Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 1097 – 1102

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM

Ms.Vimala.C[a], Dr.V.Radha[b], a*

Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women,Coimbatore,India

Abstract

The ambitious speech research for more than 50 years is having a machine to understand fluently spoken speech. This can be achieved by developing an Automatic Speech Recognition (ASR) system. Such a significant job specifically for Indian language has been focused in this paper. Different types of speech recognition systems are available for different application domains. This research work presents a speaker independent isolated speech recognition system for Tamil language. The most flexible and successful approach to speech recognition so far has been Hidden Markov Model (HMM) which is implemented in this research work. The HMM method provides a natural and highly reliable approach of recognizing speech for a broad range of applications. The experiments using HMM furnish high-quality word accuracy of 88% for trained and test utterances spoken by the speakers. The performance evaluation of the system is done based on the Word Error Rate (WER) which gives 0.88 WER for the above research work.

Keywords: Isolated Speech Recognition; Tamil language, Sphinx4; Mel Frequency Cepstral Coefficients;Hidden Markov Model;Word Error Rate;

1. Introduction

Speech is a natural mode of communication for people. People are so comfortable with speech since it does not require any special training. If people can use speech as the communication media to interact with the computers, rather than keyboards and pointing devices, then possibly they will embrace computers much adequately. Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it into written text [1]. Different types of speech recognition system can be developed based on the type of speech, speaker and vocabularies are used. These categories are used depending on the type of the

* Vimala.C. Tel.: +91-9952490710
E-mail address: vimalac.au@gmail.com.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.968

application that the people use. Today's researches mainly focus on developing speech recognition systems for Indian languages [2]. In this research work, small vocabulary speaker independent isolated speech recognition system is developed for Tamil language [3]. Tamil is a Dravidian language spoken predominantly in the state of TamilNadu in India and in Sri Lanka. It is the official language of the Indian state of TamilNadu and also has official status in Sri Lanka and Singapore. With more than 77 million speakers, Tamil is one of the widely used spoken languages of the world [3]. Hence there is special need for the ASR system to be developed for people those who speak Tamil language. This research work is developed for speaker independent isolated speech using Sphinx4 which is based on Hidden Markov Models (HMMs). It is a flexible, modular and pluggable framework to help foster new innovations in the core research of HMM recognition systems. Sphinx4 is used in this research work because of its high degree of flexibility and modularity.

The paper is organized as follows. Section 2 presents the system overview, section 3 explains about the pre-processing steps involved, section 4 gives details about the post-processing steps, section 5 investigates the experimental results and section 6 deals with the performance evaluation. Finally, the conclusion is summarized in section 7 with future work.

2. System Overview

The categories of speech recognition based on utterance include isolated speech, connected speech, continuous speech and spontaneous speech. Among these categories isolated speech recognition [4] is comparatively simple because word boundaries are obvious and the words tend to be clearly pronounced. An isolated speech recognition system requires that a speaker pause briefly between words as opposed to a continuous speech recognition system. It doesn't mean that it accepts single words, but does require a single utterance at a time where the speaker is required to wait between utterances [4].

In order to recognize speech, the system usually consists of two phases. They are called pre-processing and post-processing. Pre-processing involves feature extraction and the post-processing stage comprises of building an acoustic model, phonetic lexicon or the pronunciation dictionary and language modeling. The following figure1 explains the overview of the ASR system.

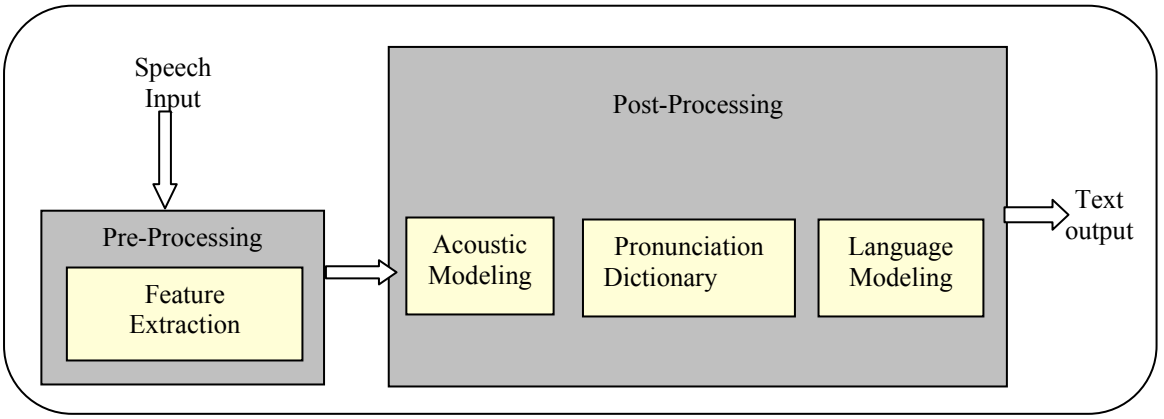

Figure 1 Steps involved in ASR system

Initially the speech waveform is processed to produce a new representation as a sequence of vectors containing values of features or parameters. The parameter values extracted from raw speech are used to build acoustic models. The Dictionary provides pronunciations for words found in the Language Model.

The pronunciations break words into sequences of sub-word units found in the Acoustic Model. The other component is called a language model, which gives the probabilities of sequences of words. The following sections explain about the entire system in detail.

3. Pre-Processing

The pre-processing or front-end of a speech recognizer is responsible for the extraction of the features from the speech waveform. The motivation of feature extraction is to convert speech waveform into a parametric representation at a lower information rate for further analysis. The Mel Frequency Cepstral Coefficients (MFCC) is the most evident example of a feature set that is extensively used in speech recognition. It approximates the human system response more closely than any other system. Sphinx-4 is typically configured with a Front-End that produces MFCCs[5]. Technique of computing MFCC [7] is based on the short-term analysis, and thus from each frame a MFCC vector is computed. The MFCC can be calculated by using the equation (1)

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \quad \text{--------- (1)}$$

The following figure 2 shows the steps involved in MFCC feature extraction [7]. These features are used for further process of post processing.

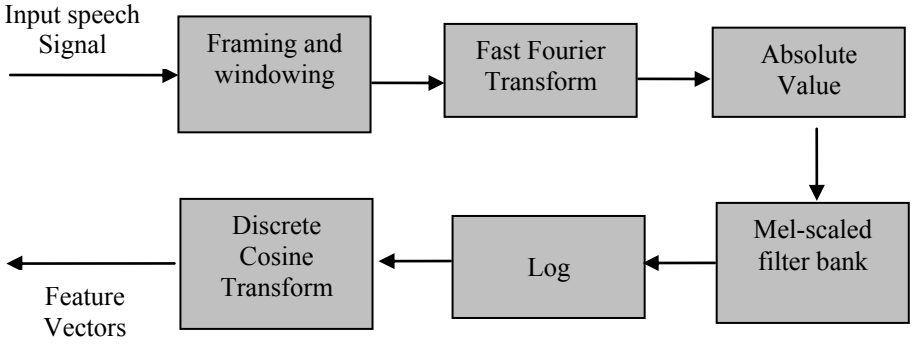

Figure 2 Feature extraction using MFCC

4. Post-Processing

Fundamentally, the problem of speech recognition can be stated as follows. When given with acoustic observation $X = X_1, X_2 \ldots X_n$, the goal is to find out the corresponding word sequence $W = W_1, W_2 \ldots W_m$ that has the maximum posterior probability P(W|X) expressed using Bayes Theorem as shown in equation (2).

$$W = \arg\max_{w} P(W/X) = \arg\max_{w} \frac{P(W)P(X/W)}{P(X)} \quad \text{---------- (2)}$$

Where P(W) is the probability of word W uttered and P(X|W) is the probability of acoustic observation of X when the word W is uttered. P (X|W) is also known as class conditioned probability distribution. P(X) is the average probability that observation X will occur. Since the maximization of equation (2) is done with variable X fixed, to find the word W it is enough to maximize the numerator alone.

In order to implement the speech recognition, any ASR system must contain three constituent parts which are also called as post- processing components. They are

- Acoustic Model
- Phonetic Lexicon or the Pronunciation Dictionary
- Language Model

These three models are needed to work together during recognition process. They are explained below in detail.

4.1. Acoustic Models

The first and foremost principle of the speech recognition that makes the system useful and powerful is the acoustic models [4]. It is a very important process because it directly affects the search speed, and accuracy. Acoustic modeling of speech typically refers to the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. It also encompasses "pronunciation modeling", which describes how a sequence or multi-sequences of fundamental speech units (such as phones or phonetic feature) are used to represent larger speech units such as words or phrases which are the object of speech recognition. The most common types of acoustic models are based on HMM [6] which is adopted for this research work.

4.2 Pronunciation Dictionary

Pronunciation Dictionary also called as lexicon contains the detail about how the words are pronounced. The lexicon should contain all the words that the speech recognition engine need to recognize. The pronunciation dictionary consists of records containing words and associated monophone sequences [4]. Some words have multiple pronunciations. Like speech corpora, a pronunciation dictionary is also a language specific resource. So, the pronunciation dictionary for every word in the vocabulary has been constructed manually for Tamil language. The sample pronunciation dictionary for Tamil speech is as follows.

திருக்குறள்	t h i r u k k u R a L
இணையம்	i n N a i y a m

4.3 Language Model

A language model is used to restrict word search. It helps a speech recognizer figure out how likely a word sequence is, independent of the acoustics. It represents previous knowledge about language and the expectations at utterances. It can be expressed in terms of which words or word sequences are possible or how frequently they occur. They are usually trained by observing sequences of words in corpora of text that contain. Every word in the language model must be in the pronunciation dictionary. The sample grammar created for Tamil isolated speech is given below.

வார்த்தை = (அமைப்பு | அழி | அனுப்பு | இணையம் | உலகச்செய்திகள்);

5. Experimental Results

The small vocabulary speaker independent isolated speech recognizer for Tamil Language is implemented using Sphinx4. Sphinx-4 is a state-of-art HMM-based speech recognition system. HMM

based speech recognition system [10] identifies speech by estimating the likelihood of each phoneme at contiguous, frames of the speech signal. A search procedure is used to determine the sequence of phonemes with the highest likelihood [8]. This search is constrained to look for phoneme sequences that correspond to words in the vocabulary, and the phoneme sequences with the highest total likelihood are identified with the spoken word.

An important requirement for developing any ASR system is having a speech corpus [7] [8]. Since speech corpora are not available for Tamil, it is created manually. The corpus containing 50 utterance of isolated speech collected from 10 females are used for training. The database consists of 5 repetitions of every word produced by each speaker. Totally (50x10x5) 2,500 words are collected for this research work.

6. Performance Evaluation

The performance of speech recognition system is generally measured in terms of word error rate, which is the ratio between misclassified words, and total number of tested words. ASR research has been focused on minimizing the recognition error to zero in real-time independent of vocabulary size, noise, speaker characteristics and accent. In this research work, out of ten speakers four speakers' data are used for testing. The system can able to recognize forty four words out of fifty words in average. The WER [9] can be computed using the equation (3).

$$\text{WER} = \frac{\text{Number of words correctly recognized}}{\text{Total number of words}} \quad \text{---------- (3)}$$

The obtained error rate for this research work is 0.88. Speech recognition is generally easy when vocabularies are small, but word error rate increases as the vocabulary size grows.

7. Conclusion

The development of technologies that enable human beings to talk naturally with computers is the dream of all researchers for many years. This can be accomplished by developing an ASR system. In this paper, the small vocabulary speaker independent isolated speech recognition system for Tamil language was developed and its performance is investigated using sphinx4. The four important components of ASR system namely feature extraction, acoustic model, pronunciation dictionary and language model are implemented using HMM. The database size used for this research work is 2,500 words and produces 88% of accuracy. As the vocabulary is small the system gives minimum word error rate for this system. In future, medium or large vocabulary isolated speech and continuous speech can be put into practice and can be experienced for Tamil language.

References

[1] Mr. R. Arun Thilak & Mrs. R. Madharaci, "Speech Recognizer for Tamil Language", Tamil Internet 2004, Singapore.

[2] M. Chandrasekar, M. Ponnavaikko, "Spoken TAMIL Character Recognition", in Electronic Journal Technical Acoustics (EJTA), ISSN 1819-2408, 2007.

[3] M. Chandrasekar, and M. Ponnavaikko, "Tamil speech recognition: a complete model", Electronic Journal Technical Acoustics (EJTA) ,ISSN 1819-2408, 2008.

[4] Gailius RAŠKINIS, "Building Medium-Vocabulary Isolated-Word Lithuanian HMM Speech Recognition System", INFORMATICA, Vol. 14, No. 1, 75–84 75, 2008.

[5] A.P.Henry Charles and G.Devaraj, "Alaigal-A Tamil Speech Recognition" , Tamil Internet 2004,Singapore.

[6] Mohammad A. M. Abushariah, Raja N.Ainon, Roziati Zainuddin, Moustafa Elshafei and Othman O. Khalifa, "Natural Speaker-Independent Arabic Speech Recognition System Based on Hidden Markov Models Using Sphinx Tools", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 201), Kuala Lumpur, Malaysia.

[7] A.Rathinavelu, G.Anupriya, A.S.Muthanantha murugavel, "Speech Recognition Model for Tamil Stops", Proceedings of the World Congress on Engineering 2007 Vol I WCE 2007, July 2 - 4, 2007, London, U.K.

[8] S.Saraswathi and T.V.Geetha, "Morphoem based language model for Tamil Speech Recognition system", The International Arab Journal of Infromation Technology, Vol.4, No.3, July 2007.

[9] H. Satori, M. Harti and N. Chenfour "Arabic Speech Recognition System Based on CMUSphinx", 1-4244-11 58-0/07/$25.OO © 2007 IEEE.

[10] R. Thangarajan, A.M. Natarajan, "Syllable Based Continuous Speech Recognition for Tamil", South Asian Language Review, VOL.XVIII. No. 1, January 2008.