Lexical Modelling in HMM based Telugu language Automatic Speech Recognition(ASR) system

A thesis submitted during 2018 to the University of Hyderabad in partial fulfillment of the award of a Ph.D. degree in Department of Computer and Information Sciences

by

Nagamani Molakatala

School of Computer & Information Sciences

University of Hyderabad
P.O. Central University, Gachibowli
Hyderabad – 500046
Telangana, India

December 2018

CERTIFICATE

This is to certify that the thesis entitled **"Lexical Modelling in HMM based Telugu language Automatic Speech Recognition(ASR) system"** submitted by **Nagamani Molakatala** bearing **Reg. No. 03MCPC04** in partial fulfillment of the requirements for the award of **Doctor of Philosophy** in **Computer Science** is a bonafide work carried out by her under my supervision and guidance.

The matter relate to plagiarism is not applicable in my case as my registration is of the year 2003 how ever this thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for award of any degree or diploma

The student has the following select publications before submission of the thesis for adjudication and has produced evidence for the same.

1. "Telugu Speech Interface Machine for University Information System", Proc. Of IEEE Advanced Computing and Communications (ADCOM), Ahmadabad, December 2004. (scopus i.e IEEE explorer)

2. "Intelligent Tutor for Telugu Language Learning – INTTELL"., Proc. Of International Conference on Image, Signal and Information Processing(ICISIP2005), January 4-7, 2005. (Scopus i.e. IEEE explorer)

3. "Implementing Phoneme Based Segmentation Algorithms to ASR system" International Conference on Advanced Computing methodologies(ICACM-2011), ISBN:978-93-81269-40-4, Reed Elsevier India private Ltd, December 2011,Hyderabad.

4. "Substitution error analysis for improving the word Accuracy in Telugu Language Automatic Speech Recognition system" IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 4 (July-Aug. 2012), PP 07-10 www.iosrjournals.org.

5. "Lexicon design for Telugu Language Automatic Speech Recognition System", 2nd International Telugu Conference (ITC 2012), 2-4th November 2012, Vizag.(Presented Paper in Telugu langauge).

6. "Environmental noise analysis for robust automatics speech recognition, Chapter Advanced Computer and Communication Engineering Technology Volume 315 of the series Lecture Notes in Electrical Engineering , 02 November 2014.

7. "Voice controlled Music Player Using Speech Recognition System" IJEEE, Issue2, Feb-Mar2016.

8. Exploring the Speech Prosody Manipulation for Dual-Language TTS System" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 4, Ver. III (Jul - Aug 2018), PP 20-22 www.iosrjournals.org DOI: 10.9790/0661-2004032022.

Further the student has passed the requisite courses towards fulfillment of course work requirement for Ph.D.

Advisor
prof. Atul Negi
School of CIS
Information Sciences
University of Hyderabad
Hyderabad – 500046, India

Dean
prof. Kavi Narayana Murthy
School of CIS
Information Sciences
University of Hyderabad
Hyderabad – 500046, India

DECLARATION

I, **Nagamani Molakatala**, hereby declare that this thesis entitled **"Lexical Modelling in HMM based Telugu language Automatic Speech Recognition (ASR) system"** submitted by me under the guidance and supervision of **Dr. Atul Negi** is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Date : **Nagamani Molakatala**

Signature of the Student
Reg. No.: 03MCPC04

Abstract

In modern times, the speech interface by way of conversion of speech to text is gaining prominence in Man-Machine interaction. The thesis is aimed at conversion of speech into text by way of Automatic Speech Recognition (ASR) system. The ASR design needs three statistical models VIZ., (i) Acoustic model (AM) for mapping the signal information to symbols using Hidden Morkov Model (HMM) (ii) Pronunciation model (PM) that maps word to symbol set and (iii) Language Model (LM) that provides prior probabilities of word sequence using linguistic knowledge. The PM is an input, which provides the interface between the outcomes of the statistical probability estimations (AM, LM). ASR helps to compile final hypothesis as system generated word sequences. It plays an important role in conversion of linguistic language to system knowledge by use of computational methods like HMM, which is core technology to develop ASR system..

Focus of thesis work is in adaptation of PM, also known as lexical model for Indian Languages. They are transcribed using phonemic symbol sequence in context of Telugu language phonemes. PM is input to the phonetic engine, to build the HMMs for given speech data by mapping words with phonetics. Crucial task is to construct sufficient training data for the development system by choosing suitable symbols to map the phonemes for building PM. It is proposed to build phone set of Telugu pronunciation phone set, say University of Hyderabad (UOH) phone set which is similar to Carnegie Mellon University (CMU) bench mark phone set for American ascent (English). It is proposed to create the Telugu language speech bench mark data set for the purpose..

Most Indian languages are syllable timed languages with one to one mapping from speech to written form. ASR system performance is complicated due

to variation in speech. In order to improve this, an adaptation is necessary in various models for standardization. Empirical speech data is recorded in quite rooms. Then manually transcribe data and compare the results with well established PM to that of the handcrafted and semi automated transcribed PM. This is done by iterating (i) training and (ii) testing of ASR system for the data. This thesis is focused on PM for both data driven and knowledge driven methods in respect of (i) Speaker Dependent (SD) and Speaker Independent (SI), (ii) vocabulary (iii) recording environments (iv) variations in dialects.

Compared with various data sets, it is observed that UOH phone set based PM, gives better performance, than other bench mark lexicons based on CMU phone set. For illustration, for isolated words, there is 10 to 25 % performance improvement in evaluation data set. With the use of UOH phone set, it is observed that significant improvement in ASR performance in terms of Word Accuracy (WA) is achieved by reducing Word Error rate (WER). Finally, a proto type model of application learning tool for Telugu language is developed, and is named as INTelligent Telugu Language Learning (INTTELL)

Keywords: Automatic Speech Recognition system, lexical model, Word Error Rate and Word Accuracy, Phoneme, Telugu language, Speaker Dependent (SD).

To my parents, ***Molakatala Nagarathnamma and Late Molakatala Hanumantha Rao****, withoutwhose support and encouragement and my beloved son* ***Undru Venkata Siva prasad Aaditya*** *with whose support this would not have been possible.*

Acknowledgements

First I would like to thank my thesis Adviser Prof. Atul Negi who accepted to be an adviser and is continuously helping and encouraging with his enlighten advices, and greatest support given pin pointed error finding in my way of written driving me into the right path of the research and finally to conclude the problem in well defined way. Aslo Advisory Committee members Prof. Kavi Narayanamuthy and Prof. Rajeev Wankar, their support both morally and subjective made me to walk though this difficult route of thesis work.

There are primarily two people who started me off on this journey, forcibly in the beginning, and continued to nag me until I was done. These are my father, **Late M. Hanumantha Rao**, whos dream was drived me and my advisor , **Atul Negi** who is a DRC member when I started and become the finally best advisor. I would like to thank them. My mother, **M. Nagarthnamma**, is probably one of the most unconventional mothers who taught me and supported all the time to drive towards goal of PhD. I am proud to have such parents and I hope my mother is the happiest person when I finally reaching to this stage. .

I lived through some really difficult times throughout the period of my PhD work – both on the professional and personal fronts. I thank my colleagues who stood by me and encouraged me to the hilt and supported me in so many various ways. I would like to thank my team of **"Hostel administrators" – Dr. Survasis Rana Dr. Rajendra Prasad lal and prof Vineet Nair, Dr. Vijayalakhmi and Prof Vasuki** – for the many wonderful help and care . I thank my **"School prof" – Prof. Chakravarthy, Prof. Arun Agarwal, Prof Girija, Dr. Rajeev Wankar, Dr. Salman, Dr. Durga bhavani,Dr.Sobha Rani, Dr**

Anupama and Dr. Rukma – for their constant support when I down. I want to thank R P Lal especially for always being there as a critic advisor and evaluator for my research papers and documentation.

My thanks to **"Dean, SCIS" Prof. Kavi Narayanamuthy** with whom my dream never come true. His kind support from the day my research domain support from his critic and technical guidance helped me to drive to reach the final goal. **Prof Chakravarthi and Prof. Arun Agarwal** who are two important people who really helped through their support in the subject of signal processing and giving timely suggestion and guide lines. Prof Arun Agarwal I would like to express my deep gratitude as, when I joined in the organization and his encouragement, care to complete my thesis and even valuable technical guidelines in the subject when I was in complete dark helped me a lot and finally boosted me to reach the goal. I also would like to thank all the members of DCIS present SCIS faculty who really helped me with their critic analysis and support during my work. I acknowledge all my students who contributed their voice for collecting the speech corpus for the thesis work. I express deep gratitude to all my family members and my friends who's support and encouragement to enforced me to achieve the goal. Especially my mother, my son Aaditya and his father supported all the days allowing me space to work. I also would like to thank my brother in-law Mr. U. Pradeep babu for his help in shaping the document. I express my heartful thanks to all my project students from KGReddy colleage and Mr. Sandeep patil who are the great helping hand to drive final goal. Finally and Almighty who has given me opportunity in my life to do the task

Nagamani Molakatala

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

Chapter 1

Introduction

In this Chapter we introduce the topic and certain background material related to the thesis topic. Human beings are most comfortable using speech, interfaces in interaction with modern digital devices. The technology behind ASR (Automatic Speech Recognition) is very complex in contrast to the ease of usage. An ASR system, transforms a human speech signal into a digital encoded form. Typically the early systems used to produce a sequence of ASCII strings. It is well known that there is tremendous linguistic diversity in the world. Not all native speakers know English. There is a tremendous need to produce ASR systems in Indic languages. Speech Technology and Language Technology together open a large domain of research challenges using the complex process of converting spoken language into written text. Systems that do this transformation are called as Automatic Speech Recognition (ASR). This process is also called as Speech to Text conversion (STT) system and it utilizes phonetic engines to implement the process.

1.1 Functional Block Diagram Of ASR System

In quite simple terms ASR is the transformation of spoken term input into feature sets as vector, and these feature vectors values are transformed as a symbolic representation of text as model. Here input spoken term is signal space and output text is the symbol space. Model transforms signal space to symbol space. One such model for signal a space is called Gaussian Mixer Model (GMM)[GHO12] and for Hidden Markov Model (HMM) [GAL98] are used for this transformation to the symbol space. The

basic principle of Speech Recognition is classically the problem of Pattern Recognition (PR)[DUD01]. In all PR systems first stage is to build a model using training data. For this purpose, Baum Walch [BAU96], re-estimation using Maximum Likely hood Linear Regression (MLLR) [DEM77][LEG95] and Maximum A Posterior MAP[GHO12] [XUE14] techniques are used. The next stage of recognizing the patterns is to map the signal to symbol space using a Viterbi search technique. Thereafter, in the application phase input patterns are recognized so as to generate appropriate symbol sequences in terms of ASCII Code encoded suitably for a language.

In speech recognition as well as in general two modules viz., Feature Extraction (FE) and Feature Matching (FM) modules are used. The task of feature extraction module is to convert the speech [ANJ06] utterance collected from the user into some type of representation for further analysis and processing. The extracted information is also called as a Feature Vector. This conversion process is done by signal processing unit, which is called as Front end module (FEM) [shown in Figure No. 1.1], as Module I. This module takes as input pre-processed speech signal, also called as a clean speech sample and the output of the Module-I is a feature vector. This feature vector is Mel Frequency Cepstral Coefficient (MFCC) [SLE79], which comprises of 39 vector values details in Chapter 2.

In the second module of feature matching, also called a classifier is to extract feature vector from unknown voice sample into their acoustic scores against acoustic model (AM), the model with maximum output score and its corresponding text is considered as a recognized word. The acoustic model is used to score the unknown voice sample. The output of front-end is given as input to the acoustic model. Generally either VQ (Vector Quantization)-code book or Gaussian Mixture Model (GMM)[KIM11] [RAB93]is used in the Acoustic model. In this work GMM is used. The Figure 1.1 shows the modularized Block diagram of ASR system[LEE94] shown below. The blocks are numbered based the work of the function. The '0' indicate the corpus collection and prepare the data set for the system '1' indicate the feature vector computing. These feature vectors are used to compute the GMM-HMM model during the training process. Here the system is learning the mapping the signal information corresponding to the symbol information. The process is done iteratively to create the patterns for a given input by user in the supervised learning mode. Here the phonetic symbols are mapped to the framed signal information. '3' indicate the computing the AM,PM and

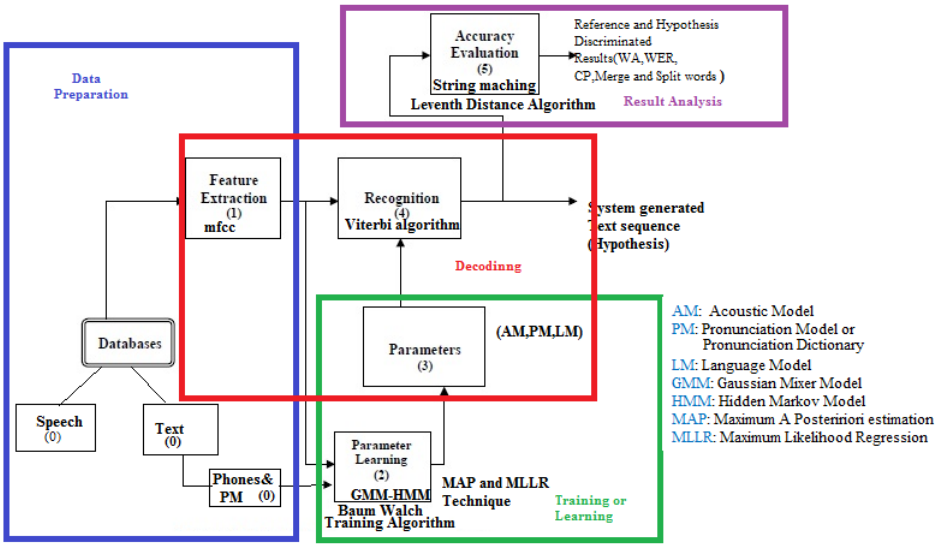


Figure 1.1: Block Diagram of modularized ASR system [RAB10]

LM[RAM16] for a word level patterns. Training process generate the HMM models for word level, and word network to lexical and language models.'4' indicate the decoding or testing process working with Viterbi search that searches text corresponding for given input probabilities. In this block using Viterbi search system compute the hypothesis text corresponding to given input speech signal in the form of feature vectors to this block. '5' is evaluates the generate text with the reference text. Reference text is user transcribed text. In this block using Levenshtein Distance measure algorithm[LIV16] computes the Word Accuracy (WA) and Word Error Rate (WER)[MOH13] which are the two measures used in this thesis for result analysis. The data flow in the system with functional description of input and output of the ASR system described with different color boxed in Figure 1.1 shown below.

1.2 Aim of thesis

In this thesis the aim is to use an Automatic ASR system for conversion of, input Telugu Language Speech utterances to output Telugu script with intermediate generation of ASCII codes, i.e, to build a Telugu Language ASR (TASR) system. The state-of-Art HMM method is used for transformation of spoken signal term into written form by using PR Technique. ASR Engine is the tool which is built on the principle of HMM to

take input as a speech and gives a text output in ASCII codes corresponding to the language of the speaker. The review articles of Prof. Raj Reddy [XUE14] and Prof. Janet M Bakar [JAN09] and team described progress over 4 decades of the ASR problem. In this thesis speech recognition for utterances in Telugu is chosen as the research problem. It is clear that a solution to this problem requires linguistic background information of the Telugu language given more details in chapter 3. Due to lack of resources in this language domain immense efforts are required to collect data and perform data annotation. The annotated text is mapped as a unit of pronunciation symbol that maps to the phoneme of a language. In Telugu Akshara [NEG01] is whole symbol to represent pronunciation. This symbol is used for building the lexical model or pronunciation dictionary. Phonological rules used as a make an attempt to incorporate the word pronunciation variations which is linguistic knowledge driven method to analyse pronunciation variation. With ASR system forced learning an attempt made to incorporate these variants into the lexical model for robust system as a data driven method [BHA11]. To strengthen both concepts, existing systems [KUM04] [AGG11] reviews of author works, and [SRE04][NAG05] are my contributions exploring with different input Telugu data in ASR implementation and their functionality and its transformation followed by use of Horse shoe model [MIS09][DUG06] [Detailed in Chap 6]. The whole transformation from ASR system [SRE04],[NAG05],[VIV06],[NAG07],[NAG16],[NAG09] to the TASR system[NAG10a],[NAG10b],[ABH10],[NAG11],[NAG12a],[NAG12b],[NAG13],[SAI15]in layered structure systematically followed with the principle of Horse shoe model software method explain in chapter 6 with Figure .6.1.

1.2.1 Contrast of SR in Human and Machine

The process of understanding speech by humans through recognition needs a large amount of knowledge in the language. This requires a command on the phonetics, phonology, lexical, semantic, grammatical and pragmatic layered principles of the language. It demands the user knowledge in context of a language must be very good so as to train the recognition task other wise communication fail. Speech communication and the components described in Figure 1.2. in layered structure of transformations from message formulation , Language coding, Neuromuscular actions to produce the acoustic waves by the articulatory system in the speech generation. Air is media to carry acoustic wav to auditory system that covert acoustic wav into the reverse process

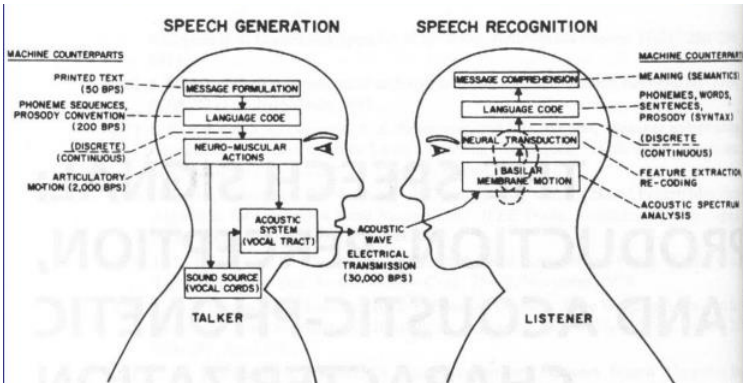


Figure 1.2: Speech Communication in Human with Speech Generation & Speech perception [RAB99].

to message computation that convey the message in the process of speech recognition modules in human. The role of acoustic module and auditory module are described in Figure 1.3. In both section important module is language code module that uses sub components of phoneme sequences, words and prosody to carry actual information. The two Figures 1.2 and 1.3 gives the importance of signal processing and language processing in speech technology. The processing in human brain is abstract level is described in detail with help of Figure 1.2. Hence the two Figures explain the necessary concepts need to understand design of ASR system. Figure 1.3 explains the physiological components that produce and perceive the signal by human brain numbered 1 and 3. '2' is the auditory system to transform signal into features set with spectral analysis of acoustic wave. It is an important block in computation technique.

Understanding of speech by a computer system is good, if it has chosen a carefully sub domain of linguistic information with well designed vocabulary in the context of command and control mode. This detailed process of Human neuro system, articulatory system to product acoustic wave, acoustic wave perception with held of auditory system and human brain is explain abstractly in Figure 1.3.

In context of speech processing Man and machines are different the way they do process.. The human ability lies in complex cognitive process that is able to map from acoustic signals to the meaning and purpose of communication. . In system speech is represented as sequence of digitial values and these values used to map in the system to make machine to recognize the speech. In this context both man and machines do the process of search for best match of text that maps to the given input signal.

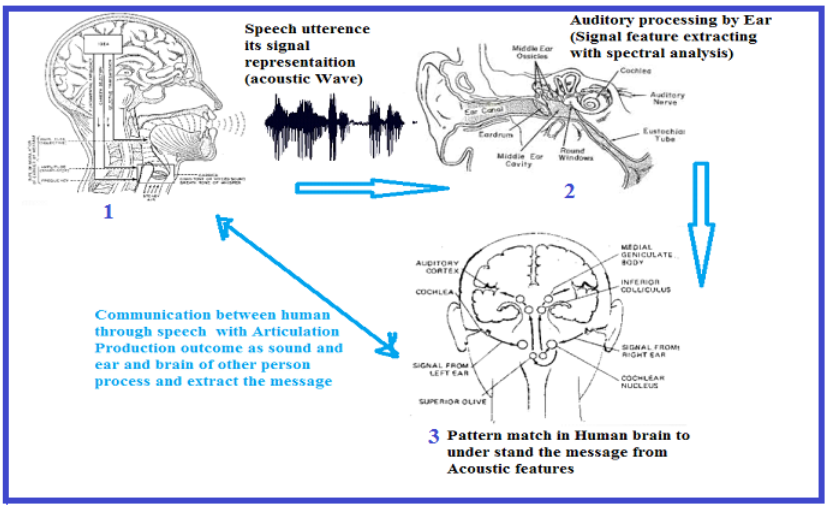


Figure 1.3: Human Speech Recognition process with articulatory and auditory system integrations [LRA89].

Using Machine learning algorithm, ASR technology compute and optimize the search to get best match[MAS13].. One of the very successful systems for ASR[XUE14] makes use of the HMM[RAB89] [JAC08]frame work. For facilitating the same probabilistic Bayes classifier of ASR system furnishes the approximate word sequence for a given acoustic signal. The probability estimation and probability distributions represented through the parametric models are used for the purpose. The N-gram model is used for estimation of probabilities of word sequences [JEL97][RAB10] as part of the language grammer and probabilities of speech signal are modelled through HMM technique. Despite a lot of technical progress human cognitive performance is presently better.

1.3 Problem Statement

In this work a HMM based ASR system is designed for the Telugu language, one of the most used Dravidian languages. The speech characteristics of previous ASR system are, built with the English Alphabetic system [BRA83]. They used American accent (English) sound units. In this work using horse shoe model re-engineering concept and well established ML algorithms for speech pattern recognition it is attempted to model the human learning process. The approach proposed here also incorporates robust adaptation methods for recognition by Speaker Independent mode of speech

into written form of the Telugu language utterance. The proposed system has an ASR engine, supervised method of learning, the use of both knowledge base and data driven procedures for adapting to the speech variations, and construct the lexical model of the Telugu language ASR system. Precisely the thesis problem states that "design of an Telugu language ASR system, using phonetic and phonological concepts from knowledge base, and pronunciation of sounds and their variants' from data driven method to produce robust system. Develop a lexical model of Telugu language with text and speech corpus with speaker variants incorporating their adaptations. Thereafter, the design is generalized for any human languages".

1.4 Motivation and Objectives

In this work its aimed design a framework for recognizing the spoken inputs in Telugu language (by a Speaker Independent model) and to convert this into its written form representation with fundamental independent sound units. The phonetic Engine (Sphinx III ASR system)[NAG05] [SRE04], is used in the target language. The approach is processed from phonemes, sub-word units as syllables, words and simple sentences which is the traditional way of learning any language. The proposed design may be useful for making a Speech to Speech interactive system for language learning, with additionally facilitating to learn an Orthography (of Telugu Akshara) through the proposed INTTELL approach [NAG05][NAG10].

1.4.1 Technical Limitation of Thesis

This section is depict constrains on the research problem chosen and description on the research problem demarcations.

- Adaptations or modeling are different kinds, and are used to improve the performance of ASR.
  - Specifically modelling pronunciation at the lexical level is about looking the variations within the word and across the words. The amount of pronunciation variation is limited in isolated words compared to the conversational speech [KUM04], hence limit to the present dissertation work on isolated words is more focused. The variation is due to limited speech variants in

isolated words, as compared to the conversational speech data i.e limited in phonetic and phonological concepts of linguistic knowledge.

– The problem of pronunciation variation at the lexical level.

- The importance is given to the speaker dependent mode, one of the focused point in the dissertation task. According to the Charles Darwin [DAR58] the best adapted individuals, with variability in favorable direction, will tend to propagate their kind. History says that at the evaluation time the vocal organs of human are adapted in such way that they are used to produce the speech. Actually their primary role in human physiology is different. (the functions of vocal cord is explained in detail in chapter 2, Human physiology section). Human perception, which is similar to the speech recognition in system also adapted to be able to process speech sounds. This task is complicated to computer system to date, in spite of 7 decades of research by many groups success rate is approximately 50% in real time domain application due the constraints of data set collection.

This work has commenced before the use of deep networks [DEN10], and in era of much less powerful computational system, the approach therefore is based on upon the strategies, where lesser computation of storage is used.

1.4.2 Applications

The speech modality can be used with latest trends in use of latest electronic gadgets like Mobile phones. The use of such miniature gadgets is leading to the health problems. Speech is the easy and simple system of inputting and out putting from digital world [HAM10] to communicate, which may help to come out of the drawbacks of current touch key press modalities. Hence, Human Language Technology, rising with combination of computer science, Information Technology and Computational Linguistic domain, is no more confined to humanities domain but now also an Engineering domain.

All these progressive technologies give full opportunity for Human interfacing system through speech Technology. This technology includes both ASR and TTS system[JAC08]. The problem addressed in this thesis i.e. lexical model for Telugu language and its adaptation using speech variant factors is going to be most important

work in both Telugu states, Telangana & Andhra Pradesh. It also gives scope for common man who does not have computer literacy, but can be able use the Technology. E- Governance will reach to every common man in the society in future.

Chapterization in thesis Structure

Chapter 1: In this chapter discussing the problem introduction, an introduction to speech recognition system, with brief explanation of ASR system functions. A motivation with social relevance of thesis problem is also presented. Here the assumptions and technical limitations for implementation are also presented. Design objectives, thesis statement and finally structure of thesis is presented.

Chapter 2: Investigation of the existing ongoing research work in the field of ASR system, linguistic aspects in context pronunciation variants that affect the word accuracy of ASR system, New language based ASR system building and their approaches and limitations. The different domains that related to and limiting factors on Recognition accuracy compared to Human Speech Recognition system. Thesis focuses topic of pronunciation variation and its adaptation techniques followed by other languages.

Chapter 3: In this chapter a discussion about the differences between phonetic and phonological information in context of ASR system building in presented. Discussion on Telugu language concepts and an attempt is to made to build a better system. Also explored ASR system and its application in Education domain and presented for proposed INTTELL (Intelligent Tutor for Telugu Language Learning) tool [NAG05][NAG10]. Existing language education tools using Speech systems are also part of the chapter description.

Chapter 4: Study of existing literature and research in pronunciation variations and their modelling in the Telugu lexicon is presented. In this chapter some of the traditional approaches that deal with pronunciation variation in ASR are discussed. The aim of this is to establish bench mark experiments in developing a lexicon model (knowledge base and data driven methods of the system) in spoken languages. The chapter concludes with directions to understand target language modeling in lexical model with the context that is suitable for creation of robust lexicon for Telugu ASR (TASR)[NAG10][NAG11][NAG12].

Chapter 5: This chapter focuses on empirical analysis of lexical modeling using ASR system frame wok. Presentation of a novel technique is applied to develop lexicon for Telugu language and analyzing the pronunciation variant factors that happen in the developed corpus. Further, the modeling pronunciation variation at the lexicon level is discussed. The benefits of including a new pronunciation dictionary for different source of speech data, in terms of gender, dialectal and language variants, during training for pronunciation variations at the lexicon level is presented [NAG13]. Successful approaches investigated in the current work and attempt is made to improve Word accuracy in ASR system. The need to develop a speech and text corpus in design aspects for developing Telugu language ASR system.

Chapter 6: Conclusion of the thesis, summarizing contributions of research carried out. Future scope for further research directions discussed. The thesis covers both the theoretical aspects as well as applied aspects of ASR system performance. Scope and definition of INTTELL system design, methodology and the architecture are carried out.

1.5 Research Contribution

As part of the PhD research contribution, papers were presented at various National, International Conferences and Journal Papers, as detail are given in chronological order at Table No.1 below.

Table 1.1: The cronological order of research contribution in ASR domain

S.No.	**Thesis relevant Contributions and references mentioned in thesis**	**Tasks focused during research**	**Context of the Task for research and its publication focus**
(i)	[NAG04] (Nagamani et.al,2004):	Speech analysis important and utilization of tools for feature extraction procedures and application proposal	Collection of speech samples in different environment and analysing the speech signals using the tools and programming scripts. Speaker Dependent and Speaker Independent system design for ASR

(ii) (iii) (iv)	[SRE04a],(G.Sreenu, Nagamani et al,2004) [GIR04] P.N. Girija, M. Nagamani et al,2004 [SRE04b], (G.Sreenu, Nagamani et al,2004)		
(vi)	[NAG05]:(Nagamani et al.,2005)	Implementation of ASR system for application development using speech technology concepts	Speech sample collection for Telugu language specific data design application with simple Telugu sentences(40 Telugu sentences), Telugu phonemes and Simple Isolated Telugu words with application proposing for language learning tool.
(vii) (viii)	[VIV06: (Vivek et al.,2006)] [NAG07]Â :(Nagamani et al.,2007)	coustic signal analysis-Missing feature concepts in context of Telugu phoneme ASR for Accuracy improvement procedure exploration task. Applications of ASR system	ASR applications and Design concepts and base system understanding for TASR
(ix) (xii)	[LAL09]Â : (R.P.Lal et al.,2009) and [NAG09] : (M. Nagamani et al.,2009)	Detailed exploration of Phonetic Engine Sphinx III in context of HMM, design and developed ASR system for Telugu speech, pronunciation of speech utterances and their variations in word level context, and their influence on Word accuracy of ASR system	Isolated Word corpus using Telugu language and implementation of the ASR system in Command Control mode for form filling applications. Also noise analysis and segmentation problem pre and post utterances to deal with Word Error Rate in ASR system

(xiii)	[NAG10a]: (M. Nagamani et al.,2010a)	Pronunciation Modeling, supportive tools for development of TASR system and corpus building tools exploration	Defining Telugu language Askhara specific phonetic symbol, use of these phonetic notation to design Telugu lexicon. Explore and propose transliteration tool for Telugu transcription and manually written Lexicon for Telugu data set used for system development
(xiv)	[NAG10b]: (M. Nagamani et al,2010b)		
(xv)	[NAG10c]: (M. Nagamani et al.,2010c)		
(xvi)	[ABH10]: (Abhijit Debbarma et al.,2010)		
(xvii)	[NAG11]: (M. Nagamani et al.,2011)		
(xviii)	[NAG12a]: (M. Nagamani et al.,2012a)		
(xix)	[NAG12b]: (M.Nagamani et al.,2012b)		
(xx)	[NAG13]: (M.Nagamani, et al,2013)	Error analysis in Telugu phoneme recognition, Speech enhancement concepts to improve accuracy of TASR, applications of ASR and Telugu Speech systems	ASR performance improvement by reducing the insertion and substitution errors by removing the noise in the signal information and also developing an application to command control based music player software through ASR system. Text to speech for Telugu language task
(xxi)	[SAI14] Sai Bala Kishore N, M, Nagamani et al.,2014)		
(xxii)	[SAI15]:(Sai N B Kishore et al,2015)		

xxiii	[NAG16]: (M.Nagamani et al.,2016)		
xxiv	[NAG18]: (M.Nagamani et al,2018)		

****

Chapter 2

Literature Review of Lexical Modelling in HMM based Telugu language ASR System

2.1 Introduction

In this chapter concepts related to thesis work "Modelling the Lexical model" using speech variations to develop Robust Telugu language ASR system are reviewed. The major contribution is an approach to building the lexicon for the TASR (Telugu language ASR). This process is used to improve word recognition accuracy by reducing the confusions causing word error rate. TASR system generates the Telugu phoneme specific phonetic units as their basic units of sound representation as the text. These phonemic specific phone set are used in lexical model. The same lexical model is adopted for generic ASR system that can work for any language. An understanding of the operation of speech recognizer is required for the purpose. In ASR system, a waveform of phonetic symbols that are derived from Telugu Aksharas is input to the recognizer system and the output is generated in terms of ASCII symbols. The entire process of Pattern Recognition is explored using HMM technique.

2.2 ASR system as a Pattern Recognition (PR) problem:

Human beings possess "capability of learning by experience" [DUD01]. This is the key concept taken for the ASR system frame work, which is used to convert a signal into

its corresponding recognized text. Spoken terms or uttered speech are transformed into a written script. A set of rules are made by understanding the nature patterns by the Humans. The major challenge in research is to incorporate this human knowledge (skill or ability), which is difficult to comprehend, into a process. Human performance of recognition has so far been superior to that of machine in this context. The underlying abstract concept is patterns, and these patterns are mapped with some underlying rules to recognize unknown patterns. Hence the scope opened by pattern recognition domain, to achieve the goal of Recognition process of "speech" that is a natural information produced by humans [BIS06]. Through decision capability by machine intelligence system, Pattern Recognition (PR) is done, through mathematical techniques. The domain of PR is the design and development of systems for recognizing patterns by analyzing the data, depicted in Figure 2.1. In most cases, the data is real world data and are gathered through sensors. Microphone or audio sensors are the data acquisition devices in this context. From the sensors signals, patterns are extracted. A Microphone is a transducer which converts acoustic pressure variations. Subsequently in the form of an electrical signal, which has to be sampled and digitized. Speech wave pressure sensing is performed by a microphone which captures human voice and converts into the electrical signal. These electrical signals are converted into pattern representations using digital signal processing techniques, sampling quantization and filtering methods. Humans do pattern processing first and then representation of data. The machines are good at data processing hence reverse process of human, i.e. representation of data and then pattern processing next [YEG94]. With the help of signal processing, the natural patterns information of human can be captured. The information is extracted in terms of data which is useful to represent the machine. Subsequently machines are taught to learn. They are able to recognize the patterns, with help of human generated rules. The data from the machines are learned similar to human. The core processing in human and machine is entirely different. Hence till to date ASR problem is a complex research task. Human are more comfortable with speech communication and systems are comfortable with script (written form) communication. To interact with the machine by human is simple, when speech modality is used with help of speech technology. The process is tedious and complex task when it is natural language rather programming language. In the universe there exists about 6700 [OMV05] spoken language. Though some language looks similar but they always have distinct features. These divergence representation

2. LITERATURE REVIEW OF LEXICAL MODELLING IN HMM BASED TELUGU LANGUAGE ASR SYSTEM

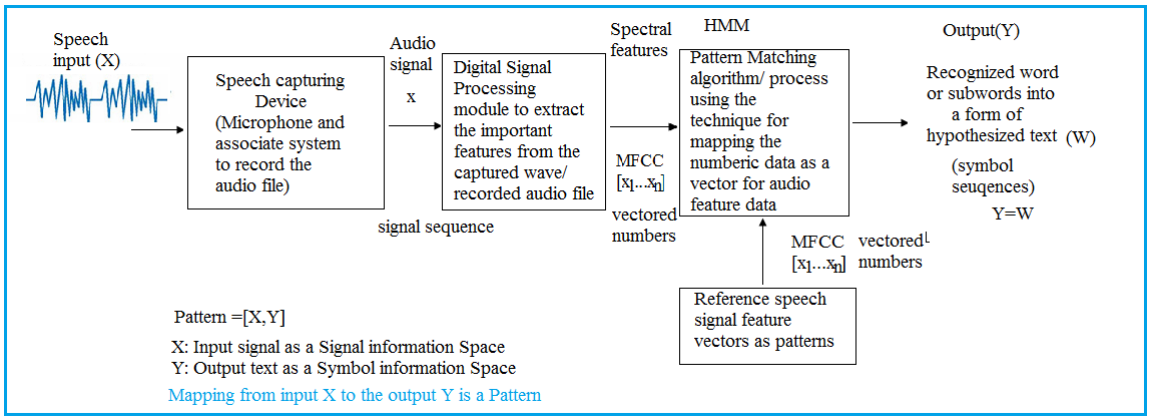


Figure 2.1: Pattern Recognition problem for ASR using the state of art HMM [DUD01] [BIS06].

by the machine is a complicated task. The ASR research dates back to 7 decades. Yet for any new human being the spoken languages implementation need to cross barriers, in spite of sufficient available technological facilities.

In the following a review specific to Telugu language ASR development is presented and finally focus is on to the speech variability and then to variations in modelling in this system. This is used to show the nuances and complexity related to ASR systems in general and specific to Telugu in particular.

The literature shows several approaches to the ASR system development. The Hidden Markov Model (HMM) , the scope in this work presented. The approach for ASR System is the HMM. HMM, has rich mathematical solutions for pattern recognition problems. The generic architecture of an ASR system is described at Figure.2.2 for Telugu language, using HMM approach for pattern generation and pattern classification for a given input speech. signal. The approaches are feature extraction, and classification.

Language specific building blocks of Figure 2.2 generic representation for any language with the following architecture of the ASR system with HMM is shown below Figure 2.3.

The ASR system researches for the Telugu language during the period from the year 2004 to 2017 is given in Table 2.1. In this table the research contribution of authors and the topic of the research is presented. Most of the cases generic ASR system with the use of HMM model is used to develop language specific ASR system.

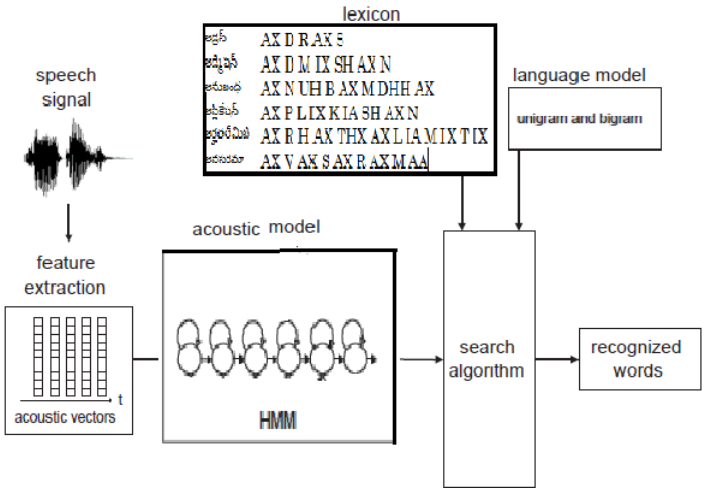


Figure 2.2: ASR system block diagram using HMM principle with Telugu Language Speech data and Lexical model using UOH symbol set.

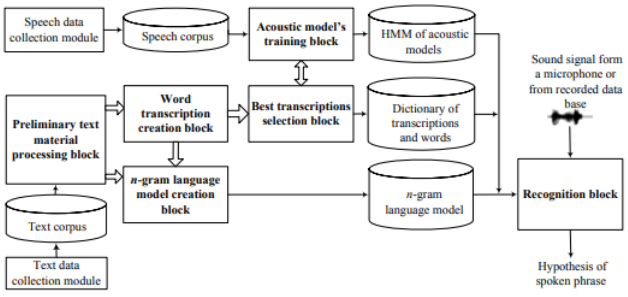


Fig. 4. The architecture of software complex of conversational Russian speech recognition system.

Figure 2.3: Generic ASR system architecture for developing language specific with use using HMM principle for Acoustic Model; Lexical Model; and Language model;[IRI13].

Hence choice of HMM in this thesis work. Definite model building compare to other contemporary model during that period. ASR is multi disciplinary system, integrating all those system development from the scratch is very difficult. This may not a proper choice as concentration of the task is the language specific system building. Hence chosen the tool that available with open source. This open source tool Sphinx III speech recognition system built with the HMM principle. The training and testing process of the system explored with the principle of PR and HMM in this system. This is second reason to be a choice of HMM in this thesis work.

Table 2.1: Review on ASR system for developing of the Telugu language by various authors in India in chronological order from 2004 to 2017.

S.No.	**Conept of Research Article in Telugu**	**Authors**	**Year of Publication**
1	A Human machine speaker dependent speech interactive system	[GIR04] P.N.Girija, M. Nagamani, M Narendra	2004
2	Improving reading and writing skills with Intelligent Tutor for Telugu Language Learning (INTTELL)	[NAG05] M. Nagamani, M Narendra Prasad , P.N.Girija	2005
3	Statistical analysis for Telugu Text Corpus	[BHA07] G.Bharadawaja Kumar, Kavi Narayana Murthy	2007
4	Real time ASR with HMM word models for Telugu	[RAM08] A.V.Ramanna, P.Laxminarayana, P.Mythilisharan.	2008
5	Intelligent Tutoring System for Telugu Language Learning	[NAG10] M. Nagamani, BSR.Krishna	2010
6	On the Design of a Tag set for Dravidian Languages	[KAV12] Kavi Narayana Murthy, Badugu Srinivasu	2012
7	Substitution Error analysis for improving the word accuracy in Telugu Language in automatic Speech Recognition System	[NAG12] M. Nagamani, P.N.Girija	2012

8	Developing efficient speech recognition system for Telugu Letter Recognition	[VEN12] Venkateshwarulu R.L.K, Teja R.R, Kumari.R.V	2012
9	Text Independent Language Recogniton system using DHM with new features	[SAD12] M.Sadanandam, V.Kamakshiprasad, V. Janaki, A.Nagesgh	2012
10	Pronounciation Variant and Substitutional errors for improving telugu language Lexical Performance in ASR system accuracy	[NAG13]M.Nagamani, P.N.Girija	2013
11	Named Entity recognition for Telugu using N-grams and Context features	[YOH13] Yohan P.M.	2013
12	Deduction of Confusion pairs on different rates of speech in Telugu	[USH13] N.Usha rani, P.N.Girija	2013
13	Morphology based POS tagging on Telugu	[SRI14]Badugu Srinivas	2014
14	Environmental Noise analysis for robust automatic Speech Recognition	[KIS15] Kishone N, Sai Bala, M Rao Venkata, M. Nagamani	2015
15	Transcription of telugu TV NEWS using ASR	[RAM15]M Ram Reddy , P Laxminarayana, A V Ramana, Markandeya J L	2015
16	Accent detection of telugu speech using prosodic and format features	[KAS15] Kasi Prasad,Mannepalli,P. Narahari Sastry	2015
17	Dependency parser for Telugu Language	[NAG16] Nagaraju G. N,Mangnathayaru and B padmaja Rani.	2016
18	Analysis of Source and System features for Speaker recognition in Emotional condition	[RAJ16] K.N.R.K Raju Alluri, V.V Vidyadhara Raju, Suryakanth.	2016
19	Study of Telugu Vowels using accoustic features	[PRU16]Pruthiv Raj Kyakala,Rajasree Nalumanchu, V.k Mittal	2016

20	Speech Recognition using arithmetic coding and MFCC for Telugu Language	[ARC16] Archek Praveen Kumar,Neeraj Kumar,Cheruku Sandesh Kumar, Ashwani Kumar	2016
21	Implicit Language identification system used on Random Forest and Support Vector Machine for Speech	[MAN17] Manish Gupta, Shambhu Shankar Bharti, Suneeta Agarwal	2017
22	Speech based emotion recogniton in Tamil and Telugu using LPCC and Hurst parameters -A comparative study using KNN and ANN classifiers	[REN17] Renjith, K.G. Manju	2017

Actual Research in the area of ASR is around 5 decades. The following Table No. 2.2. present various research contribution other than the Telugu language from the Isolated digit recognition system to spontaneous speech recognition is included. The contributions are in development of algorithm, systems, books and concept. The technology progress in chronological order presented.

Table 2.2: Research work (Non-Indian Language) carried out by various groups and researches from 1963 to 2017, few milestones in ASR with their authors and contribution years.

S.No.	Year	Authors	Concepet & Language
1	1963	[NAG63] K.Nagata;Y.Kato;S.Chiba	spoken digits;japanese
2	1978	[SAK78] H;Sakoe; S.Chiba	SpokenWord
3	1989	[RAB89]L.R.Rabiner	HMM-Review
4	1994	[MOO94] Travel;R.K moore	speech resarch and important points related to speech
5	1996	[RED96] D R Reddy	Technical Report on computer speech recognition
6	1996	[GAL96] M J F Gales;S J yong	Speech recognition using parallel model combination
7	1997	[JEL97] F Jelinek	Statistical method of speech recognition
8	1997	[VAP97] V vapnick;S Golowish;A Smola	Signal processing regression estimation
9	1998	[BUR98] Burges C	Book on ASR

10	1999	[RAN99] Lawrence Rabbiner;Biing Hwang Juuang	SVM kernel based learning method book
11	2000	[NELOO] Nello Christian,John Shawe-taylor	Natural human machine communication with ASR book
12	2000	[JUAOO] B.H Juang and S.furi	Language independent and language adaptiveacoustic modelling
13	2001	[SEH01] T.Sehultz	Named entity recognition using character level features evaluation
14	2003	[WHI03] Whitelaw;Casey and john pattrick	review article on corpus based spontaneous speech recognition
15	2005	[FUR05a] S Furui	Statistical method for recognition and speech understanding
16	2005	[RAB05] L Rabiner and B Juang	Ubiquitous speech recognition-cluster-based modeling
17	2005	[FUR05b] sadoki furuki;thomohisa Ichiba	Language survey:slavic languages speech protection knowledge in automatic speech recognition
18	2006	[SUS06] R Sussex and P Cubberley	Multilingual speech processing
19	2006	[SPE06] Spector; Simon King;Joe Frankel	deep network with local denosing critirea-stacked denosing auto encoders
20	2006	[SCH06b] Schultz, T Kirchhoff k	A review on speech recognition technique
21	2010	[VIN10] P.Vincent;H.Larochelle;I lagole	Convolution bottleneck network fetures for LVCSR
22	2010	[SAN10] Santosh K, Gaikwad;Bharti W	Sub Guassian moxture models for cross lingual
23	2011	[VES11] K. Vesel; M.Karafit and F.Grzl	Sinhala speech recognition
24	2011	[GH011] Lu L;Ghoshal A Renals	MLP training using multilingual data and their
25	2011	[NAD11] Nadungodage T,Weerasinghe R	Unsupervised cross lingual knowledge transfer in DNN

26	2012	[BREE12] N.T.V.U.W.Breiter;F.Metze	Unsupervised cross lingual knowledge transfer in DNN-based LVCSR
27	2012	[GHO12] P.Swietojanski; A.Ghoshal and S	Deep and wide: Multiple Layer in ASR
28	2012	[SW112] P.Swietojanski;A.Ghoshal and S Renals	Maximum A Posterior adaption of subspace Gaussian mixture models for cross - lingual speech recognition
29	2012	[NEL12] Nelson Morgon	isolated word ASR pasto
30	2012	[GHO12b] Lu L;Ghoshal A Renal	Multilingual training of deep neural network
31	2012	[AHM12] Ahmed Irfan;nasir Ahmad;Hazrat Ali	DNN accoustic modeling with modular multi-lingual features
32	2013	[GH013]J.Gehing;Q.B.Nguyen;F.Metze	Multilingual accoustic models using Distributed deep neural network
33	2013	[HE113]G.Heiigold;Vanhoucke	Multi-lingual hierarchical RASTA features for ASR
34	2013	[TUS13] Z.Tuske;R.Schluter and H.Ney	Database of speech recognition for comparative analysis of multi language
35	2013	[MAS13] M.Masior;M.Igras;M.Zioalko	speaker independent connect speech recognition fifth generation computer
36	2013	[FGC13] FGC article not known	indonesian isolated digit speech recognition
37	2013	[DEW13] Dewi;Ika Novita;Fahri Firdausilla	online adaption technique for spoken query system
38	2013	[SHA14] Shahnawazuddin S;Sinha R	Deep Speech:scaling up end to end speech recognition
39	2014	[HAN14] A.Hannun;C.Case ;J.Casper	state of art in statistical method for language
40	2014	[HAN14] A.Hannun;C.Case ;J.Casper	Deep convolution neutral network with layer wise context explanation
41	2016	[BEL16] J.R.BelleGrada and C.Monz	state of art in statistical method for language
42	2016	[YU_18] D.Yu;W.Xiong;J.Droppo	Deep convolution neutral network with layer wise context explanation

43	2017	[LAZ17] Lazar;Jonathan;Jinjuan Heidi Feng and Harry	research methods in human computer interaction
44	2017	[XIO17] Xion;Wayne	Microsoft 2016 Conversational speech recognition system

The main approached followed in this thesis are discussed briefly to understand structural content in the thesis. There are various approaches based on the context of data and knowledge bases and classification of the methods. The following paragraph reviews on ASR approaches.

2.3 ASR approaches

There are different approaches to the implementation ASR system following sections describes the required knowledge that used in the current thesis implementation process.

2.3.1 Knowledge based approach

The study of the Telugu language that a speaker desire to intends to express, and entail an analysis is a Linguistic context approach. The clear inclusion of expert's speech, referred as linguistic background is involved in directly to the design, in terms of the rule based system. Here the knowledge source is designed by theoretical or language parameters considered framing the knowledge base of ASR then train, and testing the system with the collected speech data to design. The knowledge based ASR system is given in Figure. 2.4.

The properties of acoustic-phonetic speech signal, is the knowledge source, this is derived though spectrogram based analysis to extract the phonetic sound units [BHA1]. The spectrogram is a 3-D ie., time, frequency, strength of a given acoustic signal. This gives connected temporal and spectral characteristics of speech signal. The spectrogram based speech utterance is shown in Figure 2.5.

The Telugu transcription of "axthxaxdu raamuxduh" (అతడు రాముడు) in spectrogram, indicates (i) the dark areas (for more strengthened), (ii) the horizontal shady bands [for the peaks created (F1,F2, F3,F4 and F5)] and (iii) red coloured (for the vocal tract resonances) [FAN74]. This visual description helps to analyze the spectrogram,

2. LITERATURE REVIEW OF LEXICAL MODELLING IN HMM BASED TELUGU LANGUAGE ASR SYSTEM

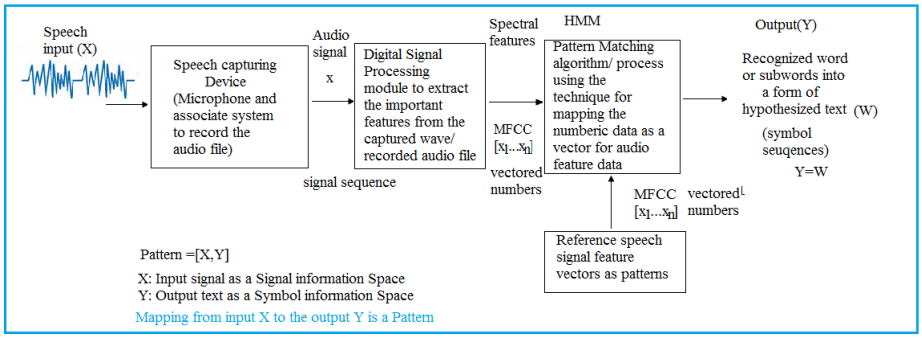


Figure 2.4: Knowledge based ASR system schematic diagram [SAK09]

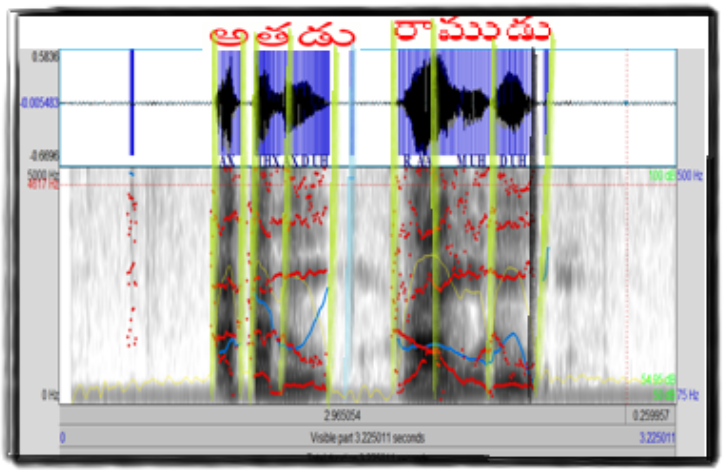


Figure 2.5: Time series speech signal represented with its spectrogram for simple Telugu sentence.

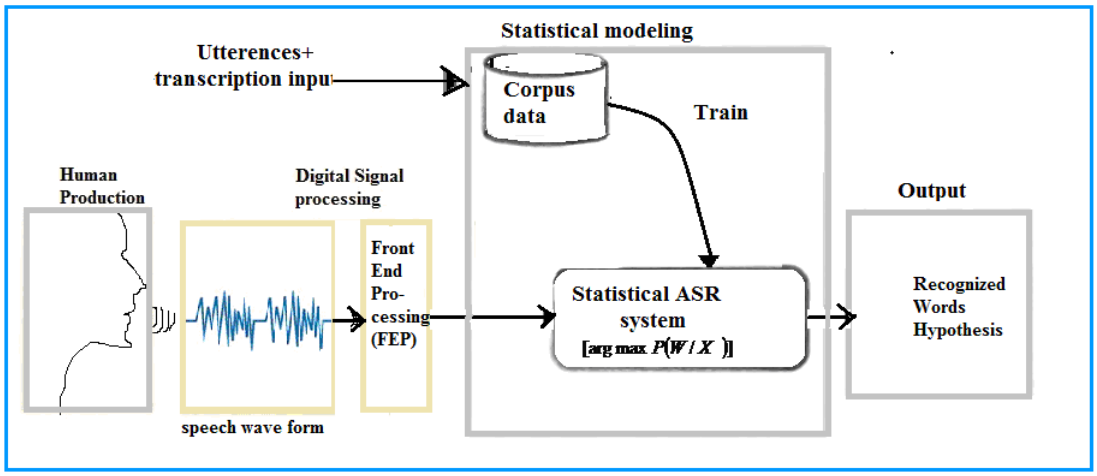


Figure 2.6: Data driven statistical ASR system approach [SAK09]

and human experts can perceive the phonetic speech signal in a language. The expert knowledge is from different subjects that include articulatory movements, phonotactics, acoustic phonetics, and linguistics [RAS16]. The knowledge sources of this kind helps to handle small vocabularies and isolated words for example: application forms, digits, phonemes etc. but it is difficult to handle the situation with LVCSR applications [ARA01]. The processing of such LVCDR data requires expensive computations, yet relatively poor system capability performance is obtained. Frequently the knowledge based approaches encounter following key obstacles:

In view of the requirement of interpretation of spectrograms by human experts, there is an uncontrolled loss of generality. Hand crafted system rules are limited due to to limitation defining knowledge for every transcribed data..Increased number of rules can lead to increased inconsistencies For SI system each speakers variation incorporation is difficult and requires huge set of phonemes for different speaker pronunciation. This knowledge base building, requires a huge amount of time[MOH05], [XIN13] [STE86].

2.3.2 Data driven Approaches

Speech and text corpus collected for the development of the system is used to deried the rules for adaptation which is shown in Figure below in Data-Driven(DD) wherein manual entering the experts' knowledge is required.

Pattern Recognition problem, is solved by using statistical approaches explain in

2. LITERATURE REVIEW OF LEXICAL MODELLING IN HMM BASED TELUGU LANGUAGE ASR SYSTEM

Figure No. and with Duda and Heart text book concepts of mathematical exploration gives sufficient knowledge for this method. herefore, the common core of the statistical classification perspective [SEH96] is "learning by examples" from a collection of data. The data driven approach shows better compared to the knowledge-based(KB) approach. In reviews [PAL03], [JUA05], [TOL01].

An HMM represents an extension of a Markov model, where each Markov state corresponds to a non-deterministic event with an associated observation probability and where the generating state sequence becomes unobserved or hidden [SAK09] [HUA01].

HMM is a double stochastic process. It is changing its state continuously with respect to time according to a set of state transition probabilities. After each transition, the process produces a symbol of the state according to an observation probability [RAB93].

Generally, the elements of an HMM are

- A set of a finite number of states $\Omega = 1, 2, \ldots N$; In this work the each state is a phonemes or sub word unit or a word. 'N' is the total number of states, and a state sequence of time length 'T'is denoted as

$$Q = q_1, q_2 \ldots q_T \tag{2.1}$$

  where Q_t is the state at time.

- A set of distinct observation symbols for every state that correspond to the physical output of the system in this work are represented by signal parameters which are modelled according to the state probabilities and state transition probabilities.

$$O = 1, 2, \ldots, M; \tag{2.2}$$

  'M'=total number of symbols, and the observed output sequence of time length is denoted as

$$X = x_1, x_2, \ldots x_t \tag{2.3}$$

  where x_t = the observed output at time t .

  in this present work there are sequence of observation on signal feature set.

- The transition probabilities from state i to the state j

$$A = a_{ij} where a_{ij} = P(q_{t=j}|q_{t-1} = i) \tag{2.4}$$

  The above explanation mapping to the speech processing the section.

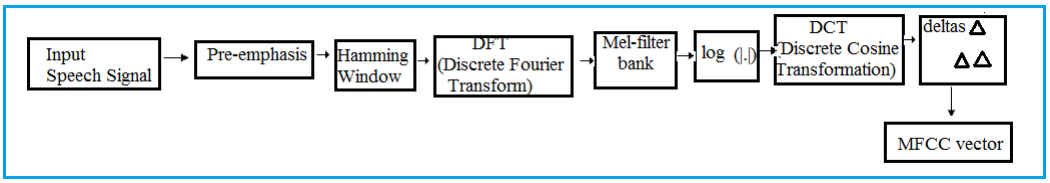


Figure 2.7: Audio data to feature extraction using DSP modules Mel-filter and deltas to space reduction in storing features set as MFCC vector of speech signal in ASR system [SLE99] [KSA98].

2.4 Speech Signal processing for pre-processing in ASR system

In this section first module of an ASR system is presented. The first module of an ASR system is the Signal processing module. The various functions that are necessary to understand the system are shown in the following sub sections.

2.4.1 Speech Signal

The observation probabilities are on signal features that produces the sounds with the movement of articulators. The position of each articular in the sound is not visible and they are observed in terms of signal parameters. The hidden parameters are indirectly observed in the observation probabilities hence the name Hidden in HMM. Extraction process of observation probabilities start with Utterance recording from a microphone and the captured signal is processed using a speech analysis tool Praat [NAG04], Audacity[ADA10] and Matlab [SID14] and shell scripts. Features are extracted with Digital Signal Processing (DSP) module of an ASR system. Components of DSP module shown in Figure. 2.7 The module gives the extraction of process from the given audio file as input speech in the form of wave file vector from given time domain input signal.

2.4.2 Speech utterance

The time domain speech signal represent in time verses signal strength given below.

The signal is represented with n frames as shown below Figure. 2.8. X(n) represents the input utterance in terms of frames. Each frame is extracted by choosing a window and sliding it over the time period. Generally Window size is 25 ms and sliding with

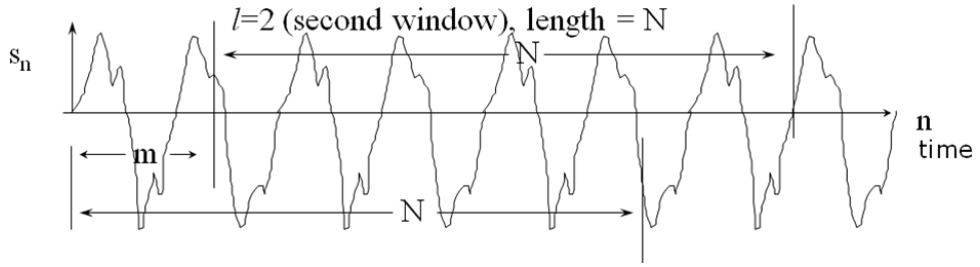


Figure 2.8: Speech signal representation in time domain with N frame length and n is time and is signal representation with 'n' no.of samples with window size of 2 seconds [RAB99].

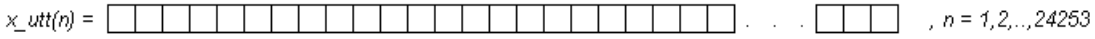


Figure 2.9: Speech utterance representation in terms of frame, 'n' represent the number of frame with frame length of 10 ms duration [RAB10].

15ms will give the total frame size as '10ms'. Hence for 1second duration signal there count for 100 frames with frame numbers 0 to 99.

2.4.3 Preprocessing

The sequence of preprocessing involves (i) reading utterance as time varying continuous speech signal, (ii) Windowing to convert the continuous signal into quasi stationary signal for short time period (iii) Transform the time domain into frequency into signal domain and (iv) to extract the different frequency component values as a vector for further processing, given below as MFCC.

The input signal for utterance representation in time domain is shown as a continuous signal as in fig.2.10:

2.4.3.1 Windowing

The windowing is a function in the signal processing block to make the continuous time varying signal into the quasi stationary signal for analysis and extract the feature of these quasi stationary signal in a short time period, as discrete values. [RAB93].

Hamming window

$$Mel\,(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5}$$

Where is the physical frequency, in Hz and Mel is the perceived frequency in Mels.

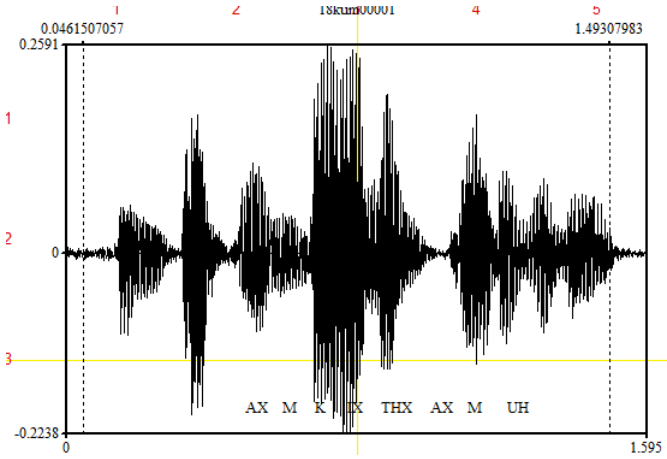


Figure 2.10: Input signal for utterance to the Front End processing unit.

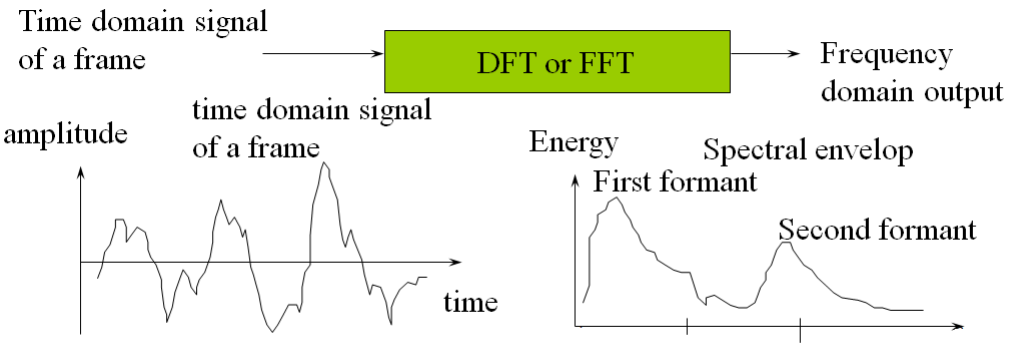


Figure 2.11: Transformation of Time domain signal into Frequency domain signal i.e transformation of DFT into FFT [GUA04].

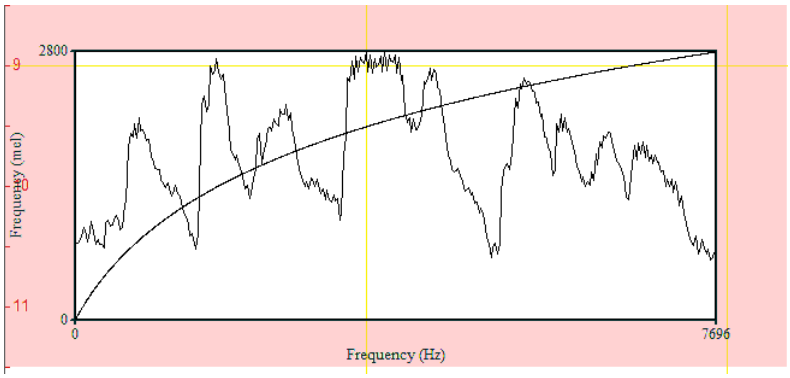


Figure 2.12: Mel Frequency response in Window function [CHE07].

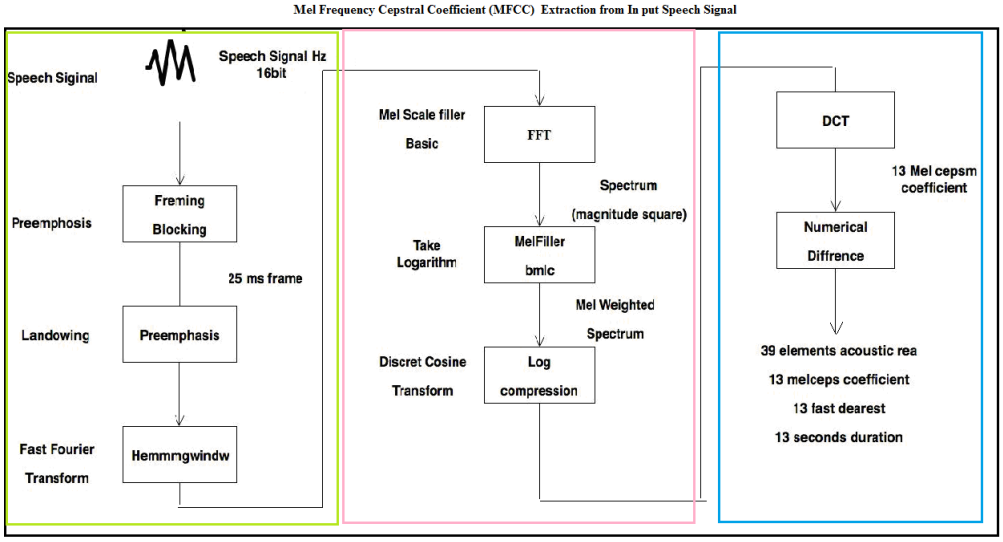


Figure 2.13: Front end processing of ASR system to extract the MFCC[SLE99].

Mel filter bank response for the word AXMKIXTHXAXMU /అంకితము/ With frame value of 25.499349884507428 value: 12 rows and 314 columns.

2.4.4 Extraction of Feature vector of Signal

The entire process of feature extraction in modular form is expressed in Figure No. 2.13

The block diagram in Figure of Front End processing of ASR theoretical and mathematical study for this thesis work is explored from the Rabiner and Sletzer articles and Rabiner Text book for Speech Recognition. From the extracted MFCC feature the acoustic scores are computed using GMM which is shown in the Figure below. 39 coefficients of MFCC are computed using statistical method and using delta and double delta coefficients for information about the speech. The computed feature vectors are again used in statistical method of extracting mean and variance these mean and variance values are represented as the acoustic score by Gaussian which are observation probabilities of articulatory moments and their features in terms of speech signal as frequency component is the main gist of concept in HMM-GMM.

Extracted GMM from the signal information is mapped to the state sequence of the HMMs during training process. Here Baum Walch re-estimation algorithm is used to compute the HMM in context Independent and Context Dependent for computing

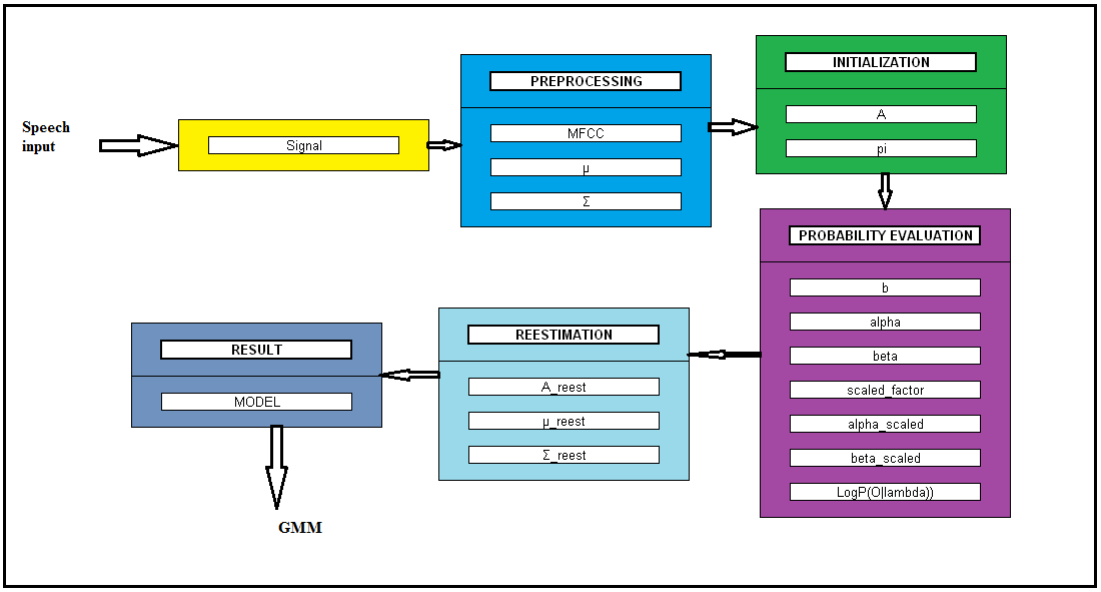


Figure 2.14: Modularised Speech signal feature extraction in terms of GMM.

the models for speech utterances. The mathematical study is from lectures of Prof Yegnanarayana and Rabiner text book gives complete mathematical explanation that required to understand the method followed in this thesis work.

2.4.5 π_i initialize the initial state distribution vector, using the left-to-right model

The initial state distribution vector is initialized with the probability, to be in state one at the beginning. This is assumed in speech recognition theory [RAB89]. It is also

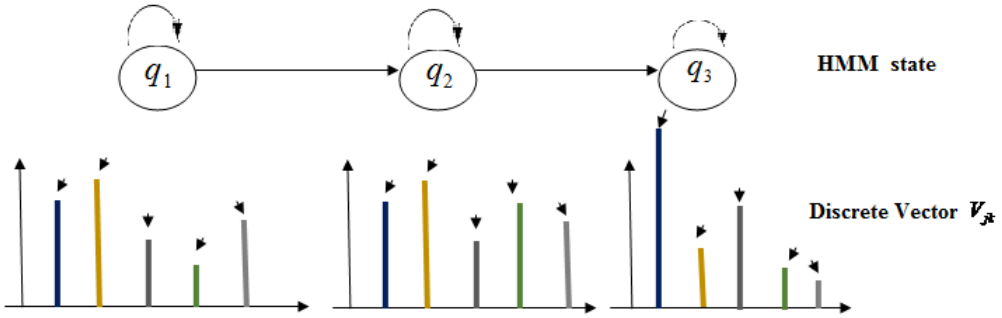


Figure 2.15: Left to right model state diagram to initialize π[RAB99]

assumed that i is equal to five states in this case.

$$\pi_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.6}$$

$, 1 \leqslant i \leqslant number of states, in case i = 5.$

Hidden Markov Model based ASR:

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 ASR. HMMs provide 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 HMMs consists of two processes namely Hidden and the Observed process, that level aliasing noise will happen. For example Signal X is sampling at 10KHZ, for, $m = N - 1$ are calculating the frequency.

2.5 Markov Process

A Markov process is a stochastic process with Markov property. The Markov property states that conditional probability[KUM13] of appearance of future state of a process, given the present and all the past states only depends on the present state and not on any of the past states. In other words, the future state is independent of the path of the process (past states of the process). If $X_1, X_2, X_3, \ldots X_N$ are the random variables which represents the occurrence of states $s_1, s_2, s_3, \ldots s_M$ then

$$P(X_i \mid X^{i-1}) = P(X \mid X_{i-1}) = P(X_i = s_p \mid X_{i-1} = s_q) \tag{2.7}$$

where $X^{i-1} = X_{i-1}; :::::; X_3, X_2, X_1, i < N$ and the random X_i and X_{i-1} represent the occurrence of state s_p and s_q where $p, q \in \begin{bmatrix} 1; & M \end{bmatrix}$ respectively. The probability $P(X_i = s_p = X_{i-1} = s_q)$ is called the transition probability between the states p and

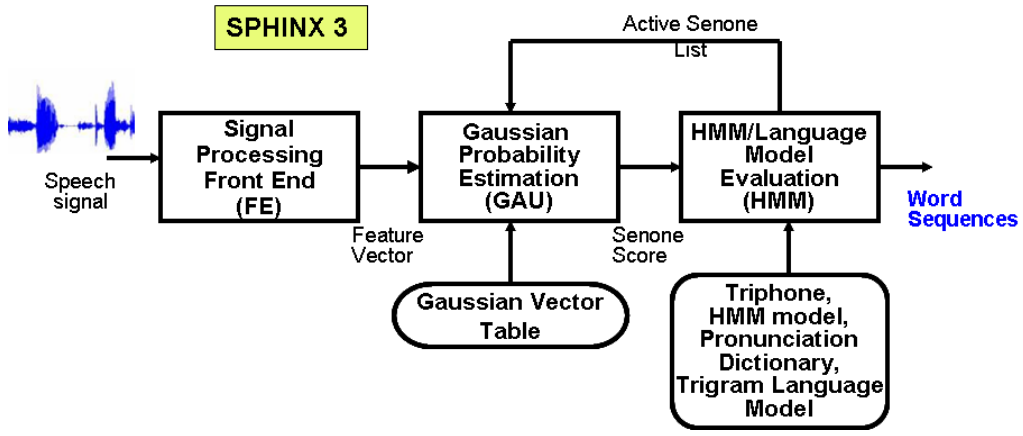


Figure 2.16: Phonetic Engine sphinx HMM for phoneme models [NAG10]

q and is represented by a_{pq} The other properties associated with Markov process are

$$\sum_{i=1}^{M} a_{ij} = 1; \tag{2.8}$$

for all i and $a_{ij} \geqslant 0$

$$\sum_{i=1}^{M} P(X_1 = s_j) = 1; \tag{2.9}$$

Both equations represent basic principles of probability theory, that the sum of the transition probabilities from one state to all of its possible next state is one and the sum of the probabilities of the initial state being in any one of the M states is also one [RAB09].

HMM for build of ASR system in training process to generate the triphone models in acoustic model. The phone indendetly as Isolated phone and Phone in context with left and right phone constitute the phone network in word using context dependent model.

In addition, the speech feature vectors are closely approximated by Gaussian distributions[NAG10]. If we assume that bw $b_w(i, x)$ is the output probability distribution of the i_{th} state of recognition class w (which could be a word, syllable, or phoneme), then we can represent $b_w(i, x)$ as GMM to map the HMM state sequences.

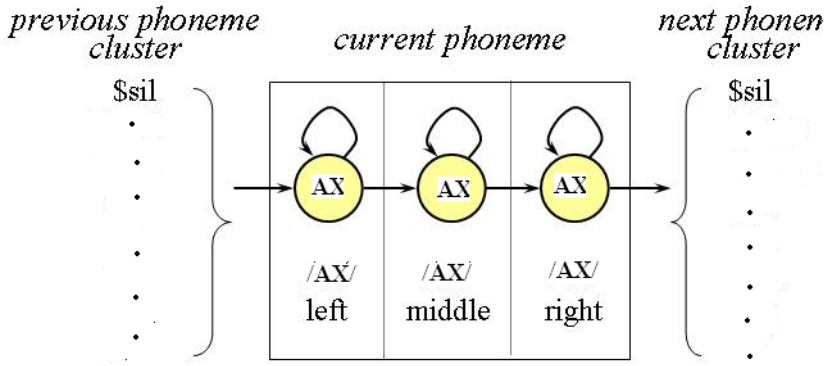


Figure 2.17: HMM phoneme model using triphone where phoneme position initial, final and middle context consider in the model building.[Nagamani, 2010]

2.5.1 Hidden Markov Models (HMM)

The state sequence of the left and right context of HMM state sequence with the observation probabilities to compute HMM-GMM is shown in Figure below and exploring the concepts from Rabiner Text book.

Nodes denoted as circles are states. O1 to O5 are observations. Observation O1 takes us to states S1. aij defines the transition probability between s_i and S_j . It can be observed that the states also have self transitions. If we are at state S1 and observation O2 is observed, we can either decide to go to state S2 or stay in state S1.

2.5.2 A hidden Markov model is defined by specifying five things:

1.

$$\textbf{A set of states s : s} = \{s_1, s_2, \ldots s_N\} \tag{2.10}$$

where N:number of states in the model where states, and state at time t, $q_t \in S$

2.

$$v = \{v_1, v_2, \ldots V_M\} \tag{2.11}$$

where M:number of observation symbols at time t, $o_t \in v$

- **State transition probability distribution :**

$$A = a_{ij} \tag{2.12}$$

$$a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i), 1 \leqslant i, j \leqslant N \tag{2.13}$$

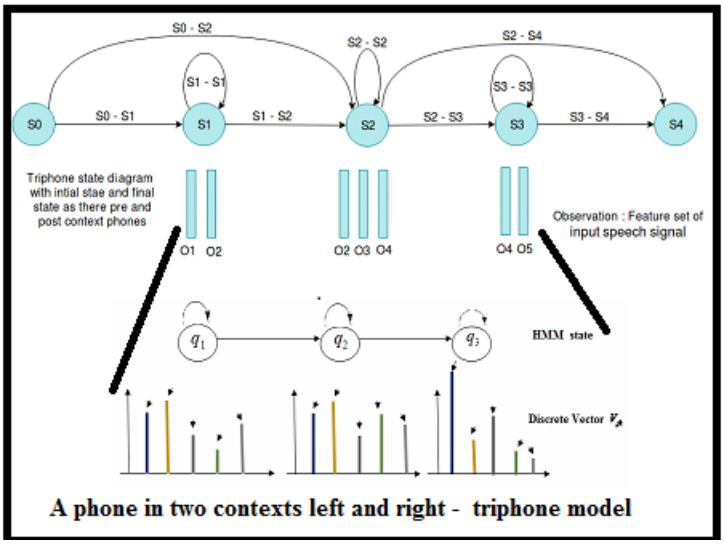


Figure 2.18: HMM-GMM state transition probabilities estimation to compute tri-phone model.

- **Observation symbol probability distribution:** B = $b_j(k)$
  Discrete observations from V are

$$b_j(k) = P(v_v att \mid q_t = s_j), 1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M \tag{2.14}$$

  Continuous observation:

$$o \in R^d \tag{2.15}$$

  Gaussian:

$$b_j(o) = N^{(\mu, \sum; o)} \tag{2.16}$$

  Mixture of Gaussian:

$$b_j(o) = \sum_{k=1}^{k} w_k N^{(\mu_k, \sum_k; o)} \tag{2.17}$$

- **Initial state distribution**

$$\pi = \{\pi_i\} : \pi_i = P(q_i = s_i), 1 \leqslant i \leqslant N \tag{2.18}$$

  Thus the HMM model is defined as :

$$\lambda = \{S, V, A, B, \pi\} or \{A, B, \pi\} \tag{2.19}$$

2.5.3 HMMs and their three problems [SAK09]

The Evaluation Problem (Scoring): Given an observation sequence $O = \{o_1, o_2, \ldots O_r\}$ and a model $\lambda = \{A, B, \pi\}$ how do we compute $P(O \mid \lambda)$, the probability of the observation sequence ?

Solution: 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 (Matching): Given an observation sequence $O = \{o_1, o_2, \ldots, O_T\}$ how do we choose a state sequence $Q = \{q_1, q_2, \ldots, q_T\}$ which is optimum in some sense?

Solution: The **Viterbi algorithm[RAB89]** 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 (Training)[RAB89]: how do we choose a state sequence $\lambda = \{A, B, \pi\}$ to maximize $P(\mathbf{o} \mid \lambda)$

Solution: The **Baum-Welch algorithm[BAU66][RAB89]** (or the forward-backward algorithm) which can find the model's parameters so that the maximum probability of generating the observations for a given model and a sequence of observations [BAU66].

2.5.4 Algorithms involved in making HMMs work.

1. For a given input speech signal estimate the conditional probabilities by using probability calculation.

2. Finding the path for a given input sequence that is closely matching with input vector to assign a state in HMM for training and then find the best path through the designed model

3. To train a model, first estimates the Gaussian parameters with statistical means and co variances and transition probabilities of a given data set for the best model.

   (a)

   Input: O **sequence of indexes**$\{1 \ldots M\}$
   of spectral vectors from a code book. (2.20)

(b) **model:** each word W in the vocabulary modeled through a $N-$ state $HMM : \lambda(W)$

(c) **Training:** $\forall W$ a training sequence O_w consisting of a number of repetitions of pronunciations of W, possibly uttered by different speakers, is employed to estimate such that:

$$\lambda_w = \arg_\lambda(O_w \mid \lambda) \tag{2.21}$$

(d) **Model Design:** Each training sequence O_w for W is segmented into states; in this way an understanding of the physical meaning of the model states is possible $\Rightarrow$ refine the model.

(e) **Recognition:** given a test observation sequence O, each model is scored and the winner determines the most likelihood word hypothesis.

$$W = \arg_w \max P(O \mid \lambda_w) \tag{2.22}$$

The development of an ASR system consists of four essential steps the are explain in the following section

2.5.5 Acoustic Modelling:

In this stage an acoustic model is created for each recognition unit (also known as sub-word units, e.g., phones). In most current state-of-the-art ASR systems, the acoustic models are based on the Hidden Markov Model (HMM) paradigm [RAB09]. HMMs model the expected variation in the signal statistically. Probability density functions for each sub-word HMM are estimated over all acoustic tokens of the recognition unit in the training material. Each basic sound in the recognizer is modelled by a HMM as shown in the beneath Figure.2.20

The stochastic models can be used to represent the words to be recognized. The word sequence $W*$ that has the highest posterior probability $P(W \mid Y)$ among all possible word sequences is given by,

$$W^* = \arg\max_w P(W \mid A) \tag{2.23}$$

where P represents the Probability. This problem can be significantly simplified by applying the Bayesian approach to find w_*

$$W^* = \arg\max_w \frac{p^{(A|W)P(W)}}{P(A)} \tag{2.24}$$

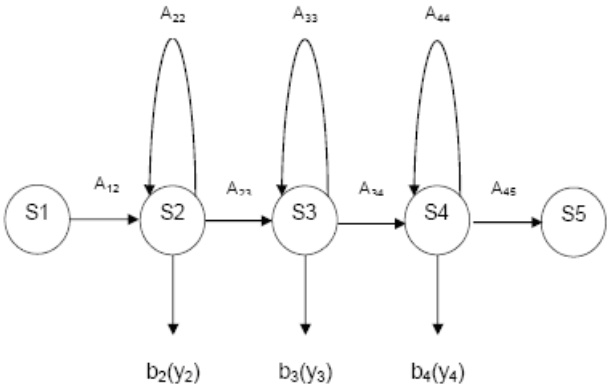


Figure 2.19: sound in the recognizer is modelled by a HMM.

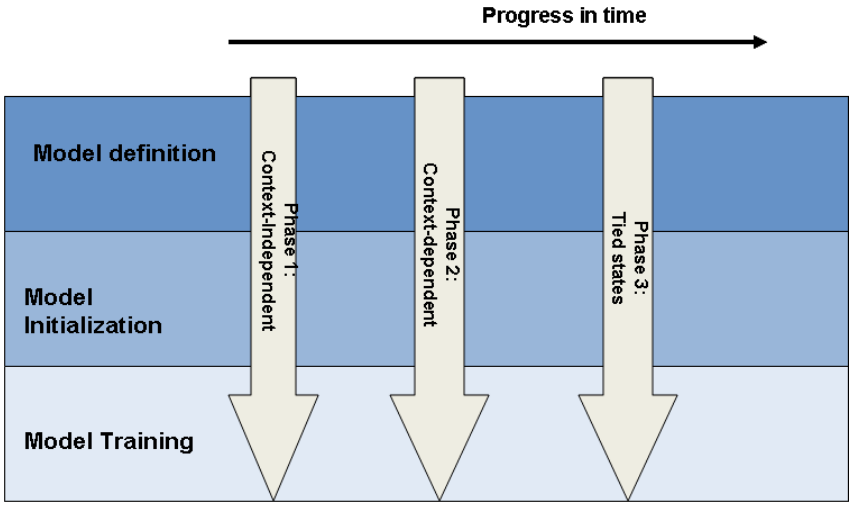


Figure 2.20: Acoustic Model with Model Initialization, definition and Training[NAG10].

Since $P(A)$ is the probability of the acoustic vectors which remains the same for all the possible sequences it can be omitted. The equation reduces to

$$W^* = \arg\max_w P(A \mid W)P(W) \tag{2.25}$$

2.5.6 Language Modelling:

An **N-gram language** model is used to in the work. Most cases Isolated word unigram, bi-gram and tri-gram is comuted for the input collected for the thesis work. Language model is not the much scope of the thesis most cases Isolated word and with in word variation of speech studied in the current work. For understanding of the ASR system simple information presented [SAM00] as

$$p_N(W) = p(w_1, w_2, \ldots, w_N) = \prod_{i-1}^{q} p(w_i \mid w_{i-N+1}) \tag{2.26}$$

Where $F(w_1, w_2, w_3)$ is the frequency of occurrence of the trigram model and $F(w_1, w_2)$ is the frequency of occurrence of the bigram model.

$$P(w_1, w_2, w_3, \ldots) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid P_2) \tag{2.27}$$

The value of N determines the number of probabilities to be estimated.

2.5.7 Search:

In ASR system HMM method using Bayes' Rule[RAB93] extracting the text for a given input speech signal. The speech the parameters extracted by FEP module as MFCC feature vector. This given feature vector ASR decoder do the search to best suitable text corresponding to the input. The knowledge source from AM,PM and LM used to map the correct word sequences using Viterbi search technique.

2.6 ASR System Design Issues:

The design issues in ASR system[LEE96] are described as

1. Speech recording and Speech enhancement: All ASR systems rely on capturing the speech signal through a microphone. A wide variety of microphones, ranging from high quality fixed-mount microphones to low cost telephone handsets, have

been used for signal capture. Speaker phones and wireless handsets are also becoming popular. The direction of the incoming speech signal and the distance between the sound source and the microphone determine the quality of the signal captured. For a real-world application, there is also a possible mismatch between the type of microphone used for training and testing. It is important to make the transducer part of the recognition system design. Hands-free signal capturing devices, such as a microphone array, have been used to track talkers and to enhance the signal to noise ratio in experimental recognition systems.

2. Recording environment and its Robustness : ASR System performance robustness is a major problem that prevents widespread deployment of ASR systems today. The pattern matching paradigm requires the training data to cover all possible acoustic variability in the operating environment. When acoustic disagreement between training and testing conditions occurs, the performance of a speech recognizer is degraded. Although there exists many techniques for dealing with some of the robustness problems, new algorithms must be developed to handle the variability caused by talkers, speaking environments, transducers, channels, speaking style, context and dialect, etc.

   It is believed that no single robust feature set will solve the robustness problems. Compensation techniques for the existing features and models are now beginning to emerge. . Alternatively, these parameters can be estimated based on blind equalization (e.g. the popular cepstral mean substration algorithm), or based on model-based equalization in which the compensation is treated as a nuisance parameter and the estimated together with the recognized string during recognition [DEN13].. In some cases, the compensation can also be introduced in the stochastic models used for recognition. Again, this compensation can be done without assuming any knowledge about possible mismatches between training and testing (e.g. minimax classification algorithm [ANU09]. We expect more algorithms to be developed in combination with adaptation techniques to improve the robustness of ASR systems.

3. Corpus building limitation: Speaker availability, application specific text corpus building, environment condition are the time consuming and human intervention tasks hence it is expensive. Speech variant condition recording is also another

challenging task.. Both acoustic and language model adaptation can also be studied by collecting a small number of application specific examples.

4. Static versus Dynamic System Design: Initial design suits for static design strategy in that all the knowledge sources needed in a system are acquired at the design phase and remain the same during use[LEE99]. Since the samples used in the design are often limited, these results in some mismatch problems. A better way is to acquire the knowledge dynamically. New information is constantly collected during development and is incorporated into the system using adaptive learning algorithms.

   Spontaneous Speech and Keyword Spotting: Spontaneous speech is different from read speech in that extraneous speech events are contained in addition to the message that is intended. False starts, disfluency, um's and ah's lip smacks and out-of-vocabulary words are a few of the examples of difficulties to be expected in spontaneous speech. The ASR system is expected to recognize the meaningful keywords embedded in fluent speech and ignore all the other speech events. High performance keyword spotting has been achieved in a ASR system used in telecommunications based on a five-word keyword recognition task [SAD05]. However, for spotting 20 keywords in fluent speech, the performance is not nearly as good. Keyword spotting in large vocabulary, continuous ASR is therefore an important research area. Accurate rejection of extraneous speech events is an important research topic and is needed to enhance our capability in dealing with spontaneous speech. Accurate detection of incorrectly recognized and partially recognized utterances (often referred to as utterance verification is also a new research area for designing more flexible and intelligent user interfaces for spoken dialogue systems.

5. Human Factors Issues: In addition to improving the quality of the ASR System technology, good GUI, intelligent command and control, speech enhancement and recovery improve the performance.. Spoken dialogue coupled with utterance confidence measures can help solve some of the problems. Research in human factors issues will help bridge the gap between the performance that can be achieved in the laboratory and what is achieved in a field application.

The speech signal is highly variable. Some of these variations are speaker-dependent and others are not. Generally, any kind of variations decreases the performance of ASR system, if they not treated carefully, a way for compensating for some of these variations through creating pronunciation variants of the words involved in an ASR system lexicon. This is done by employing a hybrid word pronunciation variation generator scheme which considers some of the phonetic deviations that occur because of intra-speaker variations. This phenomenon is a result of differences between speakers in respect to their voice quality, speaking style, dialect, etc. These factors cause partial deviations of acoustic features. These partial deviations cause systematic errors in the phone recognizer module and as a consequence, phonetic deviations occur in the recognized strings of phones. Phonetic variations occurring because of speaker varieties are a well known phenomenon arising due to different mental and physical specifications of the speakers that affect their speech generation. In ASR systems, to resolve the problem by incorporating phonetic realization in lexical model. In the literature, this method is referred to as explicit phonetic variation modelling. A distinction should be made between techniques that model pronunciation, in the sense that words may be pronounced with various allowed phonetic forms, and those that model phone recognizer errors resulting from partial pronunciation variation in acoustic features. Fig.1 shows the two mechanisms which cause phonetic deviations arising from speaker pronunciation.

2.7 Causes for speech variations

2.7.1 Intrinsic Speech Variations in ASR System

At present most important growth, in the ASR system and spoken language system are captured using technology [SAI15]. The other reason may be the feeble grammatical and semantic acquaintance. This thesis is also emphasized the shortcomings of original speech intrinsic variations. Precise feature of the speech signal of the definite population such as foreign dialect is excluded so also a few applications, pin point on the core recognition technology such as directory assistance for the reason of vivacious vocabulary. The understanding of speech is dependent on other features like (i) Geographical Area wise (ii) social and economic background and their spoken language (iii)

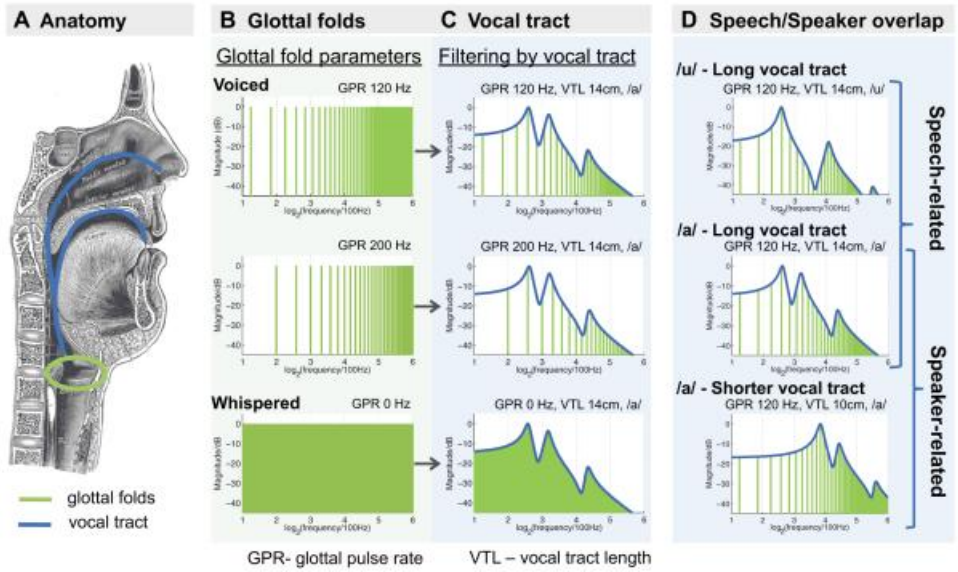


Figure 2.21: Anatomy of Vocal track length with speech and speaker variability [MEN01]
.

Surroundings of living (iv) speaker himself. The listed factors generate vast differences: (a) speaker, (b) Gender (male, female) (c) Rate at which the verbal communication is made (d) vocal effort, (e) local dialects (f) speaking method, (g) Resident, non (native).

2.7.2 Variation in Speech

Speaker length influence on speech variability with speaker specific parameters is explained in this section with analysis with physiological background shown above Figure 2.21. the involvement of glottal fold, and vocal tract factor to the speech yield. Sagittal cross section of the head and neck of human being is shown in the part A of the figure. The extent of the vocal tract is represented in blue lines and the glottal folds in green lines. The vocal tract extends is shown from the glottal folds to that of tip of nose and lip. B, the three different sound plots are representing in the part B of the figure as by glottal fold parameters. The vibration caused in the glottal folds due to lower voice speech (120 Hz GPR, top) or higher voices (200 Hz GPR, middle) is represented. The tiny glottal folds create whispered speech (0 Hz GPR, bottom). Amplitude peaks can be noticed with frequencies is presented if part C figure. The sound waves originating from the glottal folds is filtered at vocal tract and is indicated “formants”, in blue lines.

It is pertinent to mention here that the formant position is not effected by the dissimilar glottal fold parameters, but the corresponding speech and speaker vocal tract factor have impact on formants, which is represented in part D of the figure. The part D of the figure, the shorter and longer extent of vocal tract formant shifts connected with the speech sounds /u/ and /a; and, an /a/ is presented.

2.7.3 Speaker characteristics

The speech signal is related to the parameters such as the sex, age, brought up, nativity, language dialect, and physical condition influence the linguistic information so also the speaker traits, information. The speaker distinctiveness domino effect is from the multifaceted amalgamation of physiological and cultural plane [MOO94][JEL76]. Physiological divergence is for Gender and culture is for the accents respectively are significant prime factors. The physiological factors are presented in this chapter. The vocal organs versatile shapes of the speaker decide unique "timbre". The pitch and essential speaker information due to the vocal card size specific to speaker which is explain with vocal card tube experiment in [SHA05]. Where are the language specific units are independent of the pitch only depends on the articulatory system generated secondary harmonics generally called as formats. With formats the phonetic feature extraction and their variation due to the articulatory production is another dimension look into language specific variability in speech. The comparative position of the formant frequencies is constant for the specific sound of various speakers, and as a result, absolute formant positions are speaker-specific. These notes are agree with the acoustic theory be relevant to the tube resonator model, of the vocal tract. This affirms that location of the resonant frequencies is inversely proportional to the length of the vocal tract. Different techniques are developed with this principle to enhance the toughness of ASR System, to inter-speaker variability. The hold of lesser frequencies for carrying the linguistic information evaluates both by perceptual and acoustical analysis substantiates the accomplishment of the non-linear frequency scales such as Mel, Bark, Erb. Rest of other approach intend at construct acoustic features invariant to the frequency warping. Different vocal tract length normalization (VTLN)[JOS11] techniques are leading on direct application of the tube resonator model [HIS77] (i) speaker-dependent formant mapping (ii) transformation of the LPC pole modelling (iii) frequency warping, either linear or non-linear, all consists in transform the position of the formants to get

nearer to an ”average” canonical speaker. The speaker specificities are diminish in the general adaptation techniques and have a tendency to lessen the breach among speaker-dependent and speaker-independent ASR system, by adapting the acoustic models to a particular speaker [SHI01]

2.7.4 Foreign and regional accents [BEN06]

The accent is one of the important constituent of inter-speaker variability [HAM12]. The performance will be diminishing in identifying the accented speech and further diminish for non native ASR when compared to native ASR [KSA98][NAG10]. The shift is extremely changeable for the foreign accents. This is prone by the dweller language. It also depends on the intensity of the voice of speaker. Improved modelling adopted to speaker specific parameters incorporateded in lexical model. [SHR13] [DEN08] and the study is explored in the review of Ramya Rasipuram works. [RAS16]. Addition of so many systematic pronunciation variants will be harmful [SHR13]. Irrespective of the dialect data, the Non-native ASR model is unable to tackle by native speech models [HAM12]. Speakers who doesn't know the language can restore an unusual phoneme in the object language, and difficult to create phoneme inventory in native language. [BAS11][WEI89]. It is not possible to handle triphone based modelling for the alterations made with respect to the sounds restored by other sounds modeled [PRA08]. For handling the phoneme of non-native speakers multiple phonetic transcriptions is adopted [SIM06]. Multilingual phone models are explored in the anticipation of achieving language independent units. Accent categorization is learned for many years [DAV09]. The ASR technology is also applied in foreign language learning for ranking the excellence of the pronunciations.

2.7.5 Speaking rate and style

The rate of speaking is the number of syllables or sub-words units produced per second. The sub-ward may be a phoneme or morphonemes. It is crucial factor of intra-speaker variability. The timing and acoustic rate of the syllables are effected when there is increase of speaking rate due to automatic articulatory mechanism precincts. For the most part familiar methods in a sentence depend upon the assessment of the frequency of language specific unit as syllable or phoneme [BAK95], in the course of a beginning segmentation of the test utterance. In informal situations or duration parameter,

slurring pronunciations of certain phonemes certainly occur, which build on the speech redundancy, in addition to physiology. In dissimilarity, speech segment where confusability is high likely to be uttered more cautiously [RAS16] [DAV09]. When there is a noisy surrounding the speakers unintentionally adjust to increase the voice for clear communication, which is termed as the 'Lombard reflex'. The above explained are vital for speaking manner specificities as well as unprompted speech modelling. The spontaneous speech is mainly concentrated on the pronunciation studies and its increase in the precision. The dependency of pronunciation concerning the structure of the syllable is focused. [LEE01] [HUA01]. As a result, expansion of acoustic modelling reliance to the phoneme location its co-orticuation effect in sentence is reviewed in [GAL01]. The scope in this thesis is not reflected the dis-fluencies which are present, like false starts, recurrence, faltering and filled pauses.

2.7.6 Speaker Age:

The variability and mismatch is relates to the age of a person, which play vital role in ASR systems for the physiological nature [DEN08]. The vocal tract, vocal folds may vary from Children to adults, as a consequence elevated place of formants and fundamental frequency occur. Apparently, juvenile children will not have truthful pronunciation and children utter language in a diverse way. Unprompted speech will have fewer grammatical than for adults. Efforts have been put on numerous studies to deal with this predicament with the implementation of acoustic features of children speech matching to that of acoustic models prepares out of adult speech, such as vocal tract length normalization (VTLN) is reviewed in[HIS77] as well as spectral normalization in [YMA16]. Merely training a usual speech recognizer on children speech will not be adequate to accept high accuracies, Wilpon and Jacobsen [WIL96].

2.7.7 Emotions

Emotional state has substantial influence in the speech spectrum and has impact on the features take out of the speech, which has direct affect on all ASR systems including the "stressed" speech signal so also to develop man-machine communication. The intrinsic variables that relate to kind of stressed speech are (i) loud (ii) soft (iii) Lombard (iv) fast (v) angry (vi) scared and (vii) noise. Hansen work out for robust recognition

under three fields in [LEG95] [SIM06], (i) good training process, (ii) enhanced front-end processing, and (iii) improved search techniques.. There will be insensitivity and fallout due to shift from noted data to that of original training data. Then come to work out feature extraction technique in recognition both stressed and non-stressed speech so also to develop the robust recognition. Adapting to model structure in the recognition system is required to cater tom the input signal variability. With space projection, the drawbacks of the approach can be improved by bringing the training and test conditions closer.

Techniques like additional acoustic cues, adaptation at knowledge base, compensation at signal level, fine tuning the models are done to cater to the speech variation problems. The commoditization of computational resources, would lead to building better systems for handling increasingly large amounts of training data [GEL76]. This drives ASR research with multidisciplinary disciplinary angle to improve knowledge base from linguistic, physiological, Cognitive science and statistical approach to deal the problems.

2.8 Summary

ASR technology in the language specific directions is the important reviews in this work. Preliminary study made on significant advances in almost all areas of ASR to choose the right technology to incorporate the linguistic directions is major contribution. Language specific ASR especially in Telugu and its utility in e-learning, web application, e-governance domain to interact with speech are gaining importance. Not only the fundamental issues of ASR, but also variant factor in real and laboratory recorded speech parameters discussed. The research challenges that are used to bridge the gap between and a laboratory system and a real world ASR System application are: (a) building a robust ASR to cater to the speaking conditions in speaker independent system for with more data (2) Robust utterance verification to extract relevant information in Isolated speech and read speech conditions (3) Adaptive speedy ASR systems to meet the requirements of changing tasks, speakers and speaking environments.

****

Chapter 3

Linguistic concepts of ASR system

3.1 Introduction

In this chapter the importance of study of the language and its parameters like phonetics and phonology, orthography, region of the people who are using in region and their socio linguistic influence, with in the language variations, influence of other language are few notable ones, to enable ASR system in research directions is given. The history and its classifications of language is made, before analysing the target Indian language of this thesis work viz., Telugu Language. Research work is more concentration on linguistic parameters starting from language beginning layer of phonetics of speech, its main inter-disciplinary domains in computational linguistic and speech technology are critical areas of research on speech recognition. As India is multi lingual country, the area of works on signal to symbol transformation on specific languages is still at its nascent stage.

It is essential to go through the concepts on phonetics, phonology, psychophysiology, phonemic and phonetic, human physiology related to acoustic phonetics of language to comprehend the signal to symbol transformation. The focus on Telugu language Automatic Signal to symbol transformation and its refinement, pronunciation variation modelling are important subject of study. The present chapter cover the essential and over view of Telugu language, their phoneme classifications based on physiological characteristics and acoustic phonetic concepts, which are required for codification of

ASR system.

3.1.1 Language and its organization in speech context

Language is two direction written and spoken form. Mapping between the written form to spoken form helps better way of communication. By linguistic research, the mapping sound to meaning is done. The sound unit has a distinct acoustic phonetic feature values that separates it. The sounds are building blocks that correspond to the unit of written form of the language. Every distinct sound have their acoustic interpretation as their written form symbol. The speech perceptions is the process of mapping this distinct features acoustic phonetic values corresponding to the state symbol that represent to the written form of unit. This make the system to predict the corresponding state symbol from input features of the sound corresponding feature values. The phonetic or phoneme inventory enables to map the sound corresponding symbol. Hence the Distinctive features enable the phoneme inventory process in all the languages. The phoneme inventory is process of expecting the acoustic segment featured values that are distinct to its corresponding distinguishable written symbol in any languages. The sequences phonemes constitute the word with in the words sequence [BAS11] that formulated in a language to convey the information and perception of meaning is through this building block of words. Arrangement of segmentations into corresponding speech units in a language are represented as syllables. This syllables are depends on the structure of the individual language. Depending upon the language structure the syllable formation is allowed. For example few languages, syllable types are consonant (C)–vowel (V), in most of the Dravidian language, however, other languages like English, have complex syllable structure. Hence syllables specific parsing in a language with manageable chunks of speech steams for facilitating analysis are explored, reviews on language specific systems [KAV06][KAL10] language model are discussed. Prelexical phonological analysis Features is represented segmental part of the syllabic unit. The sub-unit of syllable is the morphonemes which are interfacing these units into a word. Hence in word segmentation morphonemes plays major role. As per the Psycholinguistic and neurolinguistic concern that provides empirical evidence that the morphemic structure in a language has an active role for recognition word. Next level of phonetic and phonology of the language play a role for extract information with syntactic and semantics' of language sentences and further phrases in a language. The major concern

of the thesis is service of lexical access. [MCC43]. The study of the lexical-level and prelexical level is more concern of the present chapter.

3.1.2 Artificial Intelligence in the context of Language

Present generation of the system are AI. Here the intelligence is defined" as the ability to reason and to acquire and apply knowledge as well as the abilities to perceive and manipulate physical object. These abilities require a variety of methods for representing and processing information." Artificial Intelligence seeks to process knowledge as opposed to information. Traditional computing methods only concern with calculation and storage of data but the AI emphasis on shifting the use of symbolic representations of knowledge. Hence to use this knowledge reasoning techniques must be found. AI researchers attempt to formulate this knowledge representation and reasoning techniques in process of language is the context of the chapter

3.1.3 Study of language in context of ASR System

There are around 6912 spoken languages available in the world for the purpose of communication.[NAG12]. The digitization of these spoken languages supports many human need domains such as (i)Ubiquitous information access, (ii) phone-based information access, (iii) cross-cultural human-human interaction, (iv) Human supporting multilingual communities that are in need to communicate with local people in their mother tongue will specifically demand the regional language specific ASR systems. Developing such systems which facilitates to interact in Telugu Language is the scope this thesis. Indian language context according to the professor Kavi Narayana Murthy there are around 120 spoken languages available which shown in Table 3.1

S.No	LANGUAGE	S.No.	LANGUAGE	S.No.	LANGUAGE
1	Adi	41	Khasi	81	Munnda
2	Anal	42	Khezha	82	Mundari
3	Angami	43	Kheiemnungan	83	Nahali
4	Ao	44	Khonds/Kondh	84	Nepali
5	Assamese	45	Kinnauri	85	Nicobarese
6	Baiga	46	Koch	86	Nissi/Dafla
7	Balti	47	Koda/Kora	87	Nocti
8	Banjara	48	kolami	88	Odia

9	Bengali	49	Kom	89	Paite
10	Bhili/Bilodi	50	Konda	90	Parji
11	Bhumij	51	Konkani	91	Pawi
12	Bhutia	52	Konyak	92	Phom
13	Bishnupriya	53	Korku	93	Pochuri
14	Bodo	54	korwa	94	Punjabi
15	Chakru/Chokri	55	Koya	95	Rabha
16	Chang	56	Kui	96	Rai
17	Coorgi/Kodava	57	Kurukh/Oraon	97	Rengma
18	Deori	58	Ladakhi	98	Sangtam
19	Dimasa	59	lahnda	99	sanskrit
20	Dogri	60	Lahuli	100	Santali
21	Gadaba	61	Lakher/Mara	101	Savara
22	Gangte	62	Lalung/Tiwa	102	Sema
23	Garo	63	Lepcha	103	Sherpa
24	Gondi	64	liangmiei	104	Shina
25	Gujarathi	65	Limbu	105	Simte
26	Gurung	66	Lotha	106	Sindhi
27	Halam	67	Lushai/Mizo	107	Tamang
28	Halbi	68	Mahal	108	Tamil
29	Hindi	69	Maithili	109	Tangkhul
30	Hmar	70	Malayalam	110	Tangsa
31	Ho	71	Malto	111	Telugu
32	Irula	72	Manipuri	112	Thado
33	Jatapu	73	Mao	113	Tibetan
34	Juang	74	maram	114	Tiripuri/Kokborok
35	Kabui	75	Marathi	115	Tulu
36	kannada	76	Maring	116	Urdu
37	Karbi/Mikir	77	Miri/Mishing	117	Vaiphei
38	Kashmiri	78	mishmi	118	Wancho
39	Khandeshi	79	Mogh	119	Yimchungre
40	Khari	80	Monpa	120	Zemi/Zeme

Table 3.1: Spoken language in India[According to Prof. Kavi Narayana Murthy,UOH].

3.1.4 Language usage in human society and its digitization

In the world Language Families [KAV06], the explanation of the multiplicity of languages in the world today is the combination of social separations between people and constant language change, one or many thousands of years. Languages are always changing, and populations speaking one language have, at least in the past, repeatedly

split up, with the result that the separation plus language change is at first dialects, mutually intelligible varieties of a language. With the passage of time, however, wider social separation between dialect groups and continued language change results in new languages, mutually non-intelligible varieties. Table.3.2. gives the Language families across the world and their classification in context of spoken text

S.no	Family	Three languages of family
1	Afroasiatic	Arabic ; Hebrew ; Hausa(Nigeria)
2	Amerind	Navajo(Arizona); Mayan(Mexico); Quechua(Bolivia;peru)
3	Altaic	Japanese; Turkish; Mongolian
4	Austroasiactic	Vietnamese; Thai; Khmer(Cambodian)
5	Austronesian	Indonesian/Malay; Hawaiian; Tagalog(Philippines)
6	Dravidian	Tamil(India and Sri Lanka); Malayalam(India); Telugu(India)
7	Indoeuropean	-
(a)	a.Germanic	Enshish; German; Swedish/Norwegian
(b)	b.Celtic	Irish; Welsh; Breton(France)
(c)	c.Italic	Spanish; French; Portuguese
(d)	d.Slavic	Russian; Polish; Czech
(e)	e.Indo-Iranian	Hindi; Bengali(India; Bangladesh); Persian/Farsi(Iran)
8	Niger-Khordofanian	Swahili;Yoruba(Nigeria); wolog(Sierra Leone)
9	Sino-Tibetan	-
(a)	a.Sinitic	Mandarin Chinese; Cantonese; Hunanese
(b)	b.TibetoBurman	Tibetan; Burmese; Lahu(Thailand)

Table 3.2: World languages in context of grouping in their families according to their features wikipedia languages.

The highest layer of the language structure is the Pragmatics. This is the relation between language and its context of use. With the help of Pragmatics it is possible to understand how a language works and how to convey the meanings by using language layer structure.

Written form of language is another importance study necessary to facilitate the ASR system function. The human representation of language in context of writing is represented using orthography. Based on the type of orthography languages are classified. There two (i) Deep orthography and (ii) Shallow orthography. The deep

orthography also called as a defective orthography such as in English, German and Arabic. In English voiced and voiceless phonemes are Alphonse. Another example for systematic deficiency in orthographies Arabic and Hebrew scripts which are based on their orthographic writing systems (they are abdjudic) does not have short vowels.

3.2 Phonetics and phonology in the language

ASR system engineer should have this basic knowledge and the following section explore the concepts. Speech sounds are represented in the written form as orthography. Orthography is two types shallow orthography in which every sound has equal orthographic symbol. Phonetics is dealing with shallow orthography. Deep orthography is the abstract representation of language sound. Here no equal number of speech sounds and orthography. Language specific sound is a phoneme. Range of phoneme coverage in world language approximately 40 to 60 and orthography coverage are 20 to 60. Telugu is a shallow orthographic system and English is Deep orthography

3.2.1 Phonetics context of the language

The unit sound of the language and its study is the phonetics. Mapping this phonetic unit to the written script is the major concern of the language context in the ASR system. The study of the phonetic in a language structures in hierarchical frames the ASR design. Here phonetics is independent of the language. Mapping one to one for phonetic to orthography simplifies the design. But most of the case in context of ASR varies due to the pronunciation of sounds in a particular language. Mapping using phonetic symbol to the orthography need layered understanding of the system design as well language structure which abstractly maps to the sounds corresponding to that particular language. Generally, when the phonetic symbols does not match with corresponding pronunciation, hence addition symbols are used to map the Phoneme-Grapheme [RAS13] in order to describe correct pronunciation and their correspondence to text. The language specific phonemes are different based on the language structure and their pronunciation that varies for each language between 20 and 60 [SAU87]. These phonemes are mapped to their phonetic symbols for digitization. Mostly phonemes can are chosen minimal number of phonetic symbols define the word. English is the deep orthography system, there is need for the exploration of shallow and deep orthography

in context of ASR system. The language parameters considered in the existing ASR system not sufficient to the Telugu language context as Telugu is shallow orthography language. Language specific rules that are framed to the English can not be fully modelled in Telugu. Modelling in Telugu language specific rules in lexical model of ASR system demands the study of language structure and their knowledge, to map sound unit to the text basic unit. As speech is context its variant factor in mapping also involves, incorporating this variant factor using layered structure of language and phonetic representation in the system needs to look into the parameters, speaker and supra-segmental context in pronunciation of language specific sound unit (phoneme). In continuous speech, co-articulation effect of phoneme also needs to consider to modelling the variation in pronunciation [WIT82]. In addition speaker characteristics and emotions includes to describe the variation affected phonemes in their acoustic realization. Phonemes are classified in to vowels(V) and consonants (C). In Vowels, vocal cord vibrations make them into a voiced sounds and absence or presence of vocal cord vibration in Consonants make them may be either voiced or unvoiced. Mainly consonant are classified by the manner and the place the restriction of air flow and the way air release on that position

Pronunciation in a language is made independent of the language with effort of proposing a universal represent of the phonetic sounds. International Phonetic Alphabet (IPA) comprises of large set of symbols for phonemes, supra segmental, tones /word accent contours, and diacritics to represent world languages. For an example, the fricative consonants have over twenty symbols [IPA98]. Due to its complexity in use of Greek symbols makes it difficult to use of IPA alphabet in computers. For the purpose a standard ASCII code for computer and man communication. Another method is also available using IPA symbols for 7 bit printable ASCII character is called Speech Assessment Methods - Phonetic Alphabet (SAMPA) phonetic set [ALT96]

3.2.2 English Articulatory Phonetics

English language have 50 phonemes for speech and 26 Alphabets in writing. There are about the 10 - 15 vowels and about 20-25 consonants in English phonetic symbols. English is deep orthography category in which abstract level of mapping between phonetic symbols to alphabets. Hence confusions in phonetic mapping as allophones are used to represent similar sound that are distinct in Indic languages [KRI75][BHA11].

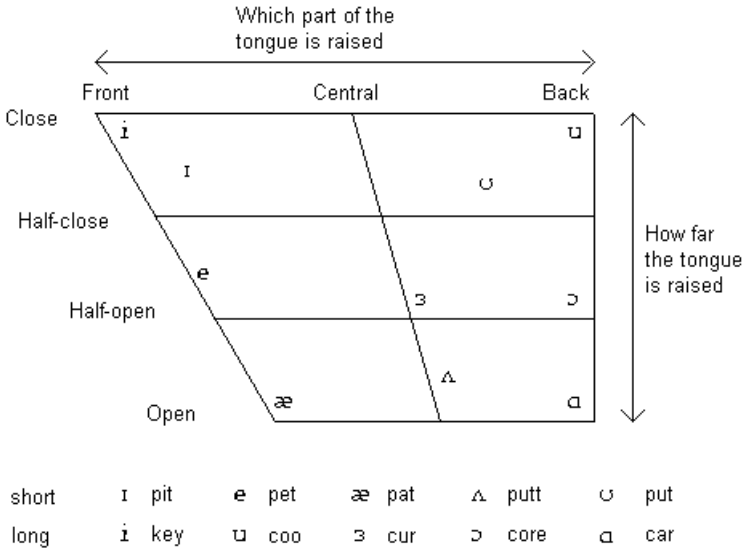


Figure 3.1: Main vowels classification in English [CAW96]

The classification of the main vowels in English [Figure 3.1], is depended on the manner of articulation (front-back) and by the shape of the mouth (open - close).

English consonants may be classified by the manner of articulation as plosives (also known as stop consonants), which is summarized in Figure 3.2.

Finally, consonants are classified based on as voiced and unvoiced.

place / manner	labial	labio-dental	dental	alveolar	palate-alveoral	palatal	velar	glottal
plosive	p b			t d			k g	
fricative		f v	q ð	s z	ʒ ʃ			h
nasal	m			n			ŋ	
liquid				r l				
semivowel	w					j		

Figure 3.2: English phoneme chart with place and manner of articulation [CAW96]

3.2.3 Phonetic transcription

Methods for phonetic transcription which is the symbolization of sounds with roman script symbols corresponding to the universal symbol sequences such as the International Phonetic Alphabet (IPA), aim at standard form of pronunciation.

The advances in algorithms, computational architectures, hardware & software platforms and signal processing has aided automatic signal to symbol transformation. This was also aided by adoption of a statistical pattern recognition paradigm, the use of stochastic acoustic and language modelling, data-driven approach with use of a rich set of speech utterances from a large population of speakers and use of dynamic programming based search methods.

Speech is natural, efficient and flexible for the human-human communication. However, the Signal to symbol transformation (SR) technology, has opened opportunities for a universal access for ubiquitous human-computer interactions. Recent studies demonstrate that machine performance is still quite far from human performance across a wide variety of effects: size vocabulary, noise environment, bandwidth, habits of speech technology use, age, social and cultural characteristics of the speaker, speech disorders, etc.[LIP97] reported that the machine performance was 43% of word error rate vs. 4% for human performance for switchboard tasks. Communicating linguistic features is just one part of the story. Faces can express emotion, a powerful and independent source of information. Eric Haseltine, Chief Scientist of Disney Interactive, has articulated the importance of emotional content in human computer interaction, argues that human communication has as much to do with speaking to the heart—the emotional content of a message—as speaking to the brain—the intellectual content[BYR06].

3.3 Phonetics and phonological concepts of Telugu language

Phonetics and phonology of a language are the basic building blocks that are used to represent in ASR system to learn as a patterns to understand the language structures. The following sections describe the phonemes of the Telugu language in Vowel and Consonants categories in detail. Basically Vowels are vocal card vibrating sounds. Consonants are the sounds produced by restricting the air passage in the articulators in vocal and nasal cavities.

Telugu Vowels : /అ/, /ఆ/, /ఇ/, /ఈ/, /ఉ/, /ఊ/, /ఎ/,/ ఏ/, /ఐ/, /ఒ/, /ఓ/, /ఔ/, /అం/

s.No	1	2	3	4	5	6	7
Telugu script	అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఋ
UOH	*AX*	*AA*	*IX*	*IY*	*UH*	*UA*	*RH*
RTS	*a*	*A*	*i*	*I*	*u*	*U*	*R*
Telugu script	ఎ	ఏ	ఐ	ఒ	ఓ	ఔ	అం
UOH	*AI*	*IA*	*AY*	*O*	*OA*	*AW*	*AM*
RTS	*e*	*E*	*ai*	*o*	*O*	*Au*	*aM*

Figure 3.3: Telugu phonemes Vowel category with their Roman representation in UOH and RTS forms

3.3.1 Vowels

Telugu Vowels :

Almost, like the five vowels of English, the Telugu has five vowel groups. For each group of Telugu vowel sounds, there is corresponding English vowel sound , as under:

3.3.2 Consonants

Consonants, the Sanskrit name is “vyaM-ja-na” (manifester). As a consonant cannot be pronounced without the help of a vowel, the vowel అ(a) is used as the default vowel that goes with consonants.

Telugu Consonants : The study of phonetics and phonology helps to understand error in ASR system hence through study on the phonemes and their properties helps in resolve ASR error hence here included details.

Telugu phoneme chart with script symbols and transcribed with roman script Consonants are now introduced below in phonetic groupings, not in the traditional alphabetical order. This study helps to understand the substitution errors in ASR system

S.No	1	3	5	7		
Telugu script	అ	ఇ	ఉ	ఋ	ఎ	ఒ
UOH	*AX*	*IX*	*UH*	*RH*	*AI*	*O*
RTS	*a*	*i*	*u*	*R*	*e*	*O*
	low-mid/mid central un rounded	high fron unrounded vowel	high back round	Tongue is retroflexed instread of firm contact		
	expanded throat utterence	similar to palatal	LABIAL	Retroflex flap		
Corresponding semi vowel		/య/, [YAX]	/వ/, [VAX]			

Figure 3.4: Corresponding English vowel sound to Telugu

The unvoiced unaspirated plosives shown in figure No.: 3.5

The unvoiced aspirated plosives ఖ, , , , (kha, cha, Tha, tha, pha) shown in figure No.: 3.6

The voiced unaspirated plosives shown in figure No.: 3.7 The voiced aspirated plosives ఘ,ఢ,ధ,భ(gha, Dha, dha, bha)shown in figure No.: 3.8

The nasals M,ఞ, N, న,మshown in figure No.: 3.9

The semi-vowelsయ, ya, and వ,va యు,ya shown in figure No. 3.10

Voiced alveolars ra,ర, r'', ర◌◌ and la,ల ర,◌ shown in figure No.: 3.11

The sibilants శ, Sa;ష, sha;స, sa shown in figure No.: 3.12

Other sounds: హ,ha. shown in figure No.: 3.13

3.3.3 Telugu Articulatory Phonetics or shallow orthography system

Characteristics of Telugu orthography The Andhra Pradesh Official Languages Commission to say that early forms of the Telugu language and its script indeed existed 2,400 years ago, and in sync with the Archaeological Survey of India (ASI). Telugu

/క/	/చ/	/ట/	/త/	/ప/
unvoiced unaspirated plosives	unvoiced unaspirated	unvoiced unaspirated plosives	unvoiced unaspirated plosives	unvoiced unaspirated plosives
Velar sound	Pre-palatal affricate	Retroflex	Dental	Bilabiable
	ఇ, ఈ, ఎ, ఏ ఈ (i, I, e, E) always palatal	tongue tip is retroflexed so that its underside touches the roof of the mouth	the tongue tip touches the teeth, not the ridge behind the teeth	

Figure 3.5: The unvoiced unaspirated plosives

The unvoiced aspirated plosives ఖ, , , , (kha, cha, Tha, tha, pha)

/ఖ/	/ఛ/	/ఠ/	/ధ/	/ఫ/
[KHAX]	[CHHAX]	[TTTAX]	[THHAX]	[FAX]
unvoiced aspirated plosives	unvoiced aspirated Palatal	unvoiced aspirated retroflex	unvoiced aspirated dental	unvoiced aspirated labial bilabial fricative

Figure 3.6: The unvoiced aspirated plosives , , , , (kha, cha, Tha, tha, pha) .

/గ/	/జ/	/ఢ/	/ద/	/బ/
voiced unaspirated plosives	voiced aspirated	voiced aspirated plosives	voiced aspirated plosives	voiced aspirated plosives
Velar	Palatal affricate	Retroflex plosive Retroflex	Dental plosives	Dental plosives
	i, I, e, E (I, ఈ, E, ఏ)	tongue tip is retroflexed so that its underside touches the roof of the mouth	the tongue tip touches the teeth, not the ridge behind the teeth	

Figure 3.7: The voiced unaspirated plosives

The voiced aspirated plosivesఘ, ఢ, ధ, భ (gha, Dha, dha, bha)

/ఘ/	/ఢ/	/ధ/	/భ/
voiced unaspirated plosives	**voiced aspirated**	**voiced aspirated plosives**	**voiced aspirated plosives**
Velar	Palatal affricate	Retroflex plosive Retroflex	Dental plosives

Figure 3.8: The voiced aspirated plosives(gha, Dha, dha, bha)

The nasals M,ఞ, N, న, మ

/ఞ/	/న/	/మ/
voiced unaspirated plosives	**voiced aspirated**	**voiced aspirated plosives**
Velar nasal	Retroflex nasal	Retroflex plosive Retroflex

Figure 3.9: The nasals M, , N, ,

The semi-vowelsయ, ya, and వ,va యa,ya

/ల/	/ర/
voiced unaspirated plosives	**voiced aspirated**
Retroflex voiced alveolar	Voiced alveolar or post-dental

Figure 3.10: The semi-vowels , a, and , a , ya

Voiced alveolars ra,ర, r'', రం and la,ళ ర,ం

/శ/	/ష/	/స/
voiced unaspirated plosives	**voiced aspirated**	**voiced aspirated plosives**
Voiced pre-palatal fricative	Palatal Voiceless retroflex fricative affricate	Voiceless alveolar or post-dental fricative

Figure 3.11: Voiced alveolars ra, , r'', '' and la, , r

The sibilants శ, Sa;ష, sha;స, sa

/హ/
Voiced unaspirated plosives
Voiced glottal fricative

Figure 3.12: The sibilants ,Sa; , sha; , sa

Other sounds: హ,ha.

/క/	/ చ /	/ ట /	/ త /	/ ప/
unvoiced unaspirated plosives	unvoiced unaspirated	unvoiced unaspirated plosives	unvoiced unaspirated plosives	unvoiced unaspirated plosives
Velar sound	Pre-palatal affricate	Retroflex	Dental	Bilabiable
	ఇ, ఈ, ఎ, ఏ ఈ (i, I, e, E) always palatal	tongue tip is retroflexed so that its underside touches the roof of the mouth	the tongue tip touches the teeth, not the ridge behind the teeth	

Figure 3.13: Other sounds: , ha.

S.no	Telugu Name-Place	English Name-Place	Description/Examples
1	Kanthya	Velar	Root of the tongue
2	Talavya	Palatal	Middle of the tongue
3	Mardhanya	Retroflex	Rounding of tounge
4	Dantya	Dental	Tip of tongue at dental
5	Oshtya	Labial	No influence of tongue position

Table 3.3: Telugu Phoneme classes based on the position and the release of air flow or restrict place of the air flow and articulatory positions.

language is a shallow orthography system. The phonemic '/ /'and phonetic symbols'[ ]' [BSA11] are one to one mapping for sound unit to the written script. It is same for most of Indian language.

3.3.4 Articulation of consonants

Articulation of consonants will be a logical combination of components in the two prayatnams. The places of articulation (passive) are classified as five listed in Table No. 3.3.

Apart from Table No. 3.3, other places are combinations of the five places has been shown in table no. 3.4

S.no	Telugu Name-Place	English Name-Place	Description/Examples
1	Dantosthya	Labio-dental	(E.g: v) teeth touches lips
2	Kantatalavya	E.g: Dipthonge	Middle of tongue touches palate
3	Kantosthya	labial-velar	(E.g:Dipthong o) lip

Table 3.4: Telugu Phoneme classes based on the position and the release of air flow or restrict place of the air flow and articulatory positions.

S.no	Telugu Name-Place	English Name-Place	Description/Examples
1	Jihvamulam	tongue root	for velar
2	Jihvamadhyam	tongue body	for palatal
3	Jihvagram	tip of tongue	for cerebral and dental
4	Adhosta	lower lip	for labial

Table 3.5: Telugu Phoneme classes based on the position and the release of air flow or restrict place of the air flow and articulatory positions.

Rounding with velam as the air restrict point, The places of articulation (active) are classified as three, they are shown in Table No.3.5

The attempt of articulation of consonants(Uccāraṇa Prayatnam) is of two types, the Place of Articulation (POA) and Manner of Articulation (MOA). Understanding of theses two method helps in analysis of errors in ASR output. Knowing this two methods it helps to reduce the substitution errors in TASR system output using Lexical modelling methods explained in Chapter 4 and executed in Chapter 5.

The air passing through the vocal card and controlled in the articulatory physiological filters defined their structure in the categories of Telugu language specific classification shown in Figure 3.4 The following structure of physiological picture describe the defined classification.

The Telugu phoneme pronunciation table with their place and manner of articulation and categorization with appropriate grapheme are shown in figure No. 3.18.

A principle underlying Telugu orthography discussed by Krishnamurti and Gwynn(1985) that , in complex words involving clusters, the ordering of the secondary form of vowel and consonant graphemes is not dictated by the way the word is articulated, as is the case with words with simple syllable structures. For instance, consider a word with

S.no	Category	Class in Telugu name	English names
1	Bahya Prayatnam(External effort)	Sprsta	Plosive
2	Bahya Prayatnam(External effort)	Ishat Sprsta	Approximant
3	Bahya Prayatnam(External effort)	Ishat Samvrta	Fricative
4	Bahya Prayatnam(External effort)	Alpapranam	Unaspirated
5	Abhyantara Prayatnam(Internal effort)	Mahapranam	Aspirated
6	Abhyantara Prayatnam(Internal effort)	Svasa	Unvoiced
7	Abhyantara Prayatnam(Internal effort)	Nadam	Voiced

Table 3.6: Telugu Phoneme classes based on the position the release of air flow.

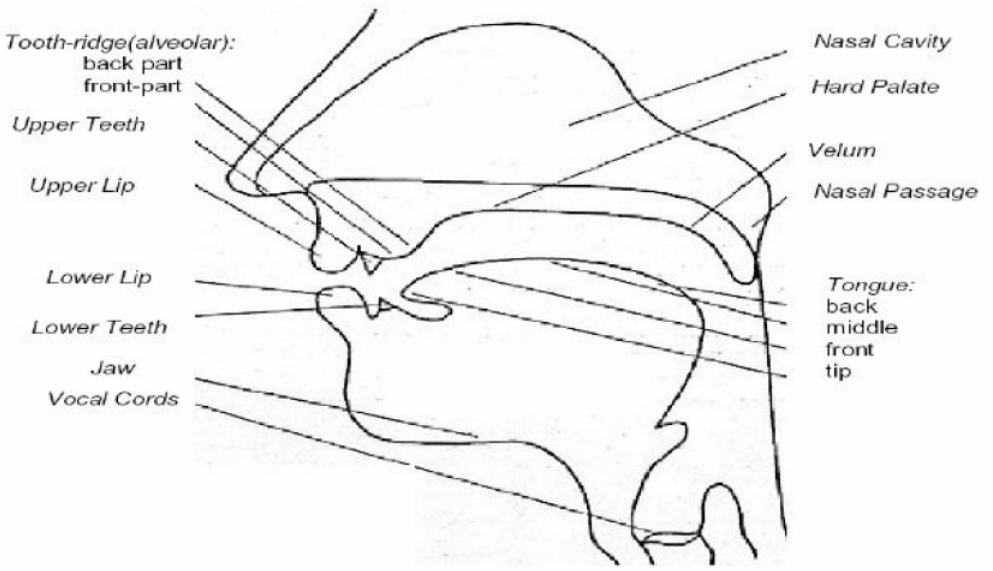


Figure 3.14: Human articulatory system for classifying phoneme based on place it articulated (Classification of Telugu aksharas based on the place of restrictions on the vocal sound hence the name of the place articulation and their components..

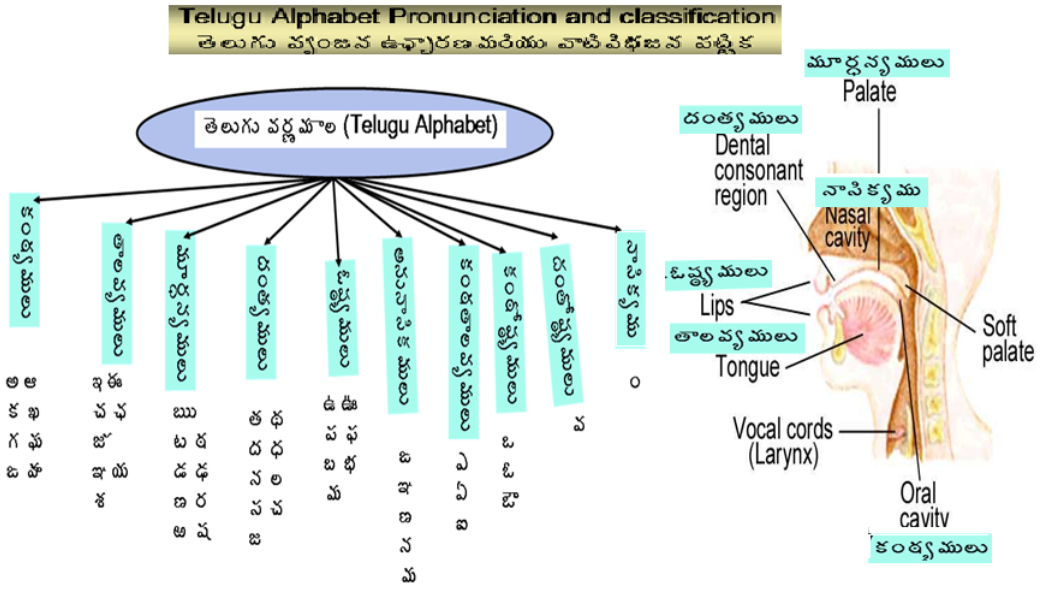


Figure 3.15: Classification of Telugu aksharas based on the place of restrictions on the vocal sound hence the name of the place articulation and their components.

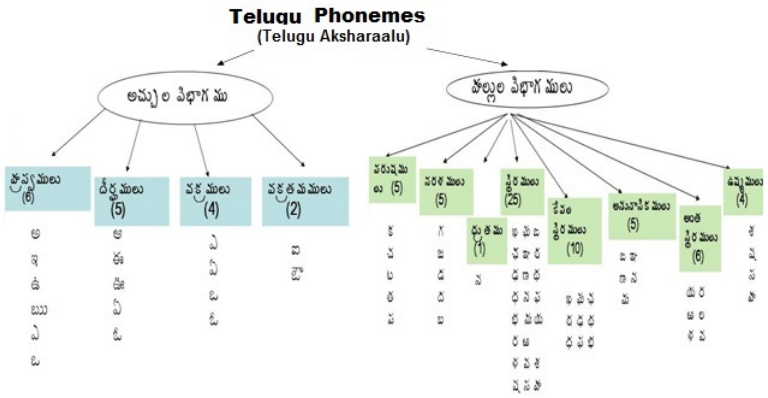


Figure 3.16: Telugu Phoneme(Telugu Varnamala) and their classification based on Vowel(Acchulu) and Consonant(Hallulu) and their categorisation based on the sound produced manner

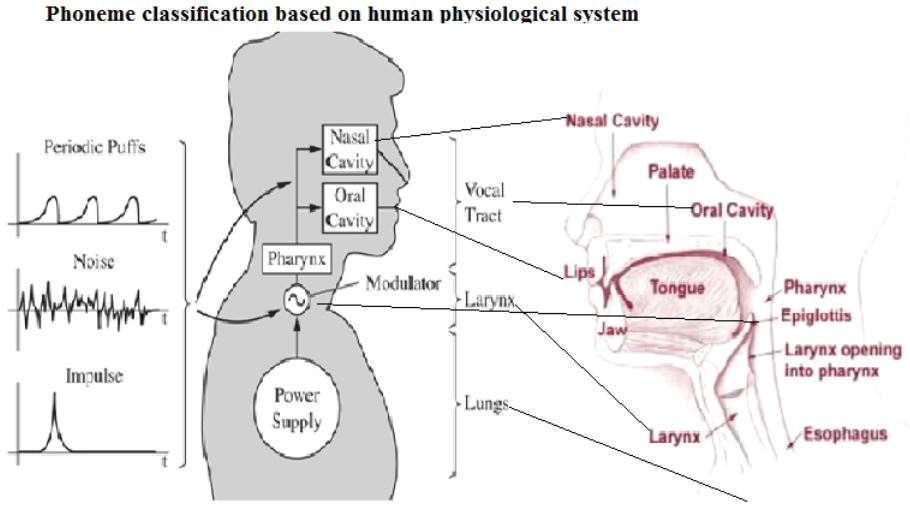


Figure 3.17: Phoneme classification and in human physiological location of speech production and their signal representation

Telugu Vyanjana Ucchārana Pattika (Consonents Pronunciation table

Prayatna Niyamāvalī	Kanthya (jihvāmūlam)	Tālavya (jihvāmadhyam)	Mūrdhanya (jihvāgram)	Dantya (jihvāgram)	Dantōṣṭya	Ōshtya (adhōsta)
Sparśa, Śvāsa, Alpaprānam	ka (క)	ca (చ)	ṭa (ట)	ta (త)	—	pa (ప)
Sparśa, Śvāsa, Mahāprānam	kha (ఖ)	cha (ఛ)	ṭha (ఠ)	tha (థ)	—	pha (ఫ)
Sparśa, Nāda, Alpaprānam	ga (గ)	ja (జ)	ḍa (డ)	da (ద)	—	ba (బ)
Sparśa, Nāda, Mahāprānam	gha (ఘ)	jha (ఝ)	ḍha (ఢ)	dha (ధ)	—	bha (భ)
Sparśa, Nādam, Alpaprānam, Anunāsikam, Dravam, Avyāhata	ṅa (ఙ)	ña (ఞ)	ṇa (ణ)	na (న)	—	ma (మ)
Antastha, Nādam, Alpaprāṇam, Drava, Avyāhata	—	ya (య)	ra (ఠ) (Lunthita)	la (ల) (Pārśvika)	va (వ)	—
Ūṣman, Śvāsa, Mahāprāṇam, Avyāhata	Visarga	śa (శ)	ṣa (ష)	sa (స)	—	—
Ūshman, Nādam, Mahāprānam, Avyāhata	ha (హ)	—	—	—	—	—

Figure 3.18: Telugu vyajana (Consonants) Uccharana Pattika (Pronunciation table) with their pronunciation in context of manner of articulation and place of articulation.

a complex structure such as CCVVCCVCVC swaardzitamస్వార్థ "self earned". The secondary forms of vowels following the two consonant groups [sw] / స్వ / and [rdz] / are a length marker [aa] / ఆ / and secondary form of the vowel [I] / ఇ /. If we go by the way the segments are articulated, [w] and [dz] should carry these markers respectively. However, the conventions of the writing system dictate that these forms be attached to the primary consonants [s] and [r]. Similarly, words containing anuswara[o] indicating a nasal before an obstruent take on different valuses depending on the preceding consonant and yet such phonological information is not encoded by the orthography

Syllabic alphabets and phonological awareness This section is to understand the phonological structure of the Telugu language taken from linguist works. This work helps to analysis the TASR output error analysis with comparison of human perception. The linguistic studies of speech segmentation abilities of illiterate and biliterate Telugu and English speaking adults[SAI97][SAI98] have shown that illiterates have considerable difficulty segmenting spoken word into syllables and that orthography strongly affects analysis of a spoken word in the literate population. The preferred syllable division by literate subjects, however, violates the sonority sequencing principle which states that from the syllable peak(vowel) onwards there must be a decline in sonority. The violation of this principle is evident in VCV /అందం/ [AMDHAM] syllables getting divided as V-CV /అం/-/దం/ [AM]-[DAM], VCCV /అత్త/ as V-CCV /అ/ - /త్త/ and VNCV as VN-CV in their production. In a later study [SAI99] argued that the reason for this is that Telugu does not permit consonants in word final position (no coda principle) except for [m]. When asked to indicate their preferred syllable division among three alternatives provided by the experimenter (the participants had to judge which division sounded right as opposed to segmenting the spoken word themselves), biliterate adults showed a preference for splitting consonants across syllables and thus accepting a coda (e.g. bhakti "devotion" was preferred over bha-kti (భ-క్తి)). However, when there is a homorganic nasal and obstruent cluster in words, they preferred VN-C over V-NCV and VNC-V options, probably because Telugu orthography has that rule (that is, the preferred division for gampa (గంప) "basket" is gam.pa(+) and not gamp.a or ga.mpa). The fact that, in Telugu, the word aksharam(అక్షరం) refers to both spoken as well as written syllables may have caused some confusion in the above experiments. Discounting that possibility, it appears from the results of Sailaja's studies that in Telugu the influence of orthography over phonology is so strong that sometimes certain phonological principles

of the language are violated in experiments dealing with speech segmentation. For preliminary evidence on the role of sonority in segmenting Telugu words with clusters presented in written form by school children in [VAS03]

The section will describe the authors on Telugu words and their phonemic division to learn. These concepts important in the thesis as part of design of language learning tool. According to authors yet to be investigated is the effect of literacy instruction in Telugu on phonological awareness in children. Since the focus of beginning literacy instruction is on getting children to master the primary graphemes, that is the 56 letters emphasizing all the primary vowels, diphongs and CV syllabograms with only the inherent[a], children, especially those who have had only one or two years of instruction, should make more mistakes then more experienced children in spelling the secondary forms of vowels and of consonants in words containing consonant clusters. The question on that the children with less instruction in Telugu orthography rely more on phonological information than on orthographic properties of words compared to older children on tasks involving use of phonological information. In short, which processing strategies – predominantly phonological vs. orthographic – underlie beginning reading in Telugu is the question mark. These were the questions motivating the authors study. Fifty pictures of referents of bisyllabic words representing common objects, body-parts, numbers, and animals were generated from a computer database with the help of a desktop publishing soft wares for design of Level 2 of INTTELL which is future scope of the thesis shown in Figure 3.6.4. A Telugu software package enabled us to print Telugu fonts corresponding either to the initial or the final syllable. The syllable structures of the target words included VCV /అంధం/ [AMDHAM], VVCV, VCCV, CVCV, CVVCV, CVCCV, CVC1C2V and CVNCV, where N is a nasal.

Pedagogical implications The study need to develop systematic instructional materials for the INTTELL [NAG05] with which can improve phonological and orthographic awareness of the words in the language in which children receive instruction during their school years. The efficacy of activity-based phonological awareness training programs for preschool children with speech interactive facility.

Telugu, one cannot assume a simple one to one correspondence between speech and orthographic syllables. Systematic activity based lessons should be prepared to help beginning readers make maximum use of their phonological awareness skills in accessing meaning. These lessons should capitalize on the orthographic principles as

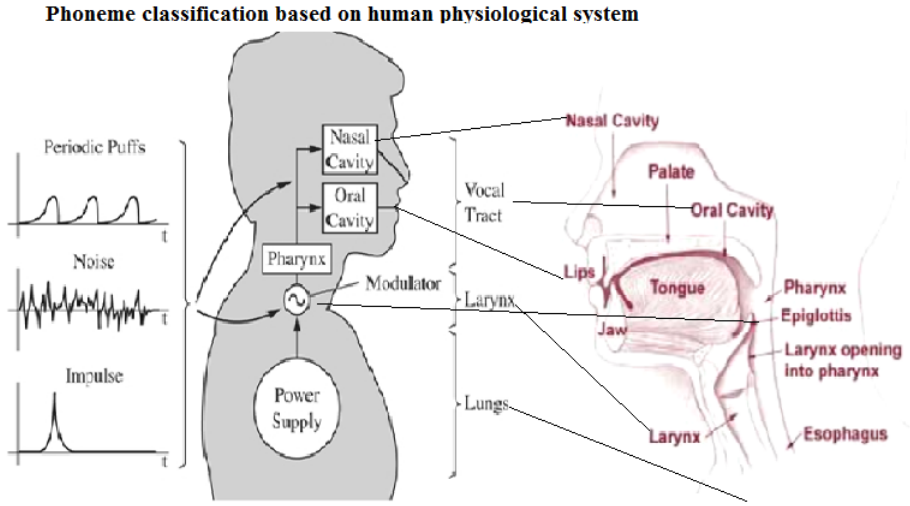


Figure 3.19: Phoneme classification and in human physiological location of speech production and their signal representation

well. Since evidence is available to show that orthographic knowledge can in turn help child grasp phonological principles underlying words, attempts should also be made to develop orthographic awareness during the primary school years. The main conclusion from the three experiments of this study is that knowledge about phonological and orthographic properties of words (at least for the highly frequent, and imageable words that were used in this study) contributes to acquisition of beginning reading skills although the strategies children use to access meaning during reading might differ at different developmental periods. Data based on a larger sample of children learning Telugu in schools where either Telugu or English is the primary medium of instruction until the high school level are needed to test the applicability.

3.4 Telugu language dialects

The dialects are the variants of the languages that have different words or word pronunciations in a language. The following figure No. 3.20 list the Telugu language dialect. Generally consider 7 different slang based on the geographical region that people live and the near by regional influence on the language. This table gives the future scope for the enhancement of the current work.

Dielects of Telugu language their their class based on the regions[ ]							
1	2	3	4	5	6	7	8
Tribal language - Telugu dialects different regions of Telangana and Andra Tribal groups							
Berad	Dasari	Dommara	Golari	Kamathi	Komta o	Kond a-Reddi	Salewari
9	10	11	12	13			
Telangala Telugu dialects							
Telangana	Warangal	Mahaboob Nagar	Gadwal	Narayana peta 4			
14	15	16	17	18			
Rayalaseema							
Kandula	Rayalaseem a	Nellooru	Prakasa m	Tirupati			
19	20	21	22				
Guntur& Krishna		**Godavari languages**					
Vijayawad a	Guntooru	Toorpu (East) Godavari	Paschim a (West) Godavar i				
23	24	25	26	27			
Uttarandra Telugu dialects							
Vadaga	Srikakula	Vadari and	Yanadi (Yenadi)	Visakhapatna m			
28	29	30					
Tamil influenced Telugu							
Vellore	Madras Telugu	Coimbatur e					

Figure 3.20: Telugu language dialects regional wise and class of languages include tribal

S.no	IPA	IPA-ASCII	SAMPA	DECtalk	Example
1	i	i	i:	iy	beet
2	I	I	I	ih	bit
3	ɛ	E	e	ey	bet
4	æ	&	{	ae	at
5	ə	@	@	ax	about
6	ʌ	V	V	ah	but

Figure 3.21: Phonetic notations used to transcribe any language script into the Roman script (machine understandable form).

Vocabulary: Utterances are the vocabulary chosen in language and they are transcribed using any of the following standards for transliteration of Telugu script.

MITalk uses a set of almost 60 two-character symbols for describing phonetic segments in it [ALT95] and it is quite common that synthesis systems use the alphabet of their own, leading to no unique accepted phonetic alphabet. Hence, system has to be automated, so as to minimise the human intervention in design and reusable to other languages. This is one of the greatest challenge in thesis and the following sections describe to bring the solution in the context of Telugu language, design of ASR is explored.

3.5 Telugu Language and ASR System

Telugu, a Dravidian language, is one of the oldest languages in the world, is third most widely spoken language primarily in South India. Native Telugu speakers are in Andhra Pradesh, with Telugu speakers across the globe. Nature of Telugu language is Syllabic as similar to the languages of India. Each symbol in Telugu script represents a complete syllable. In that sense it is a WYSIWYG form of script, and is considered to be most scientific by the linguists. This syllabic script has been achieved by the use of a set of basic symbols, a set of modifier symbols and rules for modification.

Need of building lexical model in Telugu: The following analysis gives the important task of building lexical model for Telugu as on data availability is 2014 at CIL, Mysore which is centre for Language Data Consortium.

Telugu language has eighteen Vowels, thirty six consonants and three dual symbols, yet only thirteen vowels, thirty five consonants are in common usage. Telugu script has the capacity to represent almost the entire phonetic spectrum of all Indian (and most world) languages. Telugu, has one of the best script in the world, yet maintaining an extensive sound base.[NAG12] The Telugu basic symbol pronunciation table along with the Romanization of Telugu sounds i.e RIT (Rich In Transcription) form of Telugu pronunciation using Telugu Lipi standards[SIR10]. Telugu Vuccharana (ఉచ్ఛరణ పట్టిక Pattika (Pronunciation Table) Telugu phoneme classification: There are Two ways of classification of Telugu alphabet which also known as Telugu Varnamala. One is the Manner of Articulation in which each symbol can be pronounced by considering its duration and their modulation on pronunciation

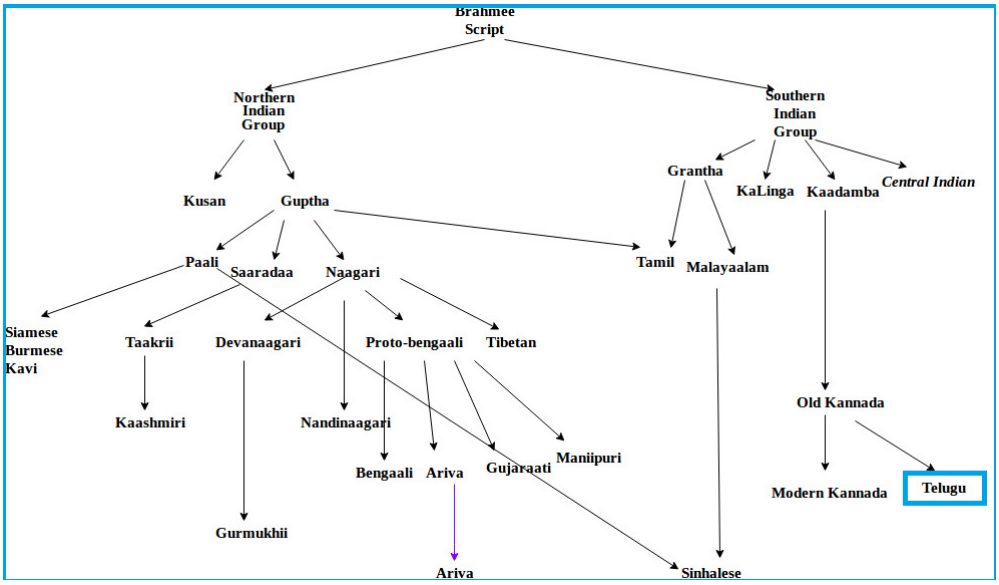


Figure 3.22: Tree shows the Bramhi script language (for Telugu Language) from the photo Dravidian languages[ ]

S.No	Phonemic	Phonetic		S.No	Phonemic	Phonetic		S.No	Phonemic	Phonetic	
	Telugu Scripts	WX-Representation	UOH SPHINX representation		Telugu Scripts	WX-Representation	UOH SPHINX representation		Telugu Scripts	WX-Representation	UOH SPHINX representation
1	/అ/	[a]	[AX]	18	/గ/	[ga]	[GAX]	33	/ద/	[xa]	[DHAX]
2	/ఆ/	[A]	[AA]	19	/ఘ/	[G]	[GHAX]	34	/ధ/	[Xa]	[DHHAX]
3	/ఇ/	[i]	[IX]	20	/ఙ/	[fa]	[NYAAX]	35	/న/	[na]	[NAX]
4	/ఈ/	[I]	[IY]	21	/చ/	[ca]	[CAX]	36	/ప/	[pa]	[PAX]
5	/ఉ/	[u]	[UH]	22	/ఛ/	[Ca]	[CHHAX]	37	/ఫ/	[Pa]	[FAX]
6	/ఊ/	[U]	[UA]	23	/జ/	[ja]	[JAX]	38	/బ/	[ba]	[BAX]
7	/ఋ/	[q]	[RH]	24	/ఝ/	[Ja]	[JHAX]	39	/భ/	[Ba]	[BHAX]
8	/ఎ/	[e]	[AI]	25	/ఞ/	[Fa]	[INYAX]	40	/మ/	[ma]	[MAX]
9	/ఏ/	[eV]	[IA]	26	/ట/	[ta]	[TAX]	41	/య/	[ya]	[YAX]
10	/ఐ/	[E]	[AY]	27	/ఠ/	[Ta]	[TTAX]	42	/ర/	[ra]	[RAX]
11	/ఒ/	[o]	[O]	28	/డ/	[da]	[DAX]	43	/ల/	[la]	[LAX]
12	/ఓ/	[O]	[OA]	29	/ఢ/	[Da]	[DXHAX]	44	/వ/	[va]	[VAX]
13	/ఔ/	[oV]	[AW]	30	/ణ/	[Na]	[NHAX]	45	/స/	[sa]	[SAX]
14	/ఀ/	[aM]	[AM]	31	/త/	[wa]	[THXAX]	46	/శ/	[Sa]	[SSHAX]
15	/ఁ/	[aH]	[AHA]	32	/థ/	[Wa]	[THHAX]	47	/ష/	[Ra]	[SHAX]
16	/క/	[ka]	[KAX]	33	/ద/	[xa]	[DHAX]	48	/హ/	[ha]	[HAX]
17	/ఖ/	[Ka]	[KHAX]	34	/ధ/	[Xa]	[DHHAX]	49	/ళ/	[lYa]	[LHAX]
18	/గ/	[ga]	[GAX]	35	/న/	[na]	[NAX]	50	/క్ష/	[kRa]	[KSHAX]
								51	/ఱ/	[rY]	[ARWAX]
								51	/ఱ/	[rY]	[ARWAX]

Figure 3.23: Thesis proposed UOH symbol set for Telugu grapheme to phoneme conversion and compatible for the SPHINX speech recognition engine compared to CMU symbol set

S.no	Languages	Hours	Minutes	Seconds
1	Assamese	105	51	38
2	Bengali	138	18	47
3	Bodo	201	10	48
4	Dogri	111	32	11
5	Gujarathi	156	23	4
6	Hindi	269	9	50
7	Indian English Bengali	34	12	57
8	Indian English Guajarati (MP3 Format)	21	40	0
9	Indian English Kannada	37	1	33
10	Kannada	198	51	3
11	Kashmiri	44	59	7
12	Konkani	195	14	47
13	Maithili	95	59	54
14	Malayalam	265	24	18
15	Manipuri	187	35	13
16	Marathi	168	13	50
17	Nepali	145	4	46
18	Oriya	165	30	5
19	Punjabi	187	53	28
20	Tamil	213	37	27
21	Telugu	50	51	36
22	Urdu	124	19	58

Table 3.7: Telugu Language Speech Corpus for scheduled languages of India available(No.21 is Telugu) and presently not freely available from the LDC IL as on date 2014.

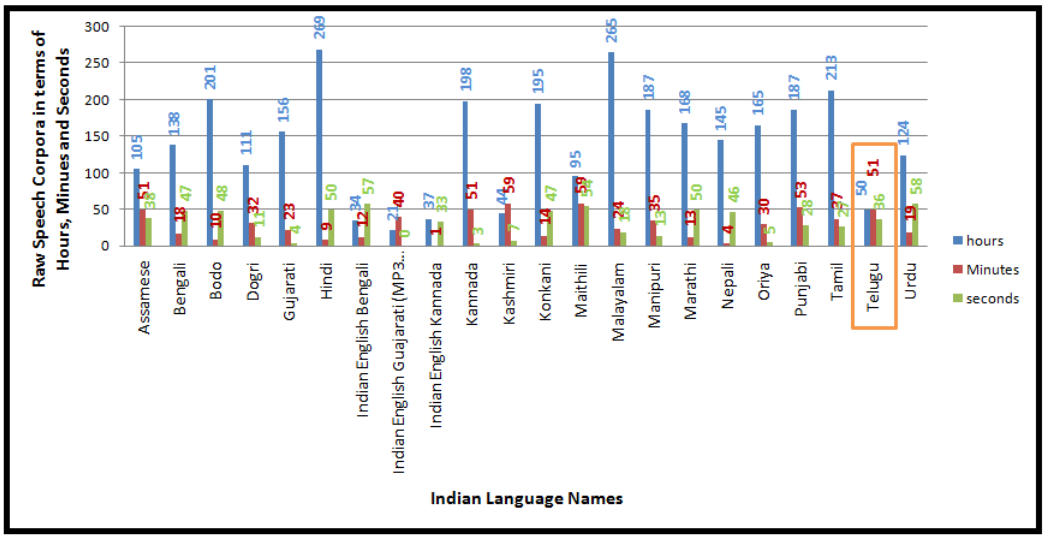


Figure 3.24: The speech corpus available at LDC-CIL 2014 for different scheduled languages in terms of its size

S.no	Language	Hour	Minutes	Seconds
1	Assamese	36	31	28
2	Bengali	21	55	46
3	Bodo	50	38	55
4	Gujarati	49	0	0
5	Hindi	45	51	32
6	Kannada	32	20	43
7	Konkani	32	29	53
8	Mlayalam	33	3	5
9	Manipuri	49	41	18
10	Nepali	23	23	35
11	Oriya	40	0	33
12	Pujabi	33	30	6
13	Tamil	48	12	10
14	Urdu	34	2	39

Table 3.8: The table and its graphical analysis for around 22 languages in which Telugu language Lexical model data is not available, as shown in this table :

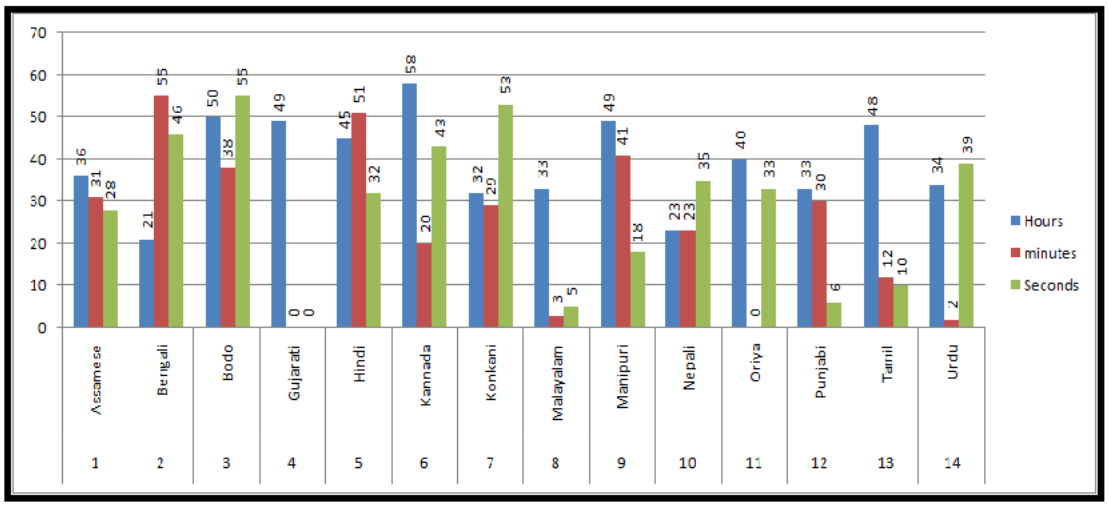


Figure 3.25: Graphical shows the language wise (hourly, minute and seconds) size of data interms its time duration.

The figure 3.31 describes the phonetic symbols used to transcribe the Telugu utterances. As the thesis proposal task for unique acceptable symbol for generic system, the GUI designed for transcribing the Telugu Utterances as well any spoken term can be used to transcribe with this tool in Telugu phonemic context.

3.6 Research problem outcome INTTELL

In this section it explains application of the problem statement of the thesis as outcome proposal for Telugu language learning tool. This proposal is to develop a language Telugu language learning without help of a human tutor. It also facilitates to assess proficiency in a language by testing the user pronunciation. To do this ASR module and its performance accuracy is important which contribution the present thesis work is. It is an e-learning method using ASR and TTS (Text to Speech System or Speech synthesis). Present role of Telugu ASR system is to first learn the Telugu specific pronunciation, then only the system is able to judge the pronunciation of the user. The entire empirical task of the present work is on this task.

INTtelligent Tutor for Telugu Language Learning (INTTELL) [NAG05] is based on cognitive psychology in Human Computer Interaction. This consists of basically two modules one is Student module which is uses ASR and another is Teacher module

S.No	/Telugu Phoneme /	#count in text	S.No	/Telugu Phoneme /	#count in text
1	అం	32	23	సు	4
2	అ	148	24	సూ	2
3	ఆ	50	25	సే	1
4	ఇం	14	26	సై	4
5	ఇ	38	27	సో	4
6	ఈ	12	28	సౌ	2
7	ఉం	22	29	స్ట	1
8	ఉ	26	30	స్టే	4
9	ఊ	14	31	స్టో	3
10	ఎం	12	32	సం	11
11	ఎ	31	33	స	62
12	ఏ	19	34	గం	1
13	ఓ	5	35	గ	7
14	క	48	36	గా	5
15	కాం	1	37	గుం	3
16	కా	38	38	గు	13
17	కుం	27	39	గొ	1
18	కూ	21	40	గో	3
19	సం	11	41	గ్రం	1
20	స	29	42	గ్రా	4
21	సా	7	43	ఘు	1
22	సీ	1			

Figure 3.26: : Text processing and phoneme/morphonemes count in a 10 pages text extracted for building corpus and lexical model for Telugu sample analysis results and phoneme/ morphonemes counts.

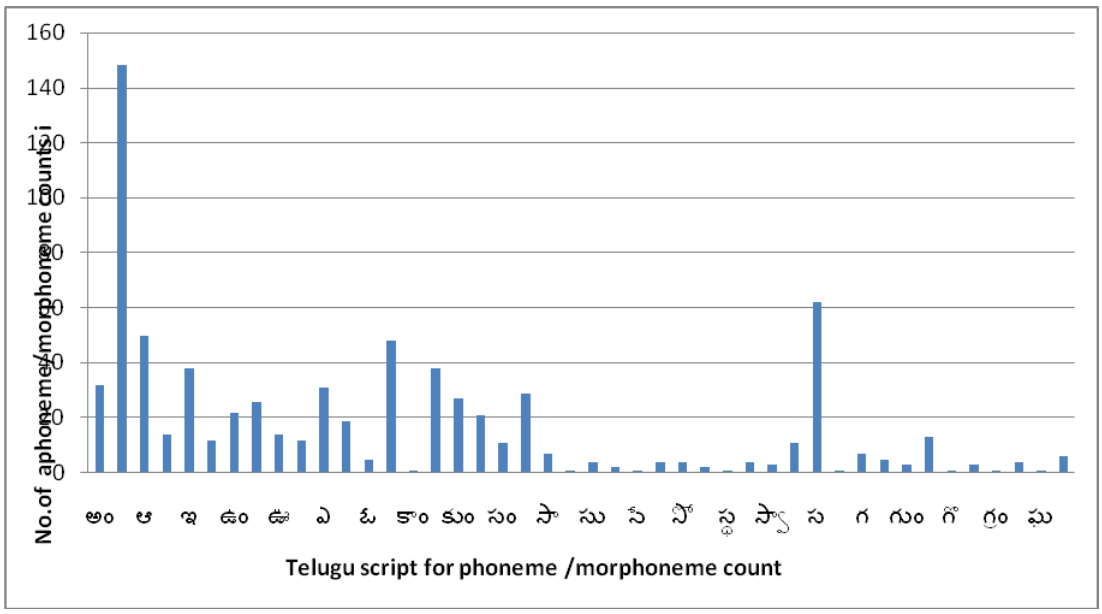


Figure 3.27: Graphical analysis for count of Telugu phoneme/morphonemes in a 10pages text corpus to develop lexical model

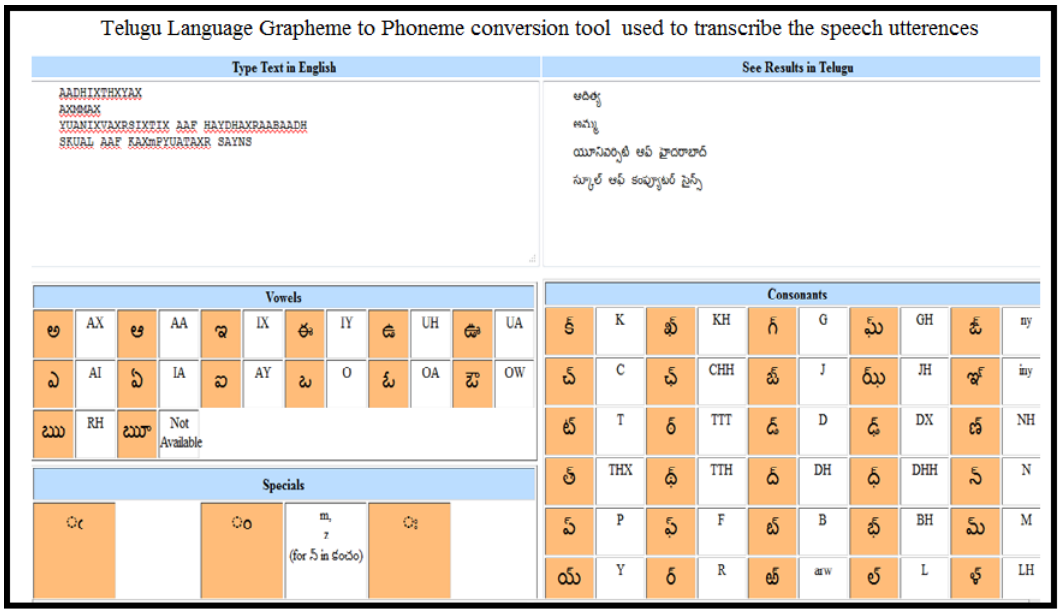


Figure 3.28: A GUI for Telugu grapheme to Phoneme mapping tool to transliterate the Telugu script.

Hand crafted Telugu phoneme specific Lexical model														
s.NO	Telugu graphemes	Phonetic transcription	Lexical Model using Telugu phoneme specific phonetic symbols											
1	అంగీకారానికి	AXNGIYKAARAANIXKIX	AX	N	G	IY	K	AA	R	AA	N	IX	K	IX
2	అంటారు	AMTAARUH	AM	T	AA	R	UH							
3	అంటారు?	AMTAARUH	AM	T	AA	R	UH							
4	అంటూ	AMTUA	AM	T	UA									
5	అంటే	AMTIA	AM	T	IA									
6	అంత	AMTHXAX	AM	THX	AX									
7	అంత	AMTHXAX	AM	THX	AX									
8	అంతకన్నా	AMTHXAXKAXNNAA	AM	THX	AX	K	AX	N	N	AA				
9	అంతహ్సరానికి	AXMTHXAXHPUHRAANIXKIX	AM	THX	AHA	P	UH	R	AA	N	IX	K	IX	
10	అంతా	AMTHXAA	AM	THX	AA									
11	అందమైన	AXMDHAXMAYNAX	AM	DH	AX	M	AY	NAX						
12	అందరి	AMDHAXRIX	AM	DH	AX	R	IX							
13	అందరికీ	AMDHAXRIXKIY	AM	DH	AX	R	IX	K	IY					
14	అందరూ	AMDHAXRUA	AM	DH	AX	R	UA							
15	అంది	AMDHIX	AM	DH	IX									
16	అందుకనే	AMDHUHKAXNIY	AM	DH	UH	K	AX	N	IY					
17	అందుకే	AMDHUHKIA	AM	DH	UH	K	IA							
18	అందులోంచి	AMDHUHLOANCIX	AM	DH	UH	L	OA	M	C	IX				

Figure 3.29: Hand crafted Lexical model for the extract text corpus from the online chandamama stories september2015

Hand crafted Telugu phoneme specific Lexical model				
S.No.	Telugu graphemes	Phonetic transcription	cananical Lexicon	Surface Lexicon
1	అంగీకారానికి	AXNGIYKAARAANIXKIX	AM G IY K AA R AA N IX K IX	AN G IY K AA R AA N IX K IX
2	అంటారు	AMTAARUH	AM T AA R UH	AN T AA R UH
3	అంటారు?	AMTAARUH	AM T AA R UH	AN T AA R UH
4	అంటూ	AMTUA	AM T UA	AN T UA
5	అంటే	AMTIA	AM T IA	AN T IA
6	అంత	AMTHXAX	AM THX AX	AN THX AX
7	అంత	AMTHXAX	AM THX AX	AN THX AX
8	అంతకన్నా	AMTHXAXKAXNNAA	AM THX AX K AX NN AA	AN THX AX K AX N N AX
9	అంతహ్సరానికి	AXMTHXAXHPUHRAANIXKIX	AM THX AHA P UH R AA N IX K IX	AN THX H P UH R AA N IX K IX
10	అంతా	AMTHXAA	AM THX AA	AN THX AA
11	అందమైన	AXMDHAXMAYNAX	AM DH AX M AY N AX	AN DH AX M AY N AX
12	అందరి	AMDHAXRIX	AM DH AX R IX	AN DH AX R IX
13	అందరికీ	AMDHAXRIXKIY	AM DH AX R IX K IY	AN DH AX R IX K IY
14	అందరూ	AMDHAXRUA	AM DH AX R UA	AN DH AX R UA
15	అంది	AMDHIX	AM DH IX	AN DH IX
16	అందుకనే	AMDHUHKAXNIY	AM DH UH K AX N IA	AN DH UH K AX N IA
17	అందుకే	AMDHUHKIA	AM D UH K IA	AN DH UH K IA
18	అందులోంచి	AMDHUHLOANCIX	AM D UH L OA M C IX	AN DH UH L OA M C IX
19	అక్కడ	AXKKAXDAX	AX K K AX D AX	AN K K AX D AX
20	అక్బర్	AXKBAXR	AX K B AX R	AX K B AX R

Figure 3.30: Canonical and surface form lexical model for extract Telugu text corpus.

Telugu aksharaalu(alphabets)- Pronunciation Tabable											
అ	a	son	క	ka	cart	త	ta	French t	శ	s'a	germanrich
ఆ	aa	master	ఖ	kha	blockhead	థ	tha	thumb	ష	sha	show
ఇ	i	if	గ	ga	goat	ద	da	then	స	sa	son
ఈ	ia	feel	ఘ	gha	ghost	ధ	dha	breathe	హ	ha	hot
ఉ	u	full	ఙ	~ma	sing	న	na	not	ళ	l'a	Retro L
ఊ	ua	fool	చ	ca	chain	ప	pa	pot			
ఋ	R	Betrri	ఛ	c'a	catch him	ఫ	pha	loophole			
ఎ	e	let	జ	ja	jet	బ	ba	ball			
ఏ	ae	late	ఝ	jha	hedgehog	భ	bha	abhor			
ఐ	ai	lie	ఞ	~na	French n	మ	ma	mother			
ఒ	o	rotate	ట	Ta	ten	య	ya	yard			
ఓ	oa	rote	ఠ	Tha	ant-hill	ర	ra	run			
ఔ	ow	now	డ	d'a	dog	ల	la	luck			
అం	am	him	ఢ	dh'a	godhood	వ	va	avert			
అ ః	a @h	half	ణ	n'a	under						

Figure 3.31: Telugu orthography with their phonetic representation with example word transliterated in English [NAR07].

which is uses TTS. Main concentration of the work is Student Module which is ASR. It tests the knowledge of Student and made them to learn how to pronounce the syllable or words. ASR accuracy is key part which gives intelligent part. But accuracy depend on many factors in which language dictionary also one. Recognition of system will be more if the number of words in dictionary is more. To build language model for any language dictionary plays key role. Previous model of ASR system used for Telugu language have 39 phone set i.e., pronunciation basic unit sound representations. It is same as the of CMU phone list. As Telugu is Dravidian language and also syllabic language and it has total 56 phones including Vowels (16) and Consonants (40) in Script. Each syllable has distinct sound so 39 phone list is not sufficient to represent the grapheme in the language. Hence the thesis proposed 51 phonetic symbols that are mapped to the phonemes of the Telugu is new phone list which are used to create the pronunciation dictionary for Telugu language. By using the new phone list(which is proposed and designed as the contribution of the thesis) for Speech tools both ASR and TTS improve the system performance. Here TTS is teacher model, it helps the user when they need to learn the pronunciation to be taught. Once the user is learnt the concepts, testing of user pronunciation by using ASR system. Enhancement in HCI with system which shows even how to write the syllable with flash multimedia animated file makes the user to learn better way without the help of external human teacher basics of the language. In the following figure shows the design and UI for INTTELL system for Telugu language learning with help of ASR system.

First level learning process of Telugu Akshra[ ]aksharas and its writing procedure while teaching is developed in INTTELL completed with the Net Beans IDE environment. Integration part of speech technology, script written in design part. ASR and TTS system working for Telugu language are the back end system for this design. Flash animation software used by taking the all phoneme images and using Action script of Flash software used to develop orthography writing procedure demonstration. Each image and its animated files are the input to the system when the system is in teaching mode these are displayed presently on click and in future enable with speech modality.

In this chapter focused linguistic requirements to develop ASR system for Telugu and transforming ASR system to the TASR system. Concepts covered from language background about spoken and written form to the classification of the language based on the script. The script of the language help to transcribe the speech correctly hence

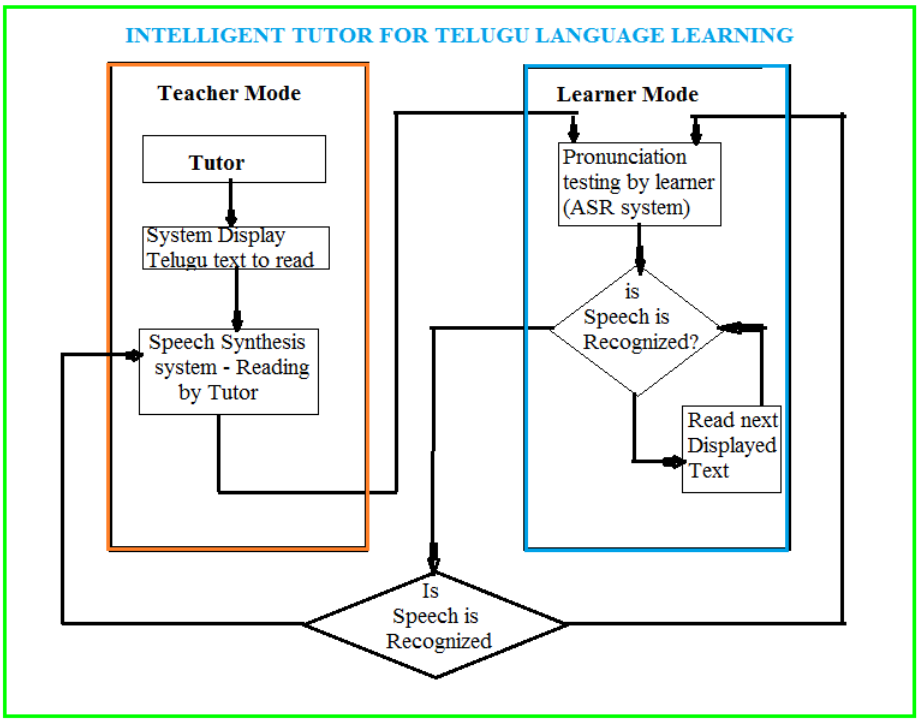


Figure 3.32: Design - Functional Flow diagram of INTTELL[NAG05]

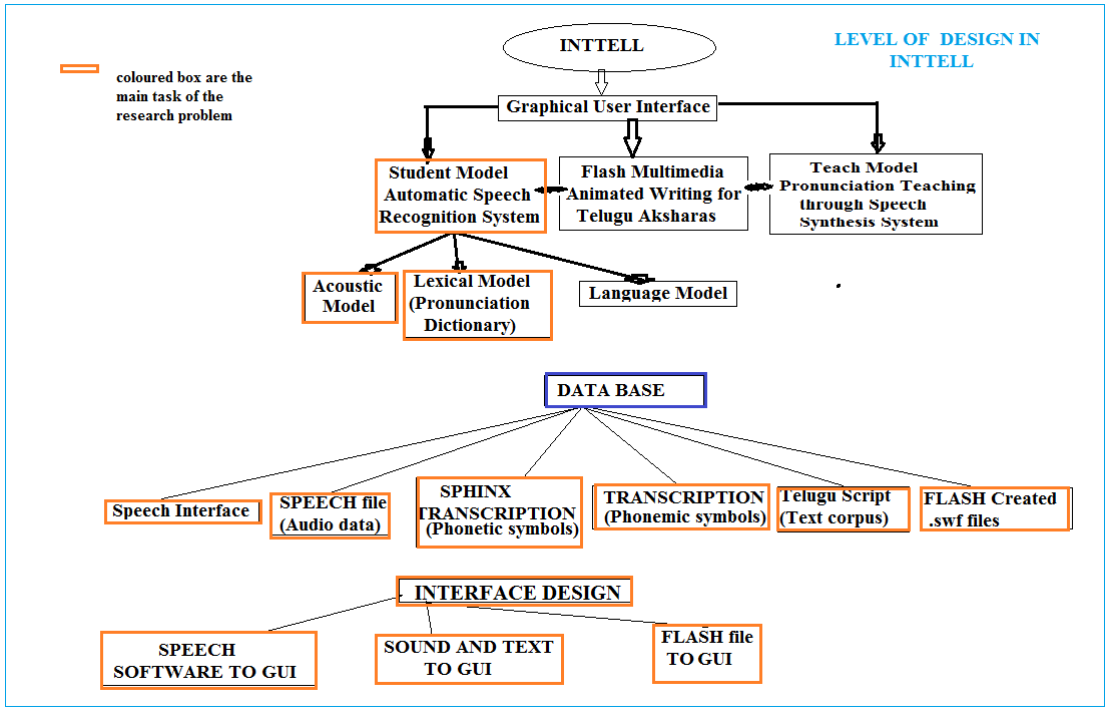


Figure 3.33: Modular Design for INTTELL system for Telugu language learning with help of ASR system.

Figure 3.34: Working model of prototype design of INTTELL, GUI system for Telugu language learning tool

Figure 3.35: Working model of prototype design of INTTELL, GUI system for Telugu language learning tool

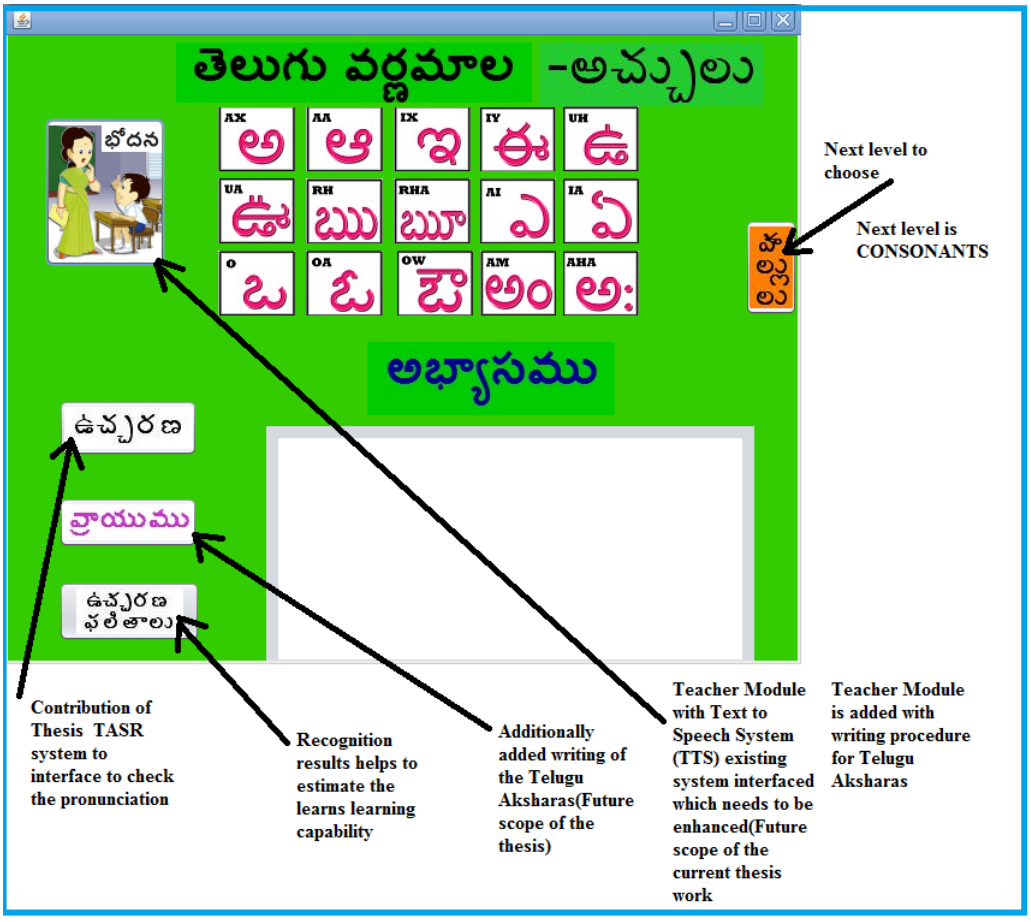


Figure 3.36: Description of Working model of prototype design of INTTELL, GUI system for Telugu language learning tool in Teacher mode and Learner Mode

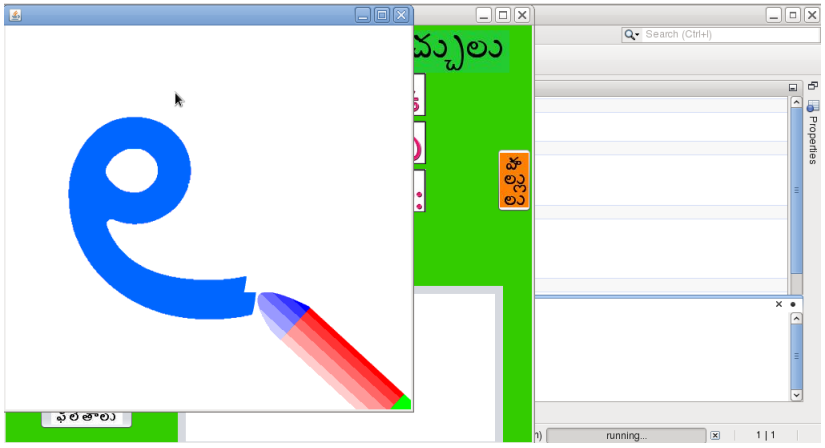


Figure 3.37: INTTELL in Teacher Mode teaching the pronunciation and writing procedure of Telugu Aksharam “AX”

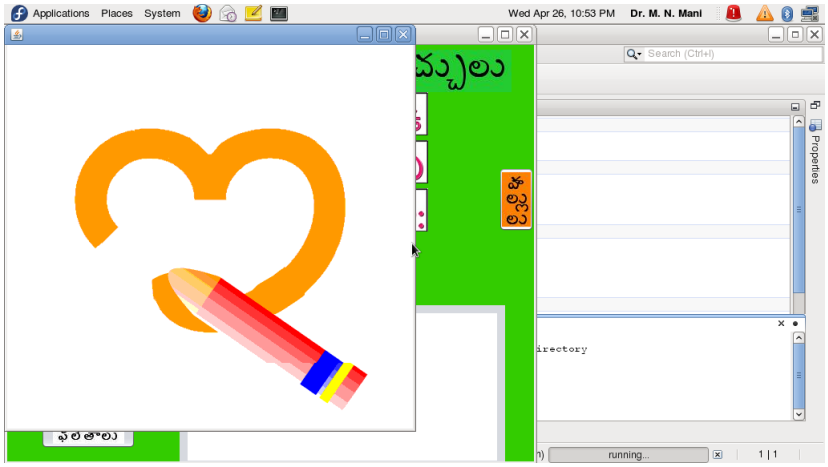


Figure 3.38: INTTELL in Teacher Mode teaching the pronunciation and writing procedure of Telugu Aksharam “IX”

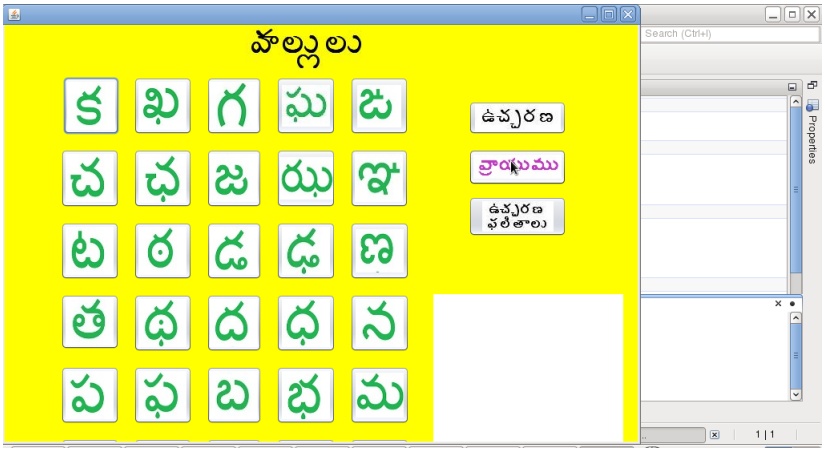


Figure 3.39: INTTELL in Teacher Mode teaching the pronunciation and writing procedure of Telugu consonants (Hallulu)

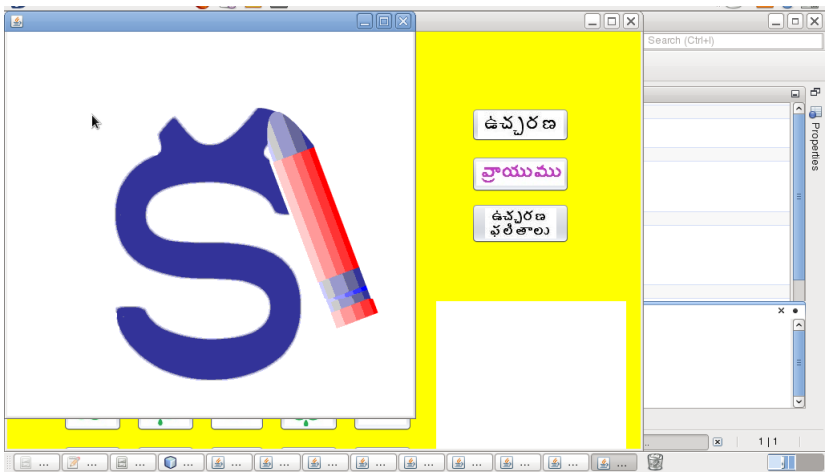


Figure 3.40: INTTELL in Teacher Mode teaching the pronunciation and writing procedure of Telugu consonants (Hallulu) and Writing and pronouncing of consonant [KAX]

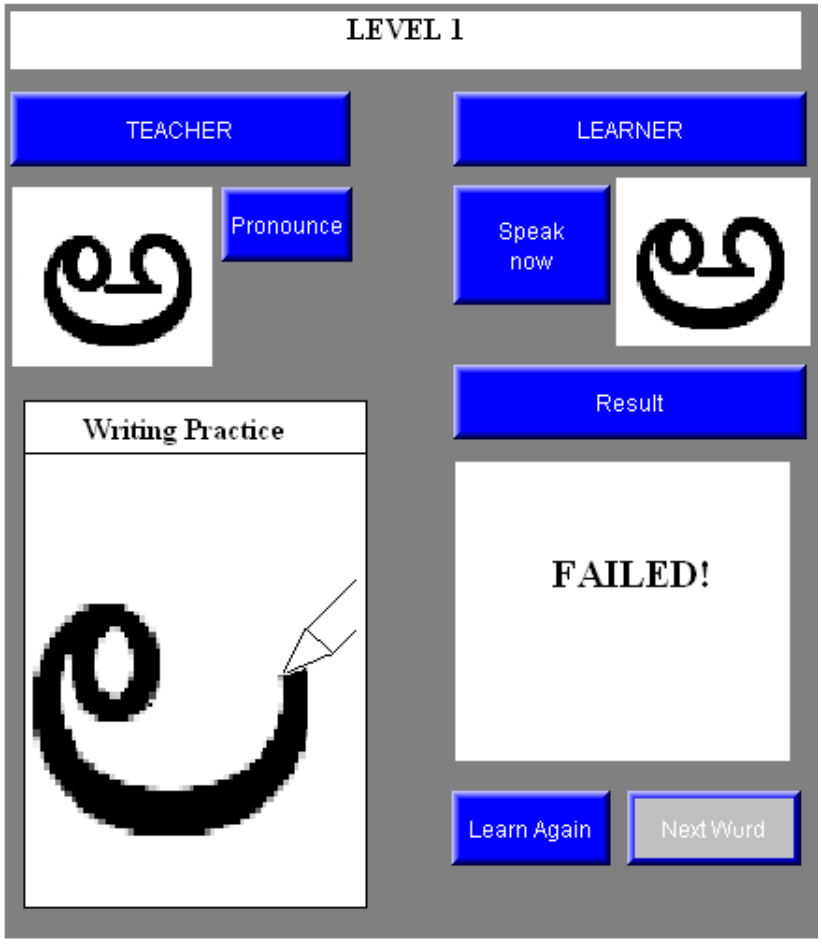


Figure 3.41: Level 1 of INTTELL learning phonemes with visual images to identify phonemes and check their pronunciation using ASR system proposed given above.

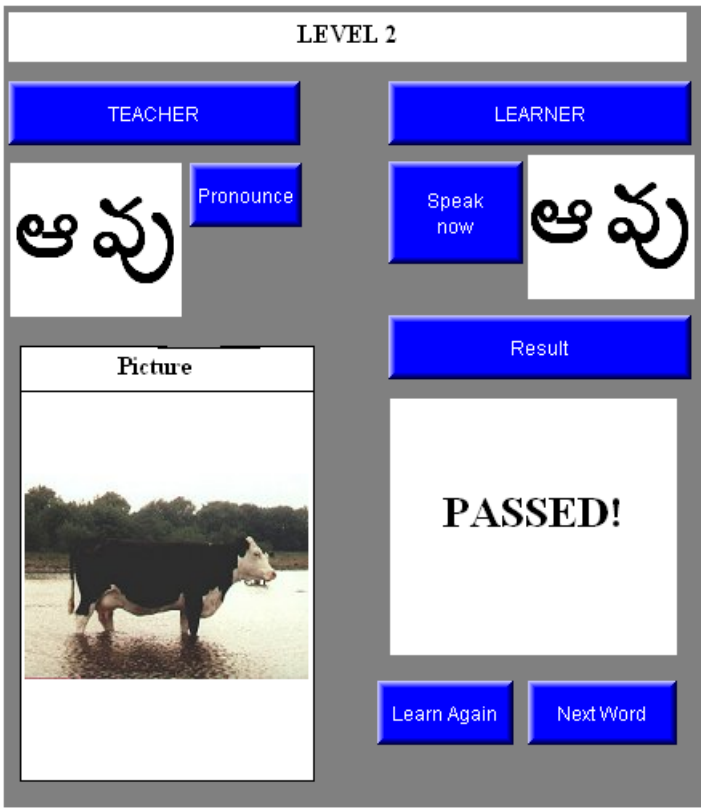


Figure 3.42: Level 2 of INTTELL learning words with visual images to identify words and check their pronunciation using ASR system proposed.

linguistic knowledge in the target language of ASR system is necessary to understand by the speech engineering. That knowledge is major description in this chapter. It also present the contribution of transcription tools designed with target language and also thesis outcome proposal of language learning tutor design methodology and preliminary implementation of INTtelligent Telugu Language Learning(INTTELL) tool and their required components covered.

****

Chapter 4

Lexical Modelling in TASR System

4.1 Introduction

In this chapter gives the process of Lexical model design, in the backdrop of the linguistic knowledge presented in chapter-3. This chapter presents the key challenges of statistical ASR system. Lexical Model, also known as a Pronunciation Model (PM) is defined as the interface of the corpus built and statistical and probabilistic estimation process of ASR system. PM is the input to an ASR system. From the pronunciation (sound) unit i.e., from the acoustic feature observation, the corresponding text is inferred, as stated in the chapter 3. The hypothesis text generated at the out put of ASR system, plays a major role in defining the Linguistic context, which is Orthography of target language-Telugu language. Linguistic context of sound production are influenced by many factors [RAS16] as reviewed in chapter 2. For the languages which do not have standard level of corpus, an attempt is made to develop a lexical model for languages to bring a standard level and thereafter generalized to any language. The Lexical model in ASR system, is build into Telugu ASR (TASR) system.

4.1.1 Lexical model or pronunciation model

Recent speech recognition literature aims at addressing the adaptation problem under lexical modeling. The lexical model affects three modules of ASR system i.e., acoustic model, lexical or pronunciation model and language model. Further, two methods are

used in the Lexical model viz., (i) knowledge driven method [HEL01] and (ii) data driven method[AMD00]. A lexical Model deals with the pronunciation of a word into elemental units of sound segments generally called as phonemes of the language. The lexical model is extremely sensitive to speech variations. is more affected by speech variations. Modeling of lexical model requires more human intervention and is also time consuming. In this chapter (i) different research works towards lexical modeling are reviewed (ii) procedures to implement lexical model in contemporary period (iii) direction towards the Telugu ASR lexical modeling and (iv) methods followed to build the system. The results are presented in successive chapter 5 related to design of Telugu Lexical (TelLex) [NAG10] using UOH phone set for Telugu ASR (TASR)[NAG12] system.

The state-of-art ASR systems knowledge base in linguistic context uses speech sound segment into their phonetic symbol as a unit of pronunciation. This sound symbol that is used to produce the unit of sound is a phone set. The hand crafted pronunciation dictionary is prepared using the phone set and mapped for the word, based on the linguistic rules (as shown in the figure 4.4). The above pronunciation dictionary can be used for Text To Speech system (TTS)[NAG14b].

4.1.1.1 Pronunciation variations in speech signal

The human vocal organs adapted to producing speech sounds, as part of the evolution [BEU98]. As explained in Figure 1.3, the human auditory & perception system evolved to process speech sounds. Speech sounds varies, based on (i) Utterance of Phoneme, sub-word, word or sequence of words is pronounced (ii) of utterance in different time periods and (iii) based on persons (a) the variation within the speaker (Intra speaker) (b) variation between speakers (Inter speaker). The variation is due to factors like height of the speaker, age, gender, regional accent, dialect, voice quality factors [FOS99]. Other factors influencing intra speaker variation are (a) speaking style which referred as stylistic variation, which depends on the read or planned speech or spontaneous speech. (b) Speaking rate [USH14] (c) co-articulation, i.e., impact of pre and post sounds pronounced (d) supra segmental features, influencing pronunciation based on stress of word/ sentence, frequency, word position in sentence, intonation, and location phonemes contained in syllable [STR99].

The ASR system has limitations in adapting to these variations, however, an attempt is made for processing by adapting to the recognition of humans speech in ASR

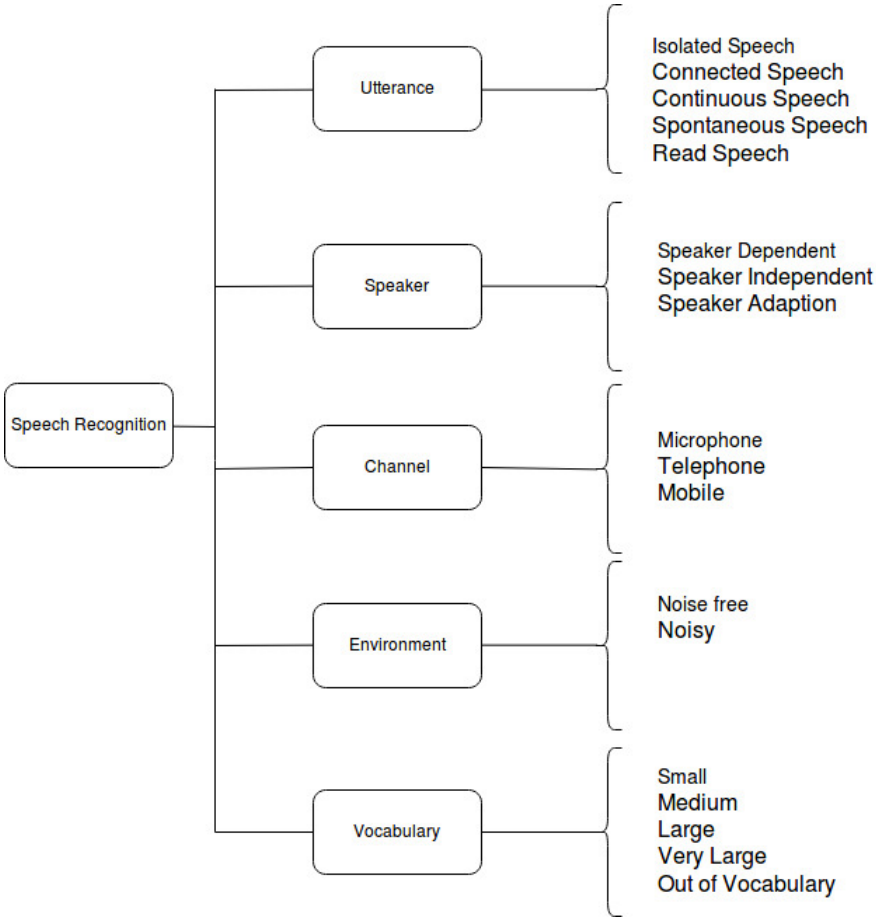


Figure 4.1: Adaptation methods in Speech recognition

system. The types of adaptation in Speech recognition are shown in Figure 4.1. The adaptions are also called as modelling.

In adaptation process, the acoustic model and language model are outputs lexical model is an input during the ASR training and decoding process [LIV05][LIV16]. The lexical model in knowledge base of ASR is an interface between text and AM. The pronunciations in the PM are transcribed sound units. The PM can be modelled in various methods viz., (i) by learning from the available data which is known as data driven method of adaptation or (ii) by understanding the language properties from theoretical knowledge which facilitates the importance of chapter 3 in this thesis work, using the linguistic knowledge build the lexical model using rules, which are known as knowledge driven method [DEN07]. These two are core concepts used in

modeling pronunciations and variations of pronunciation that are affecting the ASR system performance efficiency, measured in terms of WA and WER [RAB10] [NAG10]. Now it is proposed to discuss the factors effecting the efficiency, i.e., error that arise from pronunciation variations i.e., WER, occurring during training and decoding process of ASR system. In view of the variability in speech, the accuracy of ASR system is reduced. To improve the same, decision tree [FOS99] based modeling in lexical model with alternative word pronunciations from phonemic base forms is generated. These generated lexical model, is inputted to train the acoustic model, which would get better the word accuracy hence overall improvement of ASR system work. The above method in word recognition helps to add additional, pronunciation in phonemic layer, which is a canonical pronunciation, shown in Figure 4.4. The prediction of choosing alternate pronunciation is exhibited with GMM – HMM in training process of ASR system. The alternative pronunciation adding to defined lexical model is more suitable for deep Orthography based languages and less influence in shallow Orthography languages like Telugu language. The performance in above process is more suitable for isolated word recognition, as compared to continuous speech. The process would lead to reduction in WER [GIA07].

In Telugu language the process is simple in adding canonical layer with Akshara (Grapheme) to phoneme mapping, as shown in Figure 4.4. This alternative pronunciation used in the data and its recognition improvement, shown in the Chapter 5.

4.2 Lexical Modeling Framework

The steps in Telugu pronunciation model leading to TASR are as under:

1. **Extraction of canonical (phonemic) transcription :** from word and Orthographic layer. A standard Telugu lexical model (TelLex) [NAG10] is a Telugu phoneme specific unit, used for defining a lexical model. Base form of this is a manually written lexicon based on the phonemes of Telugu language. This was derived after much research, which is shown in Figure 4.2 below.

2. **Extraction of surface-form (phonetic) transcription :** The Orthographic text (hand labeled) spoken by language speaker has been transliterated for facilitating the lexical modeling, using Transliteration tool detailed in Chapter 3 for

Hand crafted Telugu phoneme specific Lexical model													
s.NO	Telugu graphemes	Phonetic transcription	Lexical Model using Telugu phoneme specific phonetic symbols										
1	అంగీకారానికి	AXNGIYKAARAANIXKIX	AX	N	G	IY	K	AA	R	AA	N	IX	K	IX
2	అంటారు	AMTAARUH	AM	T	AA	R	UH							
3	అంటారు?	AMTAARUH	AM	T	AA	R	UH							
4	అంటూ	AMTUA	AM	T	UA									
5	అంటే	AMTIA	AM	T	IA									
6	అంత	AMTHXAX	AM	THX	AX									
7	అంత	AMTHXAX	AM	THX	AX									
8	అంతకన్నా	AMTHXAXKAXNNAA	AM	THX	AX	K	AX	N	N	AA				
9	అంతహ్మారానికి	AXMTHXAXHPUHRAANIXKIX	AM	THX	AHA	P	UH	R	AA	N	IX	K	IX	
10	అంతా	AMTHXAA	AM	THX	AA									
11	అందమైన	AXMDHAXMAYNAX	AM	DH	AX	M	AY	NAX						
12	అందరి	AMDHAXRIX	AM	DH	AX	R	IX							
13	అందరికీ	AMDHAXRIXKIY	AM	DH	AX	R	IX	K	IY					
14	అందరూ	AMDHAXRUA	AM	DH	AX	R	UA							
15	అంది	AMDHIX	AM	DH	IX									
16	అందుకనే	AMDHUHKAXNIY	AM	DH	UH	K	AX	N	IY					
17	అందుకే	AMDHUHKIA	AM	DH	UH	K	IA							
18	అందులోంచి	AMDHUHLOANCIX	AM	DH	UH	L	OA	M	C	IX				
19	అక్కడ	AXKKAXDAX	AX	K	K	AX	D	AX						
20	అక్బర్	AXKBAXR	AX	K	B	AX	R							

Figure 4.2: Telugu Phoneme based Lexicon with handcrafted for Telugu graphemes

Hand crafted Telugu phoneme specific Lexical model				
s.NO	Telugu graphemes	Phonetic transcription	cananical Lexicon	Surface Lexicon
1	అంగీకారానికి	AXNGIYKAARAANIXKIX	AM G IY K AA R AA N IX K IX	AN G IY K AA R AA N IX K IX
2	అంటారు	AMTAARUH	AM T AA R UH	AN T AA R UH
3	అంటారు?	AMTAARUH	AM T AA R UH	AN T AA R UH
4	అంటూ	AMTUA	AM T UA	AN T UA
5	అంటే	AMTIA	AM T IA	AN T IA
6	అంత	AMTHXAX	AM THX AX	AN THX AX

Figure 4.3: Surface form representation for Telugu phoneme specific Lexical model using Telugu graphemic sequences

the Telugu UOH corpus. The variants of pronunciation added as second label of the same word, given in Figure 4.3.

3. **Alignment of canonical to surface level transcriptions :** For the alignment of transcription in training process, a dynamic programming is used with phonetic feature distance measures in GMM-HMM modeling. Most of the cases, where allophones (variation in phones) and in phoneme transformation this aspect is important. In Telugu Phonemes to phone mapping is one to one mapping. Hence, this problem is less significant in Telugu, as compared to English language. Phonetic representation is given in thesis by using "[ ]" and phonemic representation by "/ /" following approach as presented by prof. Peri Basker Rao

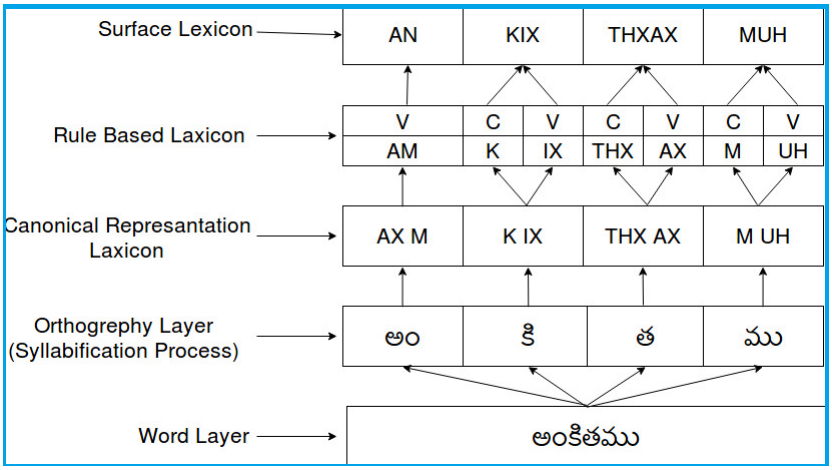


Figure 4.4: Surface form representation for Telugu phoneme specific Lexical model using Telugu Orthograph with a layered structure

in his article [BAS11][LIV07].

4. **Alignment of canonical to surface level transcriptions :** In this process, predicting the probability estimation of phonetic unit and applying the decision conditions using speech signal features. For a given word, segmented phonetic unit probabilities comparison constitute the Decision tree. Analysis on error word substitution ("Ala") in place of actual word "Alaka" is depicted as below.

5. **Alignment of canonical to surface level transcriptions :** In lexical model the dictionary-based phoneme-level recognition network is used to transform a network of surface-forms. The phoneme-level network may either be derived from linguistic knowledge driven method or from a word-lattice generated by a data driven method of recognition process.

   Modelling of ASR system in the backdrop of development of speech recognition is gaining importance in real world environment. The orthography of words which are correlated from the glossary of pronunciation vocabulary, onto a chain of sub-units called phonemes necessitate human expert knowledge is expensive and time consuming task. The automation also requires hand labeled material as back up. For this purpose in-house built Telugu phoneme specific mapped units are used. To enable the same we have developed recognition systems at our lab using

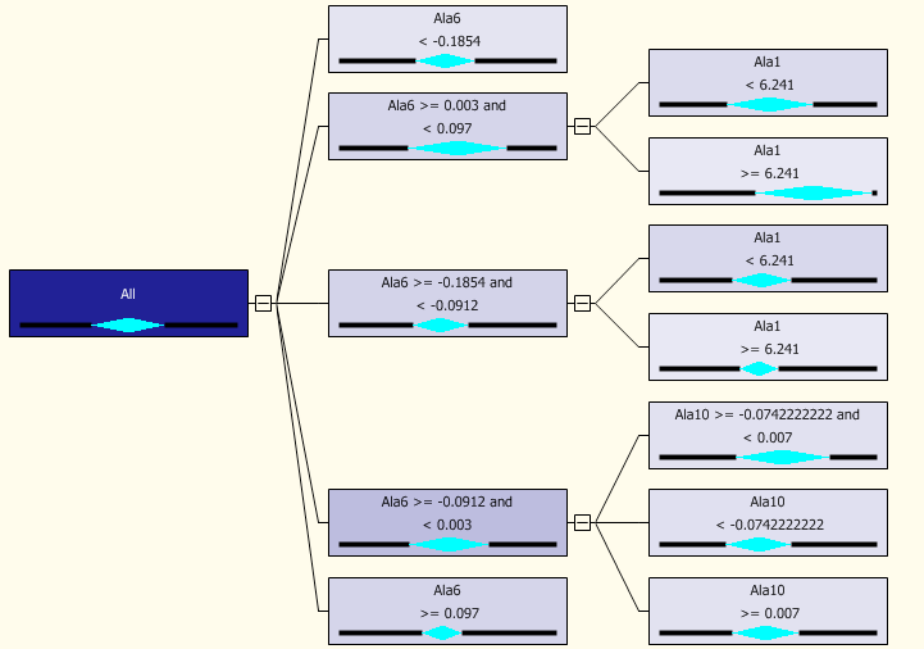


Figure 4.5: Decision tree based analysis of data (ala vs alaka) [NAG13]

Telugu phonemes as sub-units. This is simple in shallow orthography based languages. Telugu language is shallow orthography mapping in the dictionary is not difficult as each word made into syllables and alphabetic script with a rather near grapheme-to- phoneme association. With the same experiment carried out to map the Telugu unit for mapping sounds of other language and found better results in recognition. In this work, my research results into form a grapheme based Telugu recognizer trained on UOH phonelist [NAG12] with specific pronunciations. The ensuing system accomplishment was evaluated against phoneme based identification system, which was taught in CMU specific phonemic representations [SIN00][ST04]. There will be dissimilar assessment for the two context reliance models.

Phonological pronunciation variations by lexical modelling are the most popular method. Phonetic segmentations pronunciation modelling is done by the acoustic models using iterative training for given input speech. Speaker dependent parameters are better handled in lexical modeling with insertions, deletions and variation in dialects or speaking style, this model also accommodates longer contexts than acoustic modeling, permitting modeling of linguistic structure in hierarchy.

The subsequent sections deal with lexical model adaptation for pronunciation variations. The acoustic models used in recognizer not in abstract linguistic units instead pronunciation of a language [BAL99] in signal context. The mismatches in the recognized output dealt using the lexical model. The importance of lexical model in ASR knowledge demands the modelling in [ING02] by the process of changing the baseforms in to surface form based on speech variability. The newly defied lexical model is used to train the system to estimate acoustic log likelihoods which are used to prune the pronunciation variants and their framed rules are adapted.

Maximum likelihood pronunciation modeling is rule probability estimation from the point of view of automatic transcription. As per the authors [HOL99e] the frequency counts from a maximum likelihood transcription is termed as "maximum likelihood" for the adaptation in acoustic model with log likelihood estimations for given pronunciation. The thesis is the utility of decision theory to get used to the intonation disparity and pertain to implement in reduction in ambiguity/ confusability at the output of ASR system by dealing at acoustic model level [NAV14].

Pronunciation modeling using Decision theory: Applying decision theory based modeling in which ASR is working as a 'classifier is the recognizer' which is the AM, PM, LM [SAK09]. We can get classes of words from the utterance, sentence, or meaning.

The three modules of ASR are best possible discriminated. The pronunciation in lexical modelling is foremost concern to the research work and the decision theory in particular. The notations tag on as common decision theory metaphors:

The classification system for the described notation:

Classifier

$$C(.) \tag{4.1}$$

M dissimilar word sets or Classes

$$W_j \in \begin{bmatrix} W_1 & W_2 & \dots & W_M \end{bmatrix} \tag{4.2}$$

A set of baseforms $\left[B_j^{id}\right]$ for every word (Identity=id)

$$\tag{4.3}$$

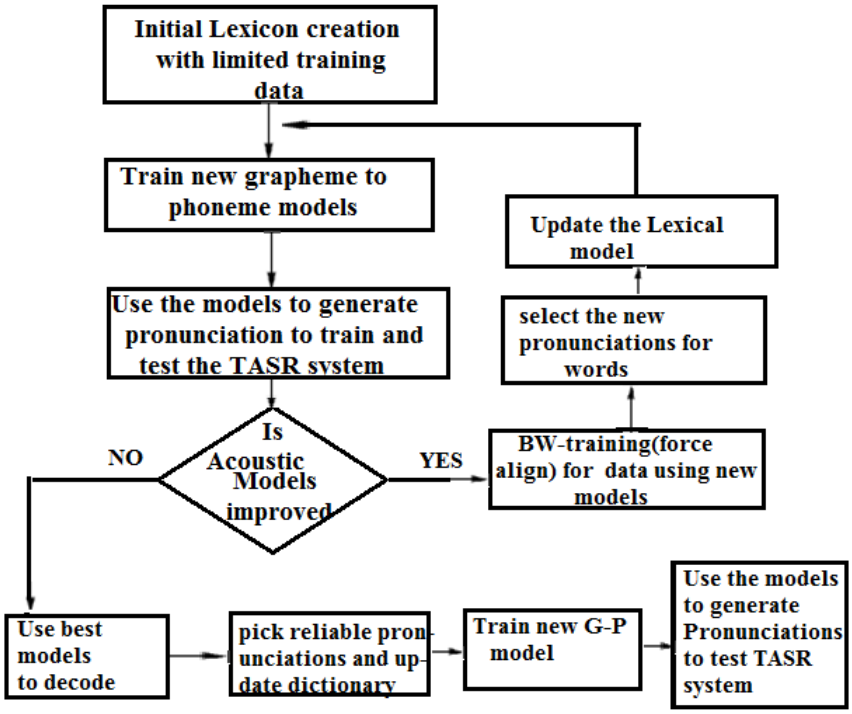


Figure 4.6: Lexical learning process through HMM training and testing for TASR system defined phone set

A group'K' training samples

$$O^{(k)} \in \left[o^{(1)} \quad , o^{(2)} \quad \ldots \quad o^{(K)}\right] \tag{4.4}$$

where, $o^{(k)}$= an array of feature vectors in prop up of an acoustic segment of word.

The baseforms B_j is used to indicate the different forms as a subscript. The Pronouncing modelling consists of two different methods with dissimilar problems. The Figure 4.9 follows the following given steps to design the lexicon for target language as per the below steps

1. **Knowledge based pronunciation modeling[RIL99]**

   In the knowledge-based method, extracting knowledge from linguistic context is done mostly by hand-crafted lexical modeling or linguistic literature. Linguistic knowledge plays major role in deriving the pronunciation. Present thesis work in Telugu language context, the structure of the phoneme arrangement is taken to formulate rules. For the consonant, rules will be based on the context which can be applicable baseline lexicon words. The consequential variant is absolutely supplemented to the lexicon. Figure No. 4.7 Phonological rules with application perspective are given in Telugu explored with the observation of data and language background. Further 5 phonological rules can be had from [KES99] and the reference to the usage. Substitution of every consonant ([K][AX]/క/, [C][AX]/చ/, [TAX]/ట/,[THX][AX]/త/, PAX/ప/) followed by oral sound of [AM] sound should be replaced with [AN]

2. **Knowledge based pronunciation modeling[RIL99]**

   a. In turn, each of the above models comprises of (i) Direct driven (ii) Indirect driven

   The new Lexica are created with the utilization of two knowledge-based and data-derived approaches to pronunciation modelling and in addition pronunciation variants are included to the baseline lexicon.

   b. Step-by-step rule derivation process is as under

   i) Reference and alternative transcription
   ii) Alignment
   iii) Rule derivation

Grapheme			Phoneme	
క	ం		KAX	AN
చ	ం		CAX	AN
ట	ం		TAX	AN
త	ం		THX	AN
అ	ం		AX	AN
ప	ం		PAX	AM

Figure 4.7: Example of Telugu transcription for Grapheme to phoneme mapping in canonical and surface level lexical model

Telugu phonemic transcription											
Simple Grapheme to Phoneme map(Canonical)						Grapheme to phoneme map using phonological rules(Surface)					
అం	క్	ఇ	త్	మ్	ఉ	అం	క్	ఇ	త్	మ్	ఉ
AM	K	IX	THX	M	UH	AN	K	IX	THX	M	UH
కం	చ్	అ				కం	చ్	అ			
KAM	C	AX				KAN	C	AX			
అం	త్	అ				అం	త్	అ			
AM	THX	AX				AN	THX	AX			
అం	ప్	అ				అం	ప్	అ			

Figure 4.8: Example of Telugu transcription for Grapheme to Phoneme in canonical and surface level lexical model

iv) Rule assessment and Pruning

c. Pronunciation variant generation assessment and pruning

With the various factors that influence for speech variability knowledge incorporate the variations and make the system to learn these variations. This may be text specific or speech specific variations.

d. Retranscription

With the ASR result analysis separate the error file and retranscribe for better performance.

e. Confusability reduction

Phonetics and phonology specific knowledge is used to reduce the phoneme level confusion in lexical model adaptation, the most used technique generic modelling technique.. Segmental variation, such as allophonic variation is modeled in acoustic model using iteratively training. With the help of inclusion, removal and disparity existing in the group of speakers the supplementary variation will be tackled at lexical point i.e, dialects, or distinctive native way of speaking. Lexical modeling deals with syllable, words and phrases where acoustic modelling only phonetic level. Pronunciation variation modeling by lexicon adaptation is presented in this thesis. Pronunciation disparity can be found in two key guiding principles, with diverse problems.

4.2.1 Based on Knowledge methods

Here best pronunciation rules are applied based on linguistic knowledge as well as phonetic. The main problem occurs if the knowledge uncovers the disparity in the model. There may be too many or too low disparity, and there is no assurance of frequent occurrence of these variations. For example: There are not many variations in Telugu language, due to scope for one to one mapping between orthography and phonemes i.e it is the shallow orthography.

4.2.2 Knowledge based pronunciation modelling

The author explored the decision trees for linguistic specific phonological rules and their use is enumerated in substitute pronunciations, Finke and Waibel [FIN97]. Here author is worked on Switch board database to classify the speaker specific character using the forced alignment to frame the rule based on the training corpus. Learning the knowledge with data driven and using decision to frame the linguistic rules is main task. The modelling based on neighborhood phonetic context and their influence, type of word used, style that which speaker followed, the rate at which speaker producing the utterances, in a word duration of the phone normally that decided at which speed utterance produced, influence of intonations using vowel stress, pitch, and computed probabilities are the parameters they studies. The decision tree is modeled with the analysis of data-verified rules and is predictable based on use of relative frequency. The ensuing change rules are understood based on speaking mode. The lexicon was extended by means of baseforms originate by forced alignment, i.e., direct modelling using intermediary step of indirect modelling. The variants originate by means of the derived rules is not augmenting the accomplishment even after picking the variants of the baseline dictionary. Other reviewer worked on the hand-labeled data to derive the knowledge based on the phonological rules [BYR98][RIL99].

Though it is the knowledge based method in the hand-labeled data, the pronunciations found employed in re-transcriptions lead to less precise in the categorizing which was illustrated [BYR98] [RIL99]. According to two reviews, the hand-labeled way of modelling lexical is the state-of-Art system. The hand-labeled models are less precise than the pronunciation model techniques with the automatic labeled data techniques in Telugu language. Some of experimental proofs are presented in Chapter 5. The variants are created using rules for unseen words with indirect pronunciation modelling. For each word the minute network of possible alternate transcription from the decision tree will be prepared. The achievement is diminished with the mentioned substitute pronunciations for recognition [FIN97] and necessity other remedy in selecting pronunciations which agree to the lexicon. In order to re-transcribe the corpus, the other substitute pronunciations from the decision trees are subsequently utilized for forced alignment. It is elucidated that there is an appreciable enhancement of the performance in use of new decision trees which are relied on the automatic transcription.

The understanding of disparity involving transcription by human perception and machine perception is noticed and found the hand crafted is costlier due to the human effort compared to automated system. The utility of explicit dictionary expansion, the weights having relatively less frequency can be placed in lexical with the Pronunciations based on the choice of handcrafted or automated.. Which implies that smaller quantity dissimilarity for the recognizer, there will be enhanced accomplishment as anticipated. Co-articulation effects connecting few words will be modeled in Multi-words. It was inferred that the cross-word context is nonessential but for few words. There is perceptible improvement in the operation with the use of defined initial decision tree based lexical modeling and with iterative training and testing improve the process.

According to Kessens and Wester [KES00] [ELDA02] (removal and one inclusion rule) studied 5 identified pronunciation rules for Dutch and there is progress in including them. Retranscribe the training data using the variants with the modification in the acoustic models proved the perceptible growth so also the Language model modification with the inclusion of pronunciation probabilities.

The comparison between the DD approach and KB was made by kessen [KES00] in his work. . The proper delete phones in the acoustic models are allowed for deletions with the adoption of deletion rules in an alternative transcription. Though the outcome of the data driven approach is lexicon, both KB approach and the DD approach provide similar results cover transcriptions to the extent of 96%. The frequency counts handle the data-driven rule source, with the basically occurring inequality are sustained. The data driven rule perspective is phone identify and that of knowledge rule perspective is restricted phone clustuers. . Even in the similar transformation KB rules. These studies help to incorporate based on the target language (Telugu language) data to bring the system performance by improving the lexicon.

4.2.3 Data-driven methods

In this approach the speech corpus source to learn the variations of speech that present in it and this knowledge is used to modelling. The limitation is that the disparity found in the given data results more exact for the same database. One of the advantages is that probability computation for the variants build in contrast of knowledge based methods. Depends on the knowledge derived from the data either directly or indirectly, the data driven method is categorized as direct or indirect method. Sufficient knowledge in the

language for a given word is required to modelled through Data-driven direct modeling, while in indirect modelling specific care must be taken to derive both direct and indirect knowledge driven method by observing word formation rule in the language specific variations. For instance "amkai- అంకై with standard manifestation [ax m k ai] and the alternative baseform [an k ai] as the nasal sound before velar [/క/,ం [KAX],/ఖ/,ం [KHAX], /గ/, [GAX],/ఘ/, [gha], palatal (/చ/, [CAX]. /ఛ/, [CHHAX],/జ/, [JAX],ఝ, [JHAX],) and dental [/ట/,[TAX], /ఠ/,[TTTAX]./డ/,[DAX],/ఢ/, [DXHAX] consonant pronounced as nasal sound with oral sound stopping with tongue closing at hard palatal place.

The rationality in Indirect DD modelling approach:

- Here the rule derived for the use of vocabulary data is unlike of test data. Rules assist in simplify the disparity observed in the adaptation data to words which are absent ("un seen words").
- The Rules repeatedly occur and depend on lesser segments than words.
- The A feasible expansion to cross-word rules would be simple.
- Little pronunciation disparity will be there in a across word boundaries [GRE99].

The different methods used for modelling pronunciation variations, is focused in the present thesis work. As different complexity, languages, and corpus are used, comparison of the performances of the different methods is difficult. The performance is assessed based on the Improvements in error rate as compared with a baseline system

4.2.4 DD Direct method of Lexical modelling

This method defining baseforms of pronunciation is important, which are chosen by either knowledge base or source of data. Most ASR lexical model are presently linguistic knowledge specific system rather optimized with ASR performance.

The lexicon is based on the adaptation data and the acoustic models, which are knowledge driven method. The DD based lexical modelling in ASR system shown in Figure 4.7. Data driven method of Telugu ASR system with TelLex (Telugu Lexical Model)[NAG12] [ING00]

Modelling pronunciations are in two ways:

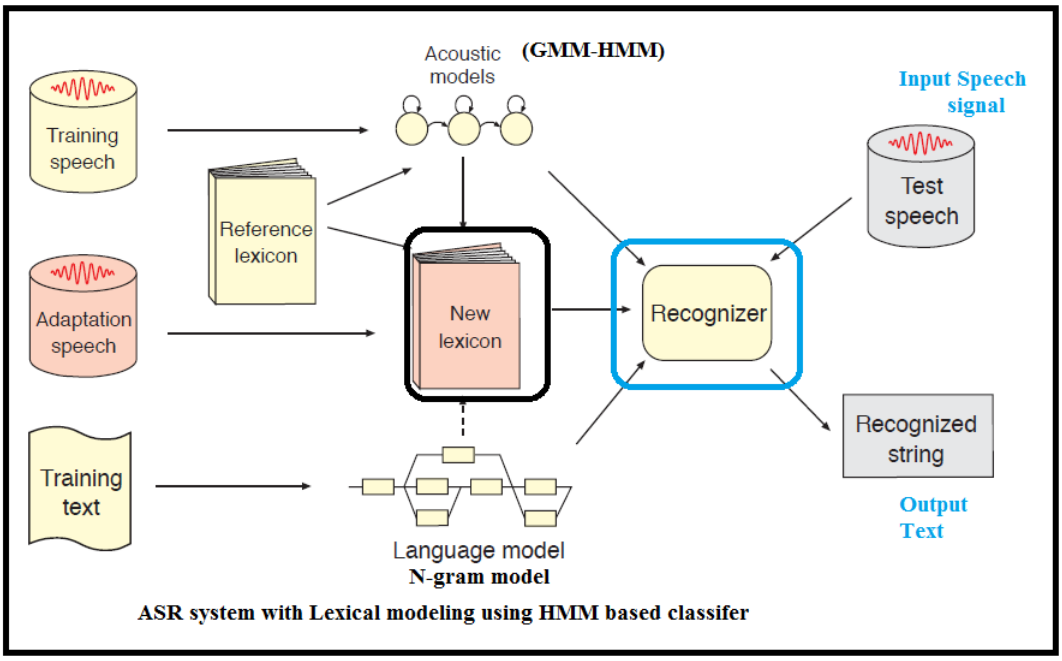


Figure 4.9: Data driven method of Telugu ASR system with TelLex (Telugu Lexical Model)[NAG12] [ING00]

1. Finding variantions in pronunciation and dealt with lexical model.

2. According to Holter and Svendsen truly DD method used with optional baseforms [HOL97a],[HOL97b] and modelled. In these experiments, only a baseline recognizer was used without rules or hand-crafted data. According to Fosler-Lussier in [FOS99a] in DD method and [FOS99b]. training set was retranscribe with frequently occurring words were added to the lexicon. The rules are derived using decision trees.,

4.2.5 Indirect DD Lexical modeling

The rules for capturing the variation between the reference pronunciation and actual pronunciation of a word are automatically derived from data. This is similar to the direct DD with the variant generation of [GIL75] and [WOL01], and also used handcrafted transcription [RIL99]. The experiments were done in indirect modelling, deriving pronunciation rules from data, was conducted by Humphries and Woodland [HUM96], the method is an entirely data-driven approach without use of hand-labelled transcription.

As may be observed from the chapter, the research describes the procedure to explore the different methods followed in review works mainly, on French and Dutch language researchers and their works. The works carried with existing system and build their language specific systems, on one hand examining their linguistic properties and on the other hand, the data is used to build the better lexical model design. Any ASR system main core knowledge source which needs human knowledge and their expertise to develop is the lexical model. Variations in speech signal according to the changes in the written script are possible only with help of lexical model. Hence, lexical modeling main core knowledge base is required for development of new language ASR system, which is the significance of this thesis. Significance in the research in lexical model is explored and evidences presented with the empirical process by presenting relevant results in chapter 5.

****

Chapter 5

Lexical Modelling of TASR – Empirical analysis

5.1 Introduction

This Chapter presents the lexical modelling for system development from ASR system to TASR through empirical process and results analysis. The results were analyzed, based on the traditional performance measures used for ASR system, viz., (i) the recognising words with context of phonetic level mapping and their accuracy in terms of WA and (ii) words that are not matched with the phonetic level mapping is WER [RAB89]. Mis-match in word due to phonetic unit that constitute as word. Mis match in words causing error are classified into three types (i) Insertion (ii) Deletion and (iii) Substitution. Word Error Rate is another parameter for result analysis used in this thesis. The word errors like Insertion and deletion are due to the split (breaking up of words) and merging of words in a test data set, while the substitution errors are due to errors in lexical matching.

5.2 Speech and Text corpus for Lexical Modelling

The initial procedure in TASR is building the isolated speech recognition system for Telugu language. Towards building the corpus (Speech and Text) of sentences and words, the isolated words like names of persons, places, numeric values, phonemes, morphonemes, Telugu words and defined command list are developed for use. The speech corpus is build from speakers' data, which would vary depending upon the age

group (child, young, elders), gender (male and female) voices list and the ecology of the utterance of words. The corpus details presented in figure 5.3, Table 5.8 and in Section 5.2 The speech data was recorded using microphone mounted on head or on the desktop. The Graphical User Interface (GUI) [Praat speech tool], Command User Interface (Linux commands), and voice recorded (on windows recorder or on mobile phones or specific voice recording device) were used in the research. In above said situations, the characteristic of speech data utilized is of mono channel, 16bit, 16kHz sampling frequency (with standard duration of one second or variable duration) as shown Figure 5.1.1. The primary level Telugu learning list (speech and text) comprises of phonemes, morphonemes, phonemic rich words, from regular usage, Telugu website etc obtained. In my research application, the names of locations, railway stations bus stations of Hyderabad, and names of persons, as a different data sets were developed for the speech interactive systems. Next the simple sentences corpus is built with average length of about 4 to 5 words. Further, the speakers' variation (both in native and non native language, gender), names, numbers words and sentences are covered. The speech corpus varies with number of utterances in application. All these collected speech samples are carefully listened and transcribed using Telugu Lipi[SIR10] notation and also new notations defined for UOH lexicon[NAG10]. Few words are covered with standard duration of 1 second for recording and others are based on the uttering the word. Most of the recordings are of male, few females and children. Average age group for recording the samples is 25, with the age ranging from 7 years to about 54 years. Some recordings were from the students of DCIS department with age range from 19 to 25 years. Various data sets were collected and collated in this work [NAG09]. The data set characteristics for building Speech Recognition system for Telugu language are "Clean speech Data samples", "Transcription", "Phone set", "Lexical model or pronunciation dictionary", "Sentence or utterances", "Filler dictionary or non speech dictionary". In addition to this human knowledge or human intervention is required to prepare input data. Thesis focused on the topic "lexical model or pronunciation model" and data prepared mostly with physically developed. The design of the lexical model is tested with the phonetic engine to generate the learned knowledge from this in-house developed data. The experiments are carried out on different type of data sets with Speaker Dependent system and Speaker Independent. Based on the utterance demarcation ASR is an Isolated word Recognition (IWR) or a Continuous Word

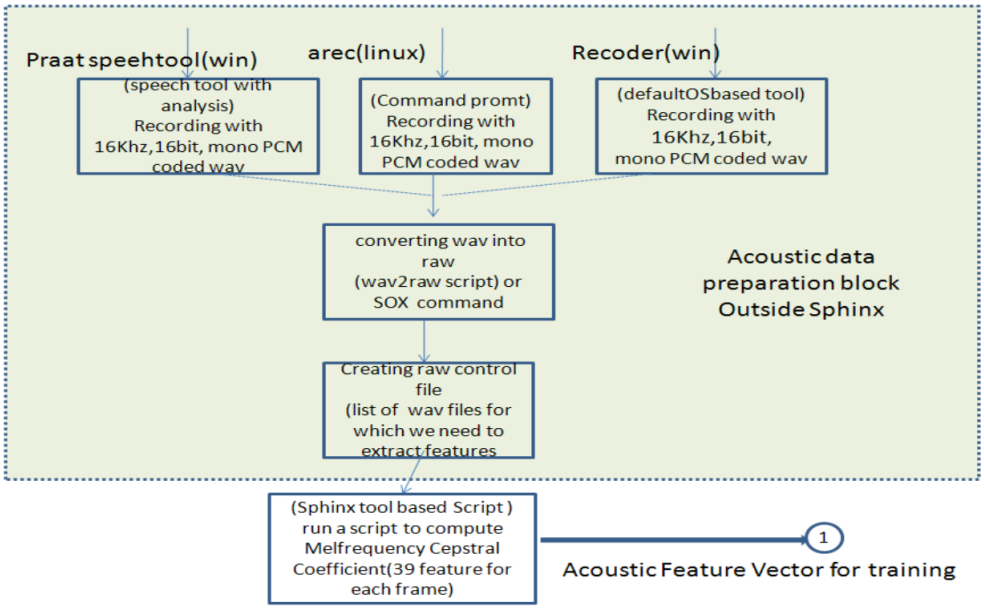


Figure 5.1: Flow diagram for speech corpus acquisition and preprocessing to generate feature vectors [NAG12]

Recognition (CWR). The performance of ASR were verified with standard parameters as WA and WER. The training process validates the data that prepared in house and testing process validate the lexical modelling task. The hypothesis generated by the phonetic engine, is compared with the physical lexical model, for checking the deviation, called as the WER. the mis-match between the reference text and hypothesis text is mainly due to three parameters viz., 1. Insertion error, 2. Deletion error and 3. Substitution error. The first two parameters are resolved by signal analysis and enhancement and third parameter can be reduced by analysis text corresponding speech signal with linguistic knowledge.

The speech corpus acquisition method in different working environment and the tools and methods followed explain above Figure 5.1.The command prompt based script for automated process of recording speech utterance and store into the specified location and also helps to make any automated system for speech interactive mode. The script written for command prompt interface enable user to capture the speech with different duration time based on the application defined. The Figure No. 5.2 gives the detailed description of the command prompt interface script for speech corpus collecting process.

The record command will vary depend on the type of operating system used. The

$ arecord -f s16_LE –d 2 -r 16000 /home/mani/TUTORIAL/wav/00001.wav	
The command description and its function	
Command	**Description of the Function**
arecord	Start the recording speech data through head phone or microphone
-f s16_LE	File format of 16 bit and MONO format
-d 2	The speech file can recorded maximum time is 2 seconds
-f 16000	Bit rate
/home/mani/TUTORIAL/wav/	Path, where to store the recording file
/mani000001	File Name(scope for covering 1lakh utterance)
.wav	File type (speech or audio for research analysis)
it is optional and it depends upon the how much data we need to record. The duration time setting in program is for system control on recording. If we don't use the duration time then there is no maximum recording time in this case we need to use the command 'Ctrl+Z' to stop the recording i.e. user control online.	

Figure 5.2: Linux command for speech utterance recording and storing in specified directory and command description [NAG09] [CMU sphinx documentation]

Linux operating systems was used for my research, and above audio recording commands were used for the corpus collection.

The clean speech data samples were collected is described in figure No. 5.3 The clean speech is devoid of the disturbance signals in manual segmentation. Further, few experiments were carried for automatic segmentation of speech and noise [SAI09] [SAI14]. The thesis is aimed more at concentration to develop speech corpus and then build Telugu ASR system for the purpose of building Language learning tool. Accordingly, the work was concentrated on collection of the speech sample, transcribing the collected sample using the phonetic form (in the Romanized script) that used for the Telugu language. In this work both existing Rich Transcription of Speech (RTS) form of Telugu Lipi [SIR10] forms are used to write the transcription, along with the phonetic form. The address of the utterance is added by using speech demarcation symbols to denote the speech information. This transcription formats are defined already for the sphinx ASR system. Same format is followed to write the transcription of all collected speech samples. The Figure 5.7 shows the Training and decoding process of the ASR system.

with CMU and UOH along with speaker and duration information.

Read Speech corpus Information									UOH	CMU
				speaker class						
data set No.	no.of word	. of speak	Type data	Male	Female	Child	speaker's type	age		
1	54	31	Isolated words	29	-	7	UOHstudent,er	25-50	√	√
2	50	50	Isolated words	-	-	-	UOHStudents	22-25		√
3	2890	1	Isolated words	-	-	-	Student	24		√
4	665	1	Isolated words	-	1	-	UOH researche	34	√	√
5	666	1	Isolated words	1- Non			Emp-Software	27		√
6	100	1	Isolated words				UOH student		√	√
7	100	1	Isolated words				Student		√	√
8	100	1	Isolated words				Student		√	√
9	100	1	Isolated words			1	Student	26	√	√
10	665	1	Isolated words	1		1	Researcher	24,35	√	√
11	15	5	Isolated words	5			student	22-25	√	√
12	466	-	Phonemes			2	Mixed	20-25	√	√
13	100	4	Isolated words				student		√	√
14	25	1	Isolated words	4		1	student		√	√
15	684	1	Isolated words	1						√
16	1509	27	Isolated words	22	5		UOHstudent,emp			√
17	1509	7	Isolated words	5		2	students			√
18	10	30	Isolated words				UOH student			√
19	40	50	Sentences	4	1		MCAstudent		√	√
20	481	50	Isolated words	37	15		MCA,students	20-34	√	√

Figure 5.3: Speech corpus used for empirical process and details of annotation of the utterance with CMU and UOH along with speaker and duration information.

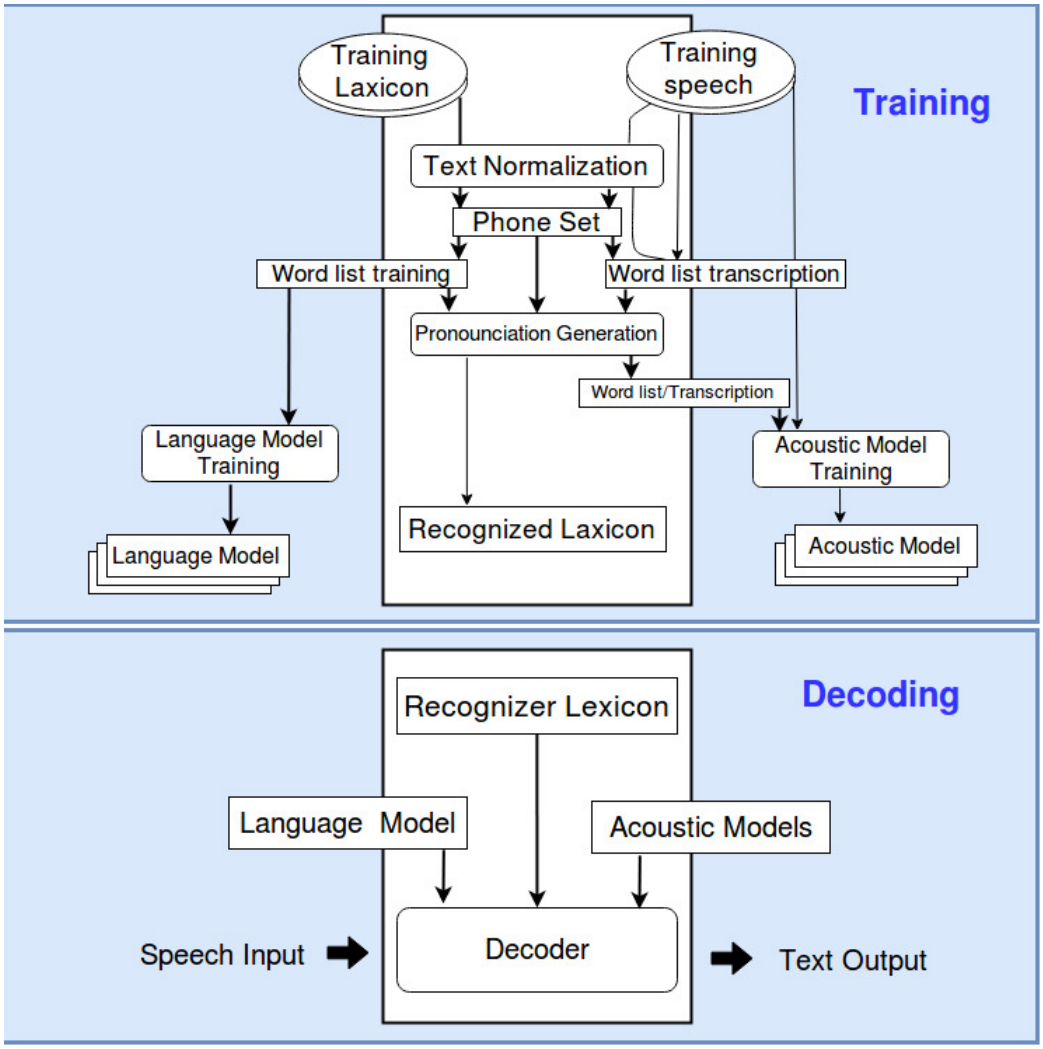


Figure 5.4: Training and decoding(Testing) process and corresponding modules and its input and output flow in ASR system interns of HMM building in acoustic and Language models.[RAB89][SLE99]

5.2.1 Training Procedure

The data set are trained with the system. The training process helps to qualify the data and make the system to learn the knowledge to pronunciation of the words with given knowledge source. Then the same data set is used to test the learned knowledge to execute the phonetic engine, which produces the hypothesized text output for a given input speech data. The reference is written annotation from the speech corpus. The reference recognizer training procedure uses Sphinx III single machine tool for phonetic engine. Three phases in this system (i) data preparation in which data is validated, pronunciations transcribed verified, poor quality speech and noisy speech removal and speech parameter extraction from audio files. These data in terms of text and audio feature file form are the input to the next training phase (ii) In training phase, the data is prepared, by removing noise, the cut audio, fail to pronounce correctly, non clear speech, phonetic letter pronunciations and characteristic files creation [VAC01]. The input to the system is wave file with format of PCM encoded. ASR System for Telugu utterances as isolated words, sentences as an input speech sample files with 16-bit, 16 kHz and mono format audio files. These audio files features are extracted using linearly, logarithmically spaced Mel filter Bank used for cepstral coefficients (MFCC) [SLE99][NAG10] as a feature set. And the size is 13 cepstral coefficients with derivative and double derivative of total 39 feature vectors along with energy coefficients. The lexical model preparation is the final step in data preparation. From the lexical model, the phonetic symbols are extracted. The lexical model is physically defined with the UOH phones and CMU phones with automated online tool "lmtool" written with perl script [ABH10]. With these inputs HMMs are built for a given context of utterance using training process. HMMs are triphone in left-to-right model, which is shown in Figure.5.20

The phone corresponding signal information are mapped by using Diagonal covariance Gaussians that are generated during the training process from the input speech signal interms of the MFCC . Initially, for the Training flat start prototype is generated. Here global mean and variance are used for boot-strap training. The phonemically rich and balanced corpus is used in boot-strapping. While training context Independent model are reestimated using Baum-Welch algorithm [BAU69] [RAB09], then demarcations of background noise and silence are taken care with SIL phone, in the tied

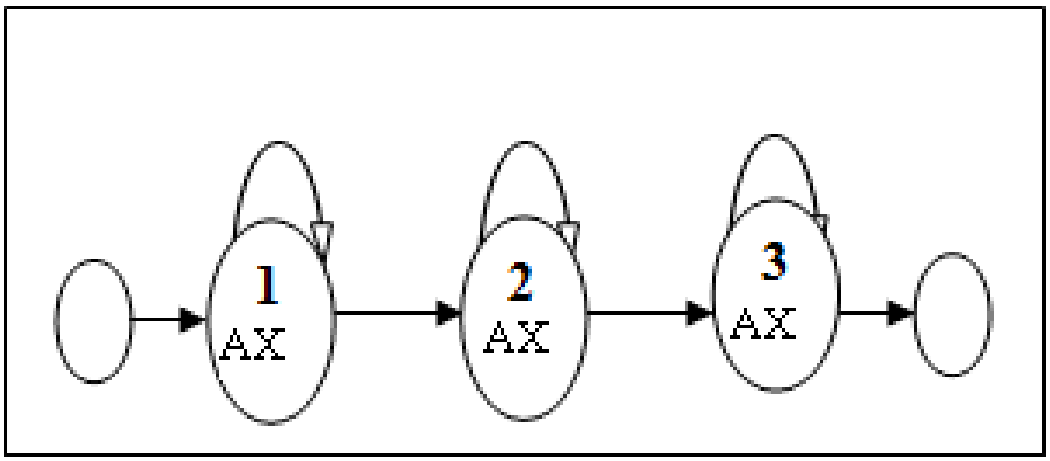


Figure 5.5: Left to Right triphone model generated in Training process of HMM

#base lft rt p attrib tmat ... state id's ...
AI - -- n/a 0 0 1 2 N
AM - -- n/a 1 3 4 5 N
AN - -- n/a 2 6 7 8 N
AX - -- n/a 3 9 10 11 N
C - -- n/a 4 12 13 14 N
D - -- n/a 5 15 16 17 N
DH - -- n/a 6 18 19 20 N
IX - -- n/a 7 21 22 23 N
K - -- n/a 8 24 25 26 N
L - -- n/a 9 27 28 29 N
M - -- n/a 10 30 31 32 N
R - -- n/a 11 33 34 35 N
SIL - --filler 12 36 37 38 N
T - -- n/a 13 39 40 41 N
THX - -- n/a 14 42 43 44 N
UH - -- n/a 15 45 46 47 N
V - -- n/a 16 48 49 50 N
AI C Ri n/a 0 51 52 53 N

#base lft rt p attrib tmat ... state id's ...
AM DH SIL e n/a 1 54 55 56 N
AN SIL DH b n/a 2 57 58 59 N
AX C Ti n/a 3 60 61 62 N
AX D SIL e n/a 3 63 64 65 N
AX DH Ri n/a 3 66 67 68 N
AX K Di n/a 3 69 70 71 N
AX K SIL e n/a 3 72 73 74 N
AX L Ki n/a 3 75 76 77 N
AX L Ri n/a 3 78 79 80 N
AX L SIL e n/a 3 81 82 83 N
AX M Di n/a 3 84 85 86 N
AX M Mi n/a 3 87 88 89 N
AX M SIL e n/a 3 90 91 92 N
AX R Ki n/a 3 93 94 95 N
AX R SIL e n/a 3 96 97 98 N
AX SIL C b n/a 3 99 100 101 N
AX SIL DH b n/a 3 102 103 104 N
AX SIL K b n/a 3 105 106 107 N

#base lft rt p attrib tmat ... state id's
AX SIL L b n/a 3 108 109 110 N
AX SIL M b n/a 3 111 112 113
AX SIL R b n/a 3 114 115 116
AX SIL THX b n/a 3 117 118 119
AX T SIL e n/a 3 120 121 122 N
AX THX SIL e n/a 3 123 124 125
C AX C i n/a 4 126 127 128 N
C C AI i n/a 4 129 130 131 N
C C AX i n/a 4 132 133 134 N
C C UH i n/a 4 135 136 137 N
D AX AX i n/a 5 138 139 140 N
D AX UH i n/a 5 141 142 143 N
DH AN AM i n/a 6 144 145 146
DH AN AX i n/a 6 147 148 149
DH AX IX i n/a 6 150 151 152 N
IX DH SIL e n/a 7 153 154 155 N
IX R SIL e n/a 7 156 157 158 N
K AX AX i n/a 8 159 160 161 N

Figure 5.6: Left to Right triphone model generated in Training process of HMM for UOH Phone (Telugu Phonemes specific phones) set for given input data

models. In forced alignment training the training set is refined by removing the failed corpusBuilding the initialized single-mixture mono phone models, then context depend (CD) models for all tri-phones. These tri-phones context is from the training set.

5.2.2 Empirical process in Lexical Modelling for TASR

The following data sets are used in the empirical process of modelling Telugu language ASR system. Each class of data set and its description given bellow.

Data set 1:I-test: Isolated digits recognition, Names, date and months name. and digits from (0-9) are manually designed data set for ASR system.

Data set II(a):S-test: Application specific simple sentence (40 Telugu sentences and 40 UIS sentences). The vocabulary contains 126, 135 words coverage in the two sets of sentences corpus. , They are small phrases which are used to represent Telugu language and polyglottal words. Both CMU and UOH specific lexical model used for the training and testing of the data set. Speaker variant observation done in these.

Data set II(b):H-test: TASR system for generalization using Hindi Isolated words and Names data are used. The size is 63 words and 100 words of Hindi using UOH phone set based Lexical model which in-house handcrafted is used for training and testing of the system.

Data set III:N- test: 461 utterances collected from the 20 speakers data. Tested in Speaker Dependent mode in TASR system and result presented. With handcrafted lexical model and transcription using UOH phone set. The train and test data set is same i.e individual speakers recognition is performed.

Data set V: P-test: Persian Isolated words of 42 utterances by the male speaker data .The test data used with application to build translation system and input is speech. Hence building ASR system for Persian language is part of the task. 42 Utterances recorded in the room environment with clean speech is used as the data set. With ASR using CMU lexicon tested the system. The same also test with TASR system.

Data set VI:S-test. MMTS and Bus STATATION NAME this data set contains the Hyderabad MMTS train station name recorded with 5 speaker data for building ASR system.

Data set VII:W-test: (Whole Phoneme coverage in the list of words) phonetically rich in Telugu words of 665 and 2890 words. The data set is used evolution, development and testing.

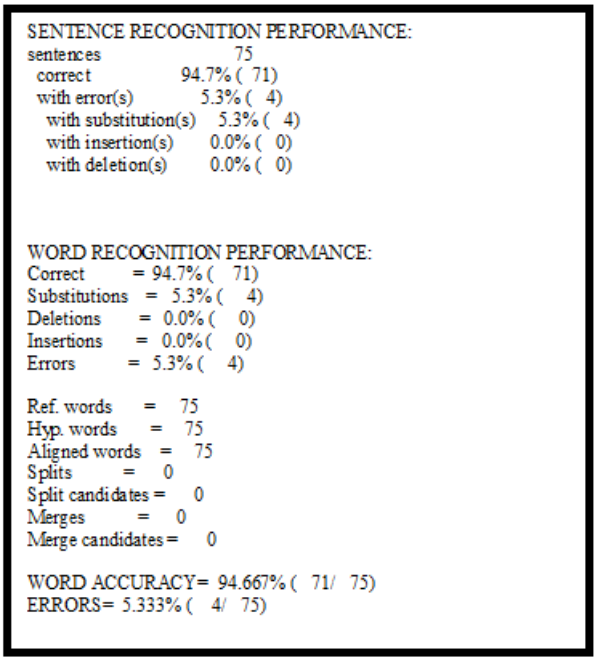


Figure 5.7: Generic ASR system with isolated word recognition system with its output as Sentence Recognition and Word Recognition with their error data analysis.

5.2.3 Empirical data analysis from the outputs of ASR and TASR

In this section recorded training and testing outcomes of ASR and TASR system and compared the results. The measures are Word Error Rates (WER) and Word Accuracy. Summaries of our test results are shown together with results reported for this thesis work in the following Figure No. 5.2 and Figures 5.1 to Figure 5.70 and Figure 5.68 that are carried and consider for the research work. Different data set described in the thesis is use as evaluation in house builds data sets and Test data set category. The following Figure. 5.7 Present the results of the ASR system with Hyderabad MMTS train station names of 15 no. of with 5 speakers data. This Speaker Independent (SI) mode with Isolated Word Recognition (IWR) as a system building with existing tools and methods. The results are taken by speech enhancement i.e., removing silence and noise used for train and test. In the present system both train and test data is same.

Data set 1 Representation of existing Lexical model with handcrafting for Telugu language pronunciation model using US phone set for Unit of sound (Phoneme). Here mapped the Telugu phonemes directly to phone set used for American Pronunciations

S.No.	SpakerID	Data: Gender	54 utterences including digits	
			WA	WER
			T.Word Recognized	T.ErrorWord (Not Recognized
1	spk1	male	47	8
2	spk2	male*	51	4
3	spk3	male	52	3
4	spk4	male	46	9
5	spk5	male*	49	13
6	spk6	male	51	6
7	spk7	male	44	11
8	spk8	male	50	5
9	spk9	male	52	4
10	spk10	male	47	7
11	spk11	female	48	8
12	spk12	female	48	8

Figure 5.8: Railway reservation form filling application data set with 12 speakers with gender variation with same age group of 54 utterances and their performance with ASR system

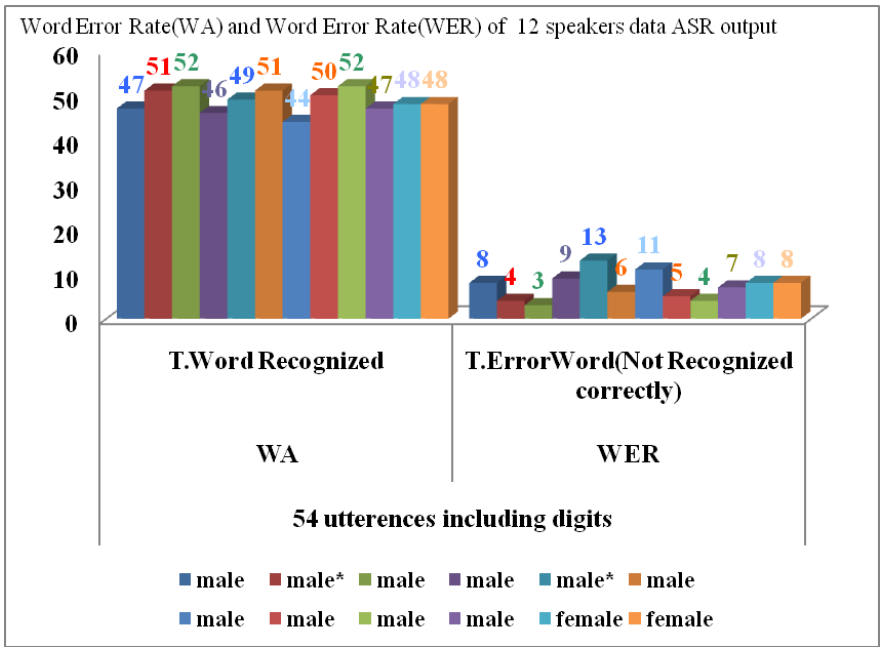


Figure 5.9: raphical comparison view of Railway reservation form filling application data set with 12 speakers with gender variation with same age group of 54 utterances and their performance with ASR system

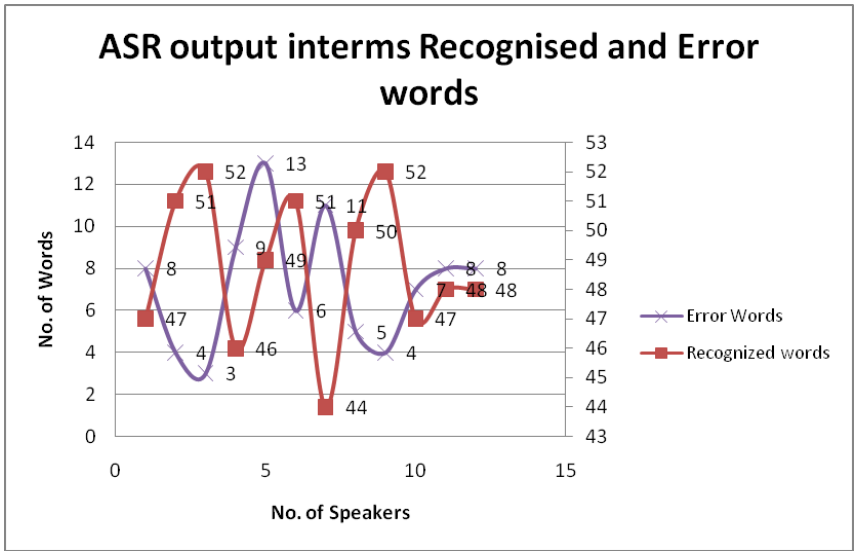


Figure 5.10: Graph on the same scale Recognized word and error words list of Data set I-54 utterances performance with the speaker Independent Mode.

S.no	Types of Error Words	Error word Count	Total #of errors words
1	CONFUSION PAIRS	122	174
2	INSERTED WORDS	14	55
3	DELETED WORDS	0	0
4	MISRECOGNIZED WORDS	49	174
5	SUBSTITUTED WORDS	39	174
6	OCCURENCES OF SPLITS	0	0
7	OCCURENCES OF SPLIT CANDIDATES	38	38
8	OCCURENCES OF MERGES	0	0
9	OCCURRENCE OF MERGE CANDIDATES	0	0

Table 5.1: Railway reservation form filling application data set with 12 speakers with gender variation with same age group of 54 utterances and their performance with ASR system.

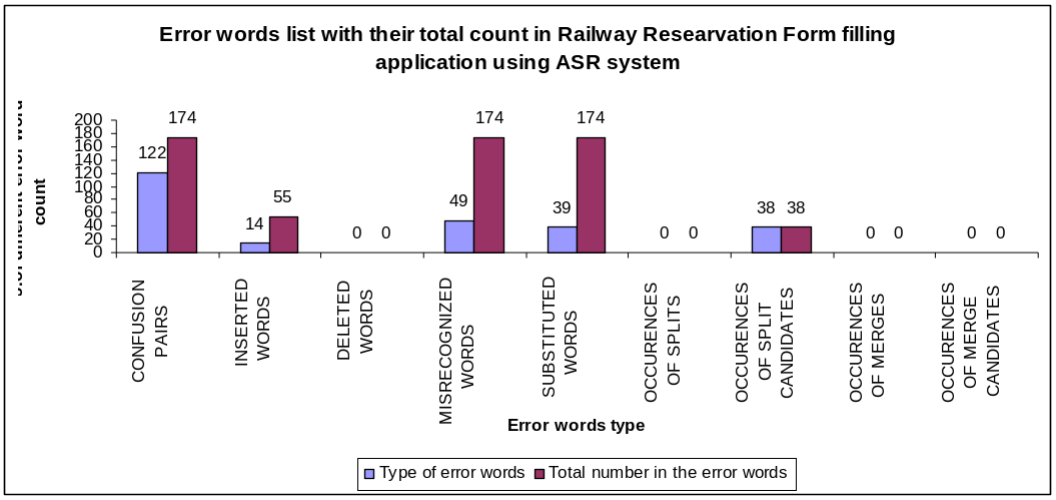


Figure 5.11: Data set VIII results in Speaker Independent mode for form filling application and their ASR output in terms of various error words types.

to build ASR system for the Telugu pronunciation.Speakers are of University of Hyderabad.

5.2.4 Continuous ASR system with Sentence Recognition Analysis:

The ASR system is trained with simple sentences covering an average no.of words in the sentence are 4 to 5 words. 40 sentences recorded in lab environment is used for ASR system and TASR system development.. The bare system with lexical model and acoustic model building using SI mode used to test the performance as far as word Accuracy (WA) and Word Error Rate (WER). WER causes analyzed using this experiment. The figure No. 5.8. Shows the 23 speakers spoken 40 sentence and their recognition performance. The three types of error counts are presented in Figure No. 5.9.

The data set No.II Telugu sentences of 40 numbers each with total of 23 speakers utterance and their ASR output performance, with existing bench mark phone set and Lexical model

The data set No.II Telugu sentences of 40 Utterances with total of 23 speakers utterance and their ASR output performance in Word recognition and their percentage in terms of total words (40 x 23) data set size.

Telugu sentences about the University Information System (UPE project 2004)				
Data set size with speakers and Utterances				
S.No	utterance	Speakers	Total Utterances	Total Words
1	40	13	520	2392
2	40	10	400	1320

Figure 5.12: 40 simple sentence (a) University Information System (b) Telugu sentences data set information

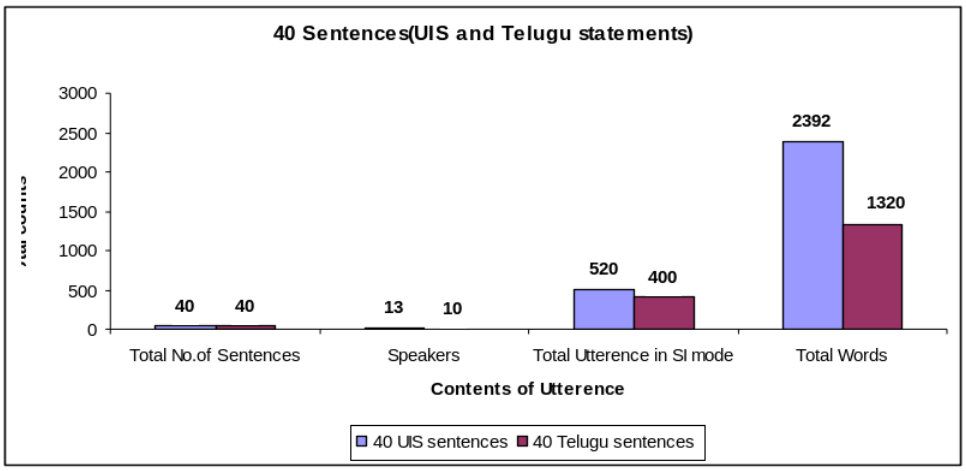


Figure 5.13: 40 simple sentence (a) University Information System (b) Telugu sentences data set information

Telugu sentences about the University Information System (UPE project 2004)				
Data set size with speakers and Utterances				
S.No	utterance	Speakers	Total Utterances	Total Words
1	40	13	520	2392
2	40	10	400	1320

Figure 5.14: 40 simple sentence (a) University Information System (b) Telugu sentences data set information

Number of sentences with WA and WER of single speaker data set	type of words	% of occurrence
correct	241	46.3
error(s)	279	53.7
substitution(s)	270	51.9
insertion(s)	41	7.9
deletion(s)	114	21.9

Table 5.2: ASR performance for SI data set with total words, correctly recognized, error words , Insertion, Deletion and sub situation Errors and over all performance of the WA and WER for 40 sentences with 23 speakers utterances

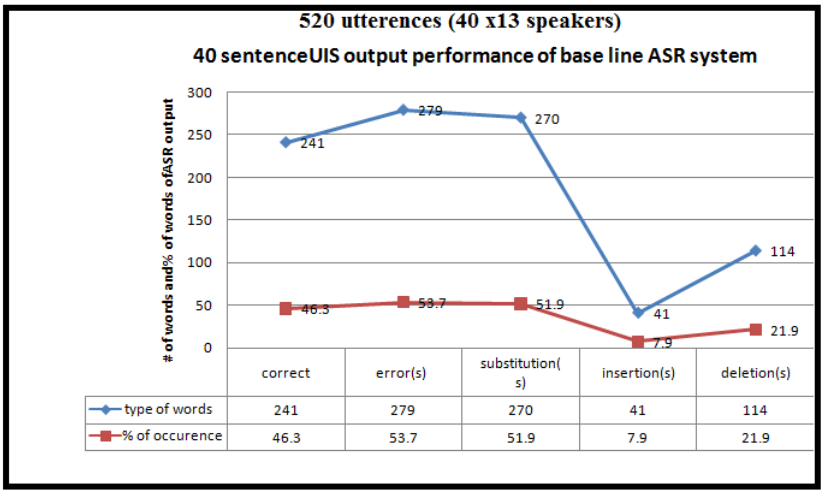


Figure 5.15: Type of words i.e Recognized and errors words distribution in ASR output for the speaker Independent speech data set –II analysis graph

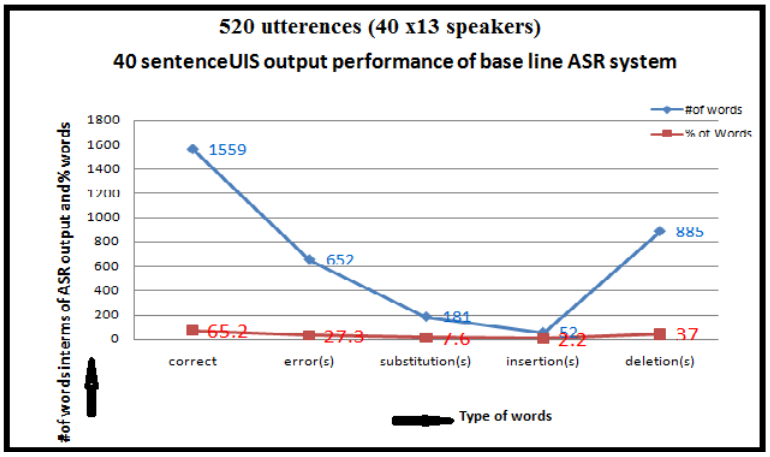


Figure 5.16: The data set No.II Telugu sentences of 40 Utterances with total of 23 speakers utterance and their ASR output performance in Word recognition and their percentage in terms of total words(40 x 23) data set size – the words recognized to the errors.

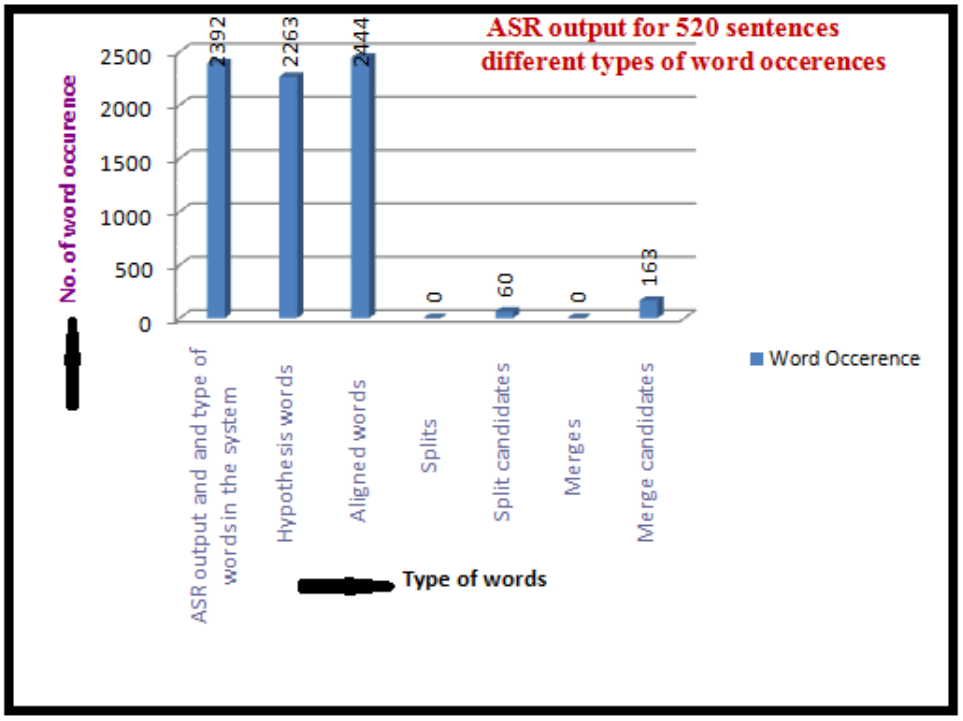


Figure 5.17: The data set No.II Telugu sentences of 40 Utterances with total of 23 speakers utterance and their ASR output performance in context of the occurrence of words.

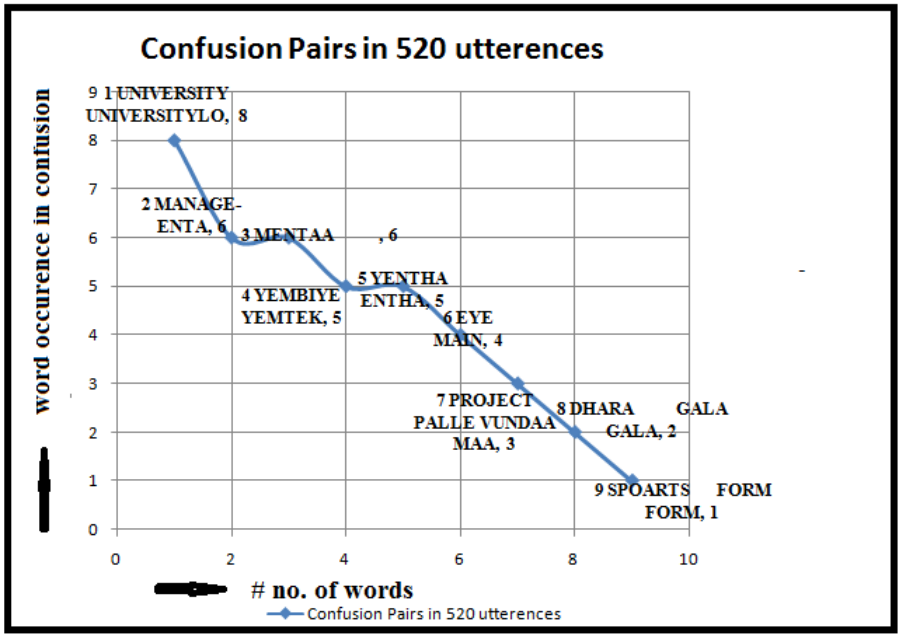


Figure 5.18: Graphical view of the data set No.II Telugu sentences of 40 Utterances with total of 23 speaker's utterance and their ASR output performance in Word recognition output in context of confusion pairs.

s.no	Type data	Word Accuracy
1	Vowel	99
2	Phonemes(CV)	99
3	Word_2 syllables	97
4	Word_3 syllables	92
5	Word_4 syllables	90
6	Word_5 syllables	89
7	Word_above5 syllables	90
8	Sentences	85

Table 5.3: INTTELL Data set peformance with the ASR system with iteratively learning with refining data set and tested

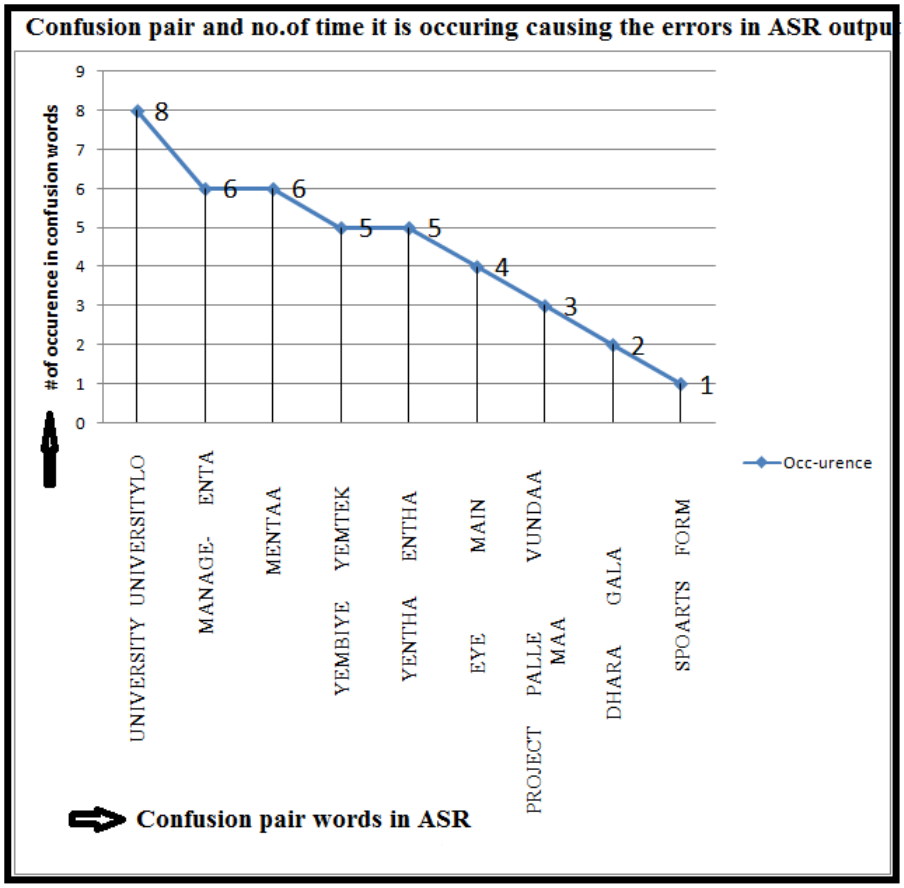


Figure 5.19: Confusion pair words in the UIS data set with 40 Polyglot Telugu sentence listing words and their count in graphical view.

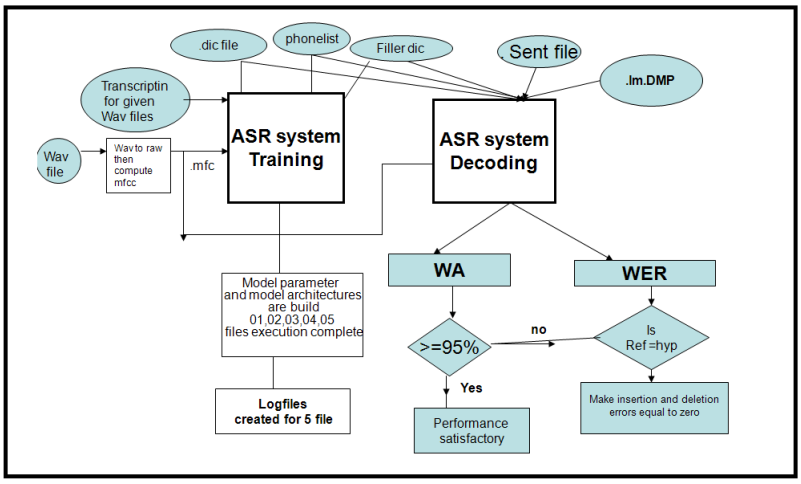


Figure 5.20: Training and Testing procedure in HMM based ASR system and thresholds to refine the data set and recognition process

		అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఋ	ఎ	ఏ	ఐ	ఒ	ఓ	ఔ	అం	అ ః
	Vowels	/AX/	/AA/	/IX/	/IY/	/UH/	/UA/	/RH/	/AI/	/IA/	/AY/	/O/	/OA/	/AW/	/AM/	/AHA/
అ	/AX/	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ఆ	/AA/	0.2	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0
ఇ	/IX/	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ఈ	/IY/	0	0	0	0.9	0	0	0	0.1	0.1	0	0	0	0	0	0
ఉ	/UH/	0	0	0	0	0.8	0.1	0	0	0	0	0	0.1	0	0	0
ఊ	/UA/	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ఋ	/RH/	0	0	0	0	0	0	0.8	0	0.1	0	0	0	0	0	0
ఎ	/AI/	0	0	0	0	0	0	0.1	0.8	0	0.3	0	0	0	0	0
ఏ	/IA/	0	0	0	0	0.1	0	0.1	0	0.9	0.1	0	0	0	0	0
ఐ	/AY/	0	0	0	0	0	0	0	0.1	0	0.8	0.1	0	0	0	0
ఒ	/O/	0	0	0	0	0.1	0	0	0	0	0	0.6	0.4	0	0	0
ఓ	/OA/	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ఔ	/AW/	0	0	0	0	0.1	0	0	0	0	0	0	0	1	0	0
అం	/AM/	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0
అ ః	/AHA/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 5.21: Confusion matrix for the phoneme (Vowel sounds) confusion in 665 words data set using UOH lexical model

Phoneme Substitutions in place Vowels in ASR decoding process

/అ/ =/ఆ/ /ఇ/ /ఉ/ /ఊ/ /ఎ/ /ఏ/ /ఒ/ /అం/

/ఆ/ = /అ/ /ఇ/ /ఉ/ /ఊ/ /ఐ/ /ఔ/ /అం/

/ఇ/ = /అ/ /ఉ/ /ఏ/

/ఈ/ = /ఆ/ /ఇ/ /ఉ/ /ఊ/

/ఉ/ = /అ/ /ఆ/ /ఇ/ /ఊ/ /ఋ/ /ఒ/ /ఓ/

/ఊ/ = /ఉ/

/ఎ/ = /అ/ /ఊ/

/ఏ/ = /ఆ/ /ఇ/ /ఈ/ /ఉ/ /ఊ/ /అం/

/ఐ/ = /ఓ/

/ఔ/ = /అ/ /ఆ/ /ఉ/ /అం/

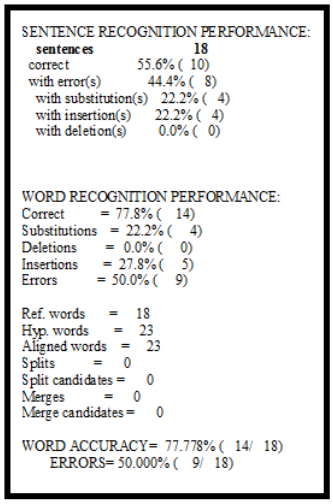
/అం/ = /ఇ/ /ఊ/

Figure 5.22: Confused phonemes in 665 words data set using UOH phoneset and lexical model

	/K/	/KH/	/G/	/GH/	/NYA/	/C/	/CHH/	/J/	/JH/	/INY/	/T/	/TTT/	/D/	/DXH/	/NH/	/THX/	/TTH/	/DH/	/DHH/	/N/	/P/	/F/	/B/	/BH/	/M/	/Y/	/R/	/L/	/V/	/SSH/	/S/	/SH/	/H/	/LH/	/KSH/	/ARW/
/K/	0.5	0.1	0	0	0	0	0.1	0	0	0	0	0.1	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
/KH/	0	0.8	0	0	0	0	0.1	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/G/	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0
/GH/	0	0	0.2	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0
/NYA/	0	0.1	0	0	0.5	0.1	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/C/	0	0	0	0	0	0.8	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/CHH/	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0
/J/	0	0	0	0	0	0.1	0	0.8	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/JH/	0	0.1	0	0	0	0	0.1	0.1	0.7	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/INY/	0	0	0	0	0	0	0	0	0	0.8	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/T/	0	0	0	0	0	0	0	0	0	0	0.9	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/TTT/	0	0	0	0	0	0	0.1	0	0	0	0.4	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/D/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/DXH/	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.8	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/NH/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0.1	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/THX/	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.8	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/TTH/	0	0	0	0	0	0	0.1	0	0	0	0.2	0	0.1	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/DH/	0	0.1	0	0	0	0	0	0	0	0	0	0	0.3	0.2	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/DHH/	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/N/	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0.3	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0
/P/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/F/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/B/	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0.7	0.2	0	0	0	0	0	0	0	0	0	0	0	0
/BH/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.8	0	0.1	0	0	0	0	0	0	0	0	0	0
/M/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
/Y/	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0
/R/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0
/L/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0.1	0
/V/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
/SSH/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0.2	0	0	0	0	0
/S/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
/SH/	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.7	0	0	0	0
/H/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.7	0	0	0
/LH/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0	0	0	0	0	0.5	0	0
/KSH/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0
/ARW/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8

Figure 5.23: Confusion matrix for the phoneme (consonant with implicit vowel) confusion in 665 words data set using UOH lexical model

Figure 5.24: Confused phonemes (Consonants) in 665 words data set using UOH phoneset and lexical model



SENTENCE RECOGNITION PERFORMANCE:
sentences 18
correct 55.6% (10)
with error(s) 44.4% (8)
with substitution(s) 22.2% (4)
with insertion(s) 22.2% (4)
with deletion(s) 0.0% (0)

WORD RECOGNITION PERFORMANCE:
Correct = 77.8% (14)
Substitutions = 22.2% (4)
Deletions = 0.0% (0)
Insertions = 27.8% (5)
Errors = 50.0% (9)

Ref. words = 18
Hyp. words = 23
Aligned words = 23
Splits = 0
Split candidates = 0
Merges = 0
Merge candidates = 0

WORD ACCURACY= 77.778% (14/ 18)
ERRORS= 50.000% (9/ 18)

Figure 5.25: Confused phonemes(Consonants) in 18words data set using UOH phoneset and lexical model

SD system Tested with 5 speaker – Vowel Recognition

Training with one speaker and tested with different speakers

S.No	Utterance	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5		%recognition
1	AX	rh	rh	rh	ai	d	0	0
2	AA	rh	rh	AA	ay	ay	1	20
3	IX	iy	chh	ai	lh	iy	0	0
4	IY	IY	rh	IY	ai	ai	2	40
5	UH	rh	iy	iy	aa	aa	0	0
6	UA	rh	rh	o	aha	aha	0	0
7	RH	RH	iy	RH	RH	iy	3	60
8	AI	ay	iy	n	rh	chh	0	0
9	IA	iy	aha	iy	aha	rh	0	0
10	AY	iy	iy	rh	AY	rh	1	20
11	O	O	O	O	O	O	5	100
12	OA	oo	iy	d	ai	rh	0	0
13	AW	rh	o	aa	ay	ay	0	0
14	AM	aa	rh	oa	rh	rh	0	0
15	AHA	AHA	aa	chh	rh	rh	1	20

Figure 5.26: Speaker Dependent Test output for Telugu phonemes recognition accuracy with Train and test mis-match and phoneme wise 5 speaker

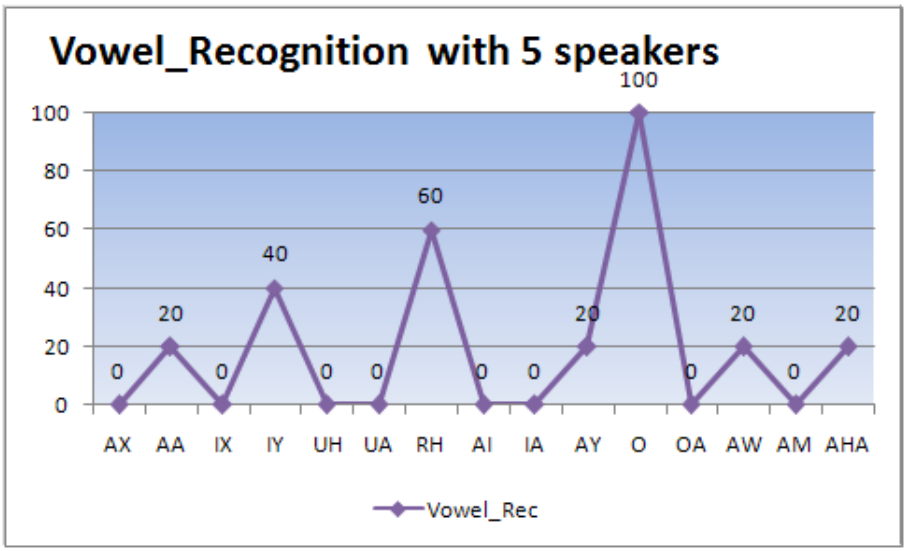


Figure 5.27: Phoneme wise only vowel sound recognition accuracy of 5 speaker's data with graphical representation

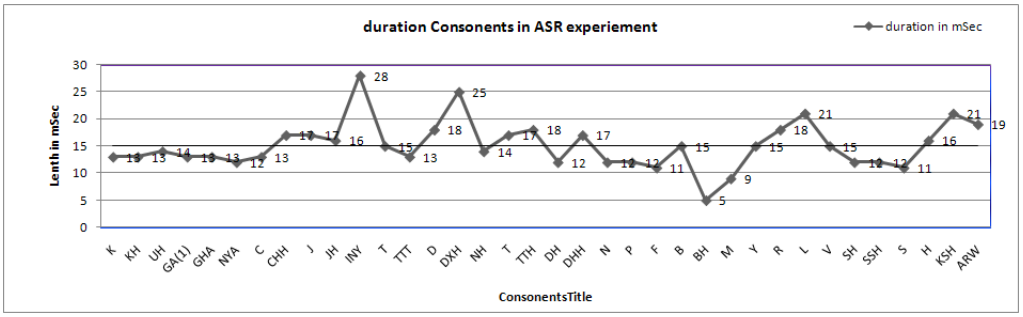


Figure 5.28: 100% recognized accuracy of Telugu phoneme in consonants and their utterence duration time in msec.

Common Substituted Phones in Speaker Dependent ASR system -								
S. No.	1	2	3	4	5	6	7	8
1	[వ్]-[వ]	[ఈ] - [ఈ]	[క్]-[క]	[అ] - [అ]	[త్] - [త]	[అ]-[అ]	[ఇ]-[ఇ]	[ఉ]-[ఉ]
2	[వ్]-[వ]	[ఊ]-[ఊ]	[ణ్]-[ణ]	[అ]-[అ]	[వ్]-[వ]	[ఉ]-[ఉ]	[ఉ]-[ఉ]	
3	[క్]-[క]	[ఆ]-[ఆ]	[క్]-[క]	[అ]-[అ]	[క్]-[క]	[ఉ]-[ఉ]		
4	[ఖ్]-[ఖ]	[ఇ]-[ఇ]	[ర్]-[ర]	[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
5	[గ్]-[గ]	[ఆ]-[ఆ]	[మ్]-[మ]	[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
6	[త్]-[త]	[[ఆ]-[ఆ]	[క్]-[క]	[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
7	[ద్]-[ద]	[ఊ]-[ఊ]	[గ్]-[గ]	[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
8	[ద్]-[ద్]	[ఈ]-[ఈ]		[ఉ]-[ఉ]	[మ్]-[మ]	[ఉ]-[ఉ]		
9	[ధ్]-[ధ]	[అ]-[అ]		[ఉ]-[ఉ]	[మ్]-[మ]	[ఉ]-[ఉ]		
10	[ధ్]-[ధ]	[ఆ]-[ఆ]		[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
11	[న్]-[న]	[అ]-[అ]		[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
12	[న్]-[న]	[అ]-[అ]		[అ]-[అ]	[అ]-[అ]	[ఉ]-[ఉ]		
13	[ణ్]-[ణ]	[అ]-[అ]		[అ]-[అ]	[మ్]-[మ]	[ఉ]-[ఉ]		
14	[ణ్]-[ణ]			[అ]-[అ]	[ఉ]-[ఉ]	[ఉ]-[ఉ]		
15	[మ్]-[మ]			[అ]-[అ]	[వ్]-[వ]	[ఉ]-[ఉ]		
16	[శ్]-[శ]			[అ]-[అ]		[మ్]-[మ]		
17	[చ్]-[చ]			[ఇ]-[ఇ]		[ఉ]-[ఉ]		
18	[చ్]-[చ]			[అ]-[అ]				
19	[వ్]-[వ]			[అ]-[అ]				
20	[మ్]-[మ]			[అ]-[అ]				
21	[ఆ]-[ఆ]			[అ]-[అ]				

Figure 5.29: Common phones in the confusion pairs causing substitution error in TASR data set -665 words with position of phonemes in the word

HMM states - FST of HMM with context phones in Telugu UOH phones										
	#	Columns		definitions						
S.No	#base	lft	Rt	p	attrib	tmat	...	state	id's	...
1	AI	-	-	-	n/a	0	0	1	2	N
2	AM	-	-	-	n/a	1	3	4	5	N
3	AN	-	-	-	n/a	2	6	7	8	N
4	AX	-	-	-	n/a	3	9	10	11	N
5	C	-	-	-	n/a	4	12	13	14	N
6	D	-	-	-	n/a	5	15	16	17	N
7	DH	-	-	-	n/a	6	18	19	20	N
8	IX	-	-	-	n/a	7	21	22	23	N
9	K	-	-	-	n/a	8	24	25	26	N
10	L	-	-	-	n/a	9	27	28	29	N
11	M	-	-	-	n/a	10	30	31	32	N
12	R	-	-	-	n/a	11	33	34	35	N
13	SIL	-	-	-	filler	12	36	37	38	N
14	T	-	-	-	n/a	13	39	40	41	N
15	THX	-	-	-	n/a	14	42	43	44	N
16	UH	-	-	-	n/a	15	45	46	47	N
17	V	-	-	-	n/a	16	48	49	50	N
18	AI	C	R	i	n/a	0	51	52	53	N

Figure 5.30: TASR system HMM state –FST triphone states of UOH phones.

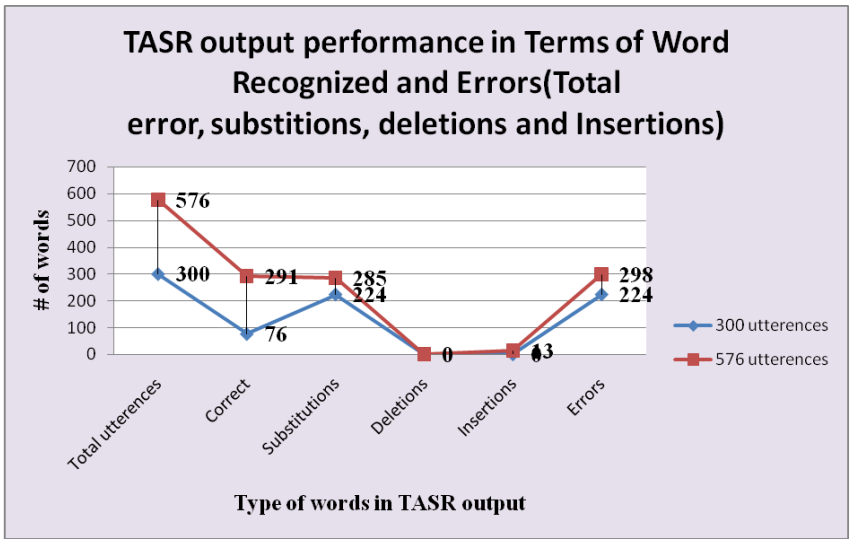


Figure 5.31: Graphical analysis of TASR system performance based on their totoal utterence, recognized word and different error words in their output.

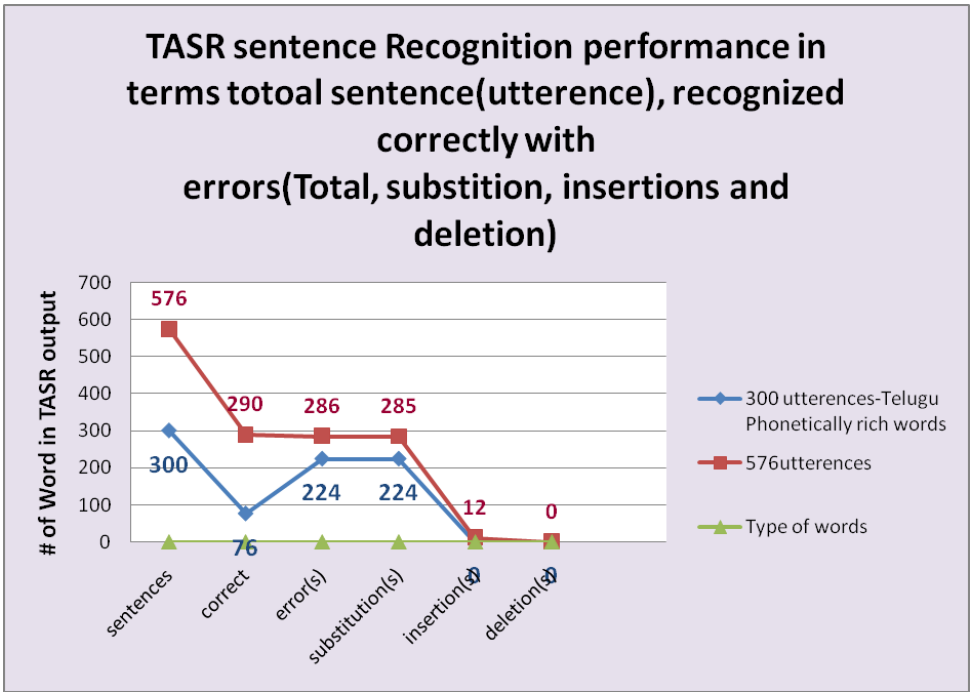


Figure 5.32: Graphical analysis of TASR system performance based on their totoal utterence, recognized word and different error words in their output.

Word Recognition Performance for 300 Telugu words				
	Total No.of words	300		
S.No.	WORD RECOGNITION	PERFORMANCE:		
1	Correct	25.30%	76	
2	Substitutions	74.70%	224	
3	Deletions	0.00%	0	
4	Insertions	0.00%	0	
5	Errors	74.70%	224	
1	Ref. words	300	Indicates No del and Insertion errors	
2	Hyp. words	300		
3	Align. words	300		
4	Splits	0	Indicates the speech utterence are clean and perfectly segmented for utterences	
5	split candidates	0		
6	Merges	0		
7	Merge candidates	0		
1	WORD ACCURACY=	25.33%	76/300	
2	ERRORS=	74.67%	224/300	

Figure 5.33: TASR system performance on 300 phonemic sequence Telugu words from 18Kwords list with its Recognized words, Error words and their types and the % of WA & WER

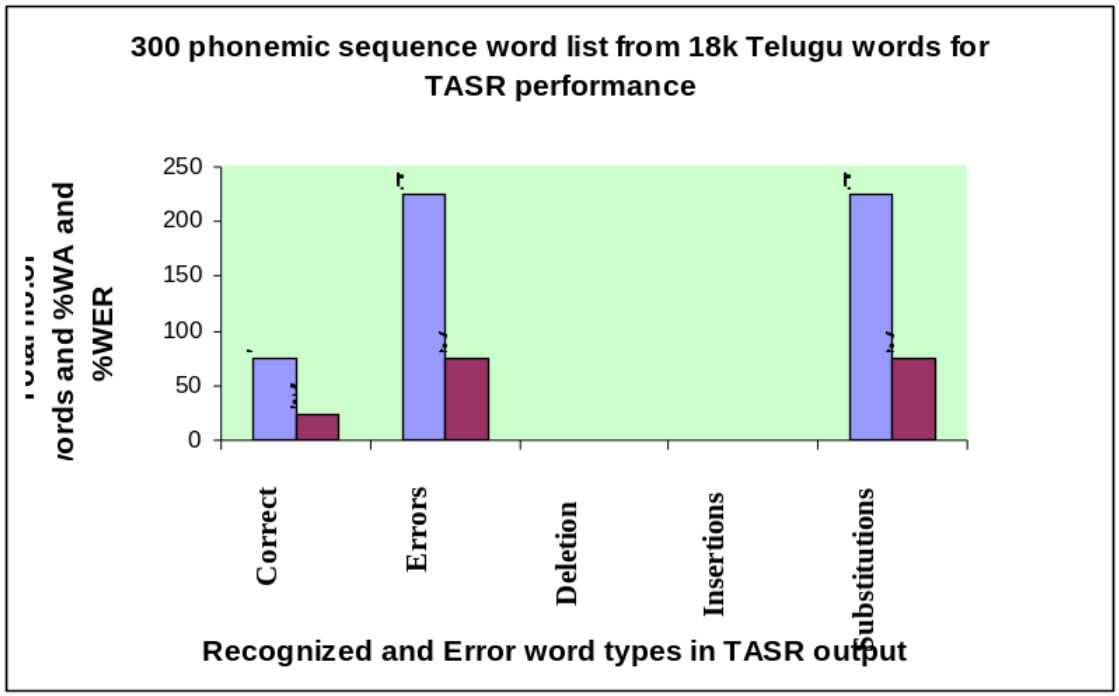


Figure 5.34: Graphical analysis of TASR system performance on 300 phonemic sequence words from 18Kwords list with recognized, error and their types words.

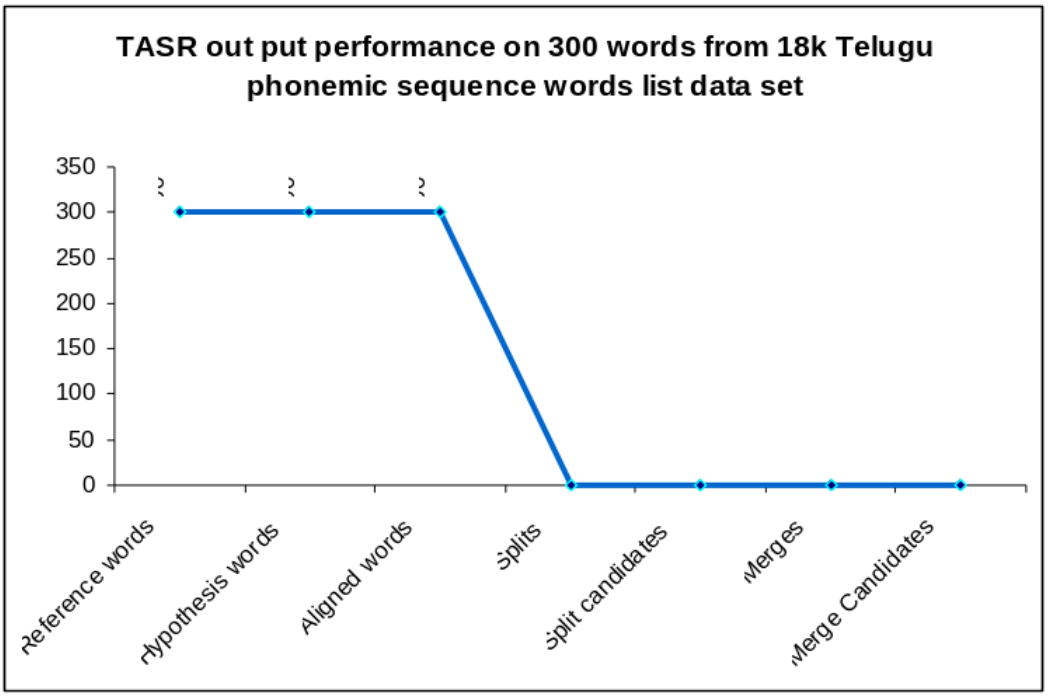


Figure 5.35: Graphical analysis of TASR system performance on 300 phonemic sequence words from 18Kwords list with its of Recognized words, Error words and their types.

INTTELL data with ASR word Accuracy		
s.no	Type data	Word Accuracy
1	Vowel	99
2	Phonemes(CV)	99
3	word_2Syllables	97
4	Word_3syllables	92
5	Word_4syllables	90
6	Word_5syllables	89
7	Word_above5 syllables	90
8	Sentences	85

Figure 5.36: INTTELL Data set peformance with the ASR system with iteratively learning with refining data set and tested

Table 5.4: Comparison of Hypothesis words with Reference words of 100 Hindi words data set with their utterance names that are correctly recognized

S.No	Reference (Transcribed speech using UOH phones)	Correctly Recognized words (System generated Text corresponding to input speech.wav)	Speech utterance file name
1	aagxnaakaarix	aagxnaakaarix	aagnakaari.wav
2	axsvaxshaxn	axsvaxshaxn	aashvaasan
3	axksh	Axksh	aksh.wav
4	axnjaxn	Axnjaxn	anjan.wav
5	axnkuhr	Axnkuhr	ankur.wav
6	axnuhcixthx	axnuhcixthx	anuchith.wav
7	axsaamaanyax	axsaamaanyax	asaamanya.wav
8	axthxaarix	axthxaarix	athaari.wav
9	axvyaxvaxsaayix	axvyaxvaxsaayix	avyavasaayi.wav
10	baans	Baans	baans.wav
11	baxccaa	Baxccaa	bachha.wav
12	baxcpaxn	Baxcpaxn	bachpan
13	baxgaavaxtth	baxgaavaxtth	bagaavath.wav
14	baxndhaxr	baxndhaxr	bandhar.wav
15	biyjgaxnhixthx	biyjgaxnhixthx	beejganith.wav
16	baithxuhkax	baithxuhkax	bethukaa.wav
17	bhaxyaxbhiythx	bhaxyaxbhiythx	bhayabheeth.wav
18	bixsthxaxr	bixsthxaxr	bisthar.wav
19	caap	Caap	chaap.wav
20	caathxaa	Caathxaa	chaatha.wav
21	caxkaxndhhaxr	caxkaxndhhaxr	chakandhar.wav
22	caxkixthx	Caxkixthx	chakith.wav
23	caxshmaa	Caxshmaa	chasma.wav
24	dhhaamaxr	dhhaamaxr	dhaamar.wav
25	dhhaxmaxniy	dhhaxmaxniy	dhamanee.wav
26	dhhaxnuhrdhhaarix	dhhaxnuhrdhhaarix	dhanurdhaari.wav
27	dhhaxnvaan	dhhaxnvaan	dhanvaan.wav
28	dhroah	Dhroah	dhroh.wav
29	dhuhraa	Dhuhraa	dhuraa.wav
30	dhuhrgaxtaxnax	dhuhrgaxtaxnax	durghatna.wav
31	faxrixsthxaa	faxrixsthxaa	faristha.wav
32	gaxdhhaa	Gaxdhhaa	gadhaa.wav
33	gaxlixyaaraa	gaxlixyaaraa	galiyaara.wav
34	gaxpshaxp	gaxpshaxp	gapshap.wav

35	guhbbaaraa	guhbbaaraa	gubbaara.wav
36	haalcaal	haalcaal	haalchaal.wav
37	haxsthxaakruhthxiy	haxsthxaakruhthxiy	hasthaakruthi.wav
38	haxsthxaakruhthxiy	haxsthxaakruhthxiy	hasthaakruthi.wav
39	haxsthxiyng	haxsthxiyng	hastheeng.wav
40	jaanvaxr	jaanvaxr	jaanvar.wav
41	jaxgaanaa	jaxgaanaa	jagaana.wav
42	jaxhaxr	jaxhaxr	jahar.wav
43	kaaryax	kaaryax	kaarya.wav
44	kaxlaabaxj	kaxlaabaxj	kalaabaj.wav
45	kaxvaxc	kaxvaxc	kavach.wav
46	kuhmaanix	kuhmaanix	khumaanee.wav
47	kowshaxl	kowshaxl	koushal.wav
48	krhuhshix	krhuhshix	krushi.wav
49	kuhnvaaraa	kuhnvaaraa	kunvaara
50	kuhthxaar	kuhthxaar	kuthaar.wav
51	laalixthxyax	laalixthxyax	laalithya.wav
52	laxngaxr	Laxngaxr	langar.wav
53	liakhaxkh	liakhaxkh	lekhak.wav
54	maanlianax	maanlianax	maanlena.wav
55	maannaa	Maannaa	maanna.wav
56	maxrhaxm	maxrhaxm	marham.wav
57	maihaxraab	maihaxraab	meharaab.wav
58	niylaamix	niylaamix	neelaamee.wav
59	uapaxr	uapaxr	oopar.wav
60	owsaxthx	owsaxthx	ousath.wav
61	paxriashaaniy	paxriashaaniy	pareshaani.wav
62	paxrixdhhaan	paxrixdhhaan	paridhaan
63	paxrixyaapthx	paxrixyaapthx	paryaapth.wav
64	paxthxaa	Paxthxaa	pathaa.wav
65	paxtth	Paxtth	path.wav
66	powdhhaa	powdhhaa	poudhaa.wav
67	praxthxiykshaa	praxthxiykshaa	pratheeksha.wav
68	praxthxixkual	praxthxixkual	prathikool.wav
69	praxthxixshoadh	praxthxixshoadh	prathishodh.wav
70	puhraalaik	puhraalaik	puraalekh.wav
71	saxhaxmaxthx	saxhaxmaxthx	sahamath.wav
72	saxmbhamdhhiy	saxmbhamdhhiy	sambandhee.wav
73	saxmpaathx	saxmpaathx	sampaath.wav
74	saxamthxuhsht	saxamthxuhsht	santushth.wav
75	saxthxaxrk	saxthxaxrk	satharkh.wav

76	shaxraxdh	shaxraxdh	sharadh.wav
77	shixlpkaar	shixlpkaar	shilpkaar.wav
78	sothxax	Sothxax	sotha.wav
79	thxaanaashax	thxaanaashax	taanaashaah.wav
80	tiajaab	tiajaab	tejaab.wav
81	thxiajoarix	thxiajoarix	tejori.wav
82	thxamgaa	thxamgaa	thamgaa.wav
83	thxuhthxuhlaanaa	thxuhthxuhlaanaa	thuthlaana.wav
84	toalix	toalix	toli.wav
85	uhdhaasiyn	uhdhaasiyn	udhaaseen.wav

S.No		Insertion Error Words	
	Reference	Hypothesis with insertion errors	Utterance name
1	AANKAXNAX********	KUHMAANIX GAXDHHAA	Aankana.wav
2	AXDDHIXVAXKTHXAX ********	SAXTHXAXRK PAXTHXAA	Adhivaktha.wav
3	********* KAXHAABAATHX	KUHTHXAAR KUHMAANIX	Kahaabath.wav

Figure 5.37: Error words due to the Insertions in TASR system recognizer out with 100 Isolated Hindi Speaker Dependent words.

S.No	Reference Word	Substituted words	Speech file	Common Phones	Distinct Phones
1	DHOBAARAX	KUHNVAARAA	(dhobaara)	AA,R	D,K,UH,O,B,V,N
2	HAXJJAAM:	AXNJAXN	(hajjam)	AX,J	H,AA,N,
3	MAYTHXRIY	VAXKIYL	(mythree)	IY	M,AY,THX,R,V,L
4	PRAXSHAXMSAA	SAXMPAATHX	(prasamsaa)	AX,P,S,M,AA	R,SH,THX
5	TOAKAXRIX	TOALIX	(tokhari)	T,OA, IX	K,R, AX
6	VAAYUH	VIXLOAM	(vaayu)	V	AA,IXUH,OA,M
7	VYAATHXAA	AXTHXAARIX	(vyathaa	AA, THX	V,Y,R,IX,AX
8	YOAGYAXTHXAX	DHUHRAA	(yogyatha	Nil	AX,UH, OA,G,THX, DH,Y,R

Figure 5.38: Hindi IWR system performance using TASR and Phoneme recognition with confusion pair analysis.

5.2.5 45 different data set comparison using ASR Result using performance measures

Table 5.5: ASR Decoding result of 45 experiments

s.No	Type of Utterence	No.of Spkrs	Utterence count	Total No.ofUtterence	Recognized words	Error words	Substituted words	Inserted words	Deleted words
1	Music player (command and control)	5	7	35	35	0	0	0	0

2	MMTS train station names for form filling application	5	15	75	71	4	4	0	0
3	Vowel sounds	1	5	5	4	1	1	0	0
4	Kok-borak words	2	10	20	20	0	0	0	0
5	Kok-borak numbers	1	21	21	20	1	1	0	1
6	kok-borak	1	23	23	20	3	3	1	1
7	kok-borok	1	11	11	10	1	1	0	1
8	kok-borok	1	21	21	7	14	14	0	0
9	Telugu wors with telangana speaker	1	665	665	496	169	162	13	7
10	Telugu words -Female	1	665	665	380	285	285	32	0
11	Telugu andra male voice	1	665	665	561	1434	108	0	1326
12	Speaker Dependent Male voice with Andra region(Guntur)	1	665	665	561	104	101	7	3
13	Speaker Dependent(SD) Male voice with Andra region(Guntur)	1	665	665	548	117	106	0	1
14	SD Male voice with Andra region(Guntur)	1	665	665	428	237	236	5	1

15	SD Female-Rayalaseema region(Anantapur)	1	665	665	529	146	135	10	1
16	SD Female-Rayalaseema region(Anantapur)	1	665	665	576	94	85	7	2
17	SD Female-Rayalaseema region(Anantapur)	1	665	665	227	438	437	59	1
18	SD Male voice with Andra region(Guntur)	1	665	665	230	456	435	21	0
19	SD Female-Rayalaseema region(Anantapur)	1	665	665	186	479	468	19	10
20	SD Female-Rayalaseema region(Anantapur)	1	665	665	186	479	468	19	10
21	SD Female-Rayalaseema region(Anantapur)	1	665	665	359	306	305	29	1
22	SD Female-Rayalaseema region(Anantapur)	1	665	665	359	306	305	29	1
23	SD Male voice with Andra region(Guntur)	1	665	665	1	664	661	7	3
24	SD Male voice with Andra region(Guntur)	1	665	665	1	671	671	0	0

25	SD Male voice with Andra region(Guntur)	1	465	465	380	85	84	28	0
26	SD Male voice with Andra region(Guntur)	1	465	465	338	127	126	20	3
27	SD Male voice with Andra region(Guntur)	1	465	465	338	127	126	20	3
28	Speaker Independent(Female Rayalaseema, Male Telangana and Female Telangana	3	100	300	76	224	224	0	0
29	SD-Male-Non-native speaker(Hindi mother tongue)	1	665	665	403	262	261	16	1
30	SD-Male-Non-native speaker(Hindi mother tongue)	1	665	665	533	132	132	5	0
31	SD-Male-Non-native speaker(Hindi mother tongue)	1	665	665	403	262	261	16	1
32	SD Male voice with Andra region(Guntur)	1	200	200	187	13	9	0	4

33	SD Male voice with Andra region(Guntur)	1	200	200	165	35	29	0	6
34	SD-Telugu Phonemes-Female	1	51	51	0	51	23	0	28
35	SD-Telugu Phonemes(Vowel sounds of 16no.s-Female	1	16	16	0	16	13	0	3
36	SD-Telugu Phonemes(Vowel sounds of 16no.s-Female-TASR	1	16	16	0	16	13	0	3
37	SD-Telugu Phonemes(Vowels only)-Female-TASR	1	15	15	13	2	2	0	0
38	SD-Telugu Phonemes(Vowel sounds of 16no.s-Female	1	16	16	16	0	0	0	0
39	SD-Telugu Phonemes(Consonents only)-Female-TASR	1	36	36	31	5	5	0	0
40	SD-Telugu Phonemes(Vowel sounds of 5no.s-Female-TASR	1	5	5	0	5	4	0	1

41	SD-Telugu Phonemes(Vowel sounds of 5no.s-Female-TASR	1	5	5	0	5	4	0	1
42	SD-Telugu Phonemes(51)-Female-TASR	1	51	51	48	3	3	0	0
43	SD-Telugu Phonemes(51)-Female-TASR	1	51	51	48	3	3	0	0
44	Telugu Phoneme for TASR system	1	51	51	51	0	0	0	0
45	SD-Telugu Phonemes-Female	1	51	51	51	0	0	0	0

TASR for Hindi words and their confusion paid word analysis used Phonemic Units and comparing the confusion words with their common and distinct phoneme unit and position of the unit sounds

S.No	WORD	Phonetic Symbols for Lexicon								Input reference mapping to the out put hypothesis	WORD	Phonetic Symbols for Lexicon						
		1	2	3	4	5	6	7	8			1	2	3	4	5	6	7
1	AANKAXNAX	AA	N	K	AX	N	AX				KUHMAANIX	K	UH	M	AA	N	IX	
2	AXDHIXVAXKTHXAX	AX	DH	IX	V	AX	K	THX	AX		SAXTHXAXRK	S	AX	THX	AX	R	K	
3	BAXRMAX	B	AX	R	M	AX					DHHAXNVAAN	DHH	AX	N	V	AA	N	
4	DHOBAARAX	DH	O	B	AA	R	AX				KUHNVAARAA	K	UH	N	V	AA	R	AA
5	HAXJJAAM	H	AX	J	J	AA	M				AXNJAXN	AX	N	J	AX	N		
6	KAXHAABAATHX	K	AX	H	AA	B	AA	THX			KUHMAANIX	K	UH	M	AA	N	IX	
7	MAYTHXRIY	M	AY	THX	R	IY					VAXKIYL	V	AX	K	IY	L		
8	PRAXSHAXMSAA	P	R	AX	SH	AX	M	S	AA		SAXMPAATHX	S	AX	M	P	AA	THX	
9	TOAKAXRIX	T	OA	K	AX	R	IX				TOALIX	T	OA	L	IX			
10	VAAYUH	V	AA	Y	UH						VIXLOAM	V	IX	L	OA	M		
11	VYAATHXAA	V	Y	AA	THX	AA					AXTHXAARIX	AX	THX	AA	R	IX		
12	YOAGYAXTHXAX	Y	OA	G	Y	AX	THX	AX			DHUHRAA	DHH	UH	R	AA			

Figure 5.39: Hindi IWR system performance using TASR and Phoneme recognition with confusion pair analysis.

beginfigure

WORD	RECOGNITION PERFORMANCE:	
	%	#
Correct	77.80%	14
Substitutions	22.20%	4
Deletions	0.00%	0
Insertions	27.80%	5
Errors	50.00%	9

Figure 5.40: Data set no. Having 18 Telugu words with starting with vowel sound"AX

I) Comparison of 665 Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words

Data set: Evaluation Data 1-665

Figure 5.2.2.1 (a): Data set 665 Telugu words

1-665 Utterances - Speaker Dependent								
Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	84	7	2	86	28	3	0	28
CMU	189	24	1	192	189	24	1	192
Description	UOH Ref words: 665 Hyp. words: 670 Align words: 672 Split candidate: 7				CMU Ref words: 666 Hyp. words: 689 Align words: 690 Split candidate: 22			

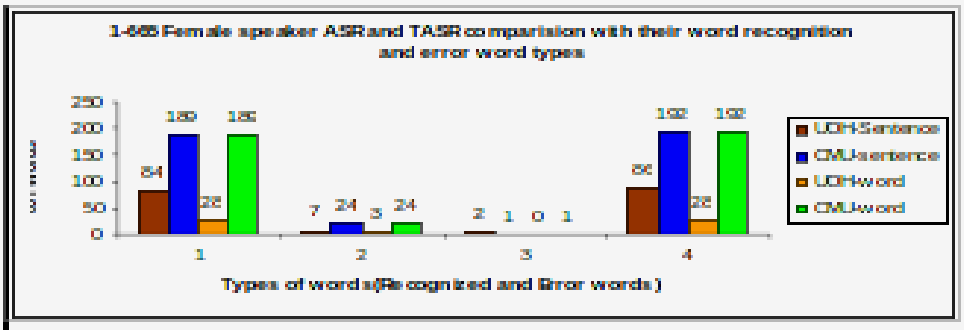


Figure 5.2.2.1(b).:Graphical view of comparison UoH and CMU –Error words

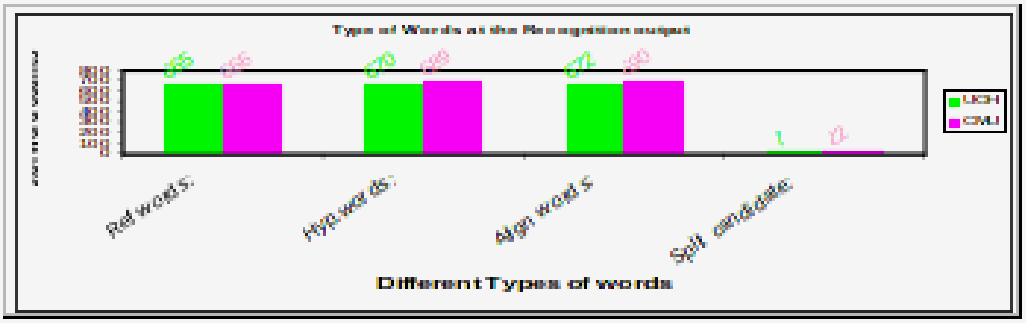


Figure No. 5.2.2.1(c).: Graphical view of comparison – Different words - Recognition output

Figure 5.41: Data set_1:Analysis

II) Comparison of 1-100 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words
Data Set: Female speaker - 1-100 Utterances with SD

Figure 5.2.2.2 (a): Data set - 1-100 Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	5	0	1	6	5	0	1	6
CMU	21	0	1	22	21	0	1	22
Description	UOH Ref words: 100 Hyp. words: 99 Align words: 100 Split candidate: 0				CMU Ref words: 100 Hyp. words: 99 Align words: 690 Split candidate: 22			

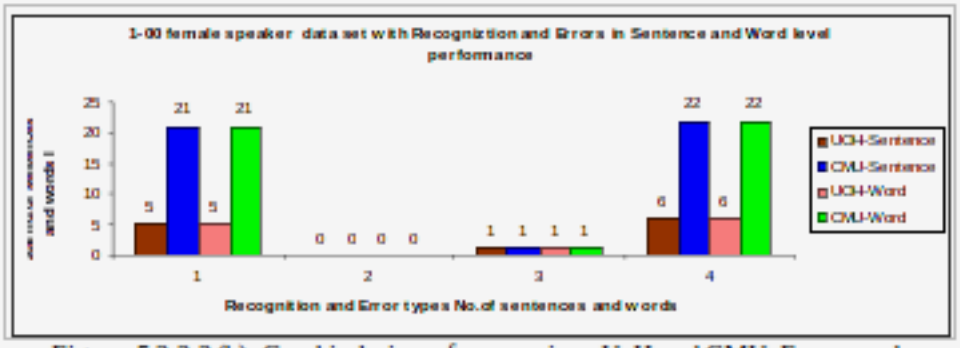


Figure. 5.2.2.2(b): Graphical view of comparison UoH and CMU–Error words

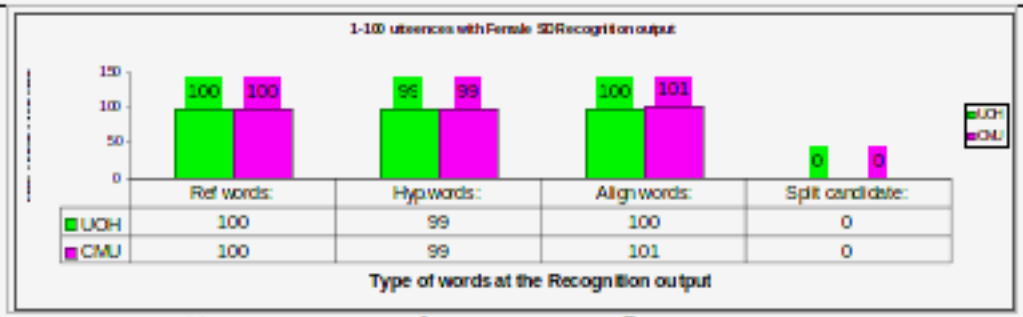


Table 5.2.2.2(c).: Graphical view of comparison – Different words - Recognition output

Figure 5.42: Data set_2:Analysis

III) Comparison of 101-200 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words
Data Set: Female speaker - 101-200 Utterances with SD

Figure 5.2.2.3. (a): 101-200 Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UOH	11	0	0	11	11	0	0	11
CMU	45	0	0	45	45	0	0	45
Description	UOH Ref words: 100 Hyp. words: 100 Align words: 100 Split candidate: 0				CMU Ref words: 100 Hyp. words: 100 Align words:100 Split candidate: 0			

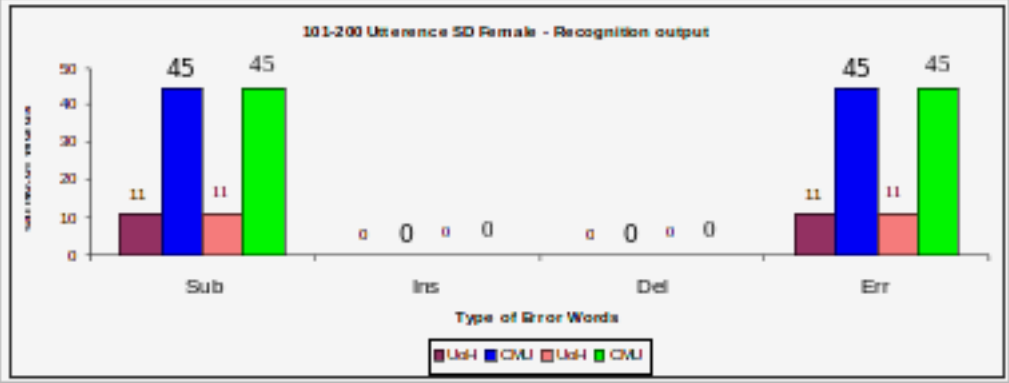


Figure 5.2.2.3(b).: Graphical view of comparison UoH and CMU– Error words

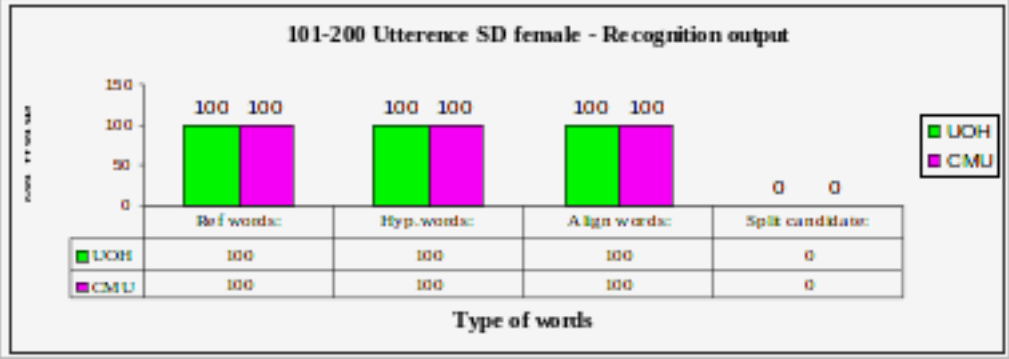


Figure 5.2.2.3(c).: Graphical view of comparison – Different words - Recognition output

Figure 5.43: Data set_3:Analysis

IV) Comparison of 301-400 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words
Data Set: Female speaker - 301-400 Utterances

Figure. 5.2.2.4 (a):301-400 Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	30	3	0	40	39	3	0	42
CMU	54	0	0	54	54	0	0	54
Description	UOH Ref words: 100 Hyp. words: 103 Align words: 103 Split candidate: 2				CMU Ref words: 100 Hyp. words: 100 Align words:100 Split candidate: 0			

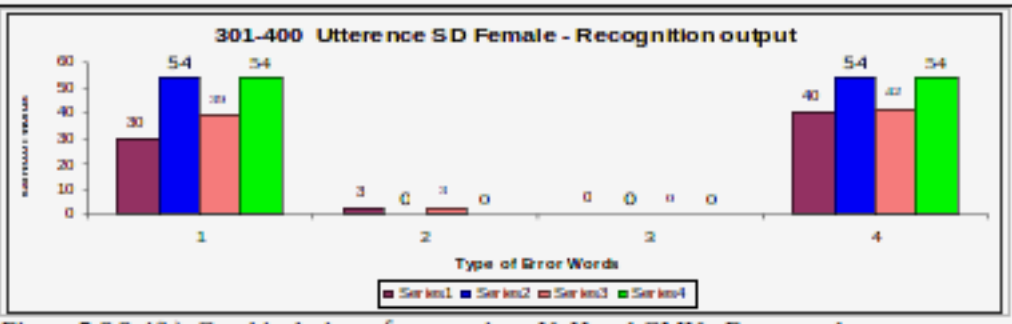


Figure 5.2.2.4(b): Graphical view of comparison UoH and CMU– Error words

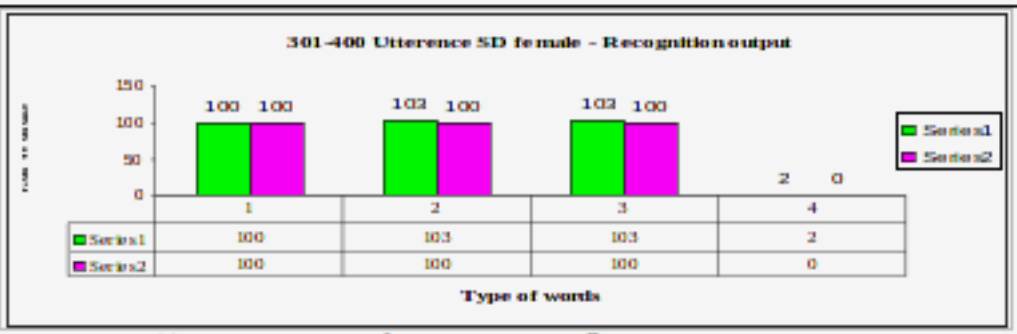


Figure 5.2.2.4(c): Graphical view of comparison – Different words - Recognition output

Figure 5.44: Data set_4:Analysis

V) Comparison of 401-500 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words

Data Set: Female speaker - 401-500 Utterances

Figure 5.2.2.5 (a): 401-500 Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	5	0	1	6	5	0	1	6
CMU	36	2	1	37	36	2	1	39
Description	**UOH** Ref words: 100 Hyp. words: 99 Align words: 100 Split candidate: 0				**CMU** Ref words: 100 Hyp. words: 101 Align words:102 Split candidate: 2			

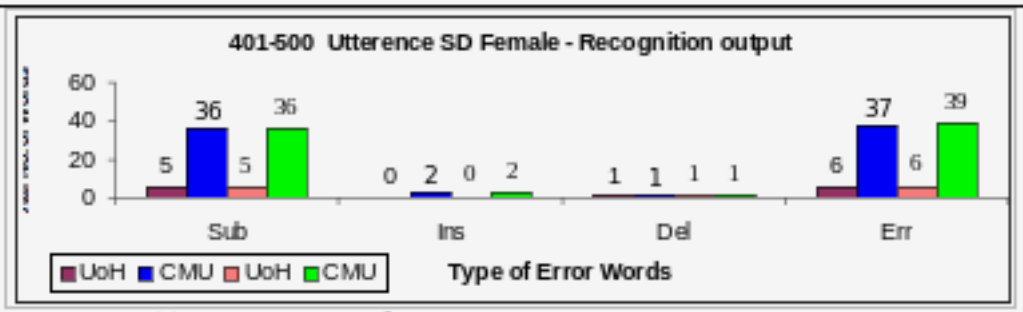


Figure. 5.2.2.5.(b): Graphical view of comparison UoH and CMU– Error words

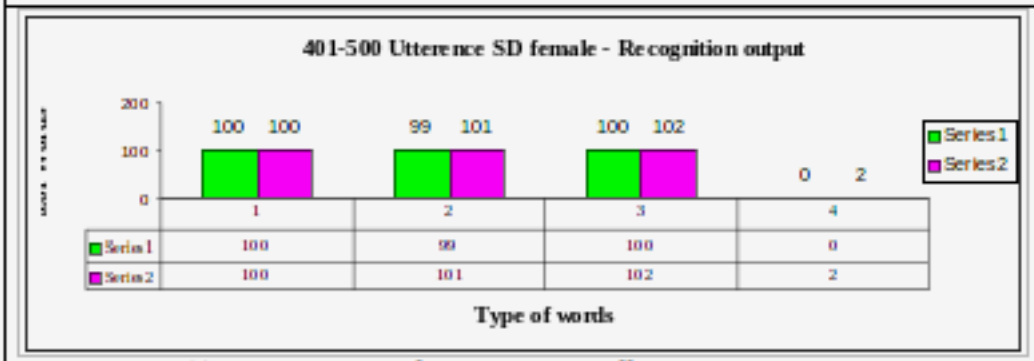


Figure. 5.2.2.5.(c): Graphical view of comparison – Different words - Recognition output

Figure 5.45: Data set_5:Analysis

VI) Comparison of 501-600 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words

Data Set: Female speaker - 501-600 Utterances

Figure 5.2.2.6 (a): Data set - 501-600 Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	7	1	0	7	7	1	0	8
CMU	15	0	0	15	15	0	0	15
Description	**UOH** Ref words: 100 Hyp. words: 101 Align words: 101 Split candidate: 1				**CMU** Ref words: 100 Hyp. words: 100 Align words:100 Split candidate: 0			

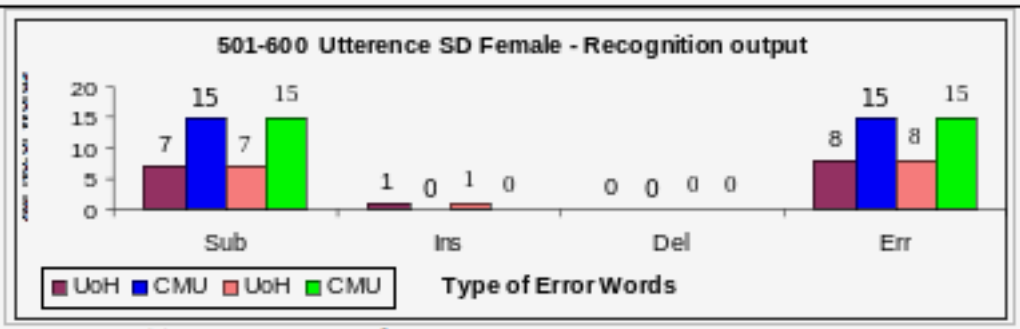


Figure. 5.2.2.6(a).: Graphical view of comparison UoH and CMU– Error words

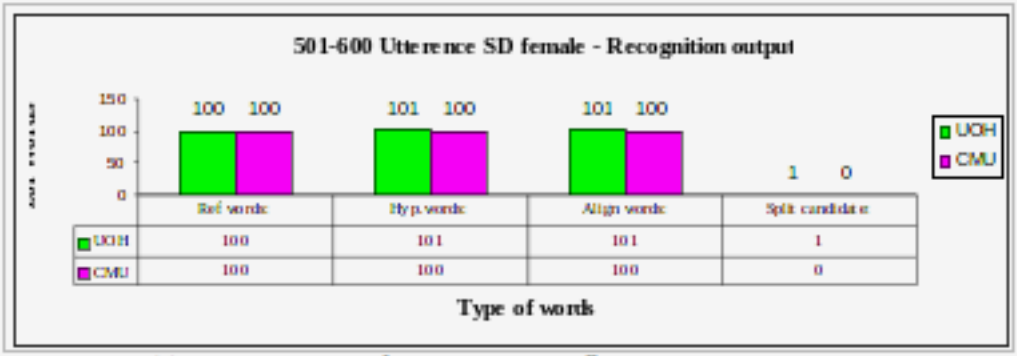


Figure. 5.2.2.6.(b): Graphical view of comparison – Different words - Recognition output

Figure 5.46: Data set_6:Analysis

VII) Comparison of 601-665 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker Phonetically rich words

Data Set: Female speaker - 601-665 Utterances

Figure 5.2.2.7 (a): 601-665 Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	2	0	1	3	2	0	1	3
CMU	10	0	0	10	10	0	0	10
Description	UOH Ref words: 65 Hyp. words: 64 Align words: 65 Split candidate: 0				CMU Ref words: 65 Hyp. words: 65 Align words:65 Split candidate: 0			

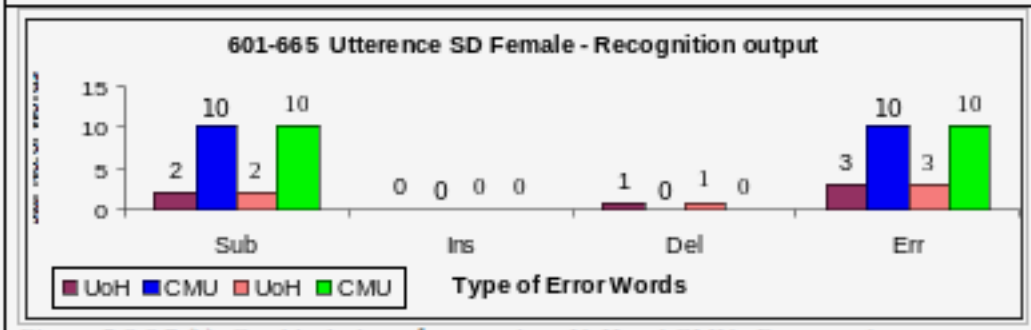


Figure. 5.2.2.7.(b): Graphical view of comparison UoH and CMU– Error words

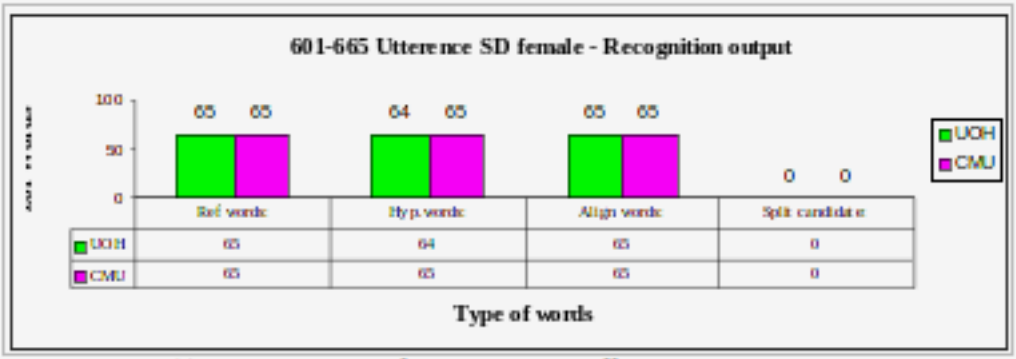


Figure. 5.2.2.7.(c): Graphical view of comparison – Different words - Recognition output

Figure 5.47: Data set_7:Analysis

VIII) Comparison of 665 Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon Female speaker (Telangana Region accent) Phonetically rich words

Data set: Evaluation Data 1-665

Figure 5.2.2.8. (a): 1-665 Utterances- Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	101	5	3	104	101	5	3	109
CMU	236	7	2	238	236	7	2	245
Description	**UOH** Ref words: 665 Hyp. words: 667 Align words: 670 Split candidate: 5				**CMU** Ref words: 665 Hyp. words: 670 Align words:672 Split candidate: 7			

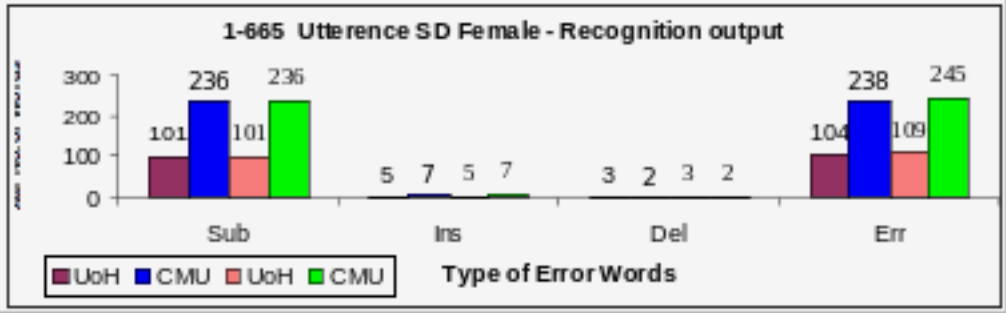


Figure. 5.2.2.8.(b): Graphical view of comparison UoH and CMU– Error words

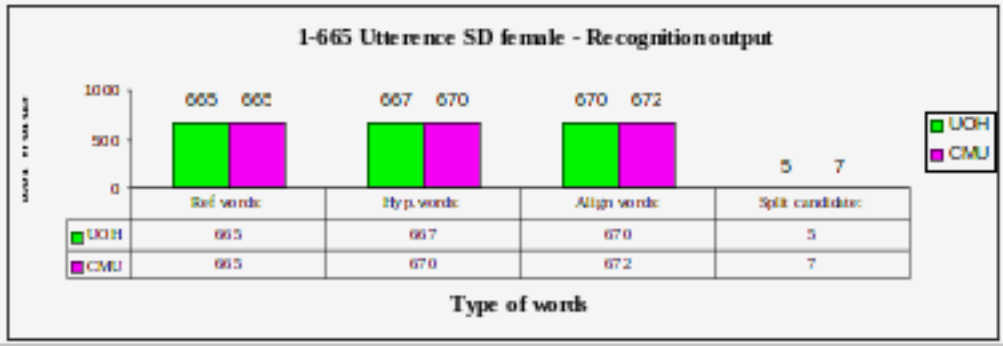


Figure. 5.2.2.8(c).: Graphical view of comparison – Different words - Recognition output

Figure 5.48: Data set_8:Analysis

IX) Comparison of 1-100 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male Speaker data set

Figure 5.2.3.1. (a): 1-100 Utterances-Male -Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	1	0	0	1	1	0	0	1
CMU	10	0	8	18	10	0	8	18
Description	**UOH** Ref words: 100 Hyp. words: 100 Align words: 100 Split candidate: 0				**CMU** Ref words: 100 Hyp. words: 92 Align words:100 Split candidate: 0			

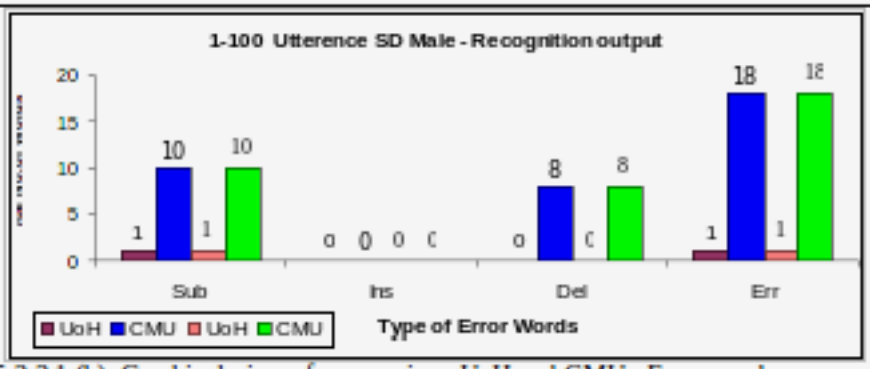


Figure. 5.2.3.1.(b): Graphical view of comparison UoH and CMU– Error words

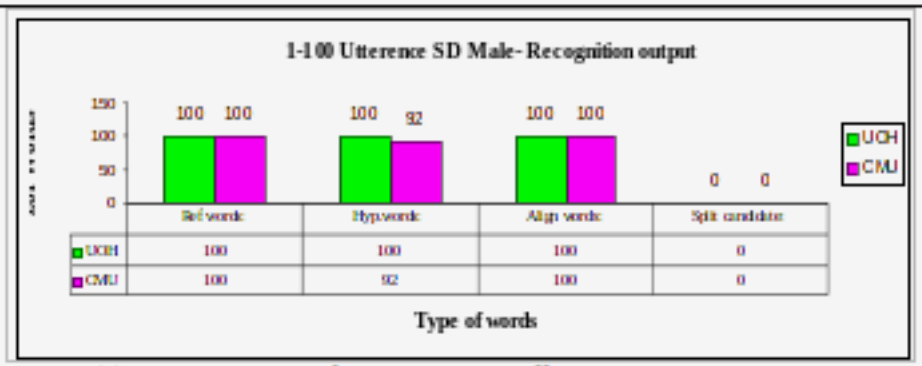


Figure. 5.2.3.1(c).: Graphical view of comparison – Different words - Recognition output

Figure 5.49: Data set_9:Analysis

X): Comparison of 101-200 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male SD data set

Figure 5.2.3.2(a): Data set- 101-200 Utterances – Male - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	3	0	0	3	3	0	0	3
CMU	12	0	4	16	12	0	4	16
Description	**UOH** Ref words: 100 Hyp. words: 100 Align words: 100 Split candidate: 0				**CMU** Ref words: 100 Hyp. words: 96 Align words:100 Split candidate: 0			

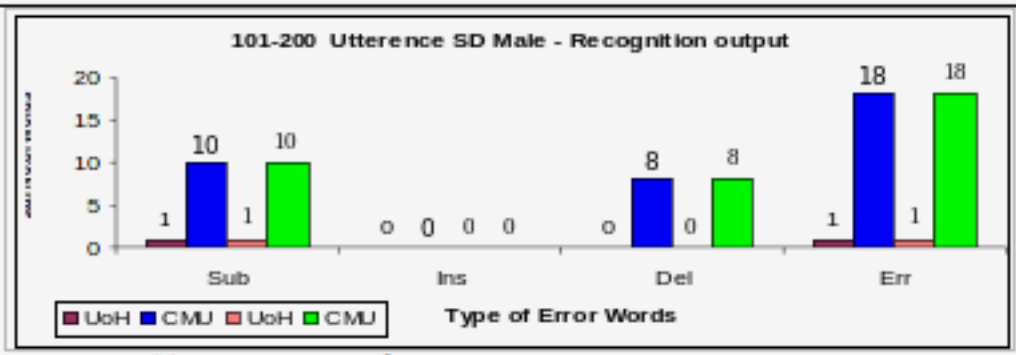


Figure. 5.2.3.2.(b): Graphical view of comparison UoH and CMU– Error words

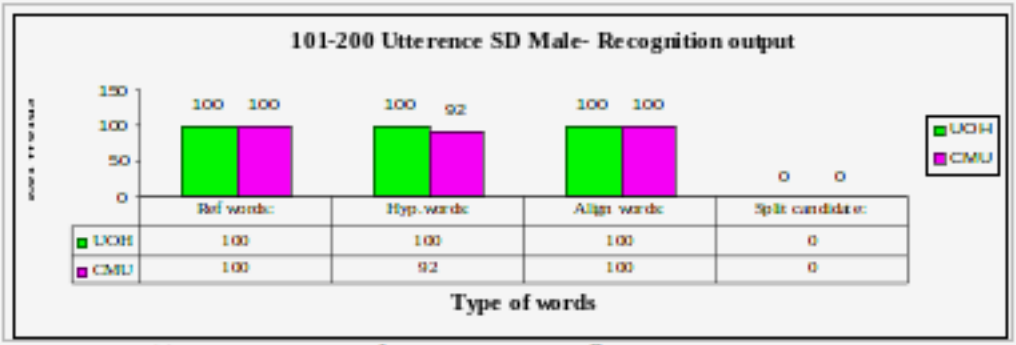


Figure. 5.2.3.2.(c): Graphical view of comparision – Different words - Recognition output

Figure 5.50: Data set_10:Analysis

XI) Comparison of 201-300 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male SD data set

Figure 5.2.3.3.(a): Data set : 201-300-Male- Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	1	0	o	1	1	0	0	1
CMU	18	0	1	19	18	0	1	19
Description	**UOH** Ref words: 100 Hyp.words: 100 Align words: 100 Split candidate: 0				**CMU** Ref words: 100 Hyp.words: 99 Align words:100 Split candidate: 0			

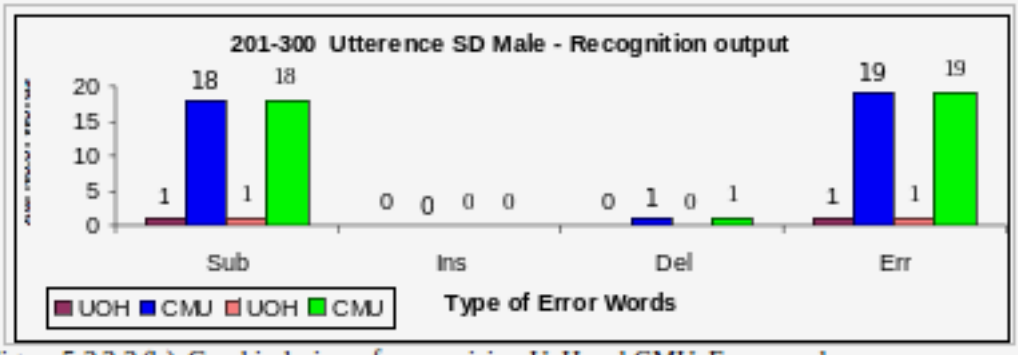


Figure 5.2.3.3.(b): Graphical view of comparision UoH and CMU–Error words

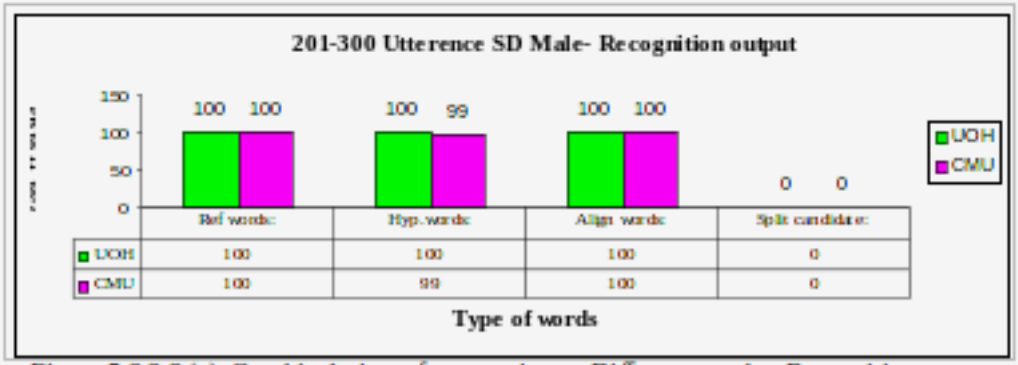


Figure 5.2.3.3.(c): Graphical view of comparison – Different words - Recognition output

Figure 5.51: Data set_11:Analysis

XII) Comparison of 301-400 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male SD data set

Figure 5.2.3.4. (a): Data set- 301-400 Utterances- Male- Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UOH	3	0	0	3	3	0	0	3
CMU	14	0	1	15	14	0	1	15
Description	UOH Ref words: 100 Hyp.words: 100 Align words: 100 Split candidate: 0				CMU Ref words: 100 Hyp.words: 99 Align words: 100 Split candidate: 0			

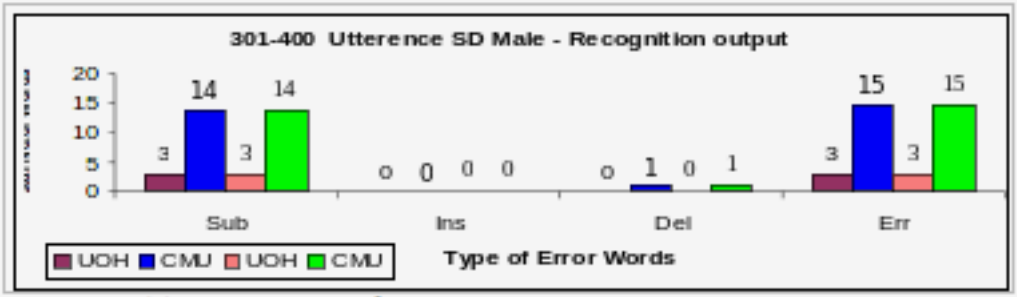


Figure 5.2.3.4.(b): Graphical view of comparison UoH and CMU–Error words

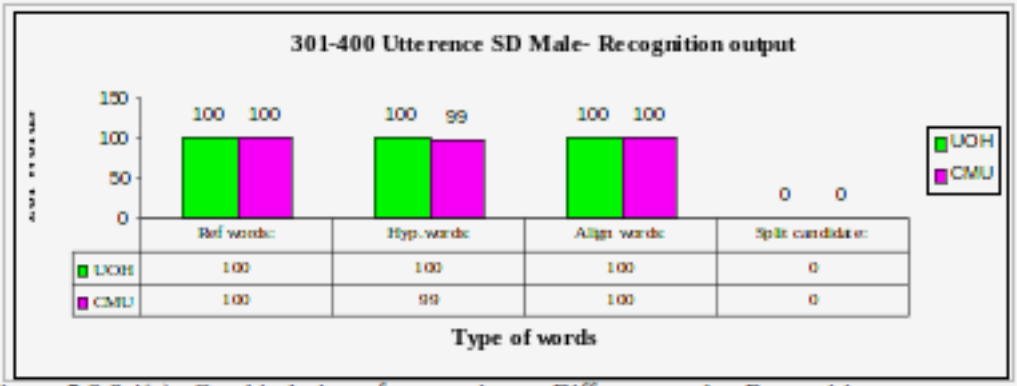


Figure. 5.2.3.4(c).: Graphical view of comparison – Different words - Recognition output

Figure 5.52: Data set_12:Analysis

XIII) Comparison of 401-500 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon on male SD data set

Figure 5.2.3.5.(a): Data set - 401-500 Utterances- Male Speaker Dependent								
Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UOH	2	0	0	2	2	0	0	2
CMU	26	0	0	26	26	0	0	26
Description	UOH Ref words: 100 Hyp. words: 100 Align words: 100 Split candidate: 0				CMU Ref words: 100 Hyp. words: 100 Align words:100 Split candidate: 0			

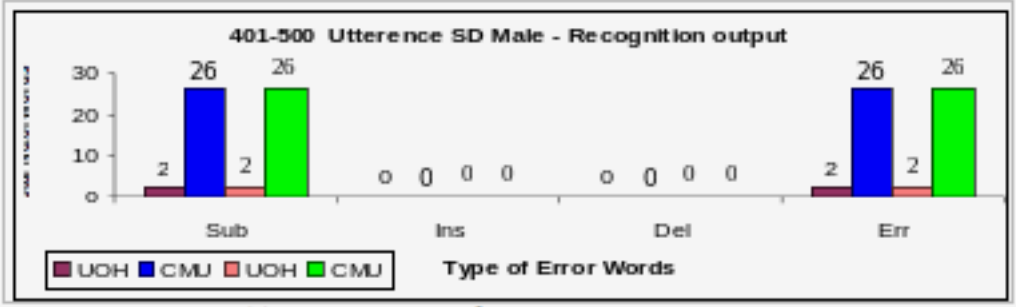


Figure. 5.2.3.5.(b): Graphical view of comparison UoH and CMU– Error words

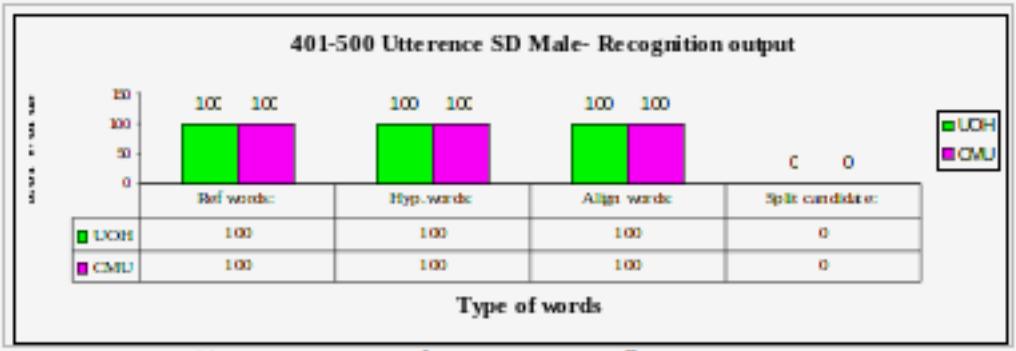


Figure. 5.2.3.5.(c): Graphical view of comparison – Different words - Recognition output

Figure 5.53: Data set_13:Analysis

XIV) Comparison of 501-600 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male SD data set

Figure 5.2.3.6.(a): Data set 501-600 - Utterances - Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UoH	2	0	0	2	2	0	0	2
CMU	15	0	1	16	15	0	1	16
Description	UOH Ref words: 100 Hyp. words: 100 Align words: 100 Split candidate: 0				CMU Ref words: 100 Hyp. words: 99 Align words:100 Split candidate: 0			

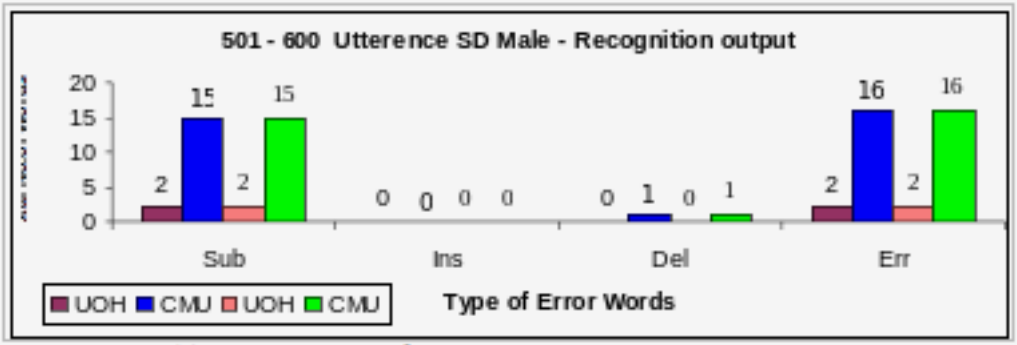


Figure. 5.2.3.6.(b): Graphical view of comparison UoH and CMU– Error words

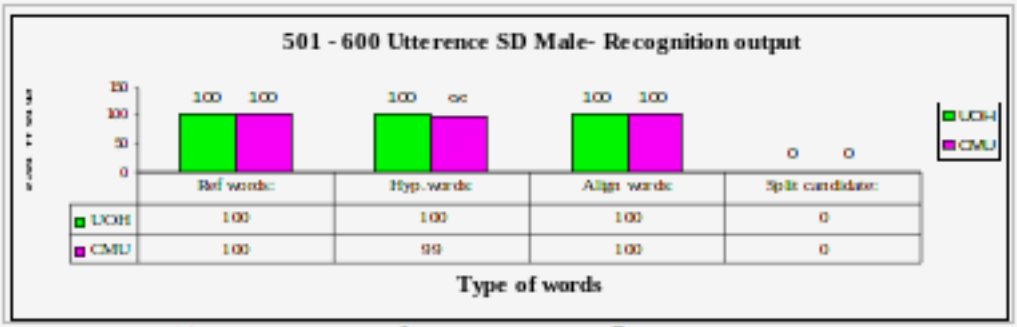


Figure. 5.2.3.6.(c): Graphical view of comparision – Different words - Recognition output

Figure 5.54: Data set_14:Analysis

XV) Comparison of 601-665 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male SD data set

5.2.3.7.(a): Data set - 601-665 - Male - Speaker Dependent								
Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UOH	2	0	0	2	2	0	0	2
CMU	7	0	0	7	7	0	0	7
Description	UOH Ref words: 65 Hyp.words: 65 Align words: 65 Split candidate: 0				CMU Ref words: 65 Hyp.words: 65 Align words:65 Split candidate: 0			

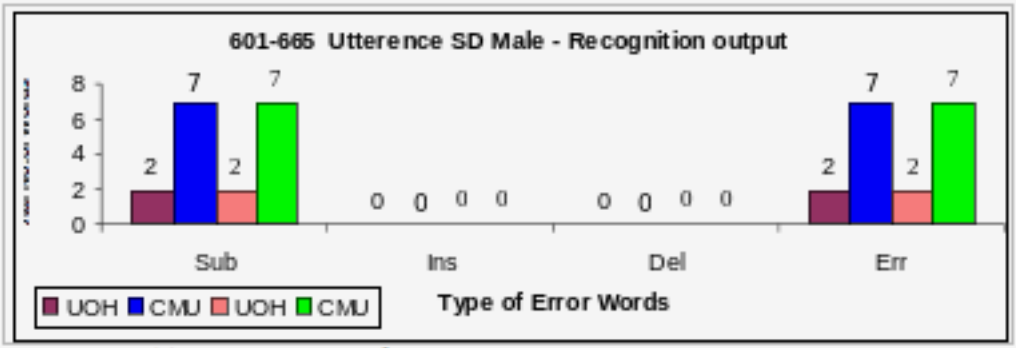


Figure. 5.2.3.7.(b): Graphical view of comparision UoH and CMU– Error words

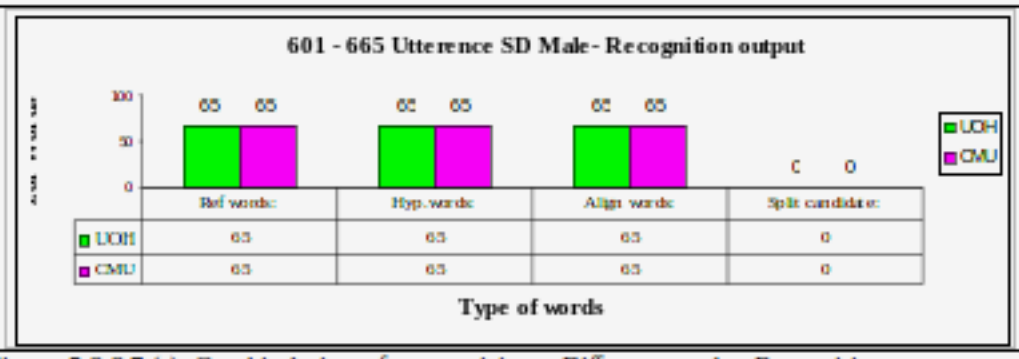


Figure. 5.2.3.7.(c): Graphical view of comparision – Different words - Recognition output

Figure 5.55: Data set_15:Analysis

XVI) Comparison of 1-665 chunk of Telugu IW data set with ASR and TASR system in context of UOH and CMU lexicon with male SD data set

Figure. 5.2.3.8.(a): Data Set - 1-665 : Male Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
UOH	87	7	0	87	87	7	0	94
CMU	232	11	1	233	232	11	1	244
Description	UOH Ref words: 666 Hyp.words: 672 Align words: 672 Split candidate: 7				CMU Ref words: 665 Hyp.words: 675 Align words:676 Split candidate: 11			

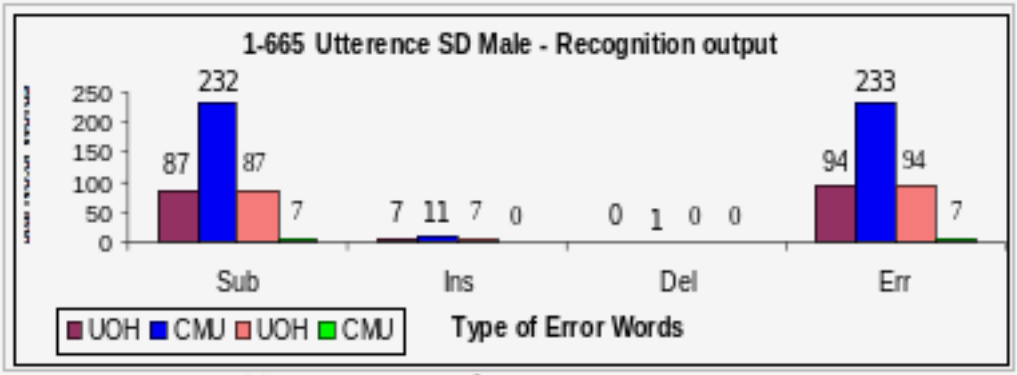


Figure. 5.2.3.8.(b): Graphical view of comparision UoH and CMU– Error words

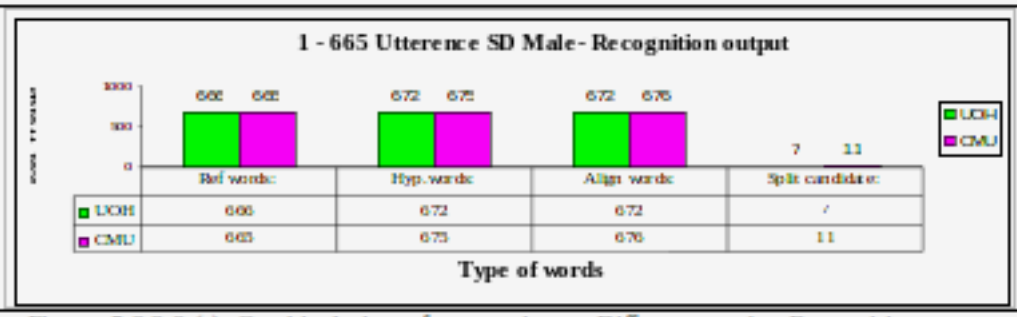


Figure. 5.2.3.8.(c): Graphical view of comparison – Different words - Recognition output

Figure 5.56: Data set_16:Analysis

XVII) Persian data set: ASR system for Persian language and their recognition output performance in terms of Sentence Recognition and Word Recognition.

42 Persian speaker data recognition accuracy with CMU lexicon, wherein 38 words are recognized in word recognition and 41 are recognized in sentence. The speech is the Persian male voice to develop and test with same data set

Figure 5.2.3.9 (a): Data set - 42 Utterances in Persian Language– Male- Speaker Dependent

Name of the file.	Sentence Recognition.				Word recognition.			
	Sub	Ins	Del	Err	Sub	Ins	Del	Err
CMU	1	3	0	5	1	3	0	4
Description	CMU Ref words: 42 Hyp.words: 46 Align words:46 Split candidate: 0							

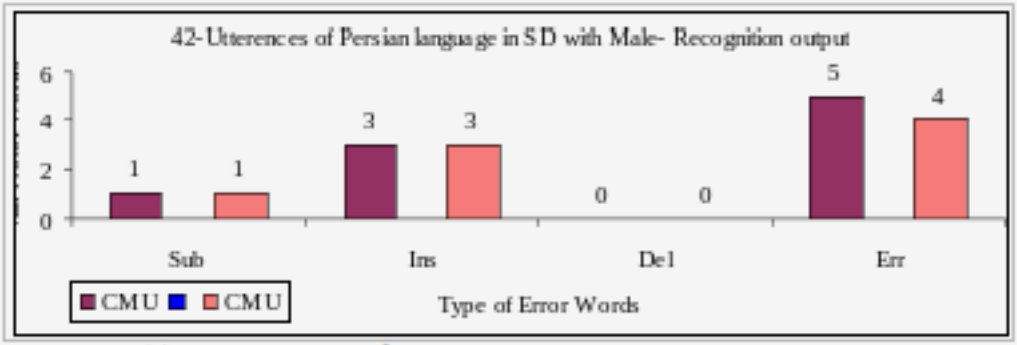


Figure. 5.2.3.9(b) : Graphical view of comparison UoH and CMU– Error words

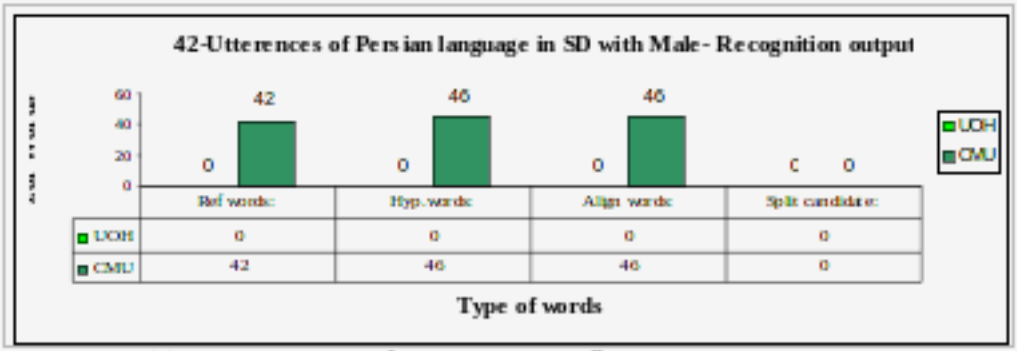


Figure. 5.2.3.9 (c) : Graphical view of comparison – Different words - Recognition output

Figure 5.57: Data set_17:Analysis

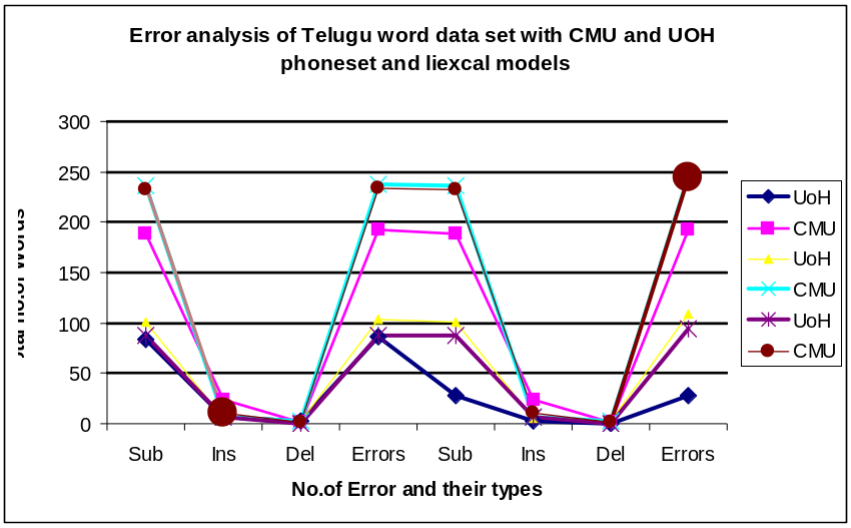


Figure 5.58: Error analysis and comparison of Recognition output in terms of sentence and word recognition for Telugu evaluation data set.

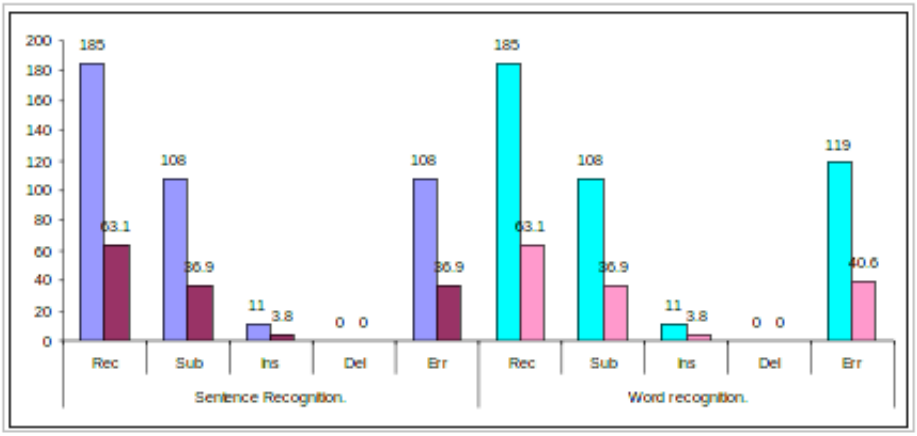
	WORD	RECOGNITION	PERFORMANCE:
S.No	Type of words	#of Words	
1	Correct	77.80%	14
2	Substitutions	22.20%	4
3	Deletions	0.00%	0
4	Insertions	27.80%	5
5	Errors	50.00%	9
S.No	Type of words	#of Words	Hyp.words= Ref.words+Insertions (18+5=23) and Alinged words =Hyp. Words means deletions in the recognized output are zero
1	Ref.words	18	
2	Hyp.words	23	
3	Aligned words	23	
4	Splits words	0	
5	Split candidates	0	
6	Merges	0	
7	Merge candidates	0	
	WORD ACCURACY	77.78%	(14/18)
	ERRORS	50.00%	(9/18)

Figure 5.59: Error analysis and comparison of Recognition output in terms of sentence and word recognition for Telugu evaluation data set.

VIII) TASR system for 300 Telugu phoneme order word list and their recognition output performance in terms of Sentence Recognition and Word Recognition using bare system

Table 5.2.3.10 (a): Data set – Male Utterances

Name of the file.	Sentence Recognition.					Word recognition.				
	Rec	Sub	Ins	Del	Err	Rec	Sub	Ins	Del	Err
UOH	185	108	11	0	108	185	108	11	0	119
%	63.1	36.9	3.8	0	36.9	63.1	36.9	3.8	0	40.6
	UOH									
	Ref words: 293									
Description	Hyp.words: 304									
	Align words:304									
	Split candidate: 11									



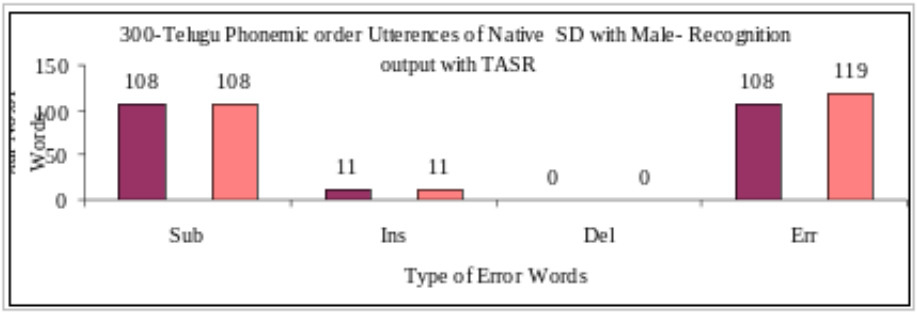


Figure 5.60: Data set – Male Utterances

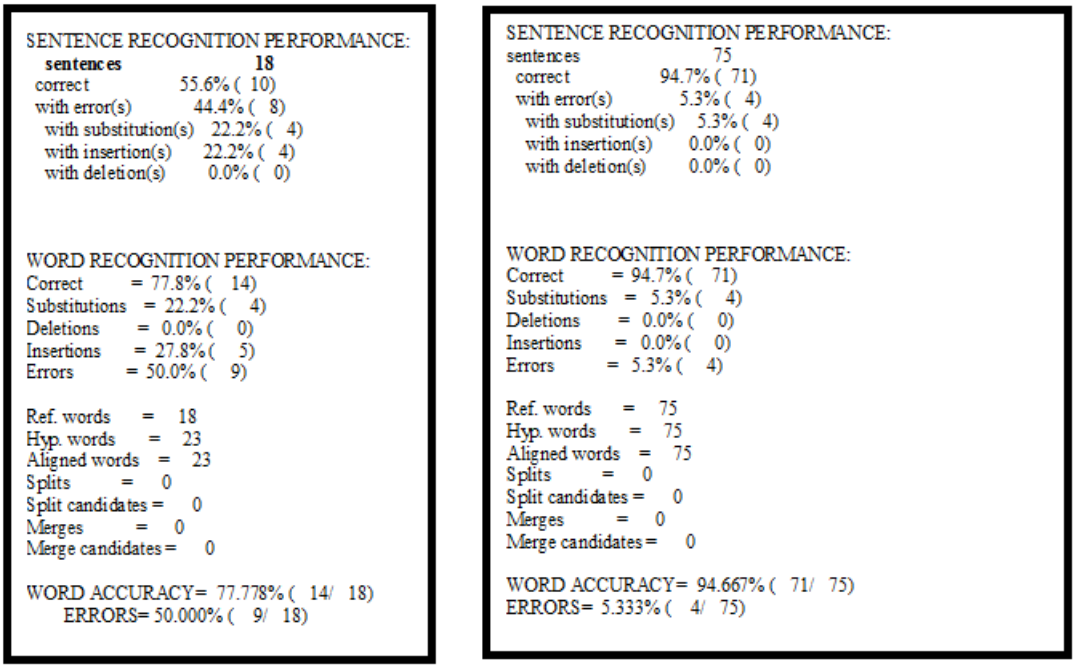


Figure 5.61: are Isolated words 18 Telugu akshara sequence ordered word list and proper names in terms of mmts train station with 5 different speakers of 15 words each as the test data set to test the TASR system the performance in terms of the Sentence and Word using WA and WER and different kind error words.

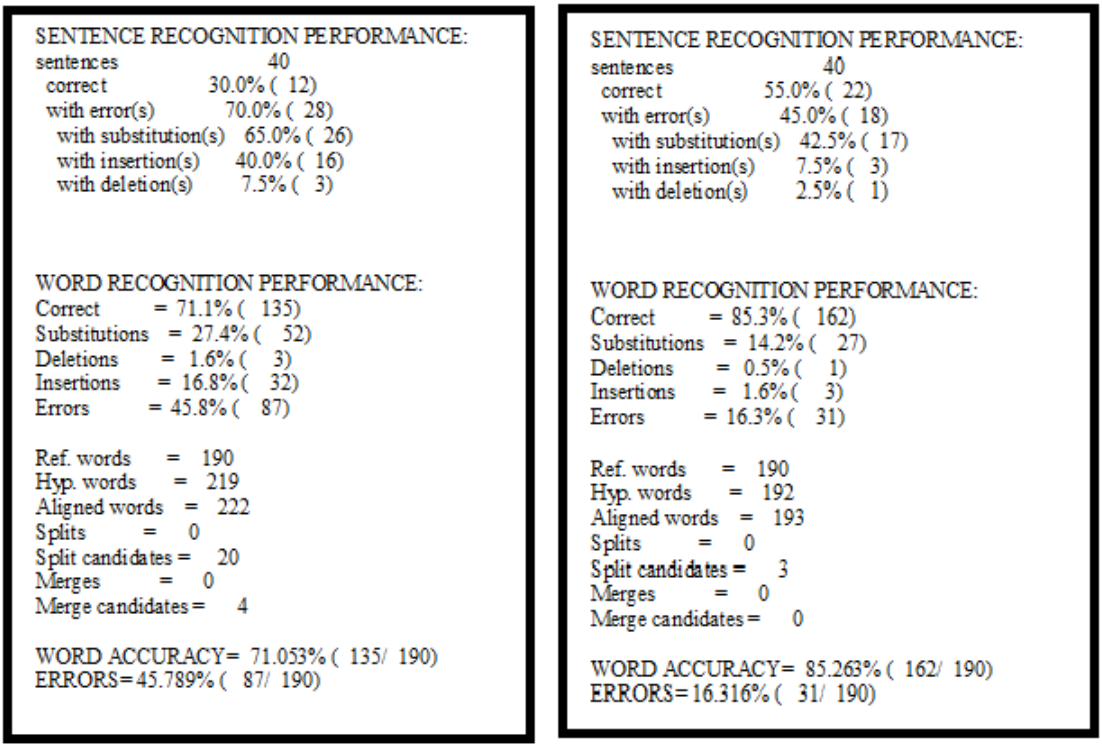


Figure 5.62: TASR system for 40 sentence of University Information system performance in terms of Sentence with recognized and error word list.

Sentence								
	UOH	CMU	UOH	CMU	UOH	CMU	UOH	CMU
Exp No.	Sub		Ins		Del		Err	
Exp 1	84	189	7	24	2	1	86	192
Exp 2	5	21	0	0	1	1	6	22
Exp 3	5	19	0	0	3	2	8	21
Exp 4	11	45	0	0	0	0	11	45
Exp 6	30	54	3	0	0	0	40	54
Exp 7	5	36	0	2	1	1	6	37
Exp 8	7	15	1	0	0	0	7	15
Exp 9	2	10	0	0	1	0	3	10
Exp 10	101	236	5	7	3	2	104	238
Exp 11	1	10	0	0	0	8	1	18
Exp 12	3	12	0	0	0	4	3	16
Exp 13	1	18	0	0	0	1	1	19
Exp 14	3	14	0	0	0	1	3	15
Exp 15	2	26	0	0	0	0	2	26
Exp 16	2	15	0	0	0	1	2	16
Exp 17	2	7	0	0	0	0	2	7
Word Recognisation								
Words	UOH	CMU	UOH	CMU	UOH	CMU	UOH	CMU
Exp No.	Sub		Ins		Del		Err	
Exp 1	28	189	3	24	0	1	28	192
Exp 2	5	21	0	0	1	1	6	22
Exp 3	5	19	0	0	3	2	8	21
Exp 4	11	45	0	0	0	0	11	45
Exp 6	39	54	3	0	0	0	42	54
Exp 7	5	36	0	2	1	1	6	39
Exp 8	7	15	1	0	0	0	8	15
Exp 9	2	10	0	0	1	0	3	10
Exp 10	101	236	5	7	3	2	109	245
Exp 11	1	10	0	0	0	8	1	18
Exp 12	3	12	0	0	0	4	3	16
Exp 13	1	18	0	0	0	1	1	19
Exp 14	3	14	0	0	0	1	3	15
Exp 15	2	26	0	0	0	0	2	26
Exp 16	2	15	0	0	0	1	2	16
Exp 17	2	7	0	0	0	0	2	7

Figure 5.63: Comparison of Errors and Error types in 17 experiments carried out on Telugu words Data set with UOH Lexical model based TASR system and CMU Lexical model based ASR system

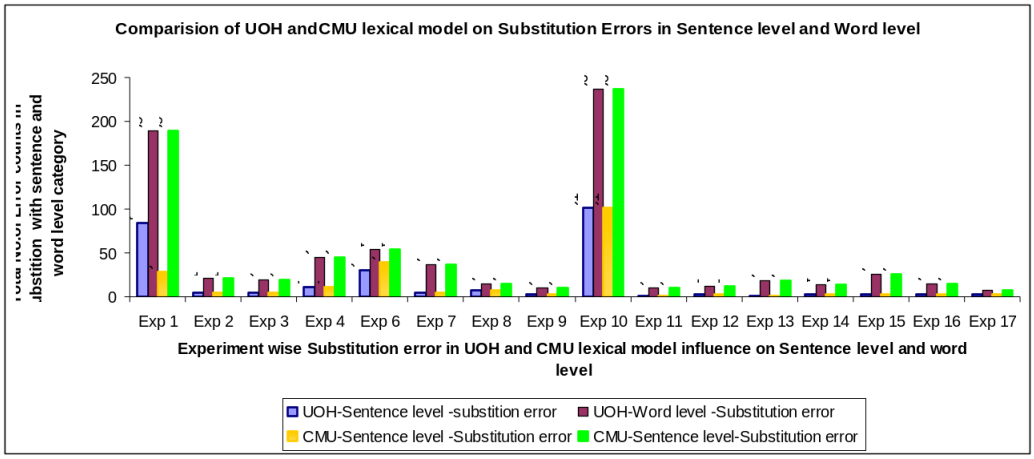


Figure 5.64: Data set of Telugu words and Substitution Errors comparison using UOH and CMU lexical model

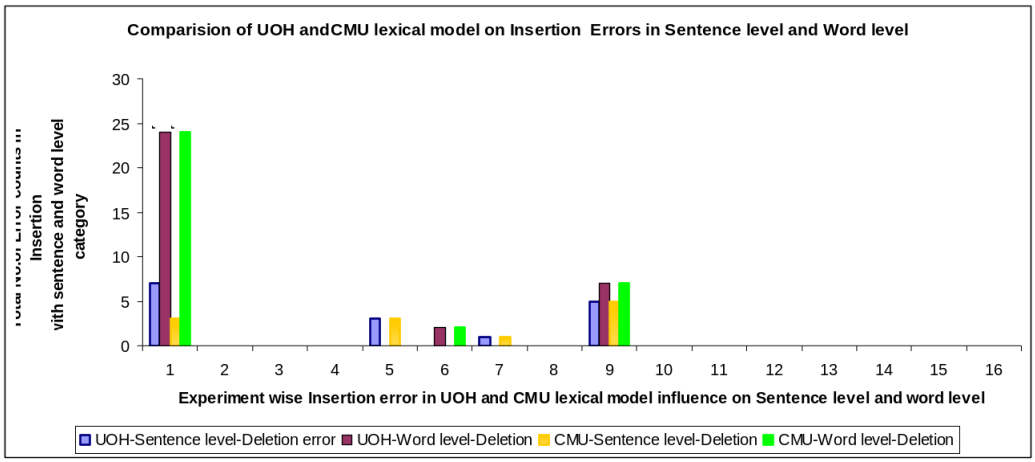


Figure 5.65: Data set of Telugu words and Insertions Errors comparison using UOH and CMU lexical model.

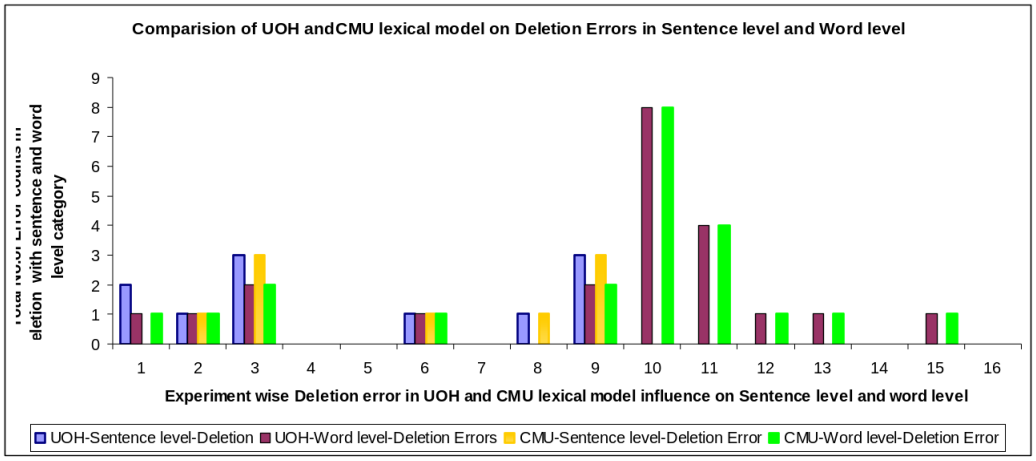


Figure 5.66: Data set of Telugu words and Deletion Errors comparison using UOH and CMU lexical model.

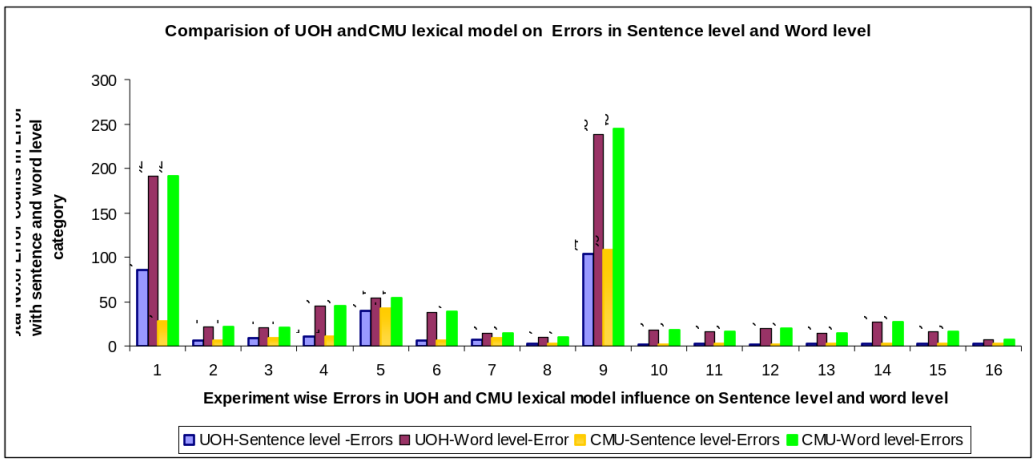


Figure 5.67: Data set of Telugu words and Word Errors comparison using UOH and CMU lexical model.

Table 5.6: 45 experiments ASR Decoding results Comparing with performance measures. i.e Accuracy, Error Rate, Recall/Sensitivity/True Positivity etc.

s.no	Datatype	Accuracy	Error-rate	Recall/ Sensitivity/ True Positivity	Specificity	Precision (positive predict value)	Matthews Correlation Coefficient	F-0.5	F1	F2
1	Exp-1	1	0	1	Infinity	1	Infinity	1.2	1	1.25
2	Exp-2	1	0	1	1	1	1	1.2	1	1.25
3	Exp-3	1	0	1	1	1	1	1.2	1	1.25
4	Exp-4	1	0	1	Infinity	1	Infinity	1.2	1	1.25
5	Exp-5	0.95	0.05	1	1	1	0.69	1.2	1	1.25
6	Exp-6	0.92	0.08	0.95	0.75	0.95	0.7	1.14	0.95	1.25
7	Exp-7	0.92	0.08	1	1	1	0.67	1.2	1	1.25
8	Exp-8	1	0	1	1	1	1	1.2	1	1.25
9	Exp-9	0.97	0.03	0.97	0.93	0.97	0.92	1.17	0.97	1.25
10	Exp-10	0.95	0.05	0.92	0.9	0.92	0.91	1.11	0.92	1.25
11	Exp-11	0.34	0.66	1	1	1	0.15	1.2	1	1.25
12	Exp-12	0.99	0.01	0.99	0.94	0.99	0.94	1.19	0.99	1.25
13	Exp-13	1	0	1	1	1	0.99	1.2	1	1.25
14	Exp-14	0.99	0.01	0.99	0.98	0.99	0.98	1.19	0.99	1.25
15	Exp-15	0.98	0.02	0.98	0.93	0.98	0.95	1.18	0.98	1.25
16	Exp-16	0.99	0.01	0.99	0.92	0.99	0.94	1.19	0.99	1.25
17	Exp-17	0.92	0.08	0.79	0.88	0.79	0.83	0.95	0.79	1.25
18	Exp-18	0.97	0.03	0.92	0.95	0.92	0.93	1.1	0.92	1.25
19	Exp-19	0.96	0.04	0.91	0.96	0.91	0.9	1.09	0.91	1.25
20	Exp-20	0.96	0.04	0.91	0.96	0.91	0.9	1.09	0.91	1.25
21	Exp-21	0.96	0.04	0.93	0.91	0.93	0.92	1.11	0.93	1.25

22	Exp-22	0.96	0.04	0.93	0.91	0.93	0.92	1.11	0.93	1.25
23	Exp-23	0.99	0.01	0.13	0.99	0.13	0.17	0.15	0.13	1.25
24	Exp-24	1	0	1	1	1	1	1.2	1	1.25
25	Exp-25	0.94	0.06	0.93	0.75	0.93	0.84	1.12	0.93	1.25
26	Exp-26	0.95	0.05	0.94	0.86	0.94	0.89	1.13	0.94	1.25
27	Exp-27	0.95	0.05	0.94	0.86	0.94	0.89	1.13	0.94	1.25
28	Exp-28	1	0	1	1	1	1	1.2	1	1.25
29	Exp-29	0.98	0.02	0.96	0.94	0.96	0.95	1.15	0.96	1.25
30	Exp-30	0.99	0.01	0.99	0.96	0.99	0.98	1.19	0.99	1.25
31	Exp-31	0.98	0.02	0.96	0.94	0.96	0.95	1.15	0.96	1.25
32	Exp-32	0.98	0.02	1	1	1	0.82	1.2	1	1.25
33	Exp-33	0.97	0.03	1	1	1	0.89	1.2	1	1.25
34	Exp-34	0.45	0.55	Infinity	1	Infinity	Infinity	Infinity	Infinity	Infinity
35	Exp-35	0.81	0.19	Infinity	1	Infinity	Infinity	Infinity	Infinity	Infinity
36	Exp-36	0.81	0.19	Infinity	1	Infinity	Infinity	Infinity	Infinity	Infinity
37	Exp-37	1	0	1	1	1	1	1.2	1	1.25
38	Exp-38	1	0	1	Infinity	1	Infinity	1.2	1	1.25
39	Exp-39	1	0	1	1	1	1	1.2	1	1.25
40	Exp-40	0.8	0.2	Infinity	1	Infinity	Infinity	Infinity	Infinity	Infinity
41	Exp-41	0.8	0.2	Infinity	1	Infinity	Infinity	Infinity	Infinity	Infinity
42	Exp-42	1	0	1	1	1	1	1.2	1	1.25

43	Exp-43	1	0	1	1	1	1	1.2	1	1.25
44	Exp-44	1	0	1	Infinity	1	Infinity	1.2	1	1.25
45	Exp-45	1	0	1	Infinity	1	Infinity	1.2	1	1.25

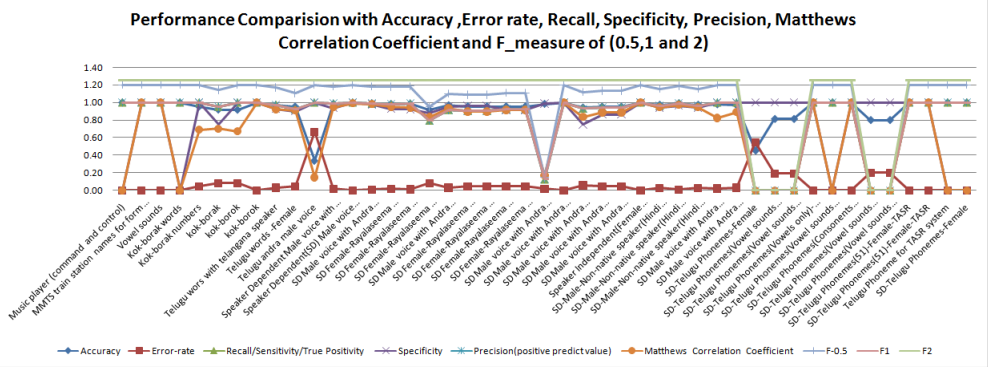


Figure 5.68: Graph of the 45 experiments ASR Decoding results Comparing with performance measures. i.e Accuracy, Error Rate, Recall/Sensitivity/True Positivity etc.

$$Accuracy = \frac{(Recognized + Substituted)}{(Recognized + Substituted + Deletion + insertion)} \tag{5.1}$$

$$Error\ Rate = \frac{(Insertion + Deletion)}{(Recognized + Substituted + Deletion + insertion)} \tag{5.2}$$

$$Recall/Sensitivity/TruepositiveRate = \frac{(Recognized)}{(Recognized + Substituted)} \tag{5.3}$$

$$Specificity = \frac{(Substituted)}{(Substituted + Inserted)} \tag{5.4}$$

$$Precision\ positive\ predictive\ value = \frac{(Recognized)}{(Recognized + Insertion)}) \tag{5.5}$$

$$Matthews\,Correlation\,Coefficient = \frac{(Rec * Sub - Ins * Del)}{\sqrt{((Rec + Ins)(Rec + Del)(Sub + Ins)(Sub + Del))}} \tag{5.6}$$

$$F\_Score_{0.5} = 1.25 * \frac{(Precision * Recall)}{(0.25 * Precision) + Recall} \tag{5.7}$$

$$F_1 - Score_1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{5.8}$$

$$F_2 - Score = 5 * \frac{(Precision * Recall)}{(4 * Precision) + Recall} \tag{5.9}$$

5.3 Transliteration tool for transcribing the Telugu script

Telugu language transcription for collected speech corpus is the major limitation in data set preparation for the target language. Design of ASR system includes the data

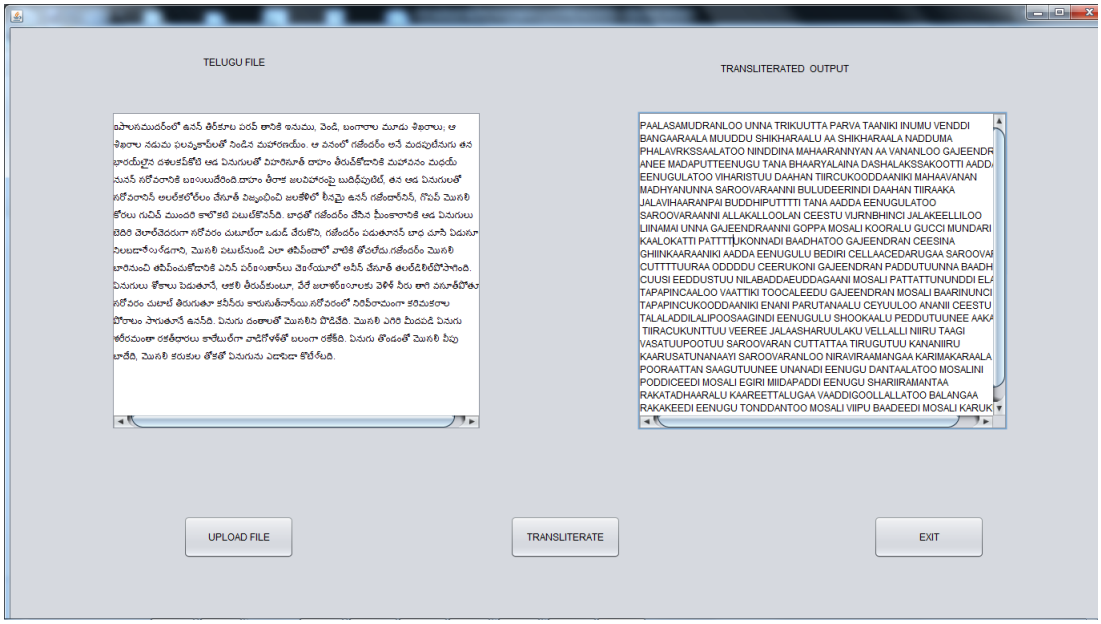


Figure 5.69: GUI for Telugu script to Roman script transliteration for Text corpus generation for building ASR system.

preparation. Transliteration using different coding unit of Akshra available but support to the system is not available. Using the syllabification and transcribe those with the Akshara specific roman script representation is the main task of the transliteration tool. The input to the system is Telugu text. Here experiment purpose large Telugu text corpus is collected from chandamama online resources. The transliteration done using simple transliteration and enhanced method done by using the Edit distance method. Design developed in NetBeans environment using java programming.The input and output of the system as snap shot shown in Figure 5.68.

Over all experiments are compared with the WA and WER. The following section is used to verify the system performance using the precision, Recall and F-measure of statistical analysis.

$$Precision = \frac{Recognized\_word}{Recognized\_word + Insertion\_error\_word} \tag{5.10}$$

OR

$$Precision = \frac{Recognized\_word}{Recognized\_word + Insertion\_error\_word + Substitution\_Error\_words} \tag{5.11}$$

.	1	2	3	4	5	6	7	8	9
Sr. No.	Total	No of correct Translit-erated words without Edit Distance	No of correct Translit-erated words with Edit Distance	No of error Translit-erated words without Edit Distance	No of error Translit-erated words with Edit Distance	Words Accu-racy Without Edit Distance	Words Accu-racy With Edit Distance	Words Error Rate Without Edit Distance	Words Error Rate Without Edit Distance
1	141	93	57	48	84	66	41	35	68
2	141	93	57	48	84	66	41	35	68
3	136	89	53	47	83	66	39	35	62
4	205	113	98	92	107	56	48	48	53
5	204	127	98	77	106	63	49	38	52
6	170	98	114	44	56	58	68	26	50
7	119	59	81	60	38	74	69	51	32
8	89	52	38	37	51	74	43	42	58
9	163	52	102	111	61	32	63	69	38
10	175	126	133	46	42	74	76	27	24

Table 5.7: transliteration tool output of 10 Experiments

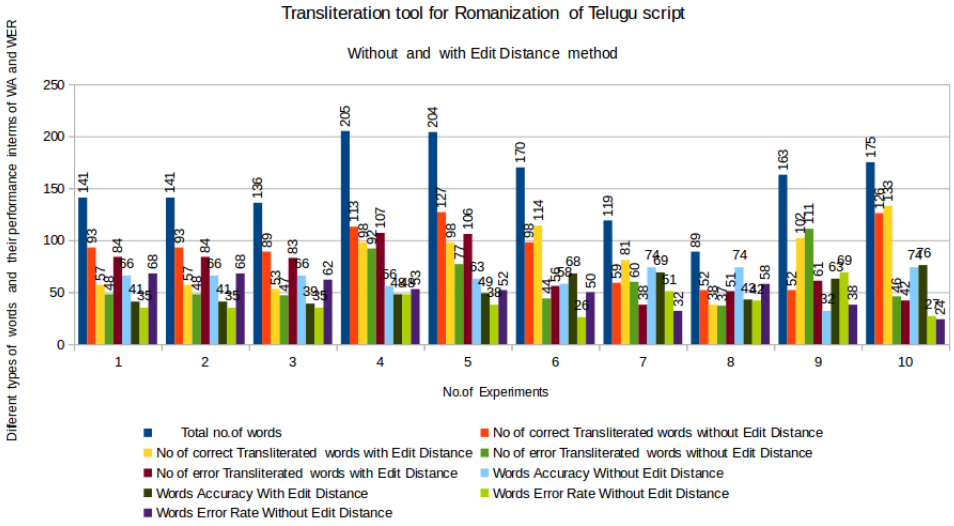


Figure 5.70: Transliteration of Telugu to English Algorithm

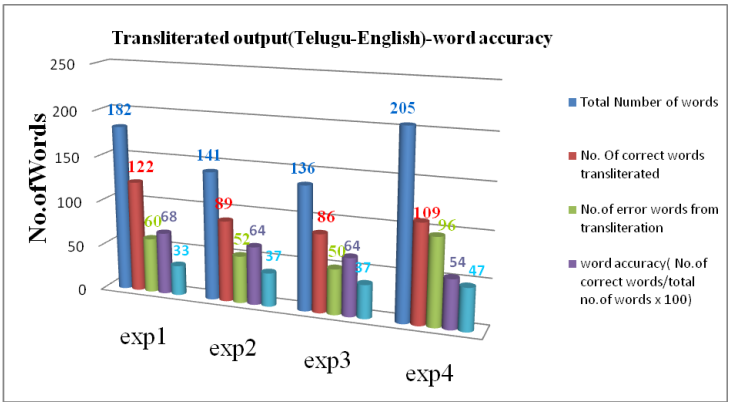
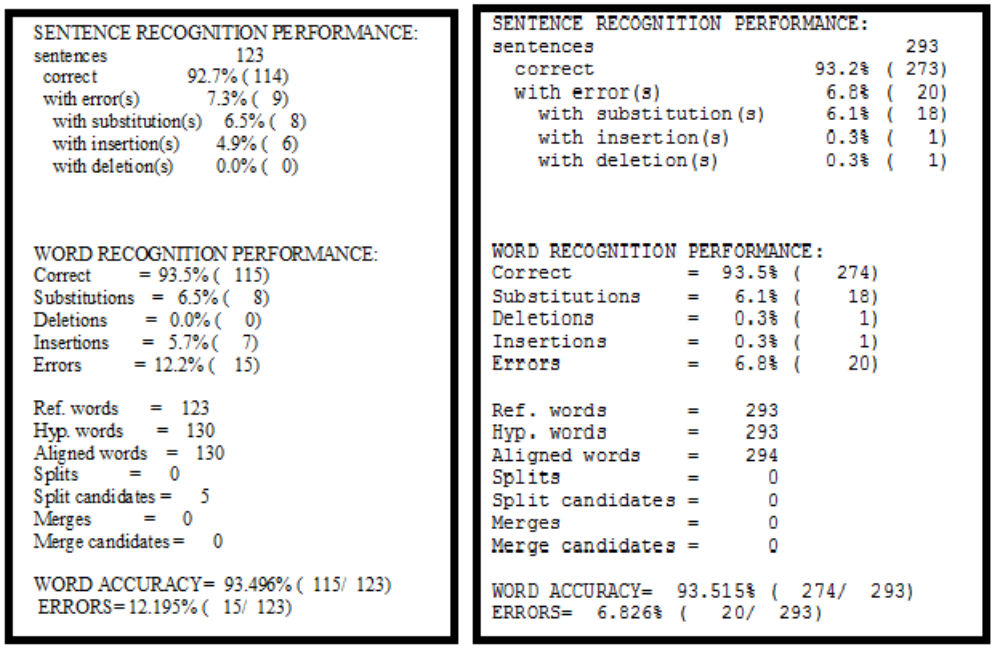


Figure 5.71: Comparison of Transliteration of Telugu to English without and with edit distance for experiments using total words transliterated with their word accuracy and error words list



SENTENCE RECOGNITION PERFORMANCE:
sentences 123
correct 92.7% (114)
with error(s) 7.3% (9)
with substitution(s) 6.5% (8)
with insertion(s) 4.9% (6)
with deletion(s) 0.0% (0)

WORD RECOGNITION PERFORMANCE:
Correct = 93.5% (115)
Substitutions = 6.5% (8)
Deletions = 0.0% (0)
Insertions = 5.7% (7)
Errors = 12.2% (15)

Ref. words = 123
Hyp. words = 130
Aligned words = 130
Splits = 0
Split candidates = 5
Merges = 0
Merge candidates = 0

WORD ACCURACY= 93.496% (115/ 123)
ERRORS=12.195% (15/ 123)

SENTENCE RECOGNITION PERFORMANCE:
sentences 293
correct 93.2% (273)
with error(s) 6.8% (20)
with substitution(s) 6.1% (18)
with insertion(s) 0.3% (1)
with deletion(s) 0.3% (1)

WORD RECOGNITION PERFORMANCE:
Correct = 93.5% (274)
Substitutions = 6.1% (18)
Deletions = 0.3% (1)
Insertions = 0.3% (1)
Errors = 6.8% (20)

Ref. words = 293
Hyp. words = 293
Aligned words = 294
Splits = 0
Split candidates = 0
Merges = 0
Merge candidates = 0

WORD ACCURACY= 93.515% (274/ 293)
ERRORS= 6.826% (20/ 293)

Figure 5.72: Isolated words of 40 sentence of UIS data set and 293 names of Data set in alphabetical ordered word list as the test data set to test the TASR system the performance in terms of the Sentence and Word using WA and WER and different kind error words

Next

$$Recall = \frac{Recognized\_word}{Recognized\_word + Deletion\_Error\_word} \tag{5.12}$$

OR

$$Recall = \frac{Recognized\_word}{Recognized\_word + Deletion\_Error\_words + Substitution\_Error\_words} \tag{5.13}$$

By using Precision and Recall, F-measure is calculated as:

$$F - measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{5.14}$$

Using above statistical measure ASR system output are compared in different condition that is used in this experimental work.

Dataset 4: Female speaker data 1-665 Telugu phonetically rich words by Telugu native with Rayalasema dialect of Telugu										
	UOH	CMU	UOH	CMU	UOH	CMU	UOH	CMU	UOH	CMU
TP(total no of words)	84	55	63	58	63	61	61	59	64	61
FN(Del)	1	0	0	0	0	0	1	0	0	0
FP(Ins)	0	0	0	0	0	0	0	0	0	0
TN(sub)	2	10	2	7	2	4	3	6	1	4
Precision	1	1	1	1	1	1	1	1	1	1
Recall	0.98824	1	1	1	1	1	0.98387	1	1	1
F-Measure	0.99408	1	1	1	1	1	0.99187	1	1	1

Figure 5.73: Statistical analysis of TASR and ASR system lexical model performance comparison using Precision, Recall and F-measures of a SD Female speaker Data set 1-100

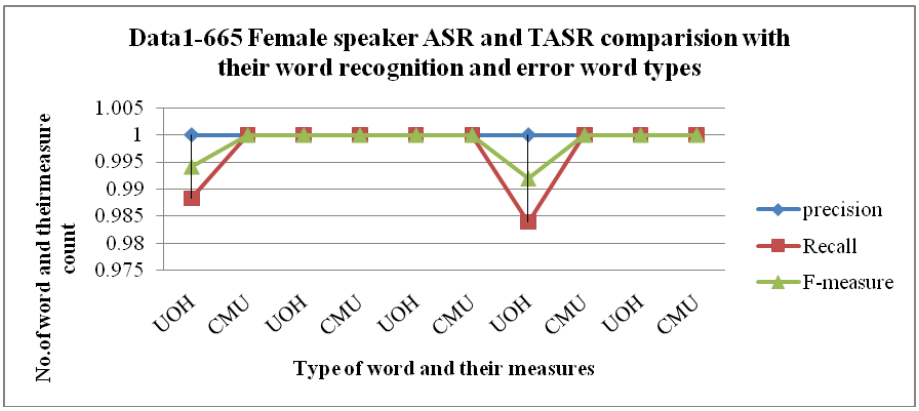


Figure 5.74: Precision, Recall and F-measure based UOH and CMU lexical model of TASR and ASR comparisons

5.4 Concluding remarks on the empirical studies:

Inference from the empiric data collected is that the newly transformed TASR is that the same is comparable with the ASR system in terms of performance. The WA and WER.

Why variation between Hypothesis and reference: Hypothesis is the system generated text i.e ASR output. The reference is the user transcribed speech corresponding text information. Comparison of reference text file with system generated text file is the performance evaluation in terms of measuring parameters for ASR system. Matching reference file text with system generated file text is considered accuracy and not matched text is considered error. Measure is taken based on human perception hence word is the default unit taken. Hence the accuracy is a match word which is generally 'WA' and not match text in terms of words is considered as Word Error. The rate at

Dataset No.1-665		Speaker: Female IWR				Native(Telugu)			
1		Sentence		Words		With substitution calculation)			
T.word		UOH	CMU	UOH	CMU	UOH	CMU	UOH	CMU
665	Error	86	192	28	192	86	192	28	192
	TP(total no of words Recognized)	579	473	637	473	579	473	637	473
	FN(Del)	2	1	0	1	2	1	0	1
	FP(Ins)	7	24	3	24	7	24	3	24
	TN(sub)	84	189	28	189	84	189	28	189
	precision	0.99	0.95	1	0.95	0.86	0.69	0.95	0.69
	Recall	1	1	1	1.00	0.87	0.71	0.96	0.71
	F-Measure	0.99	0.97	1.00	0.97	0.87	0.70	0.96	0.70

Figure 5.75: Statistical analysis of TASR and ASR system lexical model performance comparison using Precision, Recall and F-measures of a SD Female speaker Data set 1-665

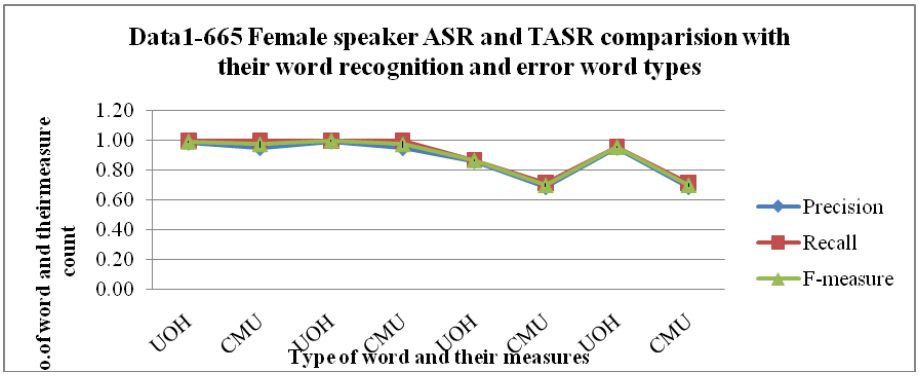


Figure 5.76: Precision, Recall and F-measure based UOH and CMU lexical model of TASR and ASR comparisons

which this word mis-match occurred is the Word Error Rate (WER). Occurrence of word error is due to the three types of errors which are occurred due to the signal level and language level.

Advantages: ASR system to TASR transformation enables to use the system for any Indian language. Telugu language covers all the world language pronunciation with their phonemic transcription due to its one to one mapping of written form to spoken form. Generating automation tool to the transcription enable to develop text corpus corresponding to the online audio data. Present developed TASR system can be used to command control mode very well. Hence developing online form filling with the present developed system enable the Interactive tool for visually challenged and physically challenged people to interact with digital media.

Limitations of the studies: The input variation amongst the different age groups, gender, dialects and accent has impact on the results. Contribution of speakers in all these ranges of variation is the major limitation. Accordingly, within the reasonable limitations, the TASR was able to deliver the expected results in terms of ASR. During the learning process clean speech improves the system performance. Implementation of speech analysis and its enhancement in the signal processing, need to explore further improvement of the system. Few contributions in worked on the present in the thesis list in chapter 6. It also need correct form of transcption to the speech, which gives good results in performance of ASR system. Initial stage intervention of human and use of their linguistic knowledge makes the system robust in writing transcription. Automating procedure requires to sufficient data set that coverers all the phonemic and morphonemic with hardware support to the working process of ASR and supporting IDEs to generate the GUI requires lot of human efforts. Integration of legacy Speech tools with the current hardware and software system requires lot of human effort to build any applications. Few tasks explored as part of the research. The entire process of empirical analysis in context of software engineering methodology explained in next chapter 6. The results for in-house data build for the research application are presented for the evidence to the hypothesis of the thesis. The generalization of the TASR also presented with other language speech data. Over all performance of the TASR compared to ASR in terms of WA and WER are around 15% to 20% improvement proves that TASR system is the better choice for Indic languages.

****

Chapter 6

Summarization and Conclusion

This chapter summarizes the empirical works carried out during the research work. The re-engineering approach followed to design the system based on demand for application prototype, with use of existing phonetic engine system, thereby transforming it into desired language (Telugu) system. Within the limitations of the research, the dissertation work was carried out, prototype designs were developed and future plans were suggested. Research work contributed in developing a phonetic engine, application of phonetic engine in interactive mode, language specific derivations for phone set, developing the lexical model, thus leading to development and utilization of Telugu ASR system.

6.1 Summarization of the Thesis

The research area of building ASR system for new languages, encourages better user interface, than any other mode of user interaction. The mother tongue based interaction with the system is now new scope of research for many applications in Information Technology. Every new language opens scope for research in the field of ASR. Lexical modeling, has local linguistic flavor, and requires human expert knowledge for development of the model. It requires development of the dataset and to implement ASR system for any application, with use of handcrafted with use of linguistic human knowledge.

The generic system development for Telugu language is presented along with entire working procedure of ASR system is explained in chapter 2. The language learning

application, developed as part of the research, along with Tools and concepts related to system design and development are given in chapter 3. The thesis also focused on Telugu language tutor, based on speech technology (ASR and TTS). As the existing ASR system is Generic User friendly applications, it is intended to develop the Lexical model, which is core part of the language model, used to develop ASRs. The Pronunciation Variants, its effect in speech signal and the variants adaptation techniques applied in ASR system with knowledge driven and Data driven methods are discussed in Chapter 4. The chapters 3 & 4, enable to drive towards building the speech corpus to justify the proposed ideas are practicable and improved the existing system. The results based on the empirical research are presented in the chapter 5.

6.2 Innovation (Novelty) in proposed system

The topic of research work, is more than 5 decades old, with researchers trying to develop the ASR system, to imitate the human speech recognition system, yet it is still at nascent stage, as regards the Indian language ASR system. As human languages in the world differ based on the region, socioeconomic culture, physiological structure and geographical effect on the human articulatory system, etc., these factors influence the sub-optimal performance of ASR system. In this work, building of ASR system is attempted, for one of southern India language, i.e Telugu (a Dravidian language). Further, its pronunciations and their variant factors, that are influencing the ASR system with negligible availability of bench mark dataset [SRE04][HUA14], has complicated the development of such systems. Hence more linguistic aspect, as well as speaker influence factors is considered to build by use of the existing system and adaptation techniques are considered for more functional design of the system [NAG12c]. Research application is to builds a system. This system teaches the language through system rather than simple human conventional method [NAG05]. The research out come of developing the tool to teach the Telugu language for traditional leaner's as well as those interested in learning the language. This research also helps to build speech interaction system for various applications [SRI04][VIV06][NAG09][NAG10b][NAG10c] and[NAG16]. Phonological ground work is taken to develop the phone set for the existing system for bringing the natural simple solutions for WER reduction and also for simple mapping techniques to design the interfacing modules for speech interactive

system. This work also attempted to bring the common training data set in Telugu language for novel users of ASR system.

6.3 Design methodology

The development of tool requires a systematic approach. There are many software engineering methods in design. The contribution of my thesis in the field of Speech Technology and Language Technology linking with the programming method is explained using Horshoe model based reengineering concept for system design. The following paragraph explain in detail the process of developing Telugu Language ASR system by designing Indian script (Telugu) based phonetic representation in Lexical model [NAG10].

6.3.1 The Horseshoe Model concepts:

Adapting the Horseshoe model software engineering concept to design proposed language i.e., Telugu language ASR system, requires transformation of algorithms, architecture and tools designed for basic structure and functionality in respect of existing generic state of art ASR system. Due to research in last 5 decades, ASR system is developed in these areas, however, it is lacking mostly in basic linguistic knowledge base, which in turn depends on social concern domains. To develop any new language ASR, the research in most native languages is still in infant stage process in the linguistic and base data set design, as it demands more human intervention rather than system utility. Other technical concept and system based design can be used by the concept of reusability and reengineering by choosing existing developed algorithms, scripts and tools to construct language dependent ASR system. The language dependency factors are considered in research problem as worldly human languages are categorized based on their properties, which are explored in chapter 4. The languages spoken by human are classified based on the rhythm which is depending on many factors [ABE67]. This is main motto for choosing the pronunciation variation adaptation in ASR system. The pronunciation of language is dependent on speaker's related factors and also geographical influence and human culture and heritages. Hence, to work in the area of pronunciation variation adaptation concepts, one should have knowledge on social sciences, physiology, psychology and cognitive science to some extent to understand

the reasons for pronunciation variations occurring during speech production process. These factors influence the making of the system, and its adaptation i.e., the existing system being restructured for developing new system. Hence, in this work the software Engineering concept of Horseshoe model is followed to design Telugu language ASR system. In this, Telugu ASR system instead of utilizing English based phones, Telugu language Phones, which are mapped to represent Telugu phonemes are used. As the task is pronunciation variation adaptation, hence lexical model was chosen. The lexical model of ASR system defines the pronunciation of uttered speech sequence. The module contains the uttered word list and their pronunciation process through phones. In this work phones are representing the alphabets of system i.e ASCII keys.

Novelty in this work is unique mapping of phonemes to phone hence less confusability and easy adaptation of any language source that can be represented by the Telugu phonemes. Important point here is almost all worldly languages sounds can be represented by Telugu phonemes so the transliteration process is easy and translation through machine learning concept also made easy. Hence the developed system can be adapted to any language with less tools and system confusability.

The building of TASR system, requires systematic adaptation of the existing ASR system. This requires the layered fashion transformation in three phases of horse shoe model [MUL00] as given in Figure 6.1 viz., (i) Reconstruction phase involves developing ASR system for target (Telugu) language using existing artifacts (ii) Transformation improvement of logical descriptions and (iii) Refining these artifacts, from architecture level to code level. TASR system thus developed can be replicated for other languages. The phases are illustrated in conceptual ”horseshoe” model of software re-engineering and evolution [BER99].

6.4 Contribution of research carried out in this thesis

The contributions that are carried out during this process of dissertation has direct relevance to the work, some are the supporting studies and few are the applications to the current work. The beginning contributions are to build, specifically using Telugu Speech Database to built the system and also understand the functionality of ASR system. Few conference and Journal papers focus only on actual implementation process

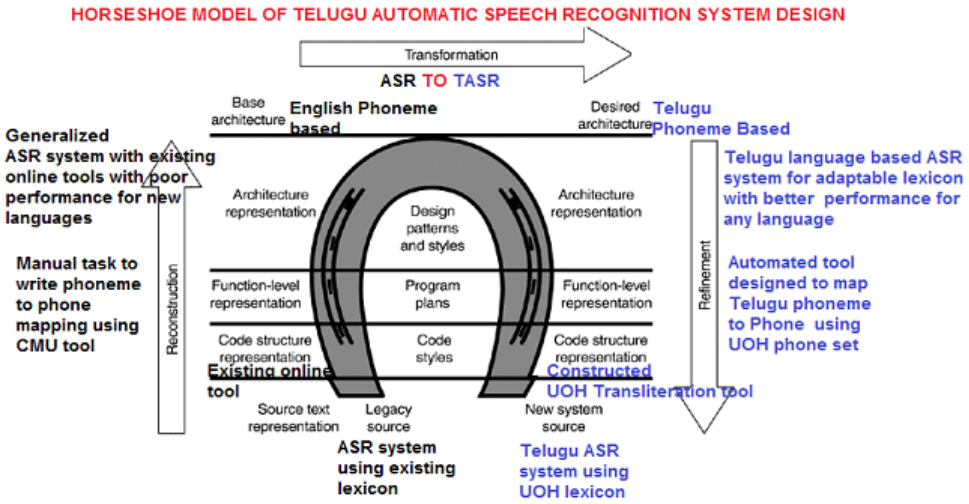


Figure 6.1: The horseshoe model for design of Telugu Language Automatic Speech Recognition System.[MUL00]

and the empirical data analysis based papers briefly explained in chapter 1 Table.no.1.1. The following research contributions of me are related to the thesis work:

i) M. Nagamani, P. N. Girija., "Investigations of Features for Audio Information Retrieval", Journal of Acoustic Society of India, Vol.32 No. 1-2, 2004. [NAG04]

ii) G. Sreenu, P.N. Girija, M. Nagamani, M. Narendra Prasad, " Telugu Speech Interface Machine for University Information System", Proc. Of IEEE Advanced Computing and Communications (ADCOM), Ahmadabad, December 2004. (scopus i.e IEEE explorer) [SRE04]

iii) P.N. Girija, G. Sreeenu, M. Nagamani, M. Narendra Prasad, " Human Machine Speech Interface Using Telugu Language – HUMSINTEL", Proc. Of IEEE International Conference on Human Machine Interfaces, Bangalore, Dec 2004. [GIR04]

iv) G. Sreenu, P. N. Girija, M. Nagamani and M. Narendra Prasad., " A Human Speech Interactive system Telugu Language" Proc. of IEEE INDICON – 2004, 20-23 December 2004. [SRE04]

v) M. Nagamani, P.N. Girija, "Intelligent Tutor for Telugu Language Learning – INTTELL"., Proc. Of International Conference on Image, Signal and Information

Processing(ICISIP2005), January 4-7, 2005. (Scopus i.e. IEEE explorer) [NAG05]

vi) Vivek, Hitesh, M. Nagamani, " Voice Driven Operating System" National Conference in Artificial Intelligence" held on 1st September 2006 at GRIET, Hyderabad. [VIV06]

vii) M. Nagamani and P.N. Girija.(Pronunciation Variation Modeling for Speech Recognition System Accuracy" by at 1st National conference of Telugu Linguists' Forum(TeLF), 6-7th Nov 2007 at University of Hyderabad. [NAG07]

viii) Anil Kumar, M. Nagaman, B.S.R. Krishna "Railway Reservation form filling throw speech interactive system" University of Hyderbad ,December 2009. [ANI09]

ix) R.P. Lal, P.S.V.S. Saiprasad, M. Nagamani "Improved Algorithm for End point Detection and Silence removal for Speech processing" Proc. of RAFIT2009 during April 9-10th 2009 at Punjab University Patiala. [LAL09]

x) M. Nagamani, P. N. Girija and B.S.R.Krishna "Pronunciation Dictionary Comparison for Telugu Language Automatic Speech Recognition system" pp. 60-69, Vol. 2, No. 1, Jan-July 2010, Proc. of "Journal of Data Engineering and Computer Science" GRIET, Hyderabad. [NAG10A]

xi) M. Nagamani, B.S.R.Krishna and Sridhar Reddy. A "Neural Networks based Visual Speech Recognition through Lip Reading" Proceeding of National Symposium on Acoustics NSA 2010 from 11th - 13th November 2010, at Rishikesh. [NAG10B]

xii) M. Nagamani and B.S.R.Krishna "Intelligent tutoring system for Telugu Language Learning" International conference on "BIOMEDICAL ENGINEERING AND ASSISTIVE TECHNOLOGIES (BEATS – 2010)" during 17th - 19th December 2010 at NIT – Jalandhar. [NAG10C]

xiii) M. Nagamani, S. Manoj, B.S.R. Krishna, "Implementing Phoneme Based Segmentation Algorithms to ASR system" International Conference on Advanced Computing methodologies(ICACM-2011), ISBN:978-93-81269-40-4, Reed Elsevier India private Ltd, December 2011,Hyderabad. [NAG11]

xiv) M. Nagamani, P.N. Girija, " Substitution error analysis for improving the word Accuracy in Telugu Language Automatic Speech Recognition system" IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 4 (July-Aug. 2012), PP 07-10 www.iosrjournals.org. [NAG12A]

xv) M. Nagamani, P. N. Girija, " Lexicon design for Telugu Language Automatic Speech Recognition System", 2nd International Telugu Conference (ITC 2012), 2-4th November 2012, Vizag.(Presented Paper in Telugu langauge). [NAG12B]

xvi) M. Nagamani, T. VenuGopal, S. Uday Bhaskar, K. Rajeev, " Image to Speech Conversion System" 2nd International Conference on Biomedical Engineering & Assistive Technologies(BEATS 2012), ISBN-13:978-81-925454-1-7, 6-7th December 2012 at NIT, Jalandar(Punjab). [NAG12C]

xvii) M. Nagamani, P.N. Girija., "Pronunciation Variants and Substitutional error analysis for improving Telugu Language Lixical performance in ASR system Accuracy",Internation ournal of Scientific Engineering Research(IJSER), ISSN 2229-5518, pp 1128-1133, Volume 4, Issue7, July 2013. [NAG13]

xviii) M. Nagamani, S. Manoj Kumar, S. Uday Basker, "Image to Speech Conversion System for Telugu Language", International Journal of Engineering and Science and Innovative Technology(IJESIT), ISSN:2319-5967, ISO9001:2008 Certified, pp.161-166, Volume2, Issue6,November 2013. [NAG13]

xix) Sai Bala Kishore N, VenkatRao M, Nagamani M.," Environmental noise analysis for robust automatics speech recognition", article in Lecture Notes in Electrical Engineering 315:801:811,22015. : 10.1007/978-3-319-07674-4_75Chapter Advanced Computer and Communication Engineering Technology Volume 315 of the series. Lecture Notes in Electrical Engineering , 02 November 2014. [SAI14]

xx) P. Navneeth Kumar, M. Nagamani, M.S. Chakravarthy, J.M. Harish Kumar., "Phoneme Prediction in ASR system using Decision Trees" Proc. Of National Conference on Advanced Computing and Pattern Recognition, 8-9th January2014. [NAV14]

xxi) Sai Bala Kishore N, VenkatRao M, Nagamani M.," Environmental noise analysis for robust automatics speech recognition, Chapter Advanced Computer and

Communication Engineering Technology Volume 315 of the series Lecture Notes in Electrical Engineering , 02 November 2014. [SAI14]

xxii) M. Nagamani, T.V. VenuGopal, S. Manoj Kumar and Uday Bhasakar S., "Image to speech conversion system for Telugu Pronunciation" ISBN.9789383038176. Proc. Of National Conference on Advanced Computing and pattern Recognition, 2014. [NAG14]

xxiii) Nagamani M , Shaik Fathima " Voice controlled Music Player Using Speech Recognition System" IJEEE, Issue2, Feb-Mar2016. [NAG16]

xxiv) M. Nagamani and syed Hameed et al., "Exploring the Speech Prosody Manipulation for Dual-Language TTS System" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 4, Ver. III (Jul - Aug 2018), PP 20-22 www.iosrjournals.org DOI: 10.9790/0661-2004032022. [NAG18]

6.5 Conclusion

The TASR system developed endogenously for Telugu language speech recognition. TASR system uses Telugu speech corpus and then drive the system towards accepting the Telugu accent phonetic symbol and finally defining Indian accent Telugu phoneme specific system. TASR system is compared with the existing ASR system and both system generated hypothesis text compared in terms of Word Accuracy and Word Error Rate for system performance testing in this thesis work. The conclusions on the research work is follows:

The ASR system, run on Sphinx III single machine TUTORIAL tool, was used for the purpose of developing a speech interactive system in Education domain applications (INTTELL). In this tool both training and testing included in single module and associative modules are developed during the work and hence its implementation for any new applications or languages is simple.

1. Speech corpus is collected for experiments taking as language learning context from basic knowledge, hence considered primary school curriculum for developing corpus and also to build the system. Most of online-Telugu tutors on web are

considered for collection of text information, also considered Peddabala shiksha, current daily Telugu news papers and common words used in our daily conversation are part of corpus. This collected information is used to develop speech tool acceptable form using another supportive tool like Telugu Lipi and Dictionary, transcription files which are build for both CMU phonelist and SLM tool kit used to prepare the input files for ASR system.

2. Transcription tool for Telugu text data is built: UOH defined phones are mapped to the Telugu script by developing phonetic writing tool. Here Telugu phonemes all represented in UOH notation phones using Unicode mapping script, which is the part of research work, developed by me. The UOH phonelist based input files are created manually and partially used the knowledge base of existing generic tool.

3. Test results presented for the system with gender, accent and age variant factors for proposed system. Results presented in comparison with existing system and also with linguistic based proposed system.

4. The contributions in developing a phonetic engine, application of phonetic engine in interactive mode, language specific derivations for phoneset, developing the lexical model, thus leading to development and utilization of Telugu ASR system.

6.6 Future Scope

Though the scope of the research is vast in Omni directional as speech technology is growing and is needed in current technology and cyber physical environment. As thesis concern few important point mainly focusing on future task explained as follows :

1. Increase the list of words: As existing data covering the syllables of all phonemes and have to include all morphemes of Telugu and all types of syllables that are covering the most the usage words in Telugu conversation. Also want to include the subject wise Telugu words to see all types of pronunciation variations. We have an idea on to implement the system for polyglotal words from different influenced languages on Telugu wants to collect and work on it.

2. Work with sentences and paragraph for proposed system: As present system only covering the Isolated words and very simple few sentences, hence we need to take sentences and paragraphs with read and spontaneous speech data to see the more pronunciation errors and want to see existing methods work for the speech interactive system

3. Design interface for front end for proposed system: The complete modules defined for implanting INTelligent TELugu Learning [INTTELL] are developed. The flash animated scripts for Telugu script writing and complete Lexicon for covering entire Telugu words in Education domain application are used to build database. Different levels of WebPages to learn modular wise language is defined and web application is developed

4. Complete architecture of proposed system INTTELL design and develop from prototype to field acceptable model to bring it in real environment usage.

5. Make the TASR in real environment use by adapting more variant factors for pronunciation like dialect, environment etc. or considered for collecting different mode of speech corpus. All these data modified and archived for future experiments as open source data sets release to the generic use of research.

****

Appendix A

Appendix

Tools installation process
ASR system: Sphinx III Running Tutorial

The procedure, how to run SphinxIII as shown below.
/**** Procedure Starts****/
1. First go to the Path, where the SphinxIII TUTORIAL is load by using the Command prompt.
[root@localhost ~]# cd /home/.../TUTORIAL/
2. Then we create new experiment name as "example" by using the below Command prompt.

[root@localhost TUTORIAL]# ./create_newexpt.csh example
3. Created one folder name 'example' in TUTORIAL folder. enter example folder.
4. enter example folder by Command prompt.

[root@localhost TUTORIAL]# cd example
   in that folder we get six folders. we delete the three link folders,
   those are feature_files,
   s3trainer and util.
4. create three folders and assign names as feature_files, wav and raw respectively.
5. Copy s3trainer, util, lm3g2dmp, wavtoraw.csh, dictophones.c folders from r TUTORIAL

```
    and paste in 'example' folder.
6. Put all wav files in wav folder(Note
: those wav files must be saved with extension of .wav)
ex: 1.wav, 2.wav, 3.wav, .....
7. In example folder  create example file/**** Procedure Starts****/
    In that file  write text form of wav file
(ex: A      AA       I)
Note:  must care about spaces don't give Space. And at one line only
one sentence
with corresponding to wav file.
8. In example folder  create example.trans file
    In that file  write textform of wav file with .wav file.
(ex: <s> A </s>  (1.wav)
  <s> AA </s>  (2.wav)
  <s> I </s>  (3.wav)  )
Note:  must care about spaces. Between <s> and A one Space, A </s>
onespace, and </s> and (1.wav) two spaces. And at one line only
one sentence with corresponding to wav file.
9.  Open wavtoraw.csh file and  change the wavdir and rawdir path
ex: like as set wavdir = /home/. . . /TUTORIAL/example/wav
set rawdir = /home/. . . /TUTORIAL/example/raw
10. Then  run wavtoraw.csh file by using the Command prompt.
[root@localhost example]# ./wavtoraw.csh
11. Then  observed r raw folder. Definately  got .raw files in r
raw folder.
12.  create one raw.ctl (control file) by using the below command.
[root@localhost example]# ls /home/eslab/TUTORIAL/example/raw/*.raw
> raw.ctl
13.  give r example.sent file to CMU file by using  link
42http://www.speech.cs.cmu.edu/tool
s/lmtool.html
 get one .tar file.  extract that file and we copy only '.dic'
 and '.lm' files and paste in r 'example' fpolder.
And give name as example.dic and example.lm respectively.
14.  open dictophones.c file in example folder. and Change the path of
*argv[]  as
```

```
*argv[] = { "","/home/eslab/TUTORIAL/example/example.dic", "" };
15. compile dictophones.c by using the command prompt.
[root@localhost example]# cc dictophns.c
16. and getting output by using command prompt.
[root@localhost example]# ./a.out
17. Copy that output from command prompt and save as example.phonelist
in r example folder.
18. sort out that exaple.phonelist file using below command.
[root@localhost example]# sort u example.phonelist
       After sorting  add SIL at end of example.phonelist file.
19. Then  enter in cscripts by using the command
[root@localhost example]# cd c_scripts/
a) open compute_mfcc.csh file and change outdirctory as
    set outdir = /home/eslab/TUTORIAL/example/feature_files/$x
b) open variables.def file and change
     set base_dir = /home/eslab/TUTORIAL/example.
20. run the compuet_mfc.csh using the command as
root@localhost c_scripts]# ./compute_mfcc.csh /home/eslab/TUTORIAL
/example/raw.ctl
21. Then we convert mfc.ctl(control) file by usin
g command as
[root@localhost c_scripts]# cd ..
[root@localhost c_scripts]# ls /home/eslab/TUTORIAL/example/
feature_files/raw/*.mfc > mfc.ctl
22. Then open mfc.ctl file and delete path and extension of .mfc
for all files before deleting it as  /home/.../TUTORIAL/raw/1.mfc
after deleting it as raw/1
/**** Training part Starts ****/
23. After that  once again open the variables.def file in cscripts
change  path as
 # Input to the trainer.
set listoffiles    = $base_dir/mfc.ctl
set transcriptfile = $base_dir/example.trans
set dictionary     = $base_dir/example.dic
set fillerdict     = $base_dir/lists/filler.dict/** Procedure Starts**/
set phonefile      = $base_dir/example.phonelist
```

```
24. After that  enter 01.cichmm folder in c_scripts folder and put
# symbol before unlimit, for all files in the folders 01.cichmm,
02.cichmm, 03.cichmm, 04.cichmm and 05.cichmm.
26. Then enter c_scripts by using command as
[root@localhost c_scripts]# cd c_scripts/
for training the data we do as two types. First one as
4327. runall .csh files by using below command.
[root@localhost c_scripts]# .runall.csh
and the second one as
[root@localhost c_scripts]# cd 01.cichmm
[root@localhost 01.cichmm]# ./slave_convg.csh
if ask y/ n enter y
[root@localhost c_scripts]#cd ../ cd 02.cichmm
[root@localhost 02.cichmm]# ./slave_convg.csh
if ask y/ n enter y
[root@localhost c_scripts]#cd ../ cd 03.cichmm
[root@localhost 03.cichmm]# ./slave_treebuilder.csh
if ask y/ n  enter y
[root@localhost c_scripts]#cd ../ cd 04.cichmm
[root@localhost 04.cichmm]# ./slave_tiestate.csh
if ask y/ n  enter y
[root@localhost c_scripts]# cd ../cd 05.cichmm
[root@localhost 05.cichmm]# ./slave_convg.csh
if ask y/ n  y
[root@localhost 05.cichmm]# cd ..
/**** Two ways are completed  and complete the training
the data is completed ****/
[root@localhost c_scripts]# cd ..
28. At present  are in exmaple in command prompt.
i.e. [root@localhost example]#
29.  convert the .dmp(lagugemodel dump) by using
[root@localhost example]# ./lm3g2dmp /home/eslab/TUTORIAL/
example/example.lm
/**** Decoder part Starts ****/
30. once again  enter variables.def file in c_scripts folder
and change as
```

```
 # Input to the decoder
set ctlfn     = $base_dir/mfc.ctl
set testref   = $base_dir/example.sent
set cepdir    = $base_dir/feature_files
set lmfile    = $base_dir/example.lm.DMP
set dictfn    = $base_dir/example.dic
set fdictfn   = $base_dir/lists/filler.dict
31. enter decoding in command prompt using command as
[root@localhost example]# cd decoding/
[root@localhost decoding]# ./compute_acc.ci.csh
32. run gaumodels file using below command
[root@localhost decoding]# ./launch_decode.ci.1gaumodels
33. compute accuracy using below command
[root@localhost decoding]# ./compute_acc.ci.csh

Then finally  get the word accuracy./**** Decoder part Ends ****/
/**** Procedure Ends***/

User Manual for festival speech synthesis system:

Installation procedure for Festival speech synthesis system
Required Software:
    1. Festival-te -0.33
    2. Festival 1.96
    3. Speech Tools 1.2.96
    4. FestVox 2.1

Installation Procedure:
1.First down load all above tools. The link to download  above tools
are http://www.cstr.ed.ac.uk/projects/festival/
http://sourceforge.net/projects/festival-te/files/
2. after that check compiler version .they should be
  C++ compiler
Note that C++ is not very portable even between different versions
of the compiler from the same vendor. Although we've tried very hard
to make the system portable, we know it is very unlikely to compile
```

without change except with compilers that have already been tested.
The currently tested systems are

1. Sun Sparc Solaris 2.5, 2.5.1, 2.6, 2.7, 2.9: GCC 2.95.1, GCC 3.2
2. FreeBSD for Intel 3.x and 4.x GCC 2.95.1, GCC 3.0
3. Linux for Intel (RedHat 4.1/5.0/5.1/5.2/6.0/7.x/8.0):
GCC 2.7.2, GCC 2.7.2/egcs-1.0.2, egcs 1.1.1, egcs-1.1.2, GCC 2.95.[123],
GCC "2.96", GCC 3.0, GCC 3.0.1 GCC 3.2 GCC 3.2.1
4. Windows NT 4.0: GCC 2.7.2 plus egcs (from Cygnus GNU win32 b19),
Visual C++ PRO v5.0, Visual C++ v6.0

3.Next create a folder and untar all the above files in that folder by using tar -xvf festival-te-0.3.3.tar.gz.
4. open terminal and enter into that folder by using ‘cd’ command.
5.cd speech tools
./speech tools
Make
6.cd festival
./festival
Make
7.cd festvox
./festvox
Make
8.cd festival-te-0.3.3
./install.sh
With this installation process ends. And to check the tool type festival in terminal
9. type festival
Festival> (voice_telugu_NSK_diphone)
This is the default telugu voice which already build in the tool.
telugu_NSK_diphone
festival>(tts "festival-te-0.3.3/examples/rams.txt" nil)
nil
Procedure to build new voice:
[root@localhost ~]# export FESTVOXDIR=/root/……/festvox
[root@localhost ~]# export ESTDIR=/root/……../speech_tools

//*** [root@localhost uniphone]# $FESTVOXDIR/src/unitsel/setup_clunits

```
hcu tlg my uniphone
-----Now replace FESTVOX and ETC in uniphone directory---
-----Create txt.done.data in etc folder ---
Txt.done.data contais the data for which you want to build the system.
 ***//

[root@localhost uniphone]#festival  -b festvox/build_clunits.scm '
(build_prompts "etc/txt.done.data")'
 [root@localhost wav]#arecord -d 10 -f cd -t wav  uniph_0001.wav

[root@localhost wav]#aplay uniph_0001.wav

[root@localhost uniphone]#bin/make_labs prompt-wav/*.wav

[root@localhost uniphone]#festival -b festvox/build_clunits.scm
'(build_utts "etc/uniphone.data")'

[root@localhost uniphone]#bin/make_pm_wave wav/*.wav

[root@localhost uniphone]#bin/make_pm_fix pm/*.pm

[root@localhost uniphone]#bin/simple_powernormalize wav/*.wav

[root@localhost uniphone]#bin/make_mcep wav/*.wav
[root@localhost uniphone]#festival -b festvox/build_clunits.scm
'(build_clunits "etc/uniphone.data")'
[root@localhost uniphone]#festival festvox/hcu_tlg_my_clunits.scm
'(voice_hcu_tlg_my_clunits)'
festival> (SayText "BArawa");
#<Utterance 0xb7199f78>
festival> (SayText "simeVnt");
#<Utterance 0xb71c1af8>
festival> (SayText "parinAmAlu");
#<Utterance 0xb71d3c28>
festival> (SayText "kAvu");
#<Utterance 0xb71ec5a8>
```

SPHINX-III

Installation and Working Procedure

```
 untar the file tutorial_single_machine.tar.gz
then we have the folder TUTORIAL/SPHINX
now we copy create_newexpr.csh file from TUTORIAL/SPHINX/
create_newexpr.csh to TUTORIAL folder.

Next  change the path in create_newexpr.csh to remove the
link and make into local host

rsync -aC /afs/cs.cmu.edu/user/robust/project/TUTORIAL/
SPHINX3/c_scripts $newexptname/
rsync -aC /afs/cs.cmu.edu/user/robust/project/TUTORIAL/
SPHINX3/decoding $newexptname/
rsync -aC /afs/cs.cmu.edu/user/robust/project/TUTORIAL/
SPHINX3/lists $newexptname/

rsync -aC /root/Desktop/Mani/TUTORIAL/SPHINX3/c_scripts
$newexptname/
rsync -aC /root/Desktop/Mani/TUTORIAL/SPHINX3/decoding
$newexptname/
rsync -aC /root/Desktop/Mani/TUTORIAL/SPHINX3/lists
$newexptname/

ln -s /net/$x/$cwd/SPHINX3/s3trainer $newexptname/s3trainer
ln -s /net/$x/$cwd/SPHINX3/util $newexptname/util

ln -s /root/Desktop/Mani/TUTORIAL/SPHINX3/s3trainer
$newexptname/s3trainer
ln -s /root/Desktop/Mani/TUTORIAL/SPHINX3/util $newexptname/util

# cepstra
ln -s /net/$x/$cwd/SPHINX3/feature_files ${newexptname}/feature_files

# cepstra
```

```
ln -s /root/Desktop/Mani/TUTORIAL/SPHINX3/feature_files
${newexptname}/feature_files

Then run the create_newexp.csh file

######################################################################
./create_newexp.csh

[root@localhost TUTORIAL]# ./create_newexpt.csh mani
hostname: Unknown host
A new directory called ./mani has been created with all the
files needed for a new experiment.  Run the new experiment in that directory.
[root@localhost TUTORIAL]#
######################################################################


#then we can find a folder in /root/Desktop/Mani/TUTORIAL/mani in
this folder we can see
#c_scripts, decoding, feature_files, lists, s3trainer, util,
#here we can delete the folders s3trainer, util and feature_files

#then copy s3trainer and util folders from SPHINX and past in mani folder

#then create 3 folders named as feature_files, wav and raw

#record the wav files usong arecord command

#Audio file recording with the prescribed format
arecord -f S16_LE -d 3 -r 16000 /root/Desktop/Mani/TUTORIAL/mani/
wav/001.wav

#then we have 2 copy some files like wav2raw.csh, dictophns.c, lm3g2dmp
     These three script are written for the purpose functioning of
     the tool with the required data preparation for the experiments.

#open wav2raw.csh file and chang the path then run
```

```
[root@localhost mani]# ./wav2raw.csh
file  1, /root/Desktop/Mani/TUTORIAL/mani/wav/001.wav
file  2, /root/Desktop/Mani/TUTORIAL/mani/wav/002.wav
file  3, /root/Desktop/Mani/TUTORIAL/mani/wav/003.wav

#create a raw ctl file

[root@localhost mani]# ls /root/Desktop/Mani/TUTORIAL/mani/raw/*.raw >
raw.ctl

#then enter into c_scripts and open compute_mfcc.csh files and change
the path

c_scripts.csh
set outdir = /root/Desktop/Mani/TUTORIAL/mani/feature_files/$x

# then run compute_mfcc.csh file

[root@localhost c_scripts]# ./compute_mfcc.csh /root/Desktop/Mani/
TUTORIAL/mani/raw.ctl
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
file = /root/Desktop/Mani/TUTORIAL/mani/raw/001.raw
outdir = /root/Desktop/Mani/TUTORIAL/mani/feature_files/raw
file 1 ,/root/Desktop/Mani/TUTORIAL/mani/raw/001.raw
file = /root/Desktop/Mani/TUTORIAL/mani/raw/002.raw
outdir = /root/Desktop/Mani/TUTORIAL/mani/feature_files/raw
file 2 ,/root/Desktop/Mani/TUTORIAL/mani/raw/002.raw
file = /root/Desktop/Mani/TUTORIAL/mani/raw/003.raw
outdir = /root/Desktop/Mani/TUTORIAL/mani/feature_files/raw
file 3 ,/root/Desktop/Mani/TUTORIAL/mani/raw/003.raw

then goto mani folder

[root@localhost c_scripts]# cd ..
[root@localhost mani]#
```

create the mfc ctl file

```
########################################################################
[root@localhost mani]# ls /root/Desktop/Mani/TUTORIAL/mani/
feature_files/raw/*.mfc > mfc.ctl
[root@localhost mani]#
```

open mfc.ctl file and replace like

```
/root/Desktop/Mani/TUTORIAL/mani/feature_files/raw/001.mfc
/root/Desktop/Mani/TUTORIAL/mani/feature_files/raw/002.mfc
/root/Desktop/Mani/TUTORIAL/mani/feature_files/raw/003.mfc
```

```
raw/001
raw/002
raw/003
########################################################################
and also create myth.send, myth.trans, myth.dic file
and we need myth.lm file we take this file from cmu
http://www.speech.cs.cmu.edu/tools/lmtool.html
here we will give the myth.sent file as input then it will give us
5 files out put but we need 1 file ".lm" file we will copy that file only
```

```
now we
Goto c_scripts and open variable.def file
change the paths
```

```
variable.def
# set base_dir = ...                  (at line no 17&18)
 set base_dir = /root/Desktop/Mani/TUTORIAL/mani

# Input to the trainer.              (at line no 35 - 40)
set listoffiles    = $base_dir/mfc.ctl
```

```
set transcriptfile = $base_dir/myth.trans
set dictionary     = $base_dir/myth.dic
set fillerdict     = $base_dir/lists/filler.dict
set phonefile      = $base_dir/myth.phonelist

# Input to the decoder               (at line no 63 - 71)
#set ctlfn     = $base_dir/decoding/rmtest.ctl
#set testref   = $base_dir/decoding/RM.test.sents
set ctlfn     = $base_dir/mfc.ctl
set testref   = $base_dir/myth.sent
set cepdir    = $base_dir/feature_files
set lmfile    = $base_dir/myth.lm.DMP
set dictfn    = $base_dir/myth.dic
set fdictfn   = $base_dir/lists/filler.dict


and open these files

/root/Desktop/Mani/TUTORIAL/mani/c_scripts/01.ci-chmm
/root/Desktop/Mani/TUTORIAL/mani/c_scripts/02.cd_untied
/root/Desktop/Mani/TUTORIAL/mani/c_scripts/03.buildtrees
/root/Desktop/Mani/TUTORIAL/mani/c_scripts/04.tiestate
/root/Desktop/Mani/TUTORIAL/mani/c_scripts/05.cd-chmm

and goto each folder and open all files and place # infront of
unlimit

#unlimit

and then runall or goto 01.ci-chmm file run the
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
[root@localhost mani]# cd c_scripts/
[root@localhost c_scripts]# cd 01.ci-chmm/
[root@localhost 01.ci-chmm]# ./slave_convg.csh
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
```

```
USAGE: ./slave_convg.csh <iteration number (def 1)>
Setting iter value to 1
Continue? (y/n) y
Cleaning up accumulator directories...
Cleaning up log directories...
Cleaning up qmanager directories...
/bin/rm: No match.
/bin/rm: No match.
Cleaning up model directories..
[root@localhost 01.ci-chmm]# cd ../02.cd_untied/
[root@localhost 02.cd_untied]# ./slave_convg.csh
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
USAGE: ./slave_convg.csh <iteration number (def 1)>
Setting iter value to 1
Continue? (y/n) y
Cleaning up accumulator directories...
Cleaning up log directories...
Cleaning up qmanager directories...
[root@localhost 02.cd_untied]# cd ../03.buildtrees/
[root@localhost 03.buildtrees]# ./slave.treebuilder.csh
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
Phone = AI, State = 0 - completed
Phone = AI, State = 1 - completed
Phone = AI, State = 2 - completed
Phone = AX, State = 0 - completed
Phone = AX, State = 1 - completed
Phone = AX, State = 2 - completed
Phone = D, State = 0 - completed
Phone = D, State = 1 - completed
Phone = D, State = 2 - completed
Phone = IX, State = 0 - completed
Phone = IX, State = 1 - completed
Phone = IX, State = 2 - completed
Phone = K, State = 0 - completed
```

```
Phone = K, State = 1 - completed
Phone = K, State = 2 - completed
Phone = N, State = 0 - completed
Phone = N, State = 1 - completed
Phone = N, State = 2 - completed
Phone = O, State = 0 - completed
Phone = O, State = 1 - completed
Phone = O, State = 2 - completed
Phone = R, State = 0 - completed
Phone = R, State = 1 - completed
Phone = R, State = 2 - completed
Phone = S, State = 0 - completed
Phone = S, State = 1 - completed
Phone = S, State = 2 - completed
Phone = T, State = 0 - completed
Phone = T, State = 1 - completed
Phone = T, State = 2 - completed
Phone = UH, State = 0 - completed
Phone = UH, State = 1 - completed
Phone = UH, State = 2 - completed
Phone = SIL, State = 0 - completed
Phone = SIL, State = 1 - completed
Phone = SIL, State = 2 - completed
[root@localhost 03.buildtrees]# cd ../04.tiestate/
[root@localhost 04.tiestate]# ./slave_tiestate.csh
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
/bin/rm: No match.
/bin/rm: No match.
[root@localhost 04.tiestate]# cd ../05.cd-chmm/
[root@localhost 05.cd-chmm]# ./slave_convg.csh
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
USAGE: ./slave_convg.csh <ngau (def 1)> <iteration number (def 1)>
Setting ngau to 1, iteration no. to 1
Continue? (y/n) : y
```

```
Assuming initial models are 1 gaussian per state
Setting iter value to 1
Cleaning up accumulator directories...
Cleaning up log directories...
Cleaning up qmanager directories...
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

we return back to mani folder and run the lm3g2dmp file

[root@localhost mani]# ./lm3g2dmp /root/Desktop/Mani/TUTORIAL/mani/mani.lm
/root/Desktop/Mani/TUTORIAL/mani/
INFO: lm3g2dmp.c(708): Reading LM file /root/Desktop/Mani/TUTORIAL/mani/mani.lm
(name "")
INFO: lm3g2dmp.c(730): ngrams 1=5, 2=10, 3=9
INFO: lm3g2dmp.c(428): Reading unigrams
INFO: lm3g2dmp.c(754):        5 = #unigrams created
INFO: lm3g2dmp.c(473): Reading bigrams
INFO: lm3g2dmp.c(766):       10 = #bigrams created
INFO: lm3g2dmp.c(767):        3 = #prob2 entries
INFO: lm3g2dmp.c(774):        4 = #bo_wt2 entries
INFO: lm3g2dmp.c(554): Reading trigrams
INFO: lm3g2dmp.c(783):        9 = #trigrams created
INFO: lm3g2dmp.c(784):        3 = #prob3 entries
INFO: lm3g2dmp.c(910): Dumping LM to /root/Desktop/Mani/TUTORIAL/mani/
/mani.lm.DMP
[root@localhost mani]#

Now we check logdir folder all files in the logdir is compiled r not
  then  we will go to decoding foleder

[root@localhost 05.cd-chmm]# cd ../../decoding/
[root@localhost decoding]# ./launch_decode.ci.1gaumodels
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
limit coredumpsize 0
unlimit datasize
```

```
onintr docleanup
set part = 1
set npart = 1
if ( 2 > 2 ) then
set startskip = 0
endif
set nlines = `wc $ctlfn | awk '{print $1}'`
wc /root/Desktop/Mani/TUTORIAL/mani/mfc.ctl
awk {print $1}
set ctloffset ctlcount
@ ctloffset = 0 + ( ( 3 * ( 1 - 1 ) ) / 1 )
@ ctlcount = ( ( 3 * 1 ) / 1 ) - 0
echo Doing 3 segments starting at number 0
Doing 3 segments starting at number 0
set result = ./result
if ( 0 == 0 ) then
set logfile = ./result/mani.cimodels-1.log
set matchfile = ./result/mani.cimodels-1.match
else
if ( ! -d ./result ) then
if ( ! -d /root/Desktop/Mani/TUTORIAL/mani/feature_files ) then
if ( ! -d /root/Desktop/Mani/TUTORIAL/mani/model_parameters/mani.ci_continuous )
then
if ( ! -f /root/Desktop/Mani/TUTORIAL/mani/myth.lm.DMP ) then
if ( ! -f /root/Desktop/Mani/TUTORIAL/mani/mfc.ctl ) then
./s3decode-anytopo.linux -logbase 1.0001 -bestpath 1 -mdeffn /root/Desktop/Mani/TUTO
-senmgaufn .cont. -meanfn
/root/Desktop/Mani/TUTORIAL/mani/model_parameters/mani.ci_continuous/means
-varfn
/root/Desktop/Mani/TUTORIAL/mani/model_parameters/mani.ci_continuous/
variances -mixwfn
/root/Desktop/Mani/TUTORIAL/mani/model_parameters/mani.ci_continuous/
mixture_weights -tmatfn
/root/Desktop/Mani/TUTORIAL/mani/model_parameters/mani.ci_continuous/
transition_matrices -langwt 9.5-feat 1s_c_d_dd -topn 32 -beam 1e-80
-nwbeam 1e-40 -dictfn
```

```
/root/Desktop/Mani/TUTORIAL/mani/myth.dic -fdictfn
/root/Desktop/Mani/TUTORIAL/mani/lists/filler.dict -lmfn
/root/Desktop/Mani/TUTORIAL/mani/myth.lm.DMP -inspen 0.2 -ctlfn
/root/Desktop/Mani/TUTORIAL/mani/mfc.ctl -ctloffset 0 -ctlcount 3 -cepdir
/root/Desktop/Mani/TUTORIAL/mani/feature_files -cepext mfc -bptblsize
400000 -matchfn
./result/mani.cimodels-1.match -agc none -varnorm no -cmn current
docleanup:
set savedstatus=0
onintr
echo ================
echo exited with status 0
exit 0

[root@localhost decoding]# ./compute_acc.ci.csh
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
ci result
/root/Desktop/Mani/TUTORIAL/mani
/net/localhost//root/Desktop/Mani/TUTORIAL/mani
WORD ACCURACY=  100.000% (    9/    9)  ERRORS=  0.000% (    0/    9)
[root@localhost decoding]#
```

Bibilography

[ABE67] Abercrombie, David. 1967. Elements of general phonetics. *Edinburgh: Edinburgh University press.*

[ABH10] Abhijit Debbarma, M. Nagamani and B.S.R.Krishna "Speech Recognition for Kokborok Language" International conference on "Biomedical Engineering And Assistive Technologies (BEATS – 2010)" during 17th - 19th December 2010 at NIT – Jalandhar.

[AGG11] Aggarval and M.Dave "Acoustic modeling problem for automatic speech recognition system: advances and refinements (Part II)," International journal of speech technology, Springer,Vol. 14, no. 4, pp, 309-320, Dec. 2011.

[AMD12] Ahmed, Irfan, Nasir Ahmad, Hazrat Ali, and Gulzar Ahmad. 2012. The development of isolated words pashto automatic speech recognition system., In Automation and Computing (ICAC), 18th International Conference on IEEE, 1-4.

[AMD00] Amdal, I., Korkmazskiy, F., Suredan, A.C., 2000., " Data-driven pronunciation modelling for non-native speakers using association strength between phones". In: ASRU2000, Vol. 1, Paris, pp. 85–90.

[ANI09] Anil Kumar, M. Nagamani, B.S.R. Krishna "Railway Reservation form filling thow speech interactive system"University of Hyderabad, December 2009.

[ANU09] M.A.Anusuya ,S.K.Katti(2009) ,Speech Recognition by Machine: A Review, International journal of computer science and Information Security. Vol. 6, No. 3.

[ARA01] Aravind Ganapathiraju , Jonathan Hamaker, Joseph Picone. (2001), Syllable-Based Large Vocabulary Continuous Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol. 9, No. 4.

[BAK75] J. K. Baker (1975), Stochastic modeling for automatic speech recognition, Speech Recognition, D. R. Reddy, Ed. New York: Academic.

[BAU66] Baum L. E., Petrie, T., (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics 37 (6): 1554–1563. doi:10.1214/aoms/1177699147. Retrieved 28 November 2011

[BEL16] J. R. Bellegarda and C. Monz, "State of the art in statistical methods for language and speech processing," Computer Speech and Language, vol. 35, pp. 163–184, Jan 2016.

[12] [PER11]Peri Bhaskararao, "Salient Phonetic features of Indian languages in Speech Technology," Sadhana, Vol. 36, Part 5, pp. 587-599, 2011.

[BIS06] M. Bishop, pattern Recognition and Machine Learing, Springer, 2006.

[BRA83] Braj B Kachru. The Indianization of English: the English language in India. Oxford University Press Oxford, 1983.

[BRE12] N. T. Vu, W. Breiter, F. Metze, and T. Schultz, "An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on asr performance," in Proceedings of the Interspeech, 2012.

[BUR98] Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).

[BYR06] W. Byrne, "Minimum Bayes risk estimation and decoding in large vocabulary continuous speech recognition," IEICE Special Issue on Statistical Modelling for Speech Recognition, 2006.

[BYR98] W. Byrne, M. Finke, S. Khundanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters G.Zavaliagkos (1998). Pronunication modelling using a hand-labelled corpus for conversational speech recognition. Proc. of ICASSP'98, pp. 313-316, Seattle.

[COH74] Cohen, P.S., Mercer, R.L.,(1974). The Phonological Component of an Automatic SpeechRecognition System. In: [23], pp. 177-187,Donovan 1996.

BIBILOGRAPHY

[DAV09] M. Davel and O. Martirosian (2009), Pronunciation dictionary development in resource-scarce environments, Interspeech. Pp 1-8.

[DEN04] L. Deng and D. Tu, "Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition" , in Processing, Jeju Island, South Korea, 2004,pp. 37-42.

[DUD01] Duda, R.O., Hart, P.B., and Stork, D.G. (2001). Pattern classification, Wiley, 2001.

[GAL96] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination, " IEEE Transactions on Speech and Audio Processing, vol. 4. 1996.

[HAM12] G. D. Hamdani, S.A.Selouani, and M. Boudraa, "Speaker-independent ASR for modern standard Arabic: Effect of regional accents," International Journal of Speech Technology, Springer, vol. 15, pp. 487- 493, 2012.

[HIS77] Hisashi Wakita(1977), Normalization of Vowels by Vocal Tract Length and Its Applications to Vowel Identification, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.25.

[IPA99] International Phonetic Association Handbook of the International Phonetic Association A Guide to the Use of the International Phonetic Alphabat A Regents pulication UK Cambridge University press1999.

[BAK09] Baker, Janet and Deng, Li and Glass, Jim and Khudanpur, Sanjeev and Lee, Chin-Hul and Morgan, Nelson and O39;Shgughnessy,Douglas"research-developments-and-directions-in-speech-recognition-and-understanding-part-1"2009.

[JEL76] F. Jelinek (1976), Continuous speech recognition by statistical methods, Proc. IEEE, vol. 64, no. 4, pp. 532–557.

[JEL97] F. Jelinek (1997), Statistical methods for speech recognition. Cambridge, MA: MIT Press.

[JOS11] Joshi, Vikas Bilgi, Raghavendra Umesh, S Benitez, Carmen García, Luz. (2011). Efficient Speaker and Noise Normalization for Robust Speech Recognition.. 2601-2604.

[JOS11a] Joshi, Vikas Bilgi, Raghavendra Umesh, S García, Luz Benitez, Carmen. (2011). Sub-Band Level Histogram Equalization for Robust Speech Recognition.. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 1661-1664.

[JUA00] B.H.Juang and S.Furui, 2000, Automatic speech recognition and understanding: A first step toward natural human machinecommunication,Proc.IEEE,88,8,pp.1142-1165.

[JUA04] Juang B.H., Lawrence R. Rabiner(2004) "Automatic Speech Recognition: A Brief History of Technology development" . www.ece.ucsb.edu/Faculty/.../354_LALI-ASRHistory-final-10-8.pdf.

[JUA05] L. Rabiner and B. Juang, Encyclopedia of Language and Linguistics, ch. Statistical methods for the recognition and understanding of speech. Netherlands: Elsevier, Amsterdam 2005, 2005.

[KAV06] Kavi Narayana Murthy and G Bharadwaja Kumar., "Language identification from small text samples". Journal of Quantitative Linguistics, 13(01):57-80, 2006.

[KSA98] K Samudravijaya, R Ahuja, N Bondale, T Jose, S Krishnan, P Poddar, PVS Rao, and R Raveendran. A feature-based hierarchical speech recognition system for Hindi. Sadhana (Academy Proceedings in Engineering Sciences), 23:313-340, 1998.

[KUM04] M. Kumar, A. Verma, and N. Rajput, "A large vocabulary speech recognition model for recognition of stop consonant-vowel (SCV) utterances," IEEE Transactionson Speech and Audio Processing. Vol. 10,no. 7, pp. 472-480, 2002.

[LEE96] Lee, Youngjik, and Kyu-Woong Hwang. "Selecting good speech features for recognition." ETRI journal 18.1 (1996): 29-40.

[LEG95] C. J. Leggetter and P. C. Woodland (1995), Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, 9, pp. 171-185 .

BIBILOGRAPHY

[LAL09] R.P. Lal, P.S.V.S. Saiprasad, M. Nagamani "Improved Algorithm for End point Detection and Silence removal for Speech processing" Proc. of RAFIT2009 during April 9-10th 2009 at Punjab University Patiala.

[LIV04] Livescu, K., and Glass, J., "Feature-based pronunciation modeling with trainable asynchrony probabilities." In Proc. ICSLP,2004.

[LIV05] Livescu, K., Feature-Based Pronunciation Modeling for Automatic Speech Recognition, Ph.D. dissertation, MIT, 2005.

[LIV07] Livescu, K., et al., "Manual transcription of conversational speech at the articulatory feature level." In Proc. ICASSP, 2007.

[LIV16] ontan, Lionel Ferrané, Isabelle Farinas, Jérôme Pinquier, Julien Aumont, Xavier. (2016). Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility. 10.21437/Interspeech.2016-431. PhD thesis, September 2005 Massachusetts Institute of Technology (1999). Https://ttic.uchicago.edu/ klivescu/papers/livescu_phd_thesis.pdf)

[MAS13] M. Masior, M. Igras, M. Ziolko, and S. Kacprzak, "Database of speech recordings for comparative analysis of multi-language phonems," Studia Informatica, vol. 34, no. 2B, pp. 79–87, 2013.

[MCC43] W. S. McCullough and W. H. Pitts,(1943), A Logical Calculus of Ideas Immanent in Nervous Activity, Bull. Math Biophysics, Vol. 5, pp. 115-133.

[MEN01] Z. Mengjie (2001), Overview of speech recognition and related machine learning techniques, http://www.mcs.vuw.ac.nz/comp/Publications/archive/CS-TR-01/CS-TR-01-15.pdf

[MOH05] Mohamed Afify, Feng Liu, Hui Jiang (2005), A New Verification-Based Fast-Match for Large Vocabulary Continuous Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4, pp 546-553.

[MOO94] R. K. Moore (1994), Twenty things we still don t know about speech , Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology .

[NAD11] Nadungodage T., Weerasinghe, R., Continuous Sinhala Speech Recognizer‖, Conference on Human Language Technology for Development, Alexandria, Egypt, May 2011.

[NAG04] Nagamani, P. N. Girija., "Investigations of Features for Audio Information Retrieval", Journal of Acoustic Society of India, Vol.32 No. 1-2, 2004.

[NAG05] M. Nagamani, P.N. Girija, "Intelligent Tutor for Telugu Language Learning – INTTELL"., Proc. Of International Conference on Image, Signal and Information Processing(ICISIP2005), January 4-7, 2005. (Scopus i.e. IEEE explorer)

[NAG07] M. Nagamani and P.N. Girija.(Pronunciation Variation Modeling for Speech Recognition System Accuracy" by at 1st National conference of Telugu Linguists' Forum(TeLF), 6-7th Nov 2007 at University of Hyderabad.

[NAG10] M. Nagamani, P. N. Girija and B.S.R.Krishna "Pronunciation Dictionary Comparison for Telugu Language Automatic Speech Recognition system" pp. 60-69, Vol. 2, No. 1, Jan-July 2010, Proc. of "Journal of Data Engineering and Computer Science" GRIET, Hyderabad

[NAG10a] M. Nagamani, B.S.R.Krishna and Sridhar Reddy. A "Neural Networks based Visual Speech Recognition through Lip Reading" Proceeding of National Symposium on Acoustics NSA 2010 from 11th - 13th November 2010, at Rishikesh.

[NAG10b] M. Nagamani and B.S.R.Krishna "Intelligent tutoring system for Telugu Language Learning" International conference on "BIOMEDICAL ENGINEERING AND ASSISTIVE TECHNOLOGIES (BEATS – 2010)" during 17th - 19th December 2010 at NIT – Jalandhar.

[NAG10c] M. Nagamani Abhijit Debbarma, M. Nagamani and B.S.R.Krishna "Speech Recognition for Kokborok Language" International conference on "BIOMEDICAL ENGINEERING AND ASSISTIVE TECHNOLOGIES (BEATS – 2010)" during 17th - 19th December 2010 at NIT – Jalandhar.

[NAG11] M. Nagamani, S. Manoj, B.S.R. Krishna, "Implementing Phoneme Based Segmentation Algorithms to ASR system" International Conference on Advanced Computing methodologies(ICACM-2011), ISBN:978-93-81269-40-4, Reed Elsevier India private Ltd, December 2011, Hyderabad.

BIBILOGRAPHY

[NAG12A] M. Nagamani, P.N. Girija, " Substitution error analysis for improving the word Accuracy in Telugu Language Automatic Speech Recognition system" IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 4 (July-Aug. 2012), PP 07-10 www.iosrjournals.org.

[NAG12b] M. Nagamani, P. N. Girija, " Lexicon design for Telugu Language Automatic Speech Recognition System", 2nd International Telugu Conference (ITC 2012), 2-4th November 2012, Vizag.(Presented Paper in Telugu langauge).

[NAG12c] M. Nagamani, T. VenuGopal, S. Uday Bhaskar, K. Rajeev, " Image to Speech Conversion System" 2nd International Conference on Biomedical Engineering Assistive Technologies(BEATS 2012), ISBN-13:978-81-925454-1-7, 6-7th December 2012 at NIT, Jalandar(Punjab).

[NAG13] M. Nagamani, P.N. Girija " Pronunciation Variants and Substitutional error analysis for Improving Telugu Language Lexical performance in ASR system Accuracy", International Journal of Scientific Engineering Research(IJSER), ISSN 2229-5518,pp. 1128-1133, Volume 4, Issue , July 2013.

[NAG14] M. Nagamani, T.V. VenuGopal, S. Manoj Kumar and Uday Bhasakar S., "Image to speech conversion system for Telugu Pronunciation" ISBN.9789383038176. Proc. Of National Conference on Advanced Computing and pattern Recognition, 2014.

[NAG16] Nagamani M , Shaik Fathima " Voice controlled Music Player Using Speech Recognition System" IJEEE, Issue2, Feb-Mar2016

[NAG18] M. Nagamani and syed Hameed et al., "Exploring the Speech Prosody Manipulation for Dual-Language TTS System" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 4, Ver. III (Jul - Aug 2018), PP 20-22 www.iosrjournals.org DOI: 10.9790/0661-2004032022.

[NAG63] K. Nagata, Y. Kato, S. Chiba, "Spoken Digit Recognizer for Japanese Language", NEC Res. Develop., No.6,1963.

[NAV14] P. Navneeth Kumar, M. Nagamani, M.S. Chakravarthy, J.M. Harish Kumar., "Phoneme Prediction in ASR system using Decision Trees" Proc. Of National Conference on Advanced Computing and Pattern Recognition, 8-9th January2014.

[NEL00] Nello Cristianini, John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[NEL12] Nelson Morgan, Fellow, IEEE, "Deep and Wide: Multiple Layers in Automatic Speech Recognition" IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012.

[PRA08] Pratyush Banerjee, Gaurav Garg, Pabitra Mitra, and Anupam Basu. Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali. 19th International Conference on Pattern Recognition, ICPR 2008., pages 1-4, 2008.

[RAB05] L. Rabiner and B. Juang, Encyclopedia of Language and Linguistics, ch. Statistical methods for the recognition and understanding of speech. Netherlands: Elsevier, Amsterdam 2005.

[RAB09] L.R. Rabiner, Biing-Hwang Juang and B. Yegnanarayana, Fundamentals of speech recognition, Chapter 2, New Delhi, India(Indian Subcontinent Adaption) : Pearson Education Inc., 2009.

[RAB10] Rabiner, L. Juang, B. H., Yegnanarayana, B.(2010), Fundamentals of Speech Recognition, Pearson Publishers.

[RAB89] Rabiner, Lawrence R. (1989) "Tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, 77 (2), pp. 257-286.

[RAB93] Rabiner.L , Juang B.H (1993), Fundamentals of speech processing, New Jersey: Prentice Hall.

[RAB99] Lawrence Rabiner, Biing Hwang Juang, "Fundamental of Speech Recognition", Copyright 1999by ATT.

[RAS16] R. Rasipuram, MM Doss.,"Articulatory featre based continuous speech recognition using probabilitistic lexical modeling" Computer Speech Language,2016 – Elsevier R Rasipuram "On learning Grapheme-to-Phoneme Relationships through the Acoustic Speech signal" The Phonetician, 2014 – infoscience.epfl.ch.

BIBILOGRAPHY

[RED96] D.R.reddy (1996), An Approach to Computer speech Recognition by direct analysis of the speechwave ,Tech.Report No.C549,Computer Science Department ,Stanford University.

[RIL99] Riley et al.,(1999). M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolie, J. McDonough, H. Nock, M. Saraclar, C. Wooters and G. Zavaliagkos, Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Comm. 29 (1999), pp. 209–224.

[SAD05] Sadaoki Furui (2005), 50 years of Progress in speech and Speaker Recognition Research , ECTI Transactions on Computer and Information Technology,Vol.1. No.2.

[SAI09] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in proc. ICASSP, pp. 4370-4373, 2009.

[SAI14] Sai Bala Kishore N, VenkatRao M, Nagamani M.," Environmental noise analysis for robust automatics speech recognition", article in Lecture Notes in Electrical Engineering 315:801:811,22015. : 10.1007/978-3-319-07674-4_75 Chapter Advanced Computer and Communication Engineering Technology Volume 315 of the series Lecture Notes in Electrical Engineering , 02 November 2014.

[SAM00] Samudravijaya, K., "Computer Recognition of Spoken Hindi",Proceeding of International Conference of Speech, Music and Allied Signal Processing, Triruvananthapuram, pages 8-13, 2000.

[SAN10] Sannella„M Speaker recognition Project Report report, From http://cs.joensuu.fi/pages/tkinnu/research/index.

[SAR10] D. Sart, A. Mueen, W. Najjar, V. Niennattrakul, E. Keogh (2010), Accelerating dynamic time warping subsequence search with GPUs and FPGAs, IEEE 10th Int. Conf. on Data Mining, pp. 1375–1378.

[SCA08] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the use of a multilingual neural network front-end," in Proceedings of the Interspeech, 2008, pp. 2711–2714.

[SCH06] Schultz, T., Kirchhoff, K., 2006. Multilingual Speech Processing. Elsevier Academic Press .

[SAK09] S Sakti, K. Markov, S. nakamura, and W. Mimker, "Statistical Speech Recognition." in Incorporating Knowledge Sources into statistical Speech Recognition, Vol.42, of lecture Notes in Electrical Engineering, pp. 19-53, USA: Springer US, 2009.

[SAK78] H. Sakoe and S. Chiba (1978), Dynamic Programming Algorithm Optimization for Spoken Word Recognition ,IEEE Trans.Acoustics, Speech, Signal Proc.,ASSP-26(1):43- 49.

[SHA05] Shamalee Deshpande, Sharat Chikkerur, and Venu Govindaraju. Accent classification in speech. Fourth IEEE Workshop on Automatic Identification Advanced Technologies., pages 139-143, 2005.

[SHA14] Shahnawazuddin, S. ; Sinha, R.," A low complexity cluster model interpolation based on-line adaptation technique for spoken query systems", 9th International Symposium on Chinese Spoken Language Processing (ISCSLP) , Pp.437 - 441, 2014.

[SHI01] K. Shinoda and C. H. Lee (2001), A structural Bayes approach to speaker adaptation, IEEE Trans. Speech and Audio Proc., 9, 3, pp. 276-287.

[SHR13] Shrikant Joshi and Preeti Rao. Acoustic models for pronunciation assessment of vowels of Indian English. International Conference on O-COCOSDA/CASLRE., pages 1-6, 2013.

[SIM06] Simon Kinga and Joe Frankel (2006), Recognition ,Speech production knowledge in automatic speech recognition , Journal of Acoustic Society of America.

[SRE04a] G. Sreenu, P.N. Girija, M. Nagamani, M. Narendra Prasad, " Telugu Speech Interface Machine for University Information System", Proc. Of IEEE Advanced Computing and Communications (ADCOM), Ahmadabad, December 2004. (scopus i.e IEEE explorer).

[SRE04b] G. Sreenu, P. N. Girija, M. Nagamani and M. Narendra Prasad., " A Human Speech Interactive system Telugu Language" Proc. of IEEE INDICON – 2004, 20-23 December 2004.

[STR99] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," Speech Commun., vol. 29, no. 2–4, pp. 225–246, 1999.

[SUS06] R. Sussex and P. Cubberley, The Slavic Languages. Cambridge Language Surveys, UK: Cambridge University Press, 2006.

[SWI12] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr," in Proceedings of the Spoken Language Technology Workshop (SLT), 2012 IEEE, IEEE. IEEE, 2012, pp. 246–251.

[TOL01] Tolba, H., O'Shaughnessy, D., 2001. Speech recognition by intelligent machines. IEEE Canadian Review 38,20-23.

[TUS13] Z. Tuske, R. Schl ¨ uter, and H. Ney, "Multilingual hier- ¨ archical mrasta features for asr," in Proceedings of the Interspeech, Lyon, France, Aug. 2013, pp. 2222–2226.

[VAP97] V. N. Vapnik (1995), The Nature of Statistical Learning Theory, Springer, New York.

[VIN10] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.- A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," J. Mach. Learn. Res., vol. 11, pp. 3371–3408, Dec. 2010.

[VIV06] Vivek, Hitesh, M. Nagamani, " Voice Driven Operating System" National Conference in Artificial Intelligence" held on 1st September 2006 at GRIET, Hyderabad.

[WEI89] A. Weibel, et. al. (1989), Phoneme recognition using time-delay neural networks, IEEE Trans. Acoustics, Speech, Signal Proc., 37, pp. 393-404.

[WES01] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 288–298, Mar. 2001.

[XIN13] Lei Xin, Andrew Senior, Alexander Gruenstein, and Jeffrey Sorensen(2013). Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices, interspeech.

[XUE14] Xuedong Huang and Li Deng, An Overview of Modern Speech Recognition. Microsoft Corporation.

[YAM16] Y Ma, MP Paulraj, S Yaacob, AB Shahriman, and SK Nataraj., "Speaker accent recognition through statistical descriptors of mel-bands spectral energy and neural network model". IEEE Conference on Sustainable Utilization and Development in Engineering and Technology, pages 262-267, 2012.