INVESTIGATION OF EFFECT OF RATE OF SPEECH(ROS) ON SPEECH RECOGNITION SYSTEM

A major project report submitted
in partial fulfillment of the requirements
for the award of the degree of

Master of Technology
in
Computer Science

By

M.NAVEEN
(08MCMT09)

Under the Guidance of

Prof. P.N.Girija

Department of Computer & Information Sciences
School of Mathematics & Computer/Information Sciences
University of Hyderabad
Hyderabad -500046, INDIA.
April – 2010

CERTIFICATE

This is to certify that the thesis entitled "Investigation of effect of Rate of Speech(RoS) on speech recognition system" being submitted by **M.NAVEEN** in partial fulfillment for the award of the degree of **Master of Technology** in Computer science, is a record of the bonafide work carried out by him under my supervision at the University of Hyderabad.

The matter embodied in this project report has not been submitted to any other university for the award of any other degree or diploma.

Prof. P.N.Girija
Supervisor,
Dept. of CIS,
University Of Hyderabad,
Hyderabad.

.

Prof. Arun Agarwal,
Head,
Dept. of CIS,
University of Hyderabad,
Hyderabad.

Prof. T. Amaranath,
Dean,
School of MCIS,
University of Hyderabad,
Hyderabad.

Acknowledgments

There are a lot of people without whose support: physical, technical and moral, neither the project nor this manuscript could possibly have neared completion. Though I would have liked very much to do so, it is unfortunately not feasible to mention all of them individually here, as that would probably occupy half of this report.

I would like to take this opportunity to thank my guide **Prof. P.N.Girija,** for her guidance, valuable suggestions, support in every task of my work, and freedom in expressing my opinions during my project work.

I would like to specially thank the Head of the department **Prof. Arun Agarwal,** who very kindly provide me the facilities required to proceed with my work.

I would like to thank the Dean of the School of MCIS, **Prof. Amaranath** for giving me valuable suggestions.

My heartful thanks to my **friends,** and thanks to **classmates** for encouragement and support which they given me in completion of my course.

Before closing, I would like to extend to special attention to my **Parents,** who gave me this position and providing me inspiration and moral support throughout my studies.

M.NAVEEN

TABLE OF CONTENTS

Abstract

The main aim of this work is to investigate the effect of Rate of Speech(RoS) on the speech recognition system and how to improve the speech recognition accuracy for fast spoken Telugu speech. The speech can be broadly classified as three categories i.e. fast, medium, and slow. In our experiments we used sphinx 3.6 speech recognition software to find recognition accuracy and we trained the system with 1200 sentences. We tried to improve recognition accuracy by performing with increasing number of iterations for unrecognized words and by changing the duration of speech signal for unrecognized words

chapter1. Introduction

1.1 Introduction

In the last few years speech recognition and human computer interaction become more and more widespread. Speech recognition become very interesting research topic in India and speech is the best natural medium for communication with computer. There are many speech recognition systems. These speech recognition systems and methods fail to show a reliable performance on rapidly spoken speech. Current research therefore aims to increase the robustness of such systems by focusing on speech rate variation of spontaneous speech.

There are many factors that effects speech recognition systems and speech recognition accuracy like Rate of speech, noise conditions, channel distortion, speaker variation, dialectal variation, environmental conditions etc. As compared to read speech spontaneous speech exhibits many hesitations and variations in the rate of speech as speakers have to plan their further speech while speaking. Also, slips of the tongue and repairs occur and the speaking style tends to become less clear. Thus, spontaneous speech shows a high degree of variation with respect to RoS. So it is challenging task to increase the recognition accuracy for fast speech.

Speech rate is the one dimension that has been known to cause a high degree of variation and causes high Word Error Rate(WER) in speech recognition. The modeling of speech rate has therefore received considerable attention. So there is a need of new methods and techniques that increases the speech recognition accuracy for fast speech and not effect other modules of speech recognition system. Training the speech recognition system by applying the new techniques and methods may help the current speech recognition systems to capture variations of the speech rate very well.

A predominant paradigm in the modeling of speech rate is the training of rate dependent

models which is achieved by a separation of the training data into discrete rate classes. This approach is not giving good results. The other methods we are proposed is changing the speech signal of fast speech and training the system and training the system for unrecognized words by taking the multiple copies of those sentences. The other method is the changing the rate of speech for fast and slow speech to normal and training the system. All these methods giving the good results for fast Telugu speech.

In general these approaches are not giving detailed knowledge about the underlying effects of speech rate on entire speech recognition system. Little is

known about the actual effects that have to be modeled for faster or slower speech and how to capture those effects. In order to provide detailed insights into the effects that the acoustic correlates of speech rate variations have on current speech recognition approaches more research is needed.

1.2 Aim of the work

The main aim of this work is to investigate the effect of Rate of Speech(RoS) on the speech recognition system and how to improve the speech recognition accuracy for fast spoken Telugu speech. The speech can be broadly classified as three categories i.e. fast, medium, and slow. In our experiments we used sphinx 3.6 speech recognition software to find recognition accuracy and we trained the system with 1200 sentences. We tried to improve recognition accuracy by performing with increasing number of iterations for unrecognized words and by changing the duration of speech signal for unrecognized words.

1.3. Organization of Report

The rest of the report is as follows

• Chapter 2: is a survey of previous work in the area of speech rate and review chapter.

• Chapter 3:design of the system and explains components of the SPHINX-3.6 CSR system relevant to this research.

• In Chapter 4:methods and their implementations with respect to RoS .

• Chapter 5:describes overall results and diagrams.

• Chapter 6 summarizes this research report and provides suggestions for future work and references.

2. Review chapter and previous work.

2.1 Previous work

There are many people who has done good work in the speech rate area for different languages. As part this project we referred several papers on how rate of speech effects the accuracy of the speech recognition for Telugu language. Jing Zheng et.al[1] studied the Variations in rate of speech produce changes in both spectral features and word pronunciations that affect automatic speech recognition (ASR) systems.

Jing Zheng et.al[1] propose to use parallel, rate-specific, acoustic models: one for fast speech, the other for slow speech. Jing Zheng et.al proposed several method for RoS those are Absolute RoS measures, such as phones per second (PPS) and inverse mean duration (IMD). In their experiments they used a relative RoS

methods where they used sub word technique word the duration distribution of its component sub word units, such as phones, are independent of each other. word's duration distribution equals the convolution of its component sub word units' distributions, which are easier to estimate from training data. They proposed a rate-dependent acoustic modeling scheme, which is able to model within-sentence speech rate variation, and does not rely on RoS estimation prior to recognition and they got a 3.4% (relative) word error rate reduction.

Jing Zheng,Andreas Stolcke[2] et .al. investigated several variants of speech-rate-dependent acoustic models for large-vocabulary conversational speech recognition, in the framework of combining rate-specific models in decoding to compensate for speech rate variation.

They studied two basic approaches to combining rate-specific models:(1) One combines models at the pronunciation level . (2) At the HMM state level.

The key idea behind this approach is to combine rate-specific acoustic models at the pronunciation level, and let the recognizer select the best matching models based on the maximum likelihood criterion during decoding.

The main requirement in this approach is that all phones in a word instance to belong to the same rate class. They conducted different experiments consists of

different number of speech rate classes and different degree of parameters sharing between classes.

The following are the conclusions that are drawn from above methods

(1) Training acoustic models conditioned on speaking rate helps reduce the word error rate of our ASR system at a level of 2% relative.

(2) Word-level combination is a better way of utilizing rate dependent models than state-level combination.

(3)Allowing models for different rate classes to share Gaussian is better than keeping them disjoint.

C.J. van Heerden and E. Barnard et .al.[3] proposed a novel approach to speech rate normalization is presented. Models are constructed to model the word speech rate variation of a specific speaker influences the duration of phonemes.

They studied refinement of predicted phone duration. In this one they used General Linear model to effect the speech rate on specific phoneme.

Speech rate variability has been found to be significant in increasing the error rate of speech recognition, especially when it deviates greatly from the training data.

GLM is their new normalization technique they proposed in this paper.

Tests on the YOHO corpus have confirmed that speech rate normalization can improve the robustness and accuracy of speech recognition also have benefit from speech rate normalization.

Jacob A.C et.al analyze the effect of speech rate variation on Afrikaans phone stability from an acoustic perspective. They introduced two techniques for the acoustic analysis of speech rate variation. There are Sources of variation include dialect differences, speaker differences, channel effects such as bandwidth and background noise, speaking style and vocabulary used. And rate of speech variation got significant in recent time and the effect of rate of speech in accuracy of speech recognition. They are specifically interested in the effect of speech rate variability on phone acoustics.

They concluded the following facts by experiments.

1. The correlation between phone duration change and acoustic change (due to speech rate variation) is weak.

2. We found that phone acoustics are affected differently for each particular phone in question.

3. There is an interesting relationship between the amount of acoustic change (between slow and fast realization of the same phone) and the difficulty of pronouncing the phone using measures such as place of articulation and rounding.

4. The Bhattacharya distance between single mixture mono phones of slow and fast speech provides some indication of the expected phone confusability as measured

by a speech recognition system.

Matthew A. Siegler et.al.[2]comcluded that It is well known that a higher-than-normal speech rate will cause the rate of recognition errors in large vocabulary automatic speech recognition (ASR) systems to increase this paper they attempt to identify and correct for errors due to fast speech. They proposed some methods and measures of speech rate which can reduce the errors of speech recognition. They identified that phone rate is meaningful measure than word rate and when data is clustered as per phone rate recognition error are reduced to some extent.

They proposed three methods. The first method is an implementation of Baum-Welch codebook adaptation. The second method is based on the adaptation of HMM state-transition-probabilities. Third method, the pronunciation dictionaries are modified using rule-based techniques and compound words are added. They improved the recognition accuracy for each method using data sets clustered according to the phone rate metric.

Results: the second method that is HMM state transition probabilities given a indeed effect recognition it reduced error rate for fast speech by relative 4 to 6%.

conclusions: modification in acoustic model can improve the accuracy of fast speech recognition. Applying schwa modifications like deletion between consonants etc. can also improve recognition of fast speech. Intra and inter word variations also used to improve fast speech recognition.

CHEN YiNing , ZHU Xuan , LIU Jia and LIU RunSheng et.al[7] they worked on duration modeling. They concluded that many insert errors occur due to the influence of non-consonant syllables. Introducing the duration model into the recognition process is a direct way to lessen these errors. But that usually could not work well as expected, for the duration is sensitive to speech rate. Hence, aiming at this problem, a novel context dependent duration distribution normalized by speech rate is proposed in this paper and applied to a speech recognition system based on the frame of improved Hidden Markov Model (HMM). They proposed a new method to estimate the speech rate of a sentence; then compute the duration probability combined with speech rate; and finally implement this duration information in the post-processing stage . The experimental results indicate that the syllable error rates decrease significantly in two different speech corpora. Especially for the insertions, the error rates reduce about sixty to eighty percent. It is clear that the insert errors increase while the speech rate rises.

2.2.Review chapter:

Based on the papers referred for this project concluded that the following are the main points to improve accuracy in speech recognition system for rapidly spoken fast Telugu speech

- Better to use phone rate instead of word rate as measure of speech rate.

- found that vowel durations are most sensitive to speech rate so ..it is better to use average vowel rate.

- code book adaptation method may improve accuracy.

- Combination of separate phone dependent and rate dependent code books may improve speech recognition accuracy.

- HMM state transition probability adaptation method can reduce error rates for fast speech by relative amount of 4 % to 6%

- Inter and intra word transformations may increase accuracy i.e by adding compound words to dictionary

- Eliminations of schwa between two consonants and eliminations of non initial non final schwa may be useful modifying dictionary.

- Using bhattacharya distance between single mixture mono phones of slow and fast provides some expected phone confusability as measured by speech

recognition system.

- Analysis of speaker correlation matrices.

- Rate dependent acoustic model within sentence will give 3.5 relative word error rate reduction in hub-5 telephone system.

- Training acoustic models conditioned on speaking rate helps to reduce WER

- Word level combination is better way to utilizing rate dependent model than state level combination

- Allowing models of different rate classes to share Gaussian is better than keeping them disjoint.

- Using decision tree clustering can tie parameters in more seamless way .

- Cross word modeling of speech rate variation could be accomplished using language model.

These are the points which are important to improve the recognition accuracy for

fast Telugu speech.

3. Speech Recognition System

A system or software capable of recognizing spoken language is known as speech recognition system. The machine or software may take the spoken language and translate it into written text, or follow the spoken instructions to perform other functions.

3.1 Speech Recognition Techniques:

There are three major types of speech recognition techniques.

First, the acoustic-phonetic approach assumes that the phonetic units are broadly characterized by the set of features, such as formant frequencies, voiced/unvoiced, and pitch. These features are extracted from the speech signal and are used to segment and label the speech.

Second, the pattern recognition approach requires no explicit knowledge of speech. This approach has two steps- namely, one training of speech patterns based on some generic spectral parameter set and another recognition of patterns via pattern comparison. The popular pattern recognition techniques include template matching, Hidden Markov Model (HMM). Intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features.

Expert systems are used widely in this approach.

3.2. Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are based on the fact that one of the difficulties of ASR is the ability to determine when a speaker starts and finishes an utterance. Most packages can fit into more than one class, depending on which mode they're using.

3.2.1 Isolated Words Recognition

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on BOTH sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

3.2.2 Connected Words Recognition

Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

3.2.3 Continuous Speech Recognition

Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation.

3.2.4 Spontaneous Speech Recognition

There appears to be a variety of definitions for what spontaneous speech actually is. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

3.3 Speech Recognition System:

SPHINX is the one of the best and most versatile Recognition systems in the world today. SPHINX-3.6 is a large vocabulary, speaker-independent, Hidden Markov Model(HMM)-based continuous speech recognition system like its predecessors, the original SPHINX system and SPHINX-2.0.

This SPHINX 3.6 developed at CMU in 1988 and it was one of the systems which

demonstrate the feasibility of accurate, speaker-independent, large vocabulary continuous speech recognition. This sphinx trainer consists of a set of programs which have been compiled for two operating systems: LINUX and ALPHA. This project uses the SPHINX-3.6 Recognition system.

SPHINX-3.6 has a more flexible structure. The user can determine whether it is continuous or semi-continuous modes to be used and the number of streams of data the system will be used and the organization of these streams. SPHINX-3.6 includes both an acoustic trainer and various decoders i.e text recognition, phoneme recognition, N-best list generation, etc.

3.3.1 Signal Processing:

Speech can be represented in terms of its message content, or information. Alternative way of characterizing speech in terms of the signal carrying the message information. That is the acoustic waveform. All the Speech recognition systems use a parametric representation of speech instead of the waveform itself which is based as the pattern recognition. These parameters carry the information related to the short-time spectrum of the signal. SPHINX-3.6 uses the mel-frequency cepstral coefficients (MFCC) as static features for speech recognition.

3.3.2 Hidden Markov Models:

Hidden Markov Models (HMM) have become the well known and widely used statistical approach to characterizing the spectral properties of frames of speech. HMM

as a stochastic modeling tool have an advantage of providing the high reliable and natural way of recognizing the speech in variety of speech based applications. HMM integrates into the systems involving information related to acoustics and syntax, currently it is predominant approach for the speech recognition.

HMM's provides a method of directly estimating the conditional probability of an observation sequence given a hypothesized identity for the sequence. An HMM is trained using example sequences and can be thought of as the extension of the Gaussian model into the temporal dimension.

HMM consists of two processes namely Hidden and the Observed process. The Hidden process consists of a collection of states connected by the transactions. And each of these transactions is described by two sets of probabilities:

Algorithms involved in making HMM's work. These are

(2) Estimating a conditional probability, calculates the probability of the input sequence given a model.

(3) Find the best path through the model, That is, find the path, which most closely matches the input sequence. This enables us to assign states to input vectors for training.

(4) Train a model, Estimate the Gaussian parameters (means and co variances) and transition probabilities to best account for a data set.

In building a recognizer with HMM we need to decide what sequences will correspond to what models. Each utterance could be considered as the HMM. One HMM for each digit in a digit recognition task. To recognize an utterance, the probability metric according to each model is computed and the models with the best fit to the utterance is chosen. To be able to recognize more than one word we need to construct models for each word. Rather than many separate models it is better to construct a network of phonemes models and have paths through the network indicate the words that are recognized[7].

A Transition probability provides the probability of making a transition from one state to another.

(i) An **output probability** density function, defines the conditional probability of observing a set of speech features when a specific transition takes place.

The goal of the decoding or recognition process in HMM is to determine the sequence of (hidden) states (transitions) that has been done in observed signal. The second goal is to define the probability of observing the particular given state of event that has been determined in the first process .

The definition of HMM ,there are three problems of interest:

The Evaluation Problem: The forward-backward algorithm is used for the finding the probability that the model generated the observations for a given model and a sequence of observations .

The Decoding Problem: The Viterbi algorithm can be found the most likely state

sequence in the model that produced the observation for a given model and the sequence of observations.

The Learning Problem: The Baum-Welch algorithm(or the forward-backward algorithm) find the model's parameters so that the maximum probability of generating the observations for a given model and a sequence of observations.

Limitations Of HMMs:

Despite their state-of-the-art performance, HMMs are are handicapped by several well-known weaknesses, namely:

- **The First-Order Assumption:** Which says that all probabilities depend solely on the current state- is false for speech applications. One consequence is that HMMs have difficulty modeling co-articulation. Because acoustic distributions are in fact strongly affected by recent state history. Another consequence is that durations are modeled inaccurately by an exponentially decaying distribution, rather than by a more accurate Poisson or other bell-shaped distribution.
- **The Independence Assumption:** Which says that there is no correlation between adjacent input frames- is also false for speech applications. In accordance with this assumption, HMMs examine only one frame of speech at a time. In order to benefit from the context of neighboring frames, HMMs must absorb those frames into the current frame.
- The HMM probability density models (discrete, continuous and semi-continuous) have suboptimal modeling accuracy. Specifically, discrete density HMMs suffer

from quantization errors, while continuous or semi-continuous density HMMs suffer from model mismatch, that is a poor match between their a priori choice of statistical model (e.g. a mixture of K Gaussians) and the true density of acoustic space.

- The Maximum Likelihood training criterion leads to poor discrimination between the acoustic models (given limited training data and correspondingly limited models). Discrimination can be proved using the Maximum Mutual Information training criterion, but this is more complex and difficult to implement properly.

3.3.3. Recognition Unit:

HMM is used to model the specific unit of speech .This Specific unit may be word, a sub word, or a complete form of sentence. For a large-vocabulary systems , HMM model the sub word units i.e. phonemes. For the small-vocabulary systems , it is used to model the word itself.

The amount of training data and storage required for word models is enormous, thats why SPHINX-III based on phonetic models. However it is inadequate to capture the variability of acoustical behaviors for the given phoneme in different contexts, for that particular contexts it will be modeled using triphones.

Training using sphinx-3.6:

The training procedure involves optimizing HMM parameters given training data. An iterative procedure, the Baum-Welch or forward-backward algorithm is employed to estimate transition probabilities, output distributions, and codebook means and variances

under the probabilistic framework.

Recognition using sphinx-3.6:

For large-vocabulary tasks in the continuous speech recognition, the search algorithm should involved the concepts of acoustic and linguistics in order to maximize the accuracy of the recognition. To apply all these, SPHINX-3.6 uses the Viterbi algorithm. The SPHINX-3.6 decoder is designed in such a way that it incorporates all the available acoustic and linguistic concepts in several phases. In initial stage, Viterbi beam search produces a single recognition hypothesis as well as a word lattice that includes word segmentations and acoustic scores.

The word lattice is then transformed into a directed acyclic graph(DAG) This DAGs are for quick search for the best hypothesis. DAGs are also used to generate N-best lists for re scoring empirically optimized parameters like the language weight and insertion penalty. This speech recognition system uses the process of learning the set of sound units which is called as the Training. The process of using the knowledge acquired to deduce the most probable sequence of units in the given signal is termed as the Decoding or simply Recognition. SPHINX system has the SPHINX trainer and the SPHINX decoder.

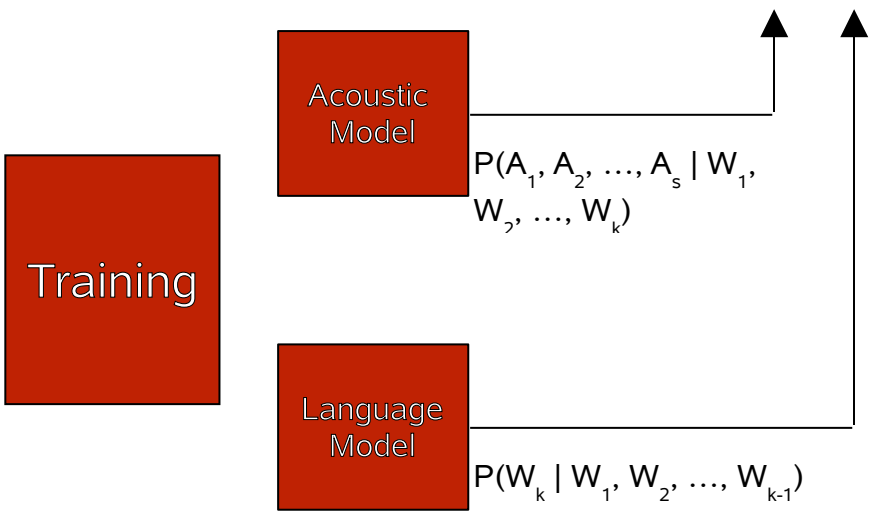

Figure 3.1: Training

Probabilistic Formulation:

Let A = {A1, A2, …, At} be a sequence of acoustic observations

Let W = {W1, W2, …, Wm} be sequence of words

Given the acoustic observations A, the probability of word sequence W.

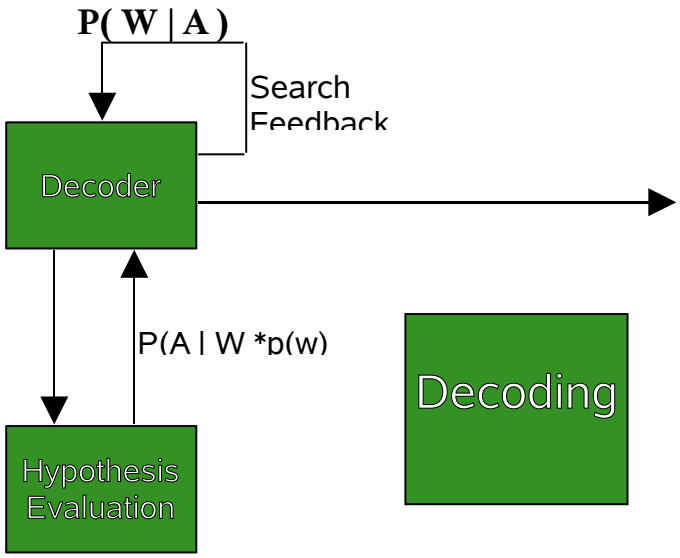

Figure 3.2: decoding

For recognition, we choose

Best = argmaxW(P(W | A)).....Eq 3.1

Bayes Rule:

argmaxW (P(W | A)) = argmaxw (P(W, A) / P(A)).....Eq 3.2

= argmaxW(P(W, A) .....Eq 3.3

= argmaxW(P(A | W) * P(W)).....Eq 3.4

A model for the probability of acoustic observations given the word sequence, P(A | W), is called an "acoustic model." A model for the probability of word sequences, P(W), is called a "language model."

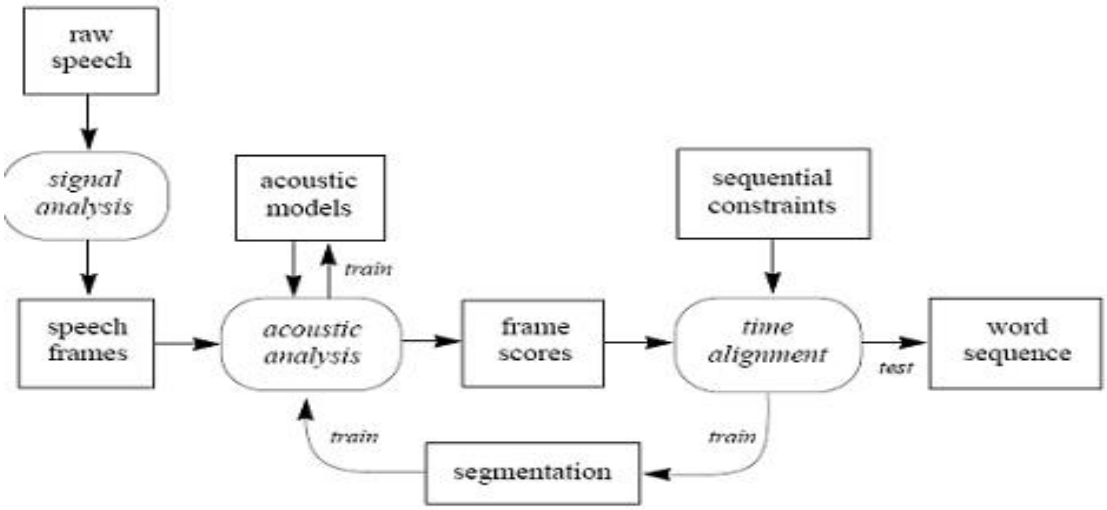

Figure 3.3: Block Diagram of the speech Recognition System

Search Problem:

Find the sequence WBest = { W1, W2, …, Wm }.....Eq 3.5

= argmaxW(P(A | W) * P(W)).....Eq 3.6

This is a search of a space of Vm word sequences, where 'V' is the vocabulary size (which may be 100,000 words or larger), and 'm' is the sentence length (which may be 20 words long or longer).

Clearly, an exhaustive search is impossible. It is necessary to have a structured, intelligent search.

The components provided for the trainer:

(i). The trainer executables

(ii). The Acoustic signals

(iii). The corresponding transcription file

(iv). A language dictionary

(v) A filler dictionary

The components provided for decoding:

(i). The decoder executable

(ii). The language dictionary

(iii).The filler dictionary

(iv).The language model

3.4 Evaluation of Speech Recognition Accuracy:

Word Error Rate (WER) is a common metric of the performance of a speech recognition system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level[9].

This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment.

Word error rate can then be computed as:

WER=(S+I+D)/N .....Eq 3.7

where

- *S* is the number of substitutions,
- *D* is the number of the deletions,
- *I* is the number of the insertions,
- *N* is the number of words in the reference.

When reporting the performance of a speech recognition system, sometimes **word recognition rate (WRR)** is used instead:

WRR=1-WER=(N-S-D-I)/N=(H-I)/N.....Eq 3.8where

(3) *H* is N-(S+D), the number of correctly recognized words.

4. Implementation

In the implementation phase, the Designs are translated into code. Computer programs are written using a conventional programming language or an Application Generator. Programming tools like Compilers, Interpreters and Debuggers are are used to generate the code. Different level programming languages like C, C++, Java, are used for coding. With respect to the type of application, the right programming language is chosen.

In this project as part of implementation we have conducted several experiments by CMU's sphinx-3.6 tool and we used the praat script. These experiments shows that how this RoS effects the speech recognition system and speech recognition accuracy. In the implementation apart from sphinx-3.6 tool we use praat script for changimg the duration of the speech signal.

4.1. SPEECH DATABASE:

Training and Testing of a Speech recognition system needs a collection of utterances forthe appropriate task. Data Collection: For implementing Speech recognition using HMM with Sphinx-3.6, Telugu speech database is recorded from different males and females using noise canceling close talking microphone. The

collected Telugu speech consists of the following.

Collected three different variants of speech

like fast, normal and slow speech.

This database consists of 40 sentences and 230 words of Telugu speech. The system is trained by 1200 of sentences recorded from 40 speakers.

Audio Properties: As sphinx supports the speech signal should be PCM 16 bit Mono, the speech signal in the data will be 16 bit Mono.

4.2. SPHINX-3.6 PROCEDURE:

4.2.1. Before Training:

1. Sphinx downloaded from http://www.cs.cmu.edu./~robust/Tutorial and download singlemachine.tar.gz or download from CD.

2. Unzip downloaded file using the command tar -zxvf singlemachine.tar.gz.

This will give the directory called TUTORIAL in this base directory it will install all the files necessary to train and test the Sphinx.

If you download from CD, you will get a sample trained directory called railway which is useful for execution of other files.

3. In TUTORIAL create new experiment by giving the command

./create_newexpt.csh file1 [file name]

(4) In this file all the files required to train and decode the system will be loaded. Copy the util directory from the Sphinx3 to this file and create another directory in file1 by giving the command

(5) mkdir train.

In this train directory paste the 5 files

create a folder named wav and paste all wav files.

create a folder named raw

To obtain raw files the path of the wav files and the path of the raw files should be given as input to wavtoraw.csh program. Then raw files will be obtained by running the program “wavtoraw.csh” .

Ex: In the way to raw.csh program

set wavedir=/home/sphinx/TUTORIAL/file1/train/wav

set rawdir=/home/sphinx/TUTORIAL/file1/train/raw

file2>./wavtoraw.csh

now you will get all raw files in the raw folder in train.

5. Write the Transcription files for the wavfiles and keep it in train .

Example of transcription file: Giving a single space in the beginning of the text as <s> vowel/consonant/words/sentences then space </s> (name01) Ex: <s> A </s> (akbar01)

If your writing transcription for sentences one should listen carefully and write transcription by checking for silence in between the words.

6. Write the sentence file by removing the names such as name01...

such as <s> A </s> and no space should be there at the end of all files.

7. Dictionary file, this will given by the CMU, by giving the sentence file to the site

www.speech.cs.cmu.edu/tool/lmtool From this we get the following files

sentence file, by right clicking save this file into train by giving name as filename.sent

dictionary file, by right clicking save this file into train by giving name as filename.dic

Language model, by right clicking save this file into train by giving name as filename.lm

phone list, this will get by running the program “dictophn.c”by giving the dictionary file as input to it.

Then compile it using

train>cc dictophn.c

Then execute it using ./a.out

Then u will get phone list at the command prompt and save that list in a new text file and give name as filename.phonelist

4.2.2. TRAINING:

1. Create the control raw file using the command

file1# ls /home/sphinx/TUTORIAL/file1/train/raw/*.raw>file.ctl

2. Copy compute_mfcc.csh file from c_scripts in the railway folder into the current working c_scripts(replace it). Run the file script in c_scripts by the command

c_scripts#] ./compute_mfcc.csh /home/sphinx/TUTORIAL/file1/train/file.ctl

3. Mfcc files are created in the folder 'raw' in the feature_files.

4. Create the control mfc file using the command

file1#] ls /home/sphinx/TUTORIAL/file1/feature_files/raw/*.mfc>mfc.ctl

5. Change the paths of 5 files in the file called variables.def contained in the c_scripts as Go to c_scripts file, change the command for all the scripts contained in the directories 01* through 05*, by keeping unlimit as #unlimit.

6. Then run the slave*.csh script(in all starting from 01* through 05*). If the input is error free, then we get number of files by running each script. The errors will be checked in the log directory contained in file1 directory.

4.3. ERRORS AND SOLUTIONS:

1. Check the paths correctly in the variables.def contained in the c_scripts.

2. There should not be any space at the end of each file in transcription file.

3. Eliminate duplication of phonemes in phone list.

4.If, input string is too large, truncate it.

5.If an error comes like “reading mfcc failed”, then remove that mfcc file and recompute mfcc for that particular wave.

4.3.1. DECODING:

1. Go to decoding directory in file1,

cd decoding/

then give the command

./launch_decode.cd.8gaumodels

2. Copy lm3g2dmp from any directory in sphinx software.

3. Weget lm3g2dmp(dump) file by giving source as .lm file path and destination as ur train then u have to write in this way

./lm3g2dmp /home/sphinx/TUTORIAL/file1/train/filename.lm

/home/sphinx/TUTORIAL/file1/train/.

4.3.2. WORD ERROR RATE:

1. Create 2 text file hyp and org in train directory. Copy the recognized text in the hyp and original text into org file, then run the program ./compute.csh contained in the decoding#] ./compute_acc.csh /home/sphinx/TUTORIAL/file1/train/org /home/sphinx/TUTORIAL/file1/train/hyp

- Then you will get output word error rate like this

WORD ACCURACY= 94.783% (109/ 115) ERRORS= 5.217% (6/ 115).

4.4. Method for changing the duration:

Normally different speakers will

use different rate of speech in their normal speech. speech rate varies from person to person. There is duration gap from fast spoken person to slow spoken person i.e there is a difference between two speakers in terms of time duration of the speaking words. In

this method mainly we changed the fast speech signal duration to threshold value and slow speech to threshold value because when people talk faster than normal, the performance of state of the art Large Vocabulary Continuous Speech Recognition (LVCSR) systems is often poor with error rates for fast speech. The principle is to normalize the phone duration by stretching the length of the utterances in order to match the "threshold" lengths of the phonetic units which are obtaining from the training corpus.

For this one we used the praat script. In this we used the standard threshold value for duration from training corpus.

4.4.1. Code to change the duration:

procedure:

```
# This script will globally change the duration and pitch of all the the selected sound by the factors given in the dialogue box.
# If the quality is bad try changing the minimum pitch or maximum pitch

form Change to Fo and duration
sentence Fo_expression self*1.0
positive Duration_factor 1.0
 comment Analysis parameters for Fo
  positive minimum_Fo 75
```

```
   positive maximum_Fo 300
   boolean Play_after_synthesis 1
   boolean Delete_Manipulation_file 1
endform
fomin = 'minimum_Fo'
fomax = 'maximum_Fo'
#find out how many Sounds have been selected
numberOfSounds = numberOfSelected  ("Sound")
#set up arrays with names and IDs of selected Sounds
#this is necessary because vtchange changes the selections
for ifile from 1 to numberOfSounds
   sound$ = selected$  ("Sound", 'ifile' )
   soundID = selected  ("Sound", 'ifile')
  ids'ifile' = soundID
   names'ifile'$ = sound$
endfor

#now go to to ,main part
for ifile from 1 to numberOfSounds
   soundID = ids'ifile'
sound$ = names'ifile'$
```

```
   call fodurnchange
endfor
#procedure fodurnchange
#initial analysis locating pitch pulses etc
select 'soundID'
durn = Get duration
#create a Pitch object
select Sound 'sound$'
To Pitch... 0.01 fomin fomax
plus Sound 'sound$'
To Manipulation
select Pitch 'sound$'
#apply the appropriate transformation to the Pitch object
Formula... 'fo_expression$'
#turn it into a PitchTier and place it into the Analysis object
Down to PitchTier
select Manipulation 'sound$'

plus PitchTier 'sound$'
Replace pitch tier
select Pitch 'sound$'
```

```
plus PitchTier 'sound$'
Remove
if duration_factor <> 1.0
	Create DurationTier... 'sound$' 0 'durn'
	Add point... 0 'duration_factor'
	select Manipulation 'sound$'
	plus DurationTier 'sound$'
	Replace duration tier
	select DurationTier 'sound$'
	Remove
endif
#resynthesise with new pitch and duration contour
select Manipulation 'sound$'
Get resynthesis (PSOLA)
if play_after_synthesis = 1
  Play
endif
Rename... 'sound$'.f'fo_expression$'.d'duration_factor'
name$ = selected$("Sound", -1)
if delete_Manipulation_file = 1
  select Manipulation 'sound$'
```

Remove

endif

select Sound 'name$'

endproc

5. RESULTS

Results :

Speech Recognition results for Telugu fast speech is obtained by conducting experiments on recorded data. The experiments are conducted on individual recorded sentences and for overall sentences i.e 1200 sentences . The results are tabulated according to the experiments.

Table1: The total sentences word recognition performance.

S.No	Type of error	% of error
1	correct	57.07
2	Insertion	7.8
3	deletion	15.6
4	substitution	27.4

Boosting is a method based on a set of weak samples to create a single strong sample. boosting applied to acoustic model training in various tasks in large vocabulary automatic speech recognition. Specifically, in this present work, by applying the AdaBoost.M2 algorithm – the training and test samples are repeatedly taken as 2 times, 4 times the level of utterances to maximize Recognition accuracy.

Speech Recognition results for Telugu fast speech is obtained for 1200 sentences collected from 40 people. taking training data as two times for unrecognized sentences,applying the repeated training with more number of iterations till the accuracy

converges.

The experiments are conducted for speaker dependent models as well as speaker independent model. Speaker dependent system are giving good results as compare with speaker dependent model. The results for different speakers as experiments are conducted for individual speakers are as follows.

Table 2: The first speakers results

S.No	Type of the error	% of the error
1	correct	97.4
2	Substitution	2.6
3	Insertion	0.0
4	Deletion	0.0
5	Total errors	2.6

The second persons sentences recognition performance is as follows.

Table 3: The second speaker results

S.No	Type of the Error	% of the Error
1	correct	84..9
2	substitution	9.3
3	Insertion	5.8
4	deletion	5.8
5	Total errors	20.9

The following are the mis recognized word and confusion pair of words .

The mis recognized words

UNTAARU

RENDUVANDALA

VANTIDI

MOKKEIVANGANIDI

THRAGADAM

RAADU

DEBBA

DASALUNDUNU

The confusion pair

UNTAARU	EKKADUNTAARU
RENDUVANDALA	KALADU
VANTIDI	MRAANEI
MOKKEIVANGANIDI	VANGUNAA
THRAGADAM	THRAGADAMKANNAA
RAADU	KUKKAKAATUKU
DEBBA	KANCHAE
DASALUNDUNU	THEMMANTUNDI

Table 4: The Third speaker results

S.No	Type of the Error	% of the Error
1	correct	82.2

2	substitution	14..4
3	Insertion	3.3
4	deletion	8.9
5	Total errors	26.7

The mis recognized word are

VAARANDARAMU

TELUGAE

MEE

PAERU

EKKADUNTAARU

MEEREKKADA

UNTAARU

HYDERAABAD

BHAARATHADESAM

NAA

MAATHRUBHUUMI

PADAVINODAMU

DEBBA

The confusion pair of words are

VAARANDARAMU-- UNTAARU

TELUGAE --PAERU

MEE-- MEERU

PAERU-- PAERAEMITI

EKKADUNTAARU-- UNTAARU

MEEREKKADA --EKKADA

UNTAARU-- EKKADUNTAARU

HYDERAABAD-- PRAANA

BHAARATHADESAM-- UNTAARU

NAA --AEDI

MAATHRUBHUUMI --THANANTHATA

PADAVINODAMU --MAATLADUDAAM

DEBBA --PUSTHAKA

Table 5:The Fourth speaker results are

S.No	Type of the Error	% of the Error
1	correct	63.0
2	substitution	29.0
3	Insertion	8.0
4	deletion	34..1
5	Total errors	71.0

The following are the mis recognized words

MEE

TELUGAE

MEEREKKADA

UNTAARU

MEERAEMI

PINDIKODDIROTTE

THRAGADAM

TELUGU

EKKADA

PITTA

KONCHEM

VAARANDARAMU

MAATLADUDAAM

PAERAEMITI

EKKADUNTAARU

VIDYALAENIVAADU

VINTHAPASUVU

NAAKU

HYDERAABAD

KILOMEETARLA

DUURAMLO

KALADU

PATNAM

NAA

MAATHRUBHUUMI

MOKKEIVANGANIDI

MRAANEI

THRAGADAMKANNAA

DARIKI

RAADU

OKKADAE

MAATHRUDAEVOBHAVA

The following are the confusion pairs

UNTAARU --EKKADUNTAARU

TELUGU-- PILLIKI

TELUGAE --OKKADAE

MEE --MEEREKKADA

MEEREKKADA-- MEERU

EKKADA-- MEERU

MEERAEMI-- AEDI

PITTA --PATNAM

KONCHEM --PALLE

PINDIKODDIROTTE --PRAANA

THRAGADAM --TELUGAE

VAARANDARAMU --RAMMANTUNDI

TELUGAE --ISTAM

MAATLADUDAAM ---BHAARATHADESAM

MEE --PAERU

PAERAEMITI --PAERU

MEE --MEERU

EKKADUNTAARU UNTAARU

MEEREKKADA --EKKADA

MEERAEMI --NEE

VIDYALAENIVAADU-- UNTAARU

VINTHAPASUVU --PUSTHAKA

NAAKU-- UNTAARU

HYDERAABAD-- UNTAARU

KILOMEETARLA --PILLIKI

DUURAMLO --ISTAM

KALADU --OKA

PATNAM --PITTA

PINDIKODDIROTTE --HYDERAABAD

NAA-- UNTAARU

MAATHRUBHUUMI --THEMMANTUNDI

MOKKEIVANGANIDI --OKKADAE

MRAANEI --PRAANA

THRAGADAMKANNAA --CHELAGAATAM

THRAGADAM --CHELAGAATAM

DARIKI --UNTAARU

RAADU-- NAAKU

OKKADAE-- KALADU

MAATHRUDAEVOBHAVA-- BHAARATHADESAM

Table 6:The results for fifth speaker

S.No	Type of the Error	% of the Error
1	correct	57.0
2	substitution	27.4
3	Insertion	7.8
4	deletion	15.6
5	Total errors	50.8

The following are the mis recognized words

ATHADU

MEEREKKADA

KILOMEETARLA

PADAVINODAMU

AEMITI

MEERAEMI

PARUGAETHTHI

KANCHAE

TELUGAE

MAATLADUDAAM

UNTAARU

VELUTHUNNADU

MRAANEI

NAMASKAARAMU

VAARANDARAMU

PAERAEMITI

MEERU

EKKADUNTAARU

VIDYALAENIVAADU

NAAKU

CHAALAA

MAAVUURINUNDI

HYDERAABAD

RENDUVANDALA

DUURAMLO

MAANAVASAEVAYAE

MAADAVASAEVA

MARYAADA

ANAEDI

PALLE

THALLILAANTIDI

PATNAM

PINDIKODDIROTTE

VANGUNAA

VAETAADENU

ANDARUU

PAALU

NILABADI

AEDI

THANANTHATA

THAANEI

The following are the confusion pair of words in recognition

ATHADU

MEEREKKADA

KILOMEETARLA

PADAVINODAMU

AEMITI

MEERAEMI

PARUGAETHTHI

KANCHAE

TELUGAE

MAATLADUDAAM

UNTAARU

VELUTHUNNADU

MRAANEI

NAMASKAARAMU

VAARANDARAMU

PAERAEMITI

MEERU

EKKADUNTAARU

VIDYALAENIVAADU

NAAKU

CHAALAA

MAAVUURINUNDI

HYDERAABAD

RENDUVANDALA

DUURAMLO

MAANAVASAEVAYAE

MAADAVASAEVA

MARYAADA

ANAEDI

PALLE

THALLILAANTIDI

PATNAM

PINDIKODDIROTTE

VANGUNAA

VAETAADENU

ANDARUU

PAALU

NILABADI

AEDI

THANANTHATA

THAANEI

Table 7:The results for the sixth speaker

S.No	Type of the Error	% of the Error
1	correct	88.7
2	substitution	11.3

3	Insertion	16.5
4	deletion	0.0
5	Total errors	27.8

The following are the mis recognized words

MEE

PAERAEMITI

PAERU

EKKADUNTAARU

MEERU

VIDYALAENIVAADU

VELUTHUNNADU

GANAM

ANDARUU

SAMAANULAE

RAADU

MAATHRUDAEVOBHAVA

The following words are confusion pair of words

MEE --GANAM

PAERAEMITI-- AEMITI

MEE --DARIKI

PAERU --RAADU

EKKADUNTAARU-- UNTAARU

MEERU --PILLIKI

VIDYALAENIVAADU --VAARANDARAMU

VELUTHUNNADU --JEEVITHAMLO

GANAM --DABBU

ANDARUU-- NAAKU

SAMAANULAE --PULI

RAADU-- UNTAARU

MAATHRUDAEVOBHAVA --ALAVAATU

Table 8: The results of seventh speaker:

S.No	Type of the Error	% of the Error
1	correct	93.9
2	substitution	5.2
3	Insertion	5.2
4	deletion	0.9
5	Total errors	11.3

The following are the mis recognized words

NAMASKAARAMU

PAERAEMITI

MEEREKKADA

PADAVINODAMU

DARIKI

ALAVAATU

The following are the confusion pair of words in recognition

NAMASKAARAMU --SANKATAM

PAERAEMITI --AEMITI

MEEREKKADA-- EKKADA

PADAVINODAMU --THRAGADAM

DARIKI --EKKADIKI

ALAVAATU-- CHELAGAATAM

Table 9: The results of eighth person

S.No	Type of the Error	% of the Error
1	correct	93.9
2	substitution	6.1
3	Insertion	6.1
4	deletion	0.0
5	Total errors	12.2

The following are the mis recognized words:

MEEREKKADA

MEERAEMI

NAAKU

MOKKEIVANGANIDI

AEDI

DEBBA

MAATHRUDAEVOBHAVA

The following are the confusion pair in recognition

MEEREKKADA ----EKKADA

MEERAEMI-- MEE

NAAKU-- NAA

MOKKEIVANGANIDI --ANAEDI

AEDI PAERU

DEBBA--- DABBU

MAATHRUDAEVOBHAVA -----PITHRUDAEVOBHAVA

Table 10 : The results of ninth person are as follows

S.No	Type of the Error	% of the Error
1	correct	92.2
2	substitution	7.0
3	Insertion	4.3
4	deletion	0.9
5	Total errors	12.2

The following are the mis recognized words

NAMASKAARAMU

PAERAEMITI

MEEREKKADA

MEERAEMI

VIDYALAENIVAADU

NAAKU

HYDERAABAD

PINDIKODDIROTTE

The following are the confusion pair of words in recognition

NAMASKAARAMU ---KALADU

PAERAEMITI -----PRAANA

MEEREKKADA ----EKKADA

MEERAEMI ---NEE

VIDYALAENIVAADU--- DAEVUDU

NAAKU ---PAALU

HYDERAABAD--- ALAVAATU

PINDIKODDIROTTE ---PAERAEMITI

Table 11: The tenth speaker results are as follows

S.No	Type of the Error	% of the Error
1	correct	76.5
2	substitution	16.5
3	Insertion	26.1
4	deletion	7.0
5	Total errors	49.6

The mis recognized words are

UNTAARU

MEERAEMI

EKKADIKI

MAAVUURINUNDI

HYDERAABAD

RENDUVANDALA

KILOMEETARLA

DUURAMLO

PALLE

PADAVINODAMU

MINNA

KANCHAE

EIKAMATYAMAE

THAANEI

SOODINCHI

DEBBA

ALAVAATU

JEEVITHAMLO

MUKYAMGAA

The confusion pair of word in recognition are

UNTAARU ---EKKADUNTAARU

MEERAEMIAN--AEDI

EKKADIKI --EKKADA

MAAVUURINUNDI-- UNTAARU

HYDERAABAD --THAANEI

RENDUVANDALA KALADU---

KILOMEETARLA NEE

DUURAMLO--- DASALUNDUNU

PALLE ---PAALU

PADAVINODAMU--- EIKAMATYAMAE

MINNA ---PRAANA

KANCHAE MANCHI

EIKAMATYAMAE--- THANANTHATA

THAANEI--- KANCHAE

SOODINCHI --SAADINCHAALI

DEBBA --DABBU

ALAVAATU-- CHELAGAATAM

JEEVITHAMLO-- NEERU

MUKYAMGAA --CHEPPU

Table 12: The eleventh speaker results are as follows

S.No	Type of the Error	% of the Error
1	correct	95.7
2	substitution	4.3
3	Insertion	8.7
4	deletion	0.0
5	Total errors	13.0

The following are the mis recognized words

EKKADUNTAARU

MEEREKKADA

MEERAEMI

PINDIKODDIROTTE

PARUGAETHTHI

The following are the confusion pair of words in the recognition

EKKADUNTAARU ---UNTAARU

MEEREKKADA--- EKKADA

MEERAEMI ---MEE

PINDIKODDIROTTE --PAERU

PARUGAETHTHI ---OKKADAE

Table 13: The twelve th speaker results are as follows

S.No	Type of the Error	% of the Error
1	correct	84..3
2	substitution	15.5
3	Insertion	19.1
4	deletion	0.0
5	Total errors	34..8

The following are the mis recognized words

PAERAEMITI

MEEREKKADA

MEERAEMI

NAAKU

ATHADU

MAAVUURINUNDI

HYDERAABAD

RENDUVANDALA

KILOMEETARLA

PINDIKODDIROTTE

MRAANEI

VANGUNAA

PADAVINODAMU

PARUGAETHTHI

NILABADI

AEDI

DARIKI

RAADU

The following are the confusion pair of words in recognition

PAERAEMITI ---EKKADIKI

MEEREKKADA ---EKKADA

MEERAEMI ---AEDI

NAAKU ---NAA

ATHADU ---BHAARATHADESAM

MAAVUURINUNDI--- ANDARUU

HYDERAABAD--- CHAALAA

RENDUVANDALA ---MEE

KILOMEETARLA ---KALADU

PINDIKODDIROTTE --KUUTHA

MRAANEI ---VAETAADENU

VANGUNAA ---NAA

PADAVINODAMU--- VAARANDARAMU

PARUGAETHTHI--- DARIKI

NILABADI--- DARIKI

AEDI ---AEMITI

DARIKI ---EKKADIKI

RAADU--- NAAKU

Table 14: thirteenth speaker results are as follows

S.No	Type of the Error	% of the Error
1	correct	94..8
2	substitution	4.3
3	Insertion	2.6
4	deletion	0.9
5	Total errors	7.8

The following words are mis recognized words:

TELUGAE

EKKADA

ATHADU

PITTA

PINDIKODDIROTTE

The following are the confusion pair of words in recognition

TELUGAE--- PAERAEMITI

EKKADA--- EKKADUNTAARU

ATHADU ---THRAGADAM

PITTA ---PATNAM

PINDIKODDIROTTE ---HYDERAABAD

Table 15: The fourteenth speaker results are as follows

S.No	Type of the Error	% of the Error
1	correct	94..8
2	substitution	4.3
3	Insertion	8.7
4	deletion	0.9
5	Total errors	13.9

The following are the mis recognized words:

MEEREKKADA

MEERAEMI

VAETAADENU

DEBBA

PITHRUDAEVOBHAVA

The following words are confusion pair of words in recognition

MEEREKKADA ---EKKADA

MEERAEMI ---PAERAEMITI

VAETAADENU ---VANTIDI

DEBBA ---PAALU

PITHRUDAEVOBHAVA ---UNTAARU

Table 16: fifteenth speaker results are as follows

S.No	Type of the Error	% of the Error
1	correct	87.0
2	substitution	13.0
3	Insertion	11.3
4	deletion	0.0
5	Total errors	24..3

The following are the mis recognized words:

NAMASKAARAMU

MEE

PAERU

AEMITI

MEEREKKADA

MEERAEMI

VANTIDI

PRIYURAALULAANTIDI

PINDIKODDIROTTE

JINKANU

NILABADI

RAADU

SAADINCHAALI

JABBU

OKKADAE

The following words are the confusion pair of words in recognition

NAMASKAARAMU ---PITTA

MEE--- AEMITI

PAERU--- DABBU

AEMITI ---PAERAEMITI

MEEREKKADA ---EKKADA

MEERAEMI--- MEE

VANTIDI ANAEDI

PRIYURAALULAANTIDI ---ANAEDI

PINDIKODDIROTTE ---DARIKI

JINKANU ---CHAENUMESTHE

NILABADI--- DARIKI

RAADU ---VAETAADENU

SAADINCHAALI--- SANKATAM

JABBU--- CHAALAA

OKKADAE--- PINDIKODDIROTTE

Table17: The total sentences word recognition performance before speech rate is not changed

S.No	Type of error	% of error
1	correct	57.07
2	Insertion	7.8
3	deletion	15.6
4	substitution	27.4

Table 18:The total sentences word recognition performance after speech rate is changed

S.No	Type of the Error	% of the Error
1	correct	62.1
2	substitution	32.0
3	Insertion	21.7
4	deletion	5.8
5	Total errors	59.6

5.1 Summary of the total results:

In the results mainly the following observations are made

1.Speaker dependent recognition will give more accuracy even for fast speech.

2.Speaker independent recognition swill not give more accuracy for fast speech.

3.Word recognition performance after the speech rate is changed is improved when compare with with out changing the speech rate.

4. the words that are end with vowels are the very sensitive to in recognition and give more errors

5.boosting and bagging technique is giving the good results in recognition of un recognized words and given good results.

6. Conclusion and future work

6.1Conclusion:

Experimental results shows that , speech recognition accuracy using SPHINX-III with HMMs by applying boosting technique for utterances given more accuracy for Telugu speech. The fast speech recognition accuracy is low when compare with normal speech but after changing the durations of the fast speech signal the results are better than the previous results. Boosting and bagging technique and durations changing and other techniques are giving good results with fast Telugu speech.

6.2 Future Directions:

The work we have done is based on Speech recognition method using HMM is giving good results. It is observed that usually the recognition accuracy is high for speaker dependent model and low for fast speech .In this work we apply boosting algorithm and speech signal duration changing method and these techniques are useful in increasing the accuracy .In this work we use only HMM based recognition method and it is recommended that HMM/SVM hybrid approach may be considered as a future work to improve the recognition accuracy and increasing the number of iterations in training process may be another approach for the future work.

6.3 References

[1].J. Zheng, H. Franco, and A. Stolcke. Rate-of-speech modeling for large vocabulary conversational speech recognition. In Proc. ISCA Tutorial and Research Workshop on Automatic Speech Recognition: Challenges for the new Millenium,
pages 145–149, Paris, 2000. ISCA.

[2]. M.A. Siegler and Richard M. Stern, “On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems,” Proc. ICASSP’95, pp. 612-615, 1995

[3].J. Zheng, H. Franco, and A. Stolcke, "Effective acoustic modeling for rate-of-speech variation in large vocabulary conversational speech recognition," in Proc. Intl. Conf. Spoken Language Processing, (Jeju, Korea), pp. 401--404, October 2004.

[4].Badenhorst, JAC and Davel, MH. 2007. Effect of speech rate variation on acoustic phone stability in Afrikaans speech recognition. PRASA 2007: Eighteenth Annual Symposium of the Pattern Recognition Association of South Africa, Pietermaritzburg, Kwazulu-Natal, South Africa, 28-30 November, pp 6

[5].J. Stadermann and G. Rigoll, “A hybrid SVM/HMM acoustic modeling approach to automatic speech recognition,” in INTERSPEECH-2004 ICSLP, 2004, pp. 661–664.

[6]. N. Mirghafori, E. Fosler and N. Morgan, “Towards Robustness to Fast Speech in ASR,” Proc. ICASSP’96, pp. I335-338, 1996

[7] Yining Chen, Xuan Zhu, Jia Liu, Runsheng Liu. Towards Robustness to Speech Rate in Mandarin All-Syllable Recognition. J. Comput. Sci. Technol., 2003: 756~761

[8] CMU Sphinx website:
http://www.speech.cs.cmu.edu/sphinxman/fr4.html

[9] Spoken Language Processing: A Guide To Theory, Algorithm And System Development

Author: Xuedong Huang, Raj reddy, Alex acero

ISBN:0130226165,ISBN-13:9780130226167,978-0130226167,Binding: Paperback

Publishing Date: Apr 2001,Publisher: Prentice Hall,Number of Pages: 1008,Language: English

[10] http://www.cs.cmu.edu/~robust/Tutorial/

[11] http://www.speech.cs.cmu.edu/tools/lmtool.html