers. But we can achieve some amount of accuracy. In order to recognize a voice, the system must first analyze the incoming audio signal and try to recover the exact sequence of words. This task requires some knowledge about *phonetics and phonology*, which gives an idea of how the words are being spoken in the colloquial world. The next task is to recognize the variations of individual words, which requires knowledge about *morphology*, which captures information about shape and behavior of words in context. Next is knowledge about the structure of various words and sentences called the *syntax* and their meanings called the *semantics.* Some of the terms involved in this speech recognition context are explained below.

Phones or Phonemes

The individual or basic units of sound are called as Phones or Phonemes. A speech recognition system needs to have a pronunciation for every word it can recognize. Phonology is the area of linguistics that describes the systematic way that sounds are differently realized in different environments, and how this system of sounds is related to the rest of the grammar. In software we have *Computational Phonology* [1], which is the study of computational mechanisms for modeling phonological rules.

Syllable

Speech recognition in English is based on Syllables. A syllable is a continuous utterance of letters separated by a break, or pause while speaking. Consonants and vowels are combined to make a syllable. A syllable is a vowel like sound together with some of the surrounding consonants that are most closely associated with it [1].

International Phonetic Alphabet (IPA)

The IPA is an evolving standard originally developed by the International Phonetic Association in 1888 with the goal of transcribing the sounds of all human languages. The IPA is not just an alphabet but also a set of principles for transcription, which differ according to the needs of the transcription. *Tamil Internet 2004, Singapore* 119

Consonants and vowels

Consonants are those sounds that are made by restricting the airflow in some manner and vowels are those that are not restricted in any way and are only produced by the movement of air. In Tamil, the consonants are nothing but the “Mei Ezhulthukal” and Vowels are nothing but the “Uyir ezhuthukal”.

3 Current Speech Recognition Engines

To design an efficient Speech Recognition Engine we need to understand the problems faced by the Speech Recognition Engines for English. One of the major drawbacks is the accent of the user. It varies with the locality. The variation in spoken words is so much that the Speech Recognition Engine for particular locality may not work at another locality. The other major problem is the ambiguity that exists in the English language itself. Some syllables have different letters but have the same pronunciation. There is no variation in some cases. Even when the grammar is well defined, the language is basically built upon letters that are given sounds and not sounds that are given letters. This is a major difference between most of the Indian languages and European languages. English and French especially have this problem where a single letter is pronounced in a different way depending upon the position in the word where it is occurring.

3.1 Speech Recognition Architecture of Current System

Speech recognition fundamentally functions as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech. 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.

4 Features of Tamil

The various unique features of Tamil that we want to exploit for the Proposed Speech Recognition Engine are given below:

o The classical language Tamil is based entirely on the phonemes. Each and every letter in the language has a distinct pronunciation that distinguishes it from the rest. The words are formed from letters and hence each word will have distinct pronunciation. This feature of Tamil can be exploited in the Speech Recognition Engine to achieve high degree of accuracy.

o The grammar for the language is well defined. The grammar for Tamil, though has lots of contents, is not ambiguous. The specifications of the grammar are clearly laid out.

o The number of words in Tamil is around 3 lakhs (approx.). Hence maintaining a large vocabulary is also difficult when the system needs to use Tamil.

o **Literary Tamil**

o There are so many dialects in standard spoken Tamil. This spoken Tamil varies from region to region within Tamil Nadu itself, mostly because of the different *Tamil Internet 2004, Singapore* 120 communities that exist within. Literary Tamil is one that is long standardized and the important fact is, it doesn't vary from region to region. Literary Tamil is the most remarkable feature of Tamil because it is completely independent of the region in which it is spoken. This is not the case in English as it varies from region to region in a wide manner. [2]

o Another important feature of Literary Tamil is that it doesn't vary by a great extent from speaker to speaker. Since the entire language is based on sounds, and the sounds are clearly defined for each letter this speaker to speaker variation is reduced. This is again not the case in English, as English varies by a great deal from person to person. Though they may use literary English or any other form of English, the difference is noticeably felt. The variation in accent is little bit more in English than in Tamil. Hence we will try to exploit these unique features of Tamil in this Speech Recognition Engine.

5 Proposed Speech Recognition Engine

There are 4 main stages in the implementation process. They are

1. Use of a new lexicon to accommodate the Tamil words and their corresponding pronunciation.
2. Top-Down Parsing technique to be adopted for predicting the next phoneme.
3. Selecting the best match for the spoken word from the lexicon and the parser above.
4. Identifying the words in context by specifying the Grammar.

The tools we have selected for this proposed implementation is Microsoft SAPI. Microsoft's® SAPI is a tool for developing applications using speech recognition and also for developing Speech Recognition Engines for other languages.

5.1 Lexicon Builder

The first step to implement this Speech Recognition Engine is to build a lexicon that will contain all the words in Tamil and also contains the corresponding pronunciation. The pronunciations are given in a series of phonemes. Each phoneme can correspond to a letter that is given in the IPA – International Phonetic Alphabet. So the system will recognize the phoneme and give its corresponding IPA code. The IPA code is a Hexadecimal number that corresponds to a particular phoneme or pronunciation of a single character. We use SAPI to build this Speech Recognition Engine. When we start building it, SAPI asks us to create a new lexicon for Tamil Language. The English characters that correspond to all the Tamil characters are identified. We have created the lexicon for some limited number of Tamil words. The set of English characters for the words in Tamil and their corresponding phonemes are identified. The lexicon is built based upon this. The phoneme representation for the other words in Tamil can also be identified and accommodated in the lexicon. When all the Tamil words are included in the lexicon, the first step of the proposed Speech Recognition Engine will be complete.

5.2 Phoneme Prediction

The next step is to predict the phoneme that the user might speak. This is used to limit what the computer might expect as the next phoneme. This is done by using a Top-Down Parsing technique. Hidden Markov Models are used to predict the probability of next *Tamil Internet 2004, Singapore* 121 phoneme in the sequence to find the word spoken and to predict the probability of word in the sentence spoken. Viterbi Algorithm can be used to compute the optimal state sequence in a Hidden Markov Model, given a sequence of observed phonemes. However, the word is not selected here. Here the system only tries to list all possibilities or combinations of the phonemes that comes after the current phoneme. This step only increases computational accuracy as the Speech Engine will know what to expect from the user. It need not store the pronunciation for entire word and search for that word in lexicon. Parsing technique will increase the speed of finding the best match for the uttered word from user.

5.2 Finding the Proper word

In Hidden Markov Model, state chains are formed and the system tries to find the best match for the uttered word. Viterbi Algorithm uses probability of occurrence of phonemes to scan for the next phoneme. This probability value helps in selecting the best match for the spoken word from the lexicon. With inputs from the previous phase and by comparing with the current phoneme that the user speaks, the Speech Recognition Engine can easily identify the word without much ambiguity. Words are distinguished from other words by a pause between them. The phonemes are distinguished by analyzing the variation in spectral components of the voice.

5.4 Grammar Specifications

The final phase will be to identify the proper words based on the grammar specified by the language. SAPI permits us to define the grammar. Though extensive, it can help in finding and placing the right words in the right tense and formation of a proper sentence.

5.4.1 Context Free Grammar

One of the techniques to reduce the computation and increase accuracy is called a "Context Free Grammar" (CFG) [4]. CFGs work by limiting the vocabulary and syntax structure of speech recognition to only those words and sentences that is applicable to the application's current state. The application specifies the vocabulary and syntax structure in a XML file following Natural Language Grammar specifications. The grammar~~s are~~ is much more complex than simple sentences. The important feature about the CFG is that it limits what the recognizer expects to hear to a small vocabulary and tight syntax. This method significantly reduces the number of generated hypothesis. When the user has finished speaking, the recognizer returns the hypothesis with the highest score, and the words that the user spoke are returned to the application. If the Speech Recognition Engine is going to be designed for a particular application then it can certainly limit the number of words in the lexicon and hence increase computational speeds.

6 Difficulties

Tamil language has more than 3 lakh words. Creating a comprehensive lexicon containing all the words along with the proper pronunciation will without doubt take time and care. We need to identify the phonemes carefully for each word and present them in IPA characters. *Tamil Internet 2004, Singapore* 122 Another major hurdle in Tamil is that there are three different la's, three different na's and two different ra's. Though they are clearly specified in Tamil upon their correct way of pronunciation it will be difficult for the system to recognize them clearly and with high accuracy. Clear recognition is possible only if we pronounce it very distinctly and clearly. This is one of the cases of ambiguity in case of use of Tamil.

We can overcome this difficulty if we can specify the entire grammar for the language, where the Speech Recognition Engine will not only try to identify the correct letter but also can judge which letter comes in this context. But this requires the Speech Recognition Engine to be supplied with lots of data about the grammar in Tamil.

7 Conclusion & Future Work

In this paper we have dealt about the features of Tamil, which will affect the performance of the Speech Recognition Engine. A method for developing the Speech Recognition Engine is also described.

If we are able to recognize the various la's, ra's and na's in Tamil, we can also try to study or develop Natural language processing abilities for the system. Here the emphasis is on the mood of the speaker that is determined by the stress and excitement in voice.

We can also develop a text to speech engine, which is basically making the system read out the phonemes that we have entered for each word. Since the pronunciations are distinct for each word and Tamil has a methodology to specify the duration for each word and letter in that word, it is possible that we can generate a real time Naturally speaking voice engine. The important feature could be that this engine need not use pre-recorded voice.

After development of a comprehensive Speech Recognition engine for literary Tamil we can also extend it for spoken Tamil. Though the number of dialects in Tamil is more we can certainly create one for each characteristic region.

We can also create a method in which the user trains the Speech Recognition Engine to adapt to the user's voice. A different voice or speech profile can be maintained that stores the alternative pronunciation for each word. This alternative pronunciation may include slight change in pitch, amplitude and duration.

Acknowledgements

We thank our Head of the Department, Principal and the Management for their constant encouragement to write this paper and to present this paper in the TI2004 conference. We also thank the review committee of this INFITT forum for selecting our paper for oral presentation. Our sincere thanks to the Internal Review Committee members of our college, for their guidance in shaping this paper more appropriately.

Bibilography

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing.* Pearson Education, 2003.
[2] Harold_F.Schiffman
URL: http://ccat.sas.upenn.edu/plc/tamilweb/book/chapter1/node1.html
Tamil Internet 2004, Singapore 123
[3] Harish E.S., Neelakantan N.S. and Govind K., [Guide: E.S.Gopi, B.E., Lecturer ECE] *Automatic Music Transcription and Raaga Recognition* (A final year project)
[4] Microsoft® SAPI 5.1 Documentation
URL: http://www.microsoft.com
URL: http://go.microsoft.com/fwlink?linkid=288&clcid=0x409