MODELLING AND ANALYSIS OF SPEECH RATE VARIATION FOR SPEAKER INDEPENDENT TELUGU AUTOMATIC SPEECH RECOGNITION SYSTEM

A thesis submitted to the University of Hyderabad
in partial fulfillment of the requirements for the award of

Doctor of Philosophy

in

Computer Science

Under the Supervision of

Prof. P. N. Girija

by

N. USHA RANI

SCHOOL OF COMPUTER AND INFORMATION SCIENCES

UNIVERSITY OF HYDERABAD

HYDERABAD – 500 046

INDIA.

2014

CERTIFICATE

This is to certify that the thesis entitled **"Modelling and Analysis of Speech Rate Variation for Speaker Independent Telugu Automatic Speech Recognition System"** submitted by **N. Usha Rani** bearing Reg. No. **06MCPC07** 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 thesis has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Prof. P.N.Girija
SUPERVISOR
School of CIS
University of Hyderabad
Hyderabad - 500046

Prof. A.K.Pujari
DEAN
School of CIS
University of Hyderabad
Hyderabad - 500046

DECLARATION

I, **N. Usha Rani**, hereby declare that this thesis entitled **"Modelling and Analysis of Speech Rate Variation for Speaker Independent Telugu Automatic Speech Recognition System"** submitted by me under the supervision of **Prof. P.N. Girija**, School of Computer and Information Sciences, University of Hyderabad, Hyderabad is a bonafide research work. I also declare that it has not been submitted previously in part or full to this University or any other University or Institution for the award of any degree or diploma.

Date: Name: N. USHA RANI

Reg. No: 06MCPC07

I would like to dedicate this Ph.D thesis to my parents

N. Srinivasulu & N. Dhana Lakshmi

ACKNOWLEDGMENTS

I'm deeply grateful to my Research Supervisor **Prof. P.N.Girija** for giving me the precious opportunity to study the challenging research topic. Without her valuable discussions and inspiring guidance, this thesis would not have been possible. I wish to express my deep sentiments of gratitude and reverence for her continuous support and encouragement throughout my research work. Words are inadequate to express my indebtedness to her for taking great effort in the completion of my research work.

I would like to express my sincere gratitude to the members of DRC, **Prof. Hrushikesh Mohanthy** and **Prof. S. Bapiraju** for continuous reviewing the research work, giving me the helpful remarks and encouragement to complete my research work.

I wish to express my gratitude to the Dean **Prof. A.K.Pujari**, for the support and encouragement to complete my research work.

I wish to express my sincere thanks to my Cousin **Rajagopal's family** for their valuable support in completion of my research work.

I convey my special thanks to my sister **N. Geetha Vani** and brother-in-law **N. Venu Gopal** for their support. And best wishes to my niece N**. Leena (Lucky)** for her sweet deeds which always refreshes me.

I am also thankful to all the people who directly and indirectly helped me in the completion of my research work.

(N. USHA RANI)

ABSTRACT

Speech is one of the most prominent and natural means of communications among the humans as well as with machines. Speaker Independent Automatic Speech Recognition (SIASR) is one of the emerging technologies playing a crucial role in our daily lives. Humans have the capability of recognizing the speech spoken at different rates of speech by different speakers in various contexts. It is becoming a complex task for the machines to recognize the speech produced by humans. A lot of effort was involved to enhance the performance of SIASR system. Still, errors are generating in current SIASR systems.

Telugu language is one of the most widely spoken South Indian languages. It is the official language for the states of Andhra Pradesh and Telangana. Speech rate is one of the factors affecting the performance of SIASR system. The present work focuses on recognition accuracy at different rates of speech in Telugu. Hence training and testing are done at different rates of speech namely Normal Speech Rate (NSR), Slow Speech Rate (SSR) and Fast Speech Rate (FSR). Recognition accuracy is more when same rate of speech is used in training and testing. Word Error Rate (WER) is more when different rates of speech are involved in training and testing. Substitution errors, deletion errors and insertion errors are analyzed that occur when different speech rates are involved in training and testing. Confusion between the words with similar phonetic transcription is the main cause for getting substitution errors. The present work focus on the reduction of confusion pairs so that substitution errors will be reduced. The proposed Pronunciation Dictionary Modification Method (PDMM) is used to update the pronunciation dictionary which reduces substitution errors. Thus

Word Recognition Rate (WRR) improves a lot. The proposed Forward and Backward Search Method (FBSM) is applied on decoder output to reduce substitution error or insertion error or deletion error. Significant improvement is observed in the Sentence Recognition Rate (SRR) when FBSM is applied on decoder output. Thus the performance of the SIASR system is improved with these two methods.

.

CONTENTS

List of Figures

List of Tables

Abbreviations

ASR	Automatic Speech Recognition
ARPA	Advanced Research Projects Agency
ASCII	American Standard Code for Information Interchange
Corr-sents	Correct-sentences
CLN	Cepstrum Length Normalization
CMU	Carnegie Mellon University
DCT	Discrete Cosine Transform
DER	Deletion Error Rate
DFT	Discrete Fourier Transform
FSR	Fast Speech Rate
FIR	Finite Impulse Response
FSN	Finite State Network
FBSM	Forward and Backward Search Method
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HYP	Hypothesis file
IER	Insertion Error Rate
MAP	Maximum a Posteriori
MFCC	Mel Frequency Cepstral Coefficient
NSR	Normal Speech Rate
PDF	Probability Density Function
PDMM	Pronunciation Dictionary Modification Method
PSOLA	Pitch Synchronous Overlap and Add
REF	Reference file
ROS	Rate of speech
Sents	Sentences
SRR	Sentence Recognition Rate
SSR	Slow Speech Rate
SIASR	Speaker Independent Automatic Speech Recognition
SDASR	Speaker Dependent Automatic Speech Recognition
SER	Substitution Error Rate

TD-PSOLA	Time Domain Pitch Synchronous Overlap and Add
VQ	Vector Quantization
WSOLA	Waveform Similarity Overlap-Add technique
WER	Word Error Rate
WRR	Word Recognition Rate

CHAPTER-1

INTRODUCTION

Speech plays a vital role in communication among humans and machines**.** Humans are very intelligent and efficient in recognizing speech but it is difficult for the machines. Automatic Speech Recognition (ASR) is the process of accepting the speech signals spoken by speakers through a recording device and converting the accepted speech into text format. Speaker Independent Automatic Speech Recognition (SIASR) system is the ability of a machine to recognize speech irrespective of any speaker. SIASR system need not be trained for every speaker before it can be used in applications. It requires training by several speakers but other speakers also can use this system. In Speaker Dependent Automatic Speech Recognition (SDASR) system, training is required for every speaker. Speaker Independent Automatic Speech Recognition (SIASR) systems are suitable for many real time applications.

Much effort has been made to achieve human-like performance in SIASR system. A great deal of research has been done to extend the capabilities of SIASR system. Research in SIASR has made significant growth in achieving development in real time applications with tremendous growth of technology. Humans can recognize speech in all contexts with background information naturally, whereas machines lack this ability. To interact with machines using speech mode, the system should be able to recognize utterances of different speakers in various contexts with some background information relating to the situation.

From a technical point of view, SIASR is a challenging task in generating a sequence of words for the given speech signal. Also it should be adaptive to operate on the characteristics of new speakers. There are many factors such as speaking style, age, gender, accent, speech rate, health conditions, emotions, environmental variation conditions, and channel variations etc., which affect the exact recognition of speech by the machine **[1]**.

In SIASR system, Hidden Markov Models (HMMs) are used to represent subword units as a sequence of states and their transition probabilities from one state to another state. Gaussian Mixture Models (GMMs) are used to model each state of an HMM. Well-organized algorithms should be used for training and decoding the SIASR system. Forward-Backward algorithm is used to evaluate probability of different HMMs generating the same observation sequence. In decoding, Viterbi algorithm is used to find the most likely state sequence for the given observation sequence. Baum-Welch algorithm is used to re-estimate the state transitions and output probabilities known as training (learning). These algorithms are explained in **Chapter-3.**

1.1. FACTORS AFFECTING THE PERFORMANCE OF SIASR

Depending on the task or the situation, humans articulate in different ways because of their own characteristics. Hence SIASR system should handle these characteristics if they are intended to be used. Inter-speaker variability such as gender and age has an impact on the accuracy of SIASR system. The average length of vocal tract of females (about 14.5 cm) is smaller than that of males (about 17.5cm), which leads to a natural variability. Recognition accuracy will be more if same gender (male/female) is used for training and testing. Recognition accuracy will be low when male voices are tested with female training models or female voices are tested with male training models **[2]**. There are many differences between adult and children's speech. Due to a shorter vocal tract and smaller vocal folds children have higher fundamental and formant frequencies than those of adults. Hence the recognition accuracy will be poor for the children when compared with adults **[3]**.

Emotional conditions such as happiness, anger, stress, confidence etc., also affect the speech recognition results. Health conditions also have an effect on the speech recognition results. The accuracy is high when the speaker is in normal health condition. If the speaker has cold or sour throat, automatically it affects the speech production since speakers cannot pronounce all the words clearly. This unclear pronunciation includes

more pauses, truncated words, heavy breathing etc. This leads to reduction of accuracy in SIASR system. However above types of speech are not considered in this work.

Pronunciation variation can cause errors in SIASR. Due to various reasons, utterances are almost pronounced differently and varied from one speaker to another in different situations. The variability is due to co-articulation, regional accents, speaking rate, speaking style, etc. Pronunciation variation can be classified into complete changes and partial changes. Complete changes or phone changes are the replacement of a phoneme by another alternative phone. Partial changes or sound changes are the variations within the phoneme such as nasalization, centralization, voiceless and voiced **[4]**.

1.2. APPROACHES TO SPEECH RECOGNITION

There are three main approaches used for speech recognition namely Acoustic-Phonetic approach, Pattern-Matching approach and Artificial Intelligence approach **[5]**.

The Acoustic-Phonetic approach to speech recognition is based on finding and providing appropriate labels to speech. The first step in this approach is a spectral analysis of speech combined with a feature detection that converts the spectral measurements to a set of features. These features describe the broad acoustic properties of different phonetic units. The second step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to every segmented region, resulting in a phoneme lattice characterization of the speech. The third step is to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling. The acoustic phonetic approach has not been widely used in most commercial applications.

The Pattern-Matching approach involves pattern training and pattern comparison. This approach uses well formulated mathematical framework and establishes consistent speech pattern representations for pattern comparison from a set of labeled training samples through an appropriate training algorithm. A speech pattern representation can

be in the form of speech template or statistical model and can be applied to a sound, a word or a phrase. In pattern comparison, direct comparison is done on unknown speech patterns and possible patterns learned during training. This is used to determine the identity of unknown speech patterns with maximum matching probability. In this, there exists two methods namely template approach and stochastic approach. In template approach, templates for all words are constructed. A collection of speech patterns are stored as reference patterns representing the dictionary of speaker's words. Recognition is done by matching an unknown spoken utterance with each of these reference templates and the best matching pattern is selected. In stochastic approach, probabilistic models are used to deal uncertain and incomplete information. Stochastic approach is the most suitable approach to speech recognition. Hidden Markov Model (HMM) is a popular stochastic approach. A HMM is a Markov chain with probability distributions of observations assigned for every state of the Markov model. The Markov chain is defined by the set of states and their transition probabilities defining a prior distribution for the state sequences. This model plays an important role in the area of SIASR which is used in the present work.

The Artificial Intelligence approach attempts to mechanize the speech recognition procedure according to the way a person applies intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. It is a hybrid of the acoustic-phonetic approach and the pattern-matching approach. It integrates phonemic, lexical, syntactic, semantic and even pragmatic knowledge for segmentation and labeling, and uses tools such as artificial neural networks for learning the relationships among phonetic events. Expert system is widely used in this approach.

1.3. APPLICATIONS OF SIASR

Applications of speech technology in daily life are growing rapidly now-a-days. This technology becomes an important part in the real world. The application of SIASR systems opens many possibilities in the world. SIASR applications are different from other applications of computer **[6].** Dictation is the common use of the SIASR. This

includes medical transcriptions, legal and business dictation. In some cases special vocabulary is used to increase the accuracy of the SIASR system. Speech systems that are designed to perform functions and actions on the system are defined as Command and Control system. Voice repertory dialer, automated Call-type, Call-distribution by voice commands, directory listing retrieval, Credit card sales validation are some of the applications that come under command and control system. Some PBX/Voice Mail systems allow callers to speak commands instead of pressing buttons to send specific tones. There is inevitable interest in the use of speech recognition in domestic appliances such as ovens, refrigerators, dishwashers and washing machines. Many people have difficulty in typing due to physical limitations such as the Repetitive Strain Injuries (RSI), muscular dystrophy, and many other disabilities. For such people, speech recognition system can be used to complete their activities.

1.4. AIM OF THE PRESENT WORK

Humans can recognize speech in different speaking styles and speaking rates (Normal, Fast and Slow), but for the computers it is a difficult task to recognize the speech. Speech recognizers implemented in computers perform relatively poorly when speech rate is very fast or very slow. In order to improve computer performance, several researchers have proposed that measuring speech rate prior to speech recognition will result in higher success rates of automatic speech recognizers and several ways to automatically measure speech rate in terms of phones or syllables per time unit have been put forward. **[7]**. Many factors such as stutters, slips of the tongue, interruptions, hesitations, lengthening, filled pauses and laughter affect speech rate **[8]**. Timing and acoustic realization of syllables or phonemes are frequently affected when the rate of speech is fast or slow. Phone-by-phone length stretching and sentence-by-sentence length stretching are used for normalization **[9]**.

The accuracy of SIASR system is severely affected when there are mismatches between training and testing conditions **[10]**. Performance of SIASR system degrades when the speech rate is varied for training and testing (decoding). In the present work, it

is observed that the accuracy was low due to mismatches between different rates of speech in training and testing conditions for SIASR system. This motivates to analyze and reduce different kinds of errors to improve the accuracy of the SIASR system for Telugu language.

The objective of this work is to improve the accuracy of the SIASR system on different rates of Telugu speech. As speech signals are non-stationary, obtaining compatibility between training and testing is difficult. The models built during the training and testing should be compatible in order to obtain better results. Some un-meaningful sounds such as lip smacking, deep breaths etc are produced in speech while recording. It is not possible to use perfect or equivalent phonetic representations for these sounds. Filled pauses normally occur during hesitations in speech while recording. So it is very difficult to represent these sounds in the pronunciation dictionary. This is one of the reasons to occur substitution errors. Substitution errors also occur due to confusion between phonetically similar words. After analyzing present experimental results it is found that there are more substitution errors than deletion and insertion errors. Hence it is necessary to reduce substitution errors to improve accuracy. Performance of SIASR system is improved with a better language model **[11]**.

The aim of the present work is to minimize the errors that occur when different speech rates are used in training and testing. Confusion between the words leads to substitution errors during recognition. Hence Pronunciation Dictionary Modification Method (PDMM) is proposed to modify the pronunciation dictionary. The modified dictionary reduces substitution errors. In this method, Levenshtein distance is used to obtain the similarity of phones in a confused pair **[12]**. The phonetic transcription of confused word is updated as the next possible phonetic transcription in the pronunciation dictionary when the Levenshtein distance is less than the length of $3/4^{th}$ of the longest word in the confusion pair. This modified dictionary (lexicon) is given to the decoder to minimize the occurrence of the confusions during recognition. An error word related to deletion or insertion or substitution in a sentence makes a greater affect on the performance of SIASR system. Hence Forward and Backward Search Method (FBSM) is proposed to reduce error word related to deletion or insertion or substitution.

1.5. ORGANIZATION OF THESIS

Chapter 2 reviews the factors affecting the performance of SIASR system and it focuses on the issues related to speech rate variations. Many previous studies related to the pronunciation dictionary are discussed in this chapter.

Chapter 3 describes HMM based acoustic model, pronunciation dictionary (lexicon) and language model used in the speech recognition system. The basic methods and algorithms used for training and testing (decoding) are discussed in this chapter.

Chapter 4 describes speech database collected for present work. Pitch Synchronous Overlap and Add (PSOLA) method is discussed in this chapter. This method is used to convert the Normal Speech Rate (NSR) into Slow Speech Rate (SSR) and Fast Speech Rate (FSR). The proposed modules in the SIASR system are also presented in this chapter.

Chapter 5 deals with the experiments. Training and testing are performed at different rates of speech. Different types of errors are analyzed that occur when different speech rates are involved in training and testing.

Chapter 6 reveals the role of pronunciation dictionary (lexicon). Substitution errors are reduced with modified dictionary using proposed Pronunciation Dictionary Modification Method (PDMM).

Chapter 7 focuses on the correction of one word error related to insertion or deletion or substitution in a sentence that obtained from the decoder of SIASR system using proposed Forward and Backward Search Method (FBSM).

Chapter 8 depicts the conclusions based on the present research work and some suggestions have been given for future research work.

CHAPTER-2
LITERATURE SURVEY

Speech rate has been identified as an important phenomenon in the recognition of speech by humans and computers. This chapter is a discussion of previous research on the effect of speech rate in SIASR system. Many previous studies are discussed here on the role of building a good pronunciation dictionary for better performance of the SIASR system.

2.1. Speech recognition errors

Speech rate has been shown to have a significant effect on recognition accuracy. Recognition accuracy degrades when the speech rate differs from training and testing. It is important to note the type of errors that occur when the speech rate is varied from training and testing **[13].** Random or indeterminate errors are caused by uncontrollable fluctuations of voice that affect parameterization and experimental results. Systematic or determinate errors are caused by instrumental, methodological or personal mistakes. Gross errors are caused by experimental carelessness or equipment failure. Based on the syntactic and semantic corrections, wrongly recognized parts of speech are usually removed on the next step of speech recognition **[14].** Errors are classified into insertions, deletions and substitutions. Decision tree is used for error analysis **[15].**

2.2. Conversion of NSR to SSR and FSR

Individual speakers also vary their speech rates in different contexts. These variations affect the acoustic patterns of the SIASR system. Variations in speech rate affect both spectral features and pronunciation of words. Speech rate is generally defined as the number of linguistic units (phones, syllables and words) spoken in unit time. The fast speech is recognized by stretching the length of the utterance in the cepstrum domain. The degree of stretching for an utterance is determined by its rate of speech which is based on a maximum likelihood. Cepstrum Length Normalization (CLN)

algorithm is used for fast speech to avoid unexpected short duration and drastic changes in the dynamic acoustic features. The objective of this algorithm is to lengthen and smooth the cepstrum. Three approaches are used to change the length of segments. The first approach is to insert/ drop frames uniformly in the speech segment. The second approach is to repeat/ delete only those frames that represent the steady state of each phone segment. This was done by searching those frames that had the minimum distortion with respect to their neighbors in a phone segment. The third approach is to create new frames by interpolating neighboring frames. This was done with approximated band-limited interpolation **[16, 17]**.

Waveform Similarity Overlap-Add technique (WSOLA) produces high quality of time-scaled speech. Normal speech is expanded to convert into slow speech and compressed to convert into fast speech. In WSOLA continuity of speech signal during modification, it is much useful for speech recognition **[18].** When people speak slowly, they can stretch the different phones but they can also insert more pauses. They can also place stress on words where it was absent at normal speech rate. These two strategies make it harder to mimic natural slow speech closely. To gain further improvement we could search where these additional pauses are added and try to replicate them in the slowed-down version. Pauses could also improve slowed-down natural fast speech where pauses are often absent. **[19]**.

2.3. Estimation of speech rate

Speaking rate is directly estimated from the speech signal without reference to lexical units. The energy envelope of the speech has rapid change when the speaking rate is high. This change should be reflected in the short term spectrum of the energy envelope. Finer (more frequency-dependent) measures could potentially provide high accuracy, but as a first attempt the wideband measure should incorporate the gross properties of speaking rate. Energy rate (enrate) measure is the first spectral moment of the energy envelope of speech and was shown to correlate with speech rate **[20].**

Speech rate variability has been found to be significant in increasing the error rate of speech recognition, especially when it deviates greatly from the training data **[21]**.

Mean and standard deviation for each triphone was calculated for the training data. Modification of the triphone duration is done according to the stretch factor. Stretch factor was obtained for all the words for every speaker based on the true word length and length of the word estimated. Duration of the vowel affects with variation of speech rate. Formant frequencies varied for different rates of speech. The long vowels occupied a more peripheral portion of the F1-F2 vowel space than the short vowels **[22, 23]**.

Rate of Speech (ROS) is computed as the average number of phones per second over an utterance. The ROS was then used to normalize the durations of the phones. There are two ways of using the ROS normalization, at the hypothesis level and at the speaker level **[24, 25]**. Word and sentence level timing in natural fast speech differs from normal speech rate. Naturally spoken fast speech is complex for modeling than artificially compressed speech. Both the segmental and the timing changes that accompany natural fast speech rates are due to articulatory restrictions, and do not serve a communicative purpose **[26]**.

The ROS for a particular sentence was calculated by dividing the number of non-silenced transcribed phones by the non-silence duration of the sentence **[27]**. Decision trees (D-trees) should be trained to predict the pronunciation of words based on information about surrounding words. D-trees are statistical classifiers that can select a set of features to improve the probability of a particular pronunciation. Thus D-tree algorithm with a substantial number of features, such as the identities and features of surrounding phones or extra-segmental features like speaking rate and word predictability. This algorithm automatically selects the best combination of these features to improve pronunciation classification **[28]**. The vowel detection algorithm provides an estimation of the actual number of vowels present in the waveform **[29]**.

Each word is given parallel pronunciations of fast and slow version phones. Both fast and slow version pronunciations are initialized from the original rate-independent version, with the simple replacement of rate-independent phones by rate-specific phones. The recognizer automatically finds the pronunciations that maximize the likelihood score during search, and thus avoids the need for ROS estimation before recognition. In

addition, the search algorithm is allowed to select pronunciations of different rates across word boundaries, thus coping with the problem of speech rate variation within a sentence **[30]**. The speaking rate in syllables per second was calculated for each utterance, removing starting and ending silence using the forced alignment **[31]**.

2.4. Role of pronunciation dictionary and language model in speech recognition

Constructing acoustic model, language model and lexicon for a particular language is important. Speech recognition accuracy can be affected by an inaccurate acoustic and language modeling and also a different pronunciation of words for each speaker **[32]**. A pronunciation dictionary for a specific language is built from a word list that may contain incorrect spellings and borrowed words from other languages **[33]**. As the number of distinct word forms can grow very large, it becomes difficult to train language models that are both effective and cover the words of the language well. Teemu Hirisimaki et al described and evaluated language models based on the segmentation of text corpora into suitable word fragments by an unsupervised machine learning algorithm **[34].**

Discriminative training optimizes the lexicon. In the discriminative training, some new words, which are frequently decoded wrong, are selected as word candidates. Through training, some significant new words are added into lexicon so that recognition errors due to these words can be eliminated. There are three kinds of new words which are frequently decoded wrong. Some words which are not consider by the linguists, domain specific words, proper noun such as person name, place name, date, number etc. **[35]**. These new words are not included in the traditional dictionary. The probabilities of these new words are estimated by the trigram of single characters. In the process of recognition, the discrimination between these new words and other similar characters is little. So they are frequently decoded wrong. Though counting the error number of decoded strings, some new words are chosen and added into dictionary to increase the discrimination.

Recognition errors are minimized by replacing error word with correct word or error phrase by correct phrase by re-speaking, typing etc. **[36]**. Some applications of

speech recognition, such as interactive telephone-based directory assistance services system require large vocabularies of surnames, first names, city names etc. Refining N-best hypotheses list provided by a speech recognizer by applying lexical rules **[37]**. The goal of lexicon optimization is to construct a lexicon with exactly those words that are most likely to appear in the text data **[38]**.

Lexical coverage indicates the ratio between the number of words from the training data that are covered in the vocabulary, and the total number of words in the training data **[39]**. Pattern recognition techniques are used for verifying the correctness of a lexicon **[40]**. Pronunciation variation degrades the performance of speech recognition system. Variation in pronunciation is observed generally for frequently uttered words. Speaking rate and word frequency influence on the pronunciation. Simply adding all the pronunciation alternatives to the lexicon may reduce the performance of speech recognition system. Probability-based pruning and count-based pruning methods are used to include the number of pronunciations for each word in the lexicon **[41]**.

Pronunciation variation is observed more in conversational speech. Without explicit models of non-verbal elements such as hesitations, artefacts, the corresponding audio segments will be recognized as keywords from the system vocabulary. The extension of the vocabulary by lexical models for each type of the hesitations and artefacts allows the speech recognition system to detect these non-verbal elements and avoid false recognition of the keyword units **[42]**. A pronunciation lexicon accounting for accented speech may be created by extending a canonical lexicon with accented pronunciations **[43]**.

A better performance is expected if a language model is adopted in a recognition system for post-processing phoneme estimates and making corrections with a set of explicit rules of the language model **[44].** In the first rule, phonological constraints are checked in three positions of the word namely beginning, middle and at the end of the word. In the second rule, vowel harmony rules could be verified to exclude some of the vowel additions and consonant replacements from the recognized words. In the third rule,

Consonant and vowel phoneme additions can be corrected by syllabification when the additional vowel phoneme will appear between two syllables after a consonant phoneme.

2.5. Confusion pairs and word frequency

Inter-word dissimilarity measure based on the Dynamic Time Warping is used to classify whether the word pairs are confusable or not confusable. Firstly, the phonetic transcriptions of the two words to compare are aligned using only phonetic information. After the alignment, the accumulated distance is obtained with a new inter-phone acoustic distance calculated between the HMMs of the phones. Distances between the phones of words are popularly known as the acoustic confusability **[45]**.

Specific words in a large corpus tend to co-occur frequently with certain other context words, and misrecognitions of those specific words will also tend to co-occur with the same context words. Co-occurrence analysis is to determine for any give word in the vocabulary whether the other words are very likely to occur near the given word and are not likely to occur elsewhere **[46]**. Stopwords are words with little intrinsic meaning. Typically, these words are found with such high frequency that they lose any usefulness as search terms. In co-occurrence analysis, stopwords are usually omitted because their over abundance in a text can affect the resulting probabilities disproportionately **[47]**.

CHAPTER -3

SPEECH RECOGNITON SYSTEM

This chapter describes the concepts of speech recognition system.

Speech recognition is the process of making the system (machine) to convert the recorded acoustic signal into a sequence of words or phrases or sentences in text format. In other words, it determines the most likely sequence of words $W = w_1,\ldots,w_n$ for the given acoustic observations extracted from the speech signal $O = o_1\ldots,o_n$. . The aim of the speech recognition is to decode the word sequence based on the acoustic observation sequence, so that the word sequence has the maximum a posteriori (MAP) probability as **[48],**

$$\widehat{W} = P(W|O) = \arg_w \max P(W|O) \quad (3.1)$$

Using Bayes rule Eq. (3.1) can be written as

$$P(W|O) = \arg_w \max \frac{P(O|W)P(W)}{P(O)} \quad (3.2)$$

P(O) can be ignored as it is independent of W. the MAP decoding rule of Eq. (3.1) is

$$\widehat{W} = \arg_w \max P(O|W)P(W) \quad (3.3)$$

The first term in Eq. (3.3), P(O|W) is called acoustic model as it estimates the probability of sequence of acoustic observations for the given word sequence. To compute P(O|W) for automatic speech recognition, build statistical models for subword speech units (phone based units, syllable based units etc.). Phone-based units are considered for the present research work. Word models are built from the subword speech unit models (using lexicon which indicates the composition of words) and then postulates word sequences and acoustic model probabilities are evaluated through concatenation. The second term in Eq. (3.3), P(W) is called the language model which describes the probability of a word sequence.

The block diagram of the speech recognition system is shown in **Figure 3.1.** All modules in the block diagram are explained in the following sections.

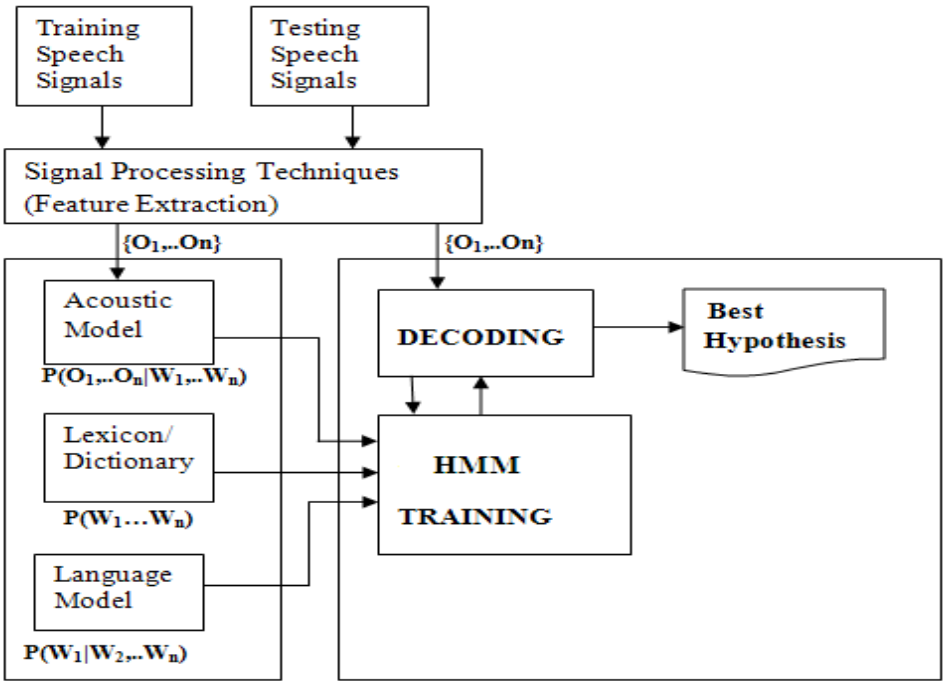

Figure 3.1: Block diagram of SIASR system

3.1. FEATURE EXTRACTION

The performance of speech recognition system depends on the extraction of features from the speech signals. The speech signal is converted into a parameterized sequence of feature vectors which are optimal for the speech recognition. Signal processing techniques are used to extract Mel Frequency Cepstral Coefficients (MFCC) **[49]**. The following **Figure 3.2** shows the steps involved in the extraction of Mel-Frequency Cepstral Coefficients (MFCC).

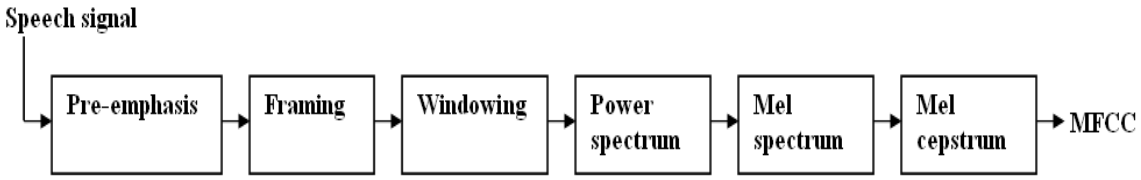

Figure 3.2: Feature Extraction process

Pre-emphasis is used to improve the speech signal quality. Pre-emphasis increases the magnitude of higher frequencies of speech signal with respect to the lower frequencies. The aim is to compensate the high frequency part of the speech signal that was suppressed during the human sound production. High pass Finite Impulse Response (FIR) filter is used for this purpose.

Speech signal is sampled and digitized into discrete samples known as frames by performing sequence of shifting. Segmenting of original speech with rectangular window causes discontinuities at the edge of the segment. In order to reduce the discontinuities of the speech signal at the edges of each segment (frame), Hamming window is preferred to be used for smoothen the transition among the frames. Frame rate is 100 frames per second and window length is 0.025625 seconds are the default values used in Sphinx-3.

Power spectrum is obtained by performing Discrete Fourier Transform (DFT). The DFT size is 512 which is a default value in Sphinx-3. It extracts the spectral information from the windowed signal. It determines the energy at each frequency band equivalent to time domain signal. Equally spaced frequency bands are generated by DFT which are highly correlated with each other. Thus the power spectrum representation is highly redundant.

Mel spectrum is obtained by applying Mel-scale filter bank on DFT power spectrum. Mel-filter concentrates more on the significant part of the spectrum to get data values. Mel-filter bank is a series of triangular band pass filters similar to the human auditory system. The filter bank consists of overlapping filters. Each filter output is the sum of the energy of certain frequency bands. Higher sensitivity of the human ear to lower frequencies is modeled with this procedure. The energy within the frame is also an important feature to be obtained. Compute the logarithm of the square magnitude of the output of Mel-filter bank. Human response to signal level is logarithm. Humans are less sensitive to small changes in energy at high energy than small changes at low energy. Logarithm compresses dynamic range of values.

Mel-cepstrum is obtained by applying Discrete Cosine Transform (DCT) on the logarithm of the mel-spectrum. DCT is used to reduce the number of feature dimensions. It reduces spectral correlation between filter bank coefficients. Low dimensionality and

uncorrelated features are desirable for any statistical classifier. The cepstral coefficients do not capture the energy. So it is necessary to add energy feature. Thus twelve (12) Mel Frequency Cepstral Coefficients plus one (1) energy coefficient are extracted. These thirteen (13) features are generally known as base features. Remaining features are not considered for speech recognition. The first order and second order differentials are used to include contextual information into features. These differentials also remove drawbacks of HMM models and improves the accuracy of speech recognition. Thus 39-dimensional features (12 MFCC + 1 energy feature + 12 delta MFCC + 1 delta energy feature + 12 double-delta MFCC + 1 double-delta energy feature) are extracted to build HMM.

3.2. ACOUSTIC MODEL

Signal processing techniques are used to represent the speech signal as a sequence of real valued vectors such as MFCCs. In general, a sequence of continuous speech feature vectors is transformed into string of vector quantization (VQ) codebooks. The aim of VQ is that each feature vector is represented by the symbol K, where K is typically ranges from 128 to1024. This is done by automatically separating the feature vectors into K groups or clusters. K-means algorithm is used for clustering generally called as VQ codebook. However this VQ lose information regarding speech signal. This missed information might be useful for speech recognition. Instead of using K possible VQ codebooks, d-dimensional Gaussians (where d is the dimension of the feature vectors) are used for generating continuous output function. Hence HMM is used for this purpose.

Continuous vectors with a particular probability are required for the continuous HMM. For getting continuous output vectors, probability density function (pdf) such as Gaussian distribution is used to assign a particular probability for each continuous feature vector. A Gaussian is defined by a mean and the variance. In one dimensional, it is bell-shaped curve and it assumes that the data must fit a Gaussian. A particular point on this bell-shaped curve is the likelihood of the corresponding vector being generated. For continuous speech recognition system, combination or mixture of Gaussians is used for output function because it is easy to estimate their means and variances using forward-

backward algorithm. The diagonal covariance matrices can be used instead of full covariance matrices. This drastically reduces the number of parameters and the computational effort. As a consequence, a higher number of mixtures can be used, and correlation of the feature vectors can thus (to a certain degree) be modelled by the combination of diagonal mixtures **[50].** In a Gaussian mixture based systems, parameter tying may occur at different levels of means, variances, mixture coefficients, states, transitions etc. In tied mixtures model, all the output functions share same set of Gaussians having its own set of mixture coefficients. This is Gaussian mixture codebook. This is similar to VQ codebook. Gaussian Mixture Models (GMMs) are used for modeling correlations in feature vectors **[48]**. The output probability is represented by set of Gaussian distributions of normal density $\mathcal{N}$ with mean vector μ_{jk} of the states and covariance matrix U_{jk} with corresponding mixture weight W_{jk} is given below:

$$b_j(o_t) = \sum_k w_{ik}\, \mathcal{N}(o_t|\mu_{jk}, U_{jk}) \tag{3.4}$$

3.2.1. Hidden Markov Model

Hidden Markov Models (HMMs) are efficient way to model both spectral and time-varying characteristics of speech signals. Acoustic model based on the HMM is widely used in speech recognition to capture the acoustic characteristics that change over the time. HMM consists of states, state transition probabilities and state emission probabilities (output probability density function). HMM based acoustic models are used in Sphinx-3.

HMMs are classified into three types based on the output probability density function (pdf). First, if the output pdf is discrete, then it is classified into discrete HMMs. Second, if the output pdf is continuous, then the classification is called as continuous HMMs. In the third type, the discrete and continuous HMMs are combined. Phone-based subword unit is a popular choice to model the speech. The production of a particular phone is influenced by both preceding and following phones. Subword units are typically a three-state left-to-right topology. In HMM, all states related to subword units are hidden

and every state emits acoustic observations as a mixture of Gaussian distribution. This representation provides continuous output distributions; so it is continuous HMMs.

HMM is defined by triple parameters as $\lambda = (A, B, \pi)$, where A = state transition probabilities a_{ij} from state i to state j, B = output probability distribution of observing sequence O being in state j at time t, $b_j(o_t)$ and π = initial state distribution. This is represented in the **Figure 3.3.**

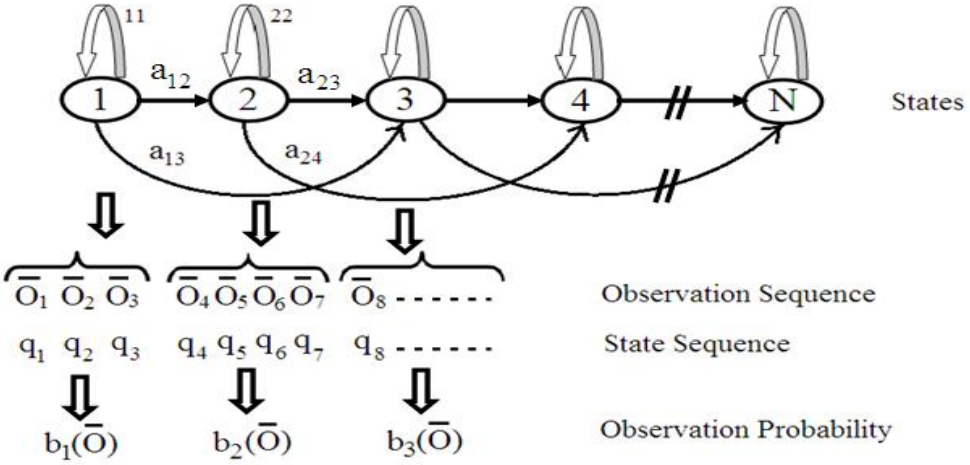

Figure 3.3: Hidden Markov Model

Once the HMM has been specified, there are three key problems that must be addressed as given below **[48]**.

1. **Evaluation problem**

   Given the observation sequence O and the model λ, how to compute efficiently P(O| λ), the probability of the observation sequence for the given model?

   The solution to this model enables us to evaluate the probability of different HMMs generating the same observation sequence. This is used to compare the probability of different models generating for the same observation sequence. If we have different HMMs for one word, then the recognized word is the one whose model has largest probability of generating the observation sequence. Forward algorithm is used for this purpose.

2. **Decoding problem**

   How to find the sequence of hidden states that most probably generated an observed sequence?

   This problem is related to find the optimal state sequence. Viterbi algorithm is used for finding best state sequence.

3. **Learning problem (Training)**

   How to adjust the model parameters to maximize the likelihood of the model λ for the given observation sequence O.

   The solution to this problem is to optimize the model parameters. Forward-Backward algorithm (Baum-Welch algorithm) is used to adjust the model parameters.

Detail explanation is given below for the problems mentioned above:

1. Evaluation problem

Consider state sequences of length T. There are N^T state sequences. Consider one state sequence $\mathbf{q} = (q_1, q_2, \ldots q_T)$, where q_1 is the initial state. The probability of the observation sequence O given the state sequence is

$$P(O|\lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda)$$

$$= b_{q1}(o_1).b_{q2}(o_2)\ldots b_{qT}(o_T) \quad (3.5)$$

The probability of such state sequence **q** is

$$P(q|\lambda) = \Pi q_1 \, a_{q1q2}a_{q2q3}\ldots.a_{qT\text{-}1qT} \quad (3.6)$$

The joint probability of O and **q** is the product of above two terms is given as

$$P(O, \mathbf{q} \mid \lambda) = P(O|\mathbf{q}, \lambda)P(\mathbf{q}|\lambda) \quad (3.7)$$

The probability of O is obtained by summing this joint probability over all possible state sequences **q** is

$$P(O|\lambda) = \sum_{\text{all } \mathbf{q}} P(O|\mathbf{q}, \lambda)P(\mathbf{q}|\lambda)$$

$$= \sum_{\mathbf{q1,q2,..qT}} \Pi q_1 \, b_{q1}(o_1), \, a_{q1q2} \, b_{q2}(o_2), \ldots a_{qT\text{-}1qT} \, b_{qT}(o_T) \quad (3.8)$$

From the above, initially the state q_1 with probability Πq_1, the observation o_1 is generated with probability $b_{q1}(o_1)$. This generating process continues in this manner until the last transition from state a_{qT}-1 to q_T with probability $a_{qT\text{-}1qT}$ generating with the symbol O_T with probability $b_{qT}(o_T)$. This is forward procedure.

Forward Procedure

Consider the forward variable defined as

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = i|\lambda) \quad (3.9)$$

probability of the partial observation sequence $o_1o_2...o_t$ and state i at time t for the given model λ. $\alpha_t(i)$ is solved inductively in three steps as follows:

1. Initialization

   $$\alpha_1(i) = \Pi_i b_i(o_1), \quad 1<i<N \quad (3.10)$$

2. Induction

   $$\alpha_1(i) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1}), \quad 1 \leq t \leq T\text{-}1, \; 1 \leq j \leq N \quad (3.11)$$

3. Termination

   $$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (3.12)$$

Step 1 initializes the forward probabilities as the joint probability of state i and initial observation o_1. Step 2 shows how the state j can be reached at time t+1 from the N possible states. Since $\alpha_t(i)$ is the probability of the joint event that $o_1o_2...o_t$ are stored, and the state at time t is i, the product $\alpha_t(i)a_{ij}$ is the probability of the joint event that $o_1o_2...o_t$ are observed and state j is reached at time t+1 through state i at time t. This product is summed over all possible states i, $1 \leq i \leq N$ at time t results in the probability of

j at time t+1 with all previous partial observations. If this is done and j is known, $\alpha_{t+1}(j)$ is obtained by observing o_{t+1}in state j. This is obtained by multiplying summed quantity by $b_j(o_{t+1})$. In step 3, $P(O|\lambda)$ is calculated as the sum of the terminal forward variables $\alpha_T(i)$. Thus it requires the order of N^2T calculations.

Backward procedure

Backward variable $\beta_t(i)$ defined as

$$\beta_t(i) = P(o_{t+1}o_{t+2}\ldots o_T|q_t = i, \lambda) \quad (3.13)$$

It means the probability of partial observation sequence from t+1 to the end for the given state i at time t for the given model λ. $\beta_t(i)$ is solved inductively in the following steps:

1. Initialization

   $$\beta_T(i) = 1, \quad 1 \le i \le N \quad (3.14)$$

2. Induction

   $$\beta_t(i) = [\sum_{i=1}^{N} a_{ij} b_j (o_{t+1})]\beta_{t+1}(j)]$$

   where t= T-1, T-2, ..,1; $1 \le i \le N$ (3.15)

In step 1, $\beta_T(i)$ is defined as 1 for all i. In step 2, in order to be in state i at time t and an observation sequence at t+1, consider all states j at t+1, transition from i to j (a_{ij}) and the observation o_{t+1} in state j. Also consider remaining partial observation sequences from state j ($\beta_{t+1}(j)$). This term is helpful in training.

3. Decoding problem

In this, optimal state sequence is to be determined for the given observation sequence. Select states q_t those are individually most likely at each time t. To do this, define posterior probability variable as

$$\gamma_t(i) = P(q_t=i|O, \lambda) \quad (3.16)$$

The above is the probability of being in state i at time t for the given observation O and model λ. $\gamma_t(i)$ is defined in terms of forward and backward variables as follows:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(j)} \quad (3.17)$$

where $\alpha_t(i)$ indicates for partial observation sequence $o_1 o_2 \ldots o_t$ and state i at time t, while $\beta_t(j)$ indicates for the remaining observation sequence $o_{t+1} o_{t+2} \ldots o_T$ for the given state q_t =i.

The most likely state q_t^* at time t using $\gamma_t(i)$ as follows

$$q_t^* = \text{argmax}[\gamma_t(i)] \quad (3.18)$$

The above maximizes the expected number of correct states (selecting most likely state for each t). But when the HMM has state transitions which have zero probability (a_{ij}=0 for some i and j), the optimal state sequence obtained cannot become a valid state sequence. To overcome this, optimal criteria should be modified by using Viterbi algorithm.

Viterbi algorithm

To find the single best state sequence **q** = $(q_1 q_2 \ldots q_T)$ for the observation sequence O = $(o_1 o_2 \ldots o_T)$, it is defined as:

$$\delta_t(i) = \max_{q1q2..qt-1} P(q_1 q_2 .. q_{t-1}, q_t = i, o_1 o_2 \ldots o_T | \lambda) \quad (3.19)$$

The above indicates that $\delta_t(i)$ is the best score (highest probability) along a single path at time t considering first t observations and ends in state i. $\delta_{t+1}(j)$ is calculated by induction as :

$$\delta_{t+1}(j) = [\max_t \delta_t(i) a_{ij}] b_j(o_{t+1}) \quad (3.20)$$

The following is the procedure for finding the best sequence as follows:

1. Initialization

$$\delta_1(i) = \Pi_i \, b_i(o_t) \qquad 1 \le i \le N \quad (3.21)$$

$$\psi_1(i) = 0 \quad (3.22)$$

2. Recursion

$$\delta_t(j) = \max_{1\le i\le N} [\delta_{t-1}(i)a_{ij}]b_j(o_t) \quad 2\le t\le T,\ 1\le j\le N \tag{3.23}$$

$$\psi_t(j) = \text{argmax}[\delta_{t-1}(i)a_{ij}] \quad t\le T,\ 1\le j\le N \tag{3.24}$$

3. Termination

$$P^* = \max_{1\le i\le N} [\delta_T(i)] \tag{3.25}$$

$$q_T^* = \text{argmax}_{1\le i\le N} [\delta_T(i)] \tag{3.26}$$

4. path(state sequence) backtracking

$$q_T^* = \psi_{t+1}(q_{t+1}^*) \tag{3.27}$$

Viterbi algorithm is similar in implementation except in backtracking. Viterbi algorithm is used to calculate maximization procedure over the previous states. But forward algorithm uses summation procedure.

4. Learning problem (Training)

The procedure for HMM re-estimation of HMM parameters, the probability of being in state i at time t and j at time t+1 for the given model and observation sequence is given as below:

$$\xi_t(i,j) = P(q_t=i,\ q_{t+1}=j|\ O, \lambda) \tag{3.28}$$

$\xi_t(i,j)$ is defined in terms of forward and backward variables as follows:

$$\xi_t(i,j) = \frac{P(q_t=i,\ q_{t+1}=j|\ O,\lambda)}{P(O|\lambda)} \tag{3.29}$$

$$= \frac{\alpha_t\ (i)a_{ij}\left(b_j\right)o_{t+1}\beta_{t+1}(j)}{p(O|\lambda)} \tag{3.30}$$

$$= \frac{\alpha_t(i)a_{ij}\,b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}\ b_j(o_{t+1})\beta_{t+1}(j)} \tag{3.31}$$

Relating $\gamma_t(i)$ to $\xi_t(i,j)$ by summing over j is given as follows:

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{3.32}$$

Let $\sum_{t=1}^{T-1}\gamma_t(i)$ = expected number of transitions from state i in O (3.33)

Let $\sum_{t=1}^{T-1} \xi_t(i,j)$= expected number of transitions from state i to state j in O (3.34)

Using above two formulas, re-estimation of parameters of HMM is given below:

$\widehat{\Pi_i}$ = expected number of times in state i at time t=1 = γ_1(i) (3.35)

$$\widehat{a_{ij}} = \frac{\text{expected number of transitions from state i to state j}}{\text{expected number of transitions from state i}} \quad (3.36)$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (3.37)$$

$$\widehat{b_j}(k) = \frac{\text{expected number of times in state j and observing symbol } v_k}{\text{expected number of times in state j}} \quad (3.38)$$

$$= \frac{\sum_{t=1, o_t = v_k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)} \quad (3.39)$$

The re-estimated model $\hat{\lambda} = (\hat{A}, \widehat{B}, \widehat{\Pi})$ is obtained. This re-estimation can be done for several times until the convergence achieved.

3.3. PRONUNCIATION DICTIONARY

The pronunciation dictionary (lexicon) consists of all possible pronunciations. A pronunciation dictionary defines how the words in the language are pronounced. The lexicon in Sphinx-3 defines the linear sequence of phones representing the pronunciation for each word in the vocabulary. The pronunciation dictionary acts as a link between the acoustic model and the language model. The pronunciation dictionary is necessary to concatenate the HMM phone models (subword units) to word models. Word pronunciations are given in terms of phones in the phoneset. In Sphinx-3, phoneset is the list of phones based on ARPAbet symbol **[51]**. ARPAbet is a phonetic transcription code developed by Advanced Research Projects Agency (ARPA) as a part of their Speech

Understanding Project (1971-1976). It represents each phoneme of general American English with a distinct sequence of American Standard Code for Information Interchange (ASCII) characters. In Sphinx-3, the pronunciation dictionary should have word with following phonetic transcription. The left hand side is the words and the right hand side is the set of phones representing the word. The delimiter for the words and the phone listing is space or tab and for the sequence of phones each of it separated by a space **[52]**. The structure of the CMU pronunciation dictionary (Carnegie Mellon University pronunciation dictionary) is shown below:

AAFIS	AA F AH S
AANDHRA	AA N D HH R AH
BUKING	B UW K AH NG
CHENNAI	CH EH N AY
CHITTOORKI	CH IH TT UW R K IY
ELA	EH L AH
ENNI	IH N IY
ETU	EH T UW
EVARINI	AH V AE R AH N IY
SIRPUR	S AH R P AH R
VUNDHI	V AH N D HH IY

3.4. LANGUAGE MODEL

The language model is used to estimate the probability of individual hypothesis sequence of words, P (W). The language mode is required to provide the posterior probability for the word sequence. In other words, it is a statistical model that fetches word and word sequence probability. This probability searches for linguistically the most probable sequence of words by assigning the probabilities compared with the unlikely sequence of words. The statistical model that gives P(W) for each word sequence W = $\{w_1,\ldots,w_N\}$ is called a language model and its parameters are estimated from a linguistic corpus **[53]**.

$P(W) = P(W) = P(W_1, W_2, \ldots, W_N)$

$= P(W_1)P(W_2|W_1)\ldots P(W_N| W_1, W_2, \ldots, W_{N-1})$

$= \prod_{n=1}^{N} P(W_N | W_1, W_2 \ldots . W_{N-1})$ (3.40)

Hence the probability of each word is only N-1 preceding words. It is complex to estimate the probability of all the possible word sequences. This model is called N-gram model.

Let $C(w_i, w_{i-1}, \ldots w_{i-N+1})$ is the number of occurrences of the word sequence $w_i, w_{i+1}, \ldots w_{i+N-1}$ in the training speech corpus. The probability of $P_N(w_i)$ for each word is shown below:

$$P_N(w_i|w_{i-1}, \ldots w_{i-N+1}) = \frac{C(W_i, W_{i-1}, \ldots W_{i-N+1})}{C(W_{i-1}, \ldots W_{i-N+1})} \quad (3.41)$$

If N = 1, it is called as Unigram Model and If N=2, it is Bigram Model and if N=3, it is Trigram Model **[53]**. Trigram language model is used in the present work.

Maximum likelihood estimates the word probabilities which are based on the counting frequency of occurrence of word sequences from the training set. And for the three words C (w_{n-2}, w_{n-1}, w_n)

$$P(w_n|w_{n-2}, w_{n-1}) = \frac{C(W_{n-2}, W_{n-1}, W_n)}{C(W_{n-2}, W_{n-1})} \quad (3.42)$$

Language model is used in Finite State Network (FSN) to integrate into the acoustic model. FSN is discussed in the next section.

3.5. Training

Every sentence in the training data consists of the waveform and its corresponding transcription. Lexicon is available which provides transcription of words in terms of set of subword units being trained. FSN is constructed using the pronunciation dictionary (lexicon) and the HMM based acoustic model to make the valid transitions between the context-dependent HMMs of subword units. Every subword unit is represented in three-

state left-to-right HMM. Word models are built by concatenating the subword HMMs **[48]**. Thus each word is represented as an FSN of subword units as shown in the **Figure 3.4**.

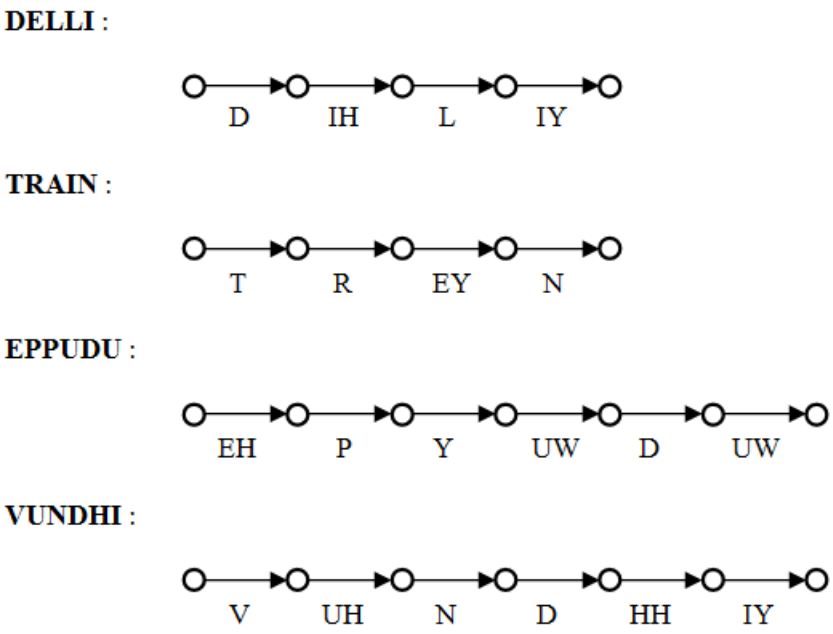

Figure 3.4: FSN for subword units

Sentence is represented as FSN of words by incorporating silence between them to generate extended HMM. Single state HMM is used for silence because silence is stationary and has no temporal structure to exploit. FSN for the sentence is shown in **Figure 3.5**. Thus the composite FSNs are created for all sentences in the training data to form the search network **[48]**. The training allows re-estimating the model parameters for each of the subunits in search network using Baum-Welch algorithm, which is discussed in **section 3.2.1.**

The following **Figure 3.5** represents composite FSN for sentence "DELLI TRAIN EPPUDU VUNDHI".

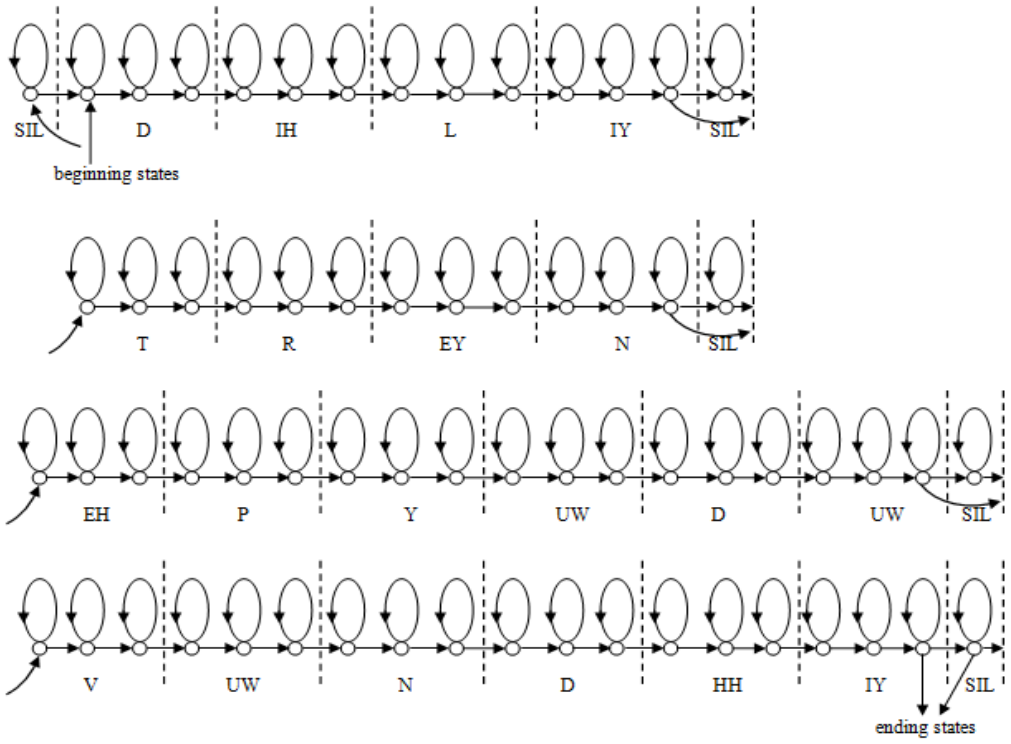

Figure 3.5: Composite FSN

3.6 DECODING

The decoder implements a search process over all the possible word sequences. The decoder utilizes composite FSN to find the best state sequence **q** for the unknown observation vector sequence. In order to determine the path probability, state space is expanded by aligning the observed feature vector with the state sequence. The path probability is obtained by computing transition and output probabilities for the given feature vector. The common method to determine the most probable state sequence is expanding the search network by computing the probabilities of all state sequences that possibly generate the observation sequence $o_1 o_2 \ldots o_T$. Computation complexity depends on the size of the lexicon and the observation sequence. Viterbi algorithm reduces the complexity by discarding some state sequences having lower probability. Thus Viterbi algorithm is used for finding the optimal state sequences for the observation sequence which is explained in the **Section 3.2.1.** Thus optimal path determines the recognized word hypothesis. This is shown in **Figure 3.6 [53]**. As this search process involves the

trigram language models, the probability of the path depends on the previous two words; more optimal paths are generated with word hypothesis. It is therefore necessary to continuously prune away some of the less likely hypotheses to maintain the search task in a manageable scale. The most important pruning method is 'beam pruning'. It is based on a simple idea of relating the partial scores of the active hypotheses to the score of the current best hypothesis at each time frame. The score of the best hypothesis acts as the reference score to which all the other scores are compared**.** Only the hypotheses whose scores are close enough to the reference score get propagated forward, the rest are pruned away **[54]**. The score threshold determining the allowed deviation is called the beam width. The result of the Viterbi search is a single recognition hypothesis as well as a word lattice that contains all the words recognized during decoding, their time segmentations and corresponding acoustic scores. For each word occurrence the word lattice contains several alternative end times, but usually only a single beginning time. Sphinx3.6 decoder is used for decoding purpose in the present work.

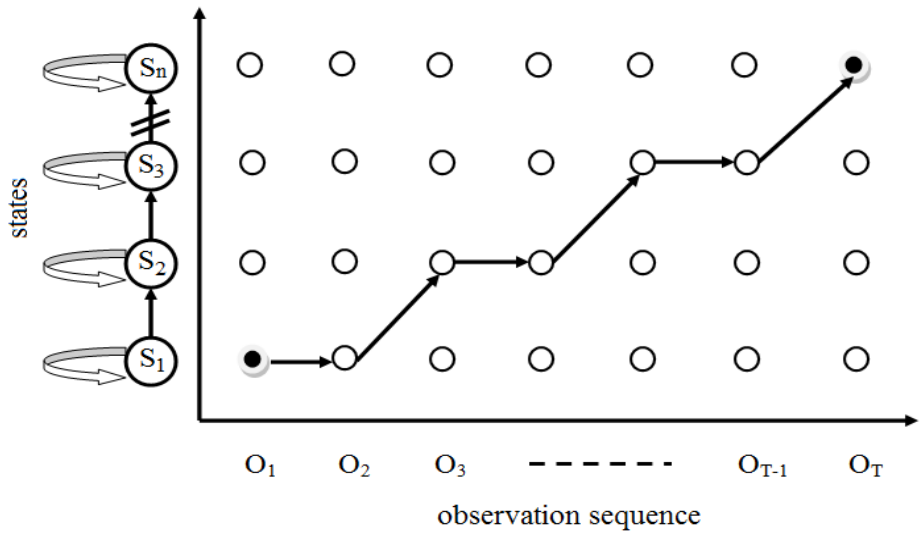

Figure 3.6: Optimal path

CHAPTER-4
SPEECH DATABASE

This chapter describes Telugu speech database collected for the present work. PSOLA method is used in the present research work to convert the Normal Speech Rate (NSR) into Slow Speech Rate (SSR) and Fast Speech Rate (FSR) **[55]**. The proposed modules used in the SIASR system are also discussed in this chapter.

Clean speech will give better results in the SIASR system. But it is hard to acquire the clean speech. Speech signal will be associated with noise if there is any background noise. This is not desirable for training and testing. Noise will be mixed even in the room environment while recording. It is necessary to listen carefully after recording is over for each speaker. Noise will be removed in noisy cases if possible. Noise removal recording devices are preferable for recording speech. If long pauses are found in the beginning or in the middle or at the end of acoustic signal, these additional lengths of pauses should be removed for better training purpose. If the acoustic signals are noisy in nature, it is better to record again with the same recording device in same environmental conditions as recorded earlier for better matching of acoustic feature vectors.

4.1. Collection of speech data

In the present work, a speech database related to railway inquiry system is created. In railway inquiry system, passengers generally ask information about the train timings, arrivals and departures of various trains, way to booking office, waiting room, parcel office and other things. It is a very good societal application which reduces cost and time of the customers. In this regard, for experimental purpose, 50 queries are prepared in Telugu and given to the selected speakers to record their voices. 20 speakers are selected in the ages between 20 and 25 years for recording. 10 male speakers and 10 female speakers are used for recording the queries related to railway inquiry system. All the speakers are in normal health condition and queries are recorded by them. Totally 1000 queries are recorded in normal speech rate by 20speakers (10 male speakers and 10

female speakers) in the room environment using head-mounted noise cancelling close talking microphone (recording device). Since it is head mounted, all speakers recorded with uniform distance from mouth to microphone.

As speech is relatively low bandwidth (100Hz – 8kHz), 8kHz (8000 samples per second) is sufficient for speech recognition. But 16 kHz (16000 samples per second) provide more accurate high frequency information. 16 bits per sample is used in speech recognition. Mono channel is used for desktop applications and stereo channel is used for telephone based applications. Hence utterances are recorded with a sampling rate of 16000 Hz and mono channel with 16 bits per sample for better training and testing.

But in reality, speakers may vary in their speech rate in asking the queries at the railway inquiry system. But this variation will be relatively slow or fast, but should not be too slow or fast. While inquiring there is a possibility of occurring different types of errors such as substitution, deletion and insertion errors **[15]**. These errors degrade recognition accuracy. Efforts have been taken to reduce such errors. Linearly time-compressed speech has a temporal and a segmental processing advantage over naturally produced fast speech **[26]**. Hence Normal Speech Rate (NSR) which is collected from the 20 speakers is converted into Slow Speech Rate (SSR) and Fast Speech Rate (FSR) using PSOLA method. This is discussed in the next **Section 4.1.1.** Thus in total 3000 queries are used for training and testing in the present work.

4.1.1. PSOLA Method

In our approach NSR, SSR and FSR are separately taken for the purpose of the training and decoding. Time Domain Pitch Synchronous Overlap and Add (TD-PSOLA) method is used to convert NSR into SSR and FSR for training and testing (decoding) **[55, 56]**. In this method, first the acoustic signal is segmented into small overlapping frames using Hanning Window. These windowed frames can then be recombined by placing their centres at the original epoch positions and adding the overlapping regions. Lengthening is achieved by duplicating frames. For a given set of frames, certain frames are duplicated, inserted back into sequence, and then overlap-add. The result is a longer speech signal. Shortening is achieved by removing frames. For a given set of frames,

certain frames are removed and the remaining ones are overlap-added. The result is a shorter speech signal **[56].** Thus TD-PSOLA is applied on all NSR sentences to convert into SSR and FSR.

The following **Figure 4.1** shows the duration of recorded sentence '***THIRUMALA EKSPRES EPPUDU VASTHUNDHI'*** in NSR**.**

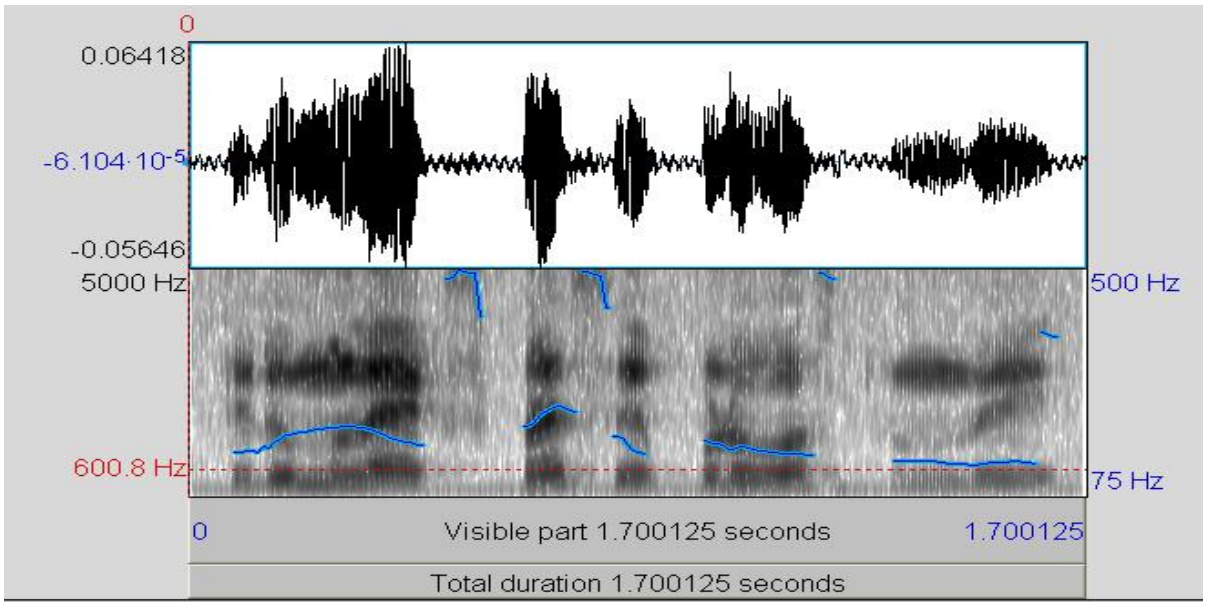

Figure 4.1: NSR sentence

TD-PSOLA is applied on the above sentence to stretch the sentence in order to increase the duration. This is shown in **Figure 4.2**. TD-PSOLA is applied on the above sentence to compress the sentence in order to decrease the duration. This is shown in **Figure 4.3**.

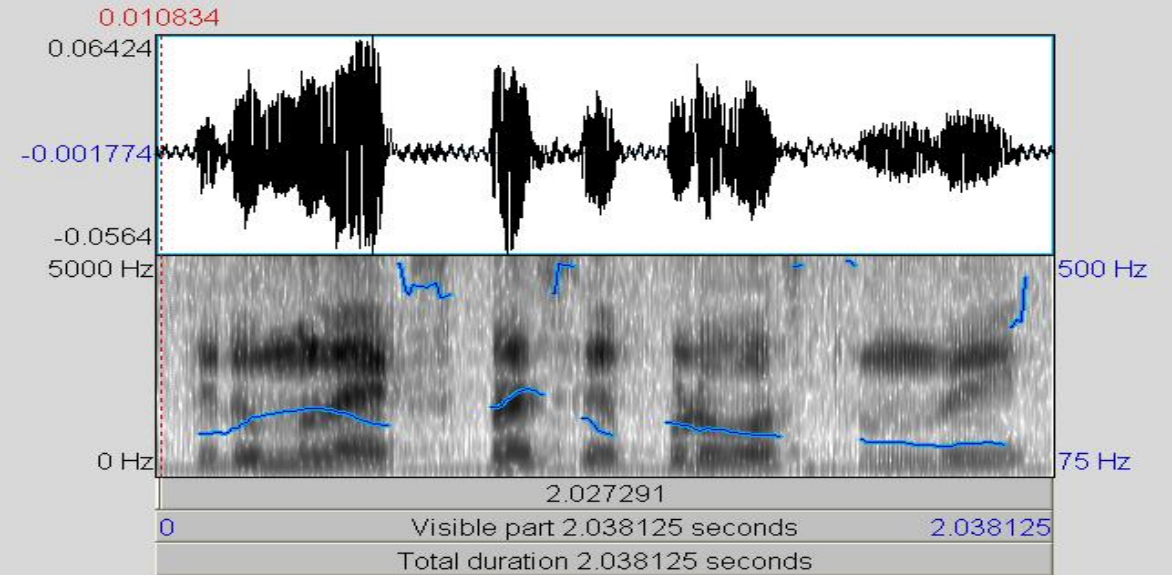

Figure 4.2: SSR sentence after TD-PSOLA

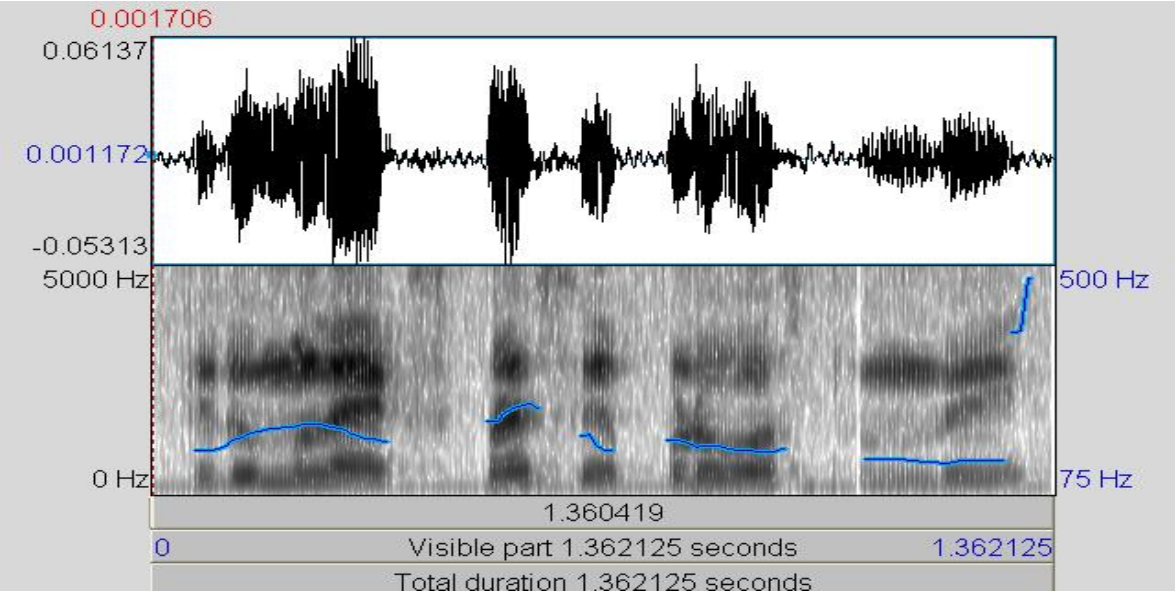

Figure 4.3: FSR sentence after TD-PSOLA

4.2. **Proposed modules**

The following **Figure 4.4** illustrates the proposed modules in green in the SIASR system.

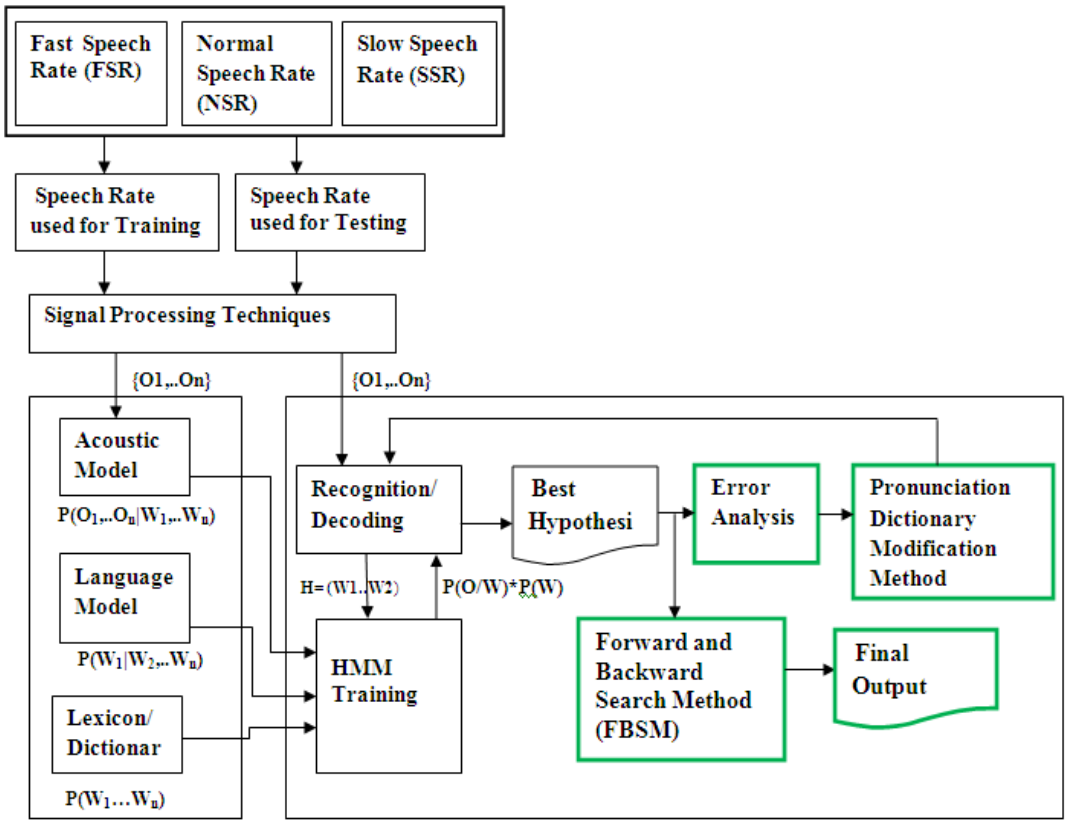

Figure 4.4: Proposed modules in SIASR system

Collection of speech database was explained in previous **Section 4.1.1**. Feature extraction, training and decoding algorithms used in speech recognition are already explained in **Chapter 3**. In the present work, training and testing are performed at different rates of speech. Sphinx-3 speech recognition system developed in Carnegie Mellon University which is used for training and testing in the present work. It is well suited for SIASR. Baum-Welch algorithm is used for training in Sphinx-3. Sphinx-3.6 decoder based on Viterbi algorithm and Beam search is used for decoding. It is a collection of open source tools and resources which are helpful for performing training and testing for the present research work **[57]**. Substitution, insertion and deletion errors

are observed from the decoder results of SIASR system. These results are tabulated in **Chapter 5**. These errors are analyzed to notice the type of errors that occur during testing. Thus error analysis is necessary for incorporating methods to improve the recognition accuracy.

Substitution errors are reduced through Pronunciation Dictionary Modification Method (PDMM) which is explained in detail in **Chapter 6**. Results obtained with modified dictionary using PDMM are tabulated in Chapter 6. Forward and Backward Search Method (FBSM) is applied on the decoder results which are obtained in Chapter 6. This method reduces one insertion or one deletion or one substitution error. FBSM is explained in **Chapter 7**. This method improves sentence recognition accuracy. These results are tabulated in **Chapter 7**.

CHAPTER-5
SPEECH RECOGNITION ACCURACY AT DIFFERENT RATES OF SPEECH

The speech database discussed in the **Section 4.1** of **Chapter 4** is used for the experiments performed in this chapter. Training and testing are performed at different rates of speech namely Normal Speech Rate (NSR), Slow Speech Rate (SSR) and Fast Speech Rate (FSR). The main aim of this chapter is to observe the performance of SIASR system when different speech rates are used for training and testing. It is also important to examine the type of errors that occur when speech rate, number of speakers and number of sentences vary in training and testing.

In the present work, separate training models are developed for NSR, SSR and FSR. Each of the training models is tested by NSR, SSR and FSR. Here six cases are involved to examine the type of errors when the speech rates and number of speakers vary in training and testing. When the number of speakers varies, automatically the number of sentences in training and testing will also vary. The six cases for training and testing are explained in **Section 5.1**. Thus eighteen experiments are performed and results are tabulated in this **Section 5.1**.

The most popular metrics for the performance evaluation of the SIASR system are Sentence Recognition Rate (SRR), Word Recognition Rate (WRR) and Word Error Rate (WER). These metrics are based on the alignment of the decoder output or hypothesis file (HYP) with the reference file (REF). Based on the alignment, errors are classified into substitutions, deletions and insertions. SRR, WRR and WER are calculated for different rates of speech which are tabulated in the **Section 5.1 [58].** Different types of errors are discussed in **Section 5.2**. Errors are analyzed in **Section 5.3**. This analysis is needed to improve the recognition accuracy by incorporating methods.

The following are the evaluation metrics used in the present work.

Let correct-sentences (corr-sents) = Number of sentences correctly recognized in hypothesis file

M = Total number of sentences (**sents**) in references

The Sentence Recognition Rate (SRR) is calculated as follows:

$$\text{SRR} = \frac{\text{correct} - \text{sentences}}{\text{M}} \tag{5.1}$$

Let 1000 sentences are in reference file and 800 sentences correctly recognized which are in hypothesis file, then SRR = 800/1000 = 0.8. Thus the percentage of correct sentences (corr-sents) in hypothesis file is 80%.

WRR and WER are calculated as follows:

Let N = Number of words in the reference file

C = Number of correctly recognized words (number of words in hypothesis file and reference file are correctly aligned)

S = Number of substitution errors

D = Number of deletion errors

I = Number of insertion errors

$$\text{Word Recognition Rate (WRR)} = \frac{\text{C}}{\text{N}} \tag{5.2}$$

$$\text{Word Error Rate (WER)} = \frac{(\text{S}+\text{D}+\text{I})}{\text{N}} \tag{5.3}$$

$$\text{Substitution Error Rate (SER)} = \frac{\text{S}}{\text{N}} \tag{5.4}$$

$$\text{Deletion Error Rate (DER)} = \frac{\text{D}}{\text{N}} \tag{5.5}$$

$$\text{Insertion Error Rate (IER)} = \frac{I}{N} \quad (5.6)$$

The following sample is taken for calculating WER.

REF: KERALA ekspres EKKADA NUNDI start avuthundhi

HYP: KRISHNAA ekspres EKKADIKI start avuthundhi

In the above sample, N=6, C=3, S=2, D=1, I=0; S+D+I = 3, Thus WER = 0.5 and WRR = 0.5. The percentage of error words in a sentence is 50%. The percentage of words correctly recognized is 50% **[58]**.

Thus SRR, WRR, SER, DER, IER and WER are calculated in all test cases involved in all training models. The above notations are used in all the tables in **Chapter 5, Chapter 6 and Chapter 7.**

5.1. EXPERIMENTAL RESULTS AT DIFFERENT RATES OF SPEECH IN TRAINING AND TESTING

5.1.1. NSR Training

5.1.1.1. NSR is tested with NSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 sentences in NSR are used for training and the same sentences are used for testing. 1000 NSR sentences consisting of 4740 words are used for training and testing. The SRR is 97.4%. 31 substitution errors, 11 deletion errors and 10 insertion errors are occurred in this case. The WER in this case is 1.1%. Thus WRR is 99.1%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 sentences in NSR are used for training and the remaining voice of 1 speaker (1 Female speaker) with 50 sentences in NSR is used for testing. 950 NSR sentences consisting of 4503 words are used for training. 50 NSR sentences consisting of 237 words are used for testing. 1 substitution error and 2 deletion errors are occurred in

this case. Insertion errors are 0 (zero) in this case. Thus WRR is 98.7% and WER is 1.3%. Thus 96% of SRR is noticed in this case.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 sentences in NSR are used for training and remaining voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 sentences in NSR are used for testing. 900 NSR sentences consisting of 4266 words are used for training. 100 NSR sentences consisting of 474 words are used for testing. 2 substitution errors and 5 deletion errors are occurred here. Insertion errors are 0 (zero) in this case. As the total errors are 7, WER is 1.5%. The WRR obtained in this case is 98.5%. Thus SRR is 97% in this case.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 sentences in NSR are used for training and remaining voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 NSR sentences are used for testing. 800 NSR sentences consisting of 3792 words are used for training. And 200 NSR sentences consisting of 948 words are used for testing. Here 4 substitution errors and 10 deletion errors are noticed. Here also insertion errors are 0. WER of 1.5% is observed here. Thus SRR is 96.5%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 sentences in NSR are used for training and remaining voices of five speakers (3 Male speakers and 2 Female speakers) with 250 sentences in NSR are used for testing. 750 NSR sentences consisting of 3555 words are used for training. 250 NSR sentences consisting of 1185 words are used for testing. The total errors are 22 in which 15 substitution errors and 7 deletion errors are noticed. Here 0 (zero) insertion errors are noticed. Thus WRR is 98.1% and WER is 1.9%. Thus SRR is 96% in this case.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 sentences in NSR are used for training and the remaining voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 sentences in NSR are used for testing. 2370 words in 500 NSR sentences are used for training and the remaining 2370 words are used for testing. The total errors are 93 in which 63 substitution errors, 14 deletion errors

and 16 insertion errors are noticed. The WER is 3.9% and WRR is 96.8%. Thus the SRR is 91%.

The above discussed results are tabulated in the following **Table 5.1.**

Table 5.1: Training with NSR and Testing with NSR

S.No	Training Data in NSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of Words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	974 97.4%	4698 99.1%	31 0.7%	11 0.2%	10 0.2%	52 1.1%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	48 96%	234 98.7%	1 0.4%	2 0.8%	0 0%	3 1.3%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	97 97%	467 98.5%	2 0.4%	5 1.1%	0 0%	7 1.5%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	193 96.5%	934 98.5%	4 0.4%	10 1.1%	0 0%	14 1.5%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	240 96%	1163 98.1%	15 1.3%	7 0.6%	0 0%	22 1.9%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	455 91%	2293 96.8%	63 2.7%	14 0.6%	16 0.7%	93 3.9%

5.1.1.2. SSR is tested with NSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 NSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 SSR sentences are used for testing. 1000 NSR sentences consisting of 4740 words are used for training. 1000 SSR sentences consisting of 4740 words are used for testing. Here WRR is 98.8%. 47 substitution errors, 8 deletion errors and 15 insertion errors are obtained in this case. The WER is 1.5%. Thus SRR is 97% in this case.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 NSR sentences are used for training and voice of 1 speaker (1 female speaker) with 50 SSR sentences is used for testing. 950 NSR sentences consisting of 4503 words are used for training. 50 SSR sentences consisting of 237 words are used for

testing. Here insertion errors are 0 (zero) but 8 substitution errors and 3 insertion errors are seen. The WRR is 96.6% and SRR is 90%. Thus WER is 4.6% in this case.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 NSR sentences used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 SSR sentences are used for testing. 900 NSR sentences consisting of 4266 words are used for training. 100 SSR sentences consisting of 474 words are used for testing. Here 9 substitution errors, 2 insertion errors and only 1 deletion error are obtained. Thus WER is 2.5%. The SRR is 93% and WRR is 97.9%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 NSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 SSR sentences are used for testing. 800 NSR sentences consisting of 3792 words are used for training. 200 SSR sentences consisting of 948 words are used for testing. The SRR obtained in this case is 95.5%. The WRR is 98.4%. 13 substitution errors, 4 insertion errors and 2 deletion errors are found here. The WER obtained is 2% in this case.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 NSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with SSR sentences are used for testing. 750 NSR sentences consisting of 3555 words are used for training. And 250 SSR sentences consisting of 1185 words are used for testing. The SRR is 94.8% and WRR is 97.8%. 17 substitutions, 7 deletions and 2 insertions are noticed in this case. Thus WER obtained here as 2.2%.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 NSR sentence are used for training and voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 SSR sentences are used for testing. 500 sentences of NSR consisting of 2370 words are used for training. 500 sentences of SSR consisting of 2370 words are used for testing. In this case, total errors are 105 in which 72 substitution errors, 17 deletion errors and 16 insertion errors are noticed. Thus WRR is 96.2% and WER is 4.4%. In this case, 88.6% of SRR is noticed.

The above discussed results are tabulated in the following **Table 5.2.**

Table 5.2: Training with NSR and Testing with SSR

S.No	Training Data in NSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of Sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	970 97%	4685 98.8%	47 1%	8 0.2%	15 0.3%	70 1.5%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	45 90%	229 96.6%	8 3.4%	0 0%	3 1.3%	11 4.6%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	93 93%	464 97.9%	9 1.9%	1 0.2%	2 0.4%	12 2.5%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	191 95.5%	933 98.4%	13 1.4%	2 0.2%	4 0.4%	19 2%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	237 94.8%	1161 97.8%	17 1.4%	7 0.6%	2 0.2%	26 2.2%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	443 88.6%	2281 96.2%	72 3.5%	17 0.7%	16 0.7%	105 4.4%

5.1.1.3. FSR is tested with NSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 NSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with FSR sentences are used for testing. 1000 NSR sentences consisting of 4740 words are used in training. 1000 FSR sentences consisting of 4740 words are used for testing. The total errors are 135 in which 96 substitutions, 28 deletions and 11 insertions are noticed. The WRR is 97.4% and WER is 2.8%. Here SRR noticed in this case is 92.2%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 NSR sentences are used for training and voice of 1 speaker (1 Female speaker) with 50 FSR sentences is used for testing. 950 NSR sentences consisting of 4503 words are used for training and 50 FSR sentences consisting of 237 words are used for testing. The total errors are 23 in which 14 are substitutions, 8 deletions and 1 insertion are noticed. Thus WRR is 90.7% and WER is 9.7%. Here SRR is 80%.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 NSR sentences are used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 FSR sentences are used for testing. 900 NSR sentences consisting of 4266 words are used for training. 100 FSR sentences consisting of 474 words are used for testing. 93.2% of WRR is obtained in this case. 25 substitution errors, 7 deletion errors and 1 insertion errors are obtained here. Thus WER is 7%. The SRR found in this case is 78%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 NSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 FSR sentences are used for testing. 800 NSR sentences consisting of 3792 words are used for training. 200 FSR sentences consisting of 948 words are used for testing. 86% of SRR is obtained in this case. The total errors are 49 in which 33 substitutions, 14 deletions and 2 insertion errors are noticed. Thus WER is 5.2% and WRR is 95%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 NSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 FSR sentences are used for testing. 750 NSR consisting of 3555 words are used for training. 250 sentences consisting of 1185 words are used for testing. The WRR is 94.8%. Here 40 substitutions, 22 deletions and 2 insertions are noticed. The WER is 5.4% in this case. Thus 85.6% of SRR is obtained in this case.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 NSR sentences are used for training and voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 FSR sentences are used for testing. 500 NSR sentences consisting of 2370 words are used for training. 500 FSR sentences consisting of 2370 words are used for testing. Here total errors are 185 in which 144 substitutions, 38 deletions and 3 insertion errors are noticed. Thus WER found in this case is 7.8%. The WRR is 92.3%. Here SRR found in this case is 79.2%.

The above discussed results are tabulated in the following **Table 5.3.**

Table 5.3: Training with NSR and Testing with FSR

S.No	Training Data in NSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of Sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	922 92.2%	4616 97.4%	96 2%	28 0.6%	11 0.2%	135 2.8%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	40 80%	215 90.7%	14 5.9%	8 3.4%	1 0.4%	23 9.7%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	78 78%	442 93.2%	25 5.3%	7 1.5%	1 0.2%	33 7%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	172 86%	901 95%	33 3.5%	14 1.5%	2 0.2%	49 5.2%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	214 85.6%	1123 94.8%	40 3.4%	22 1.9%	2 0.2%	64 5.4%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	396 79.2%	2188 92.3%	144 6.1%	38 1.6%	3 0.1%	185 7.8%

The following observations are noticed from the above results mentioned in **Tables 5.1**, **5.2** and **5.3**:

(i) S>D>I in 4 cases; D>S>I in two cases; S>I>D in one case when NSR is tested with NSR.

(ii) S>I>D in all cases when SSR is tested with NSR.

(iii) S>D>I in all cases when FSR is tested with NSR.

It is found that more errors occurred when FSR is tested with NSR. SER is more when FSR is tested with NSR. Thus WER is more in this case when compared with SSR and NSR.

From the **Figure 5.1**, it is noticed that SRR is more for NSR when compared with other speech rates such as SSR and FSR. From the **Figure 5.2**, it is found that WRR is more when NSR is tested with NSR.

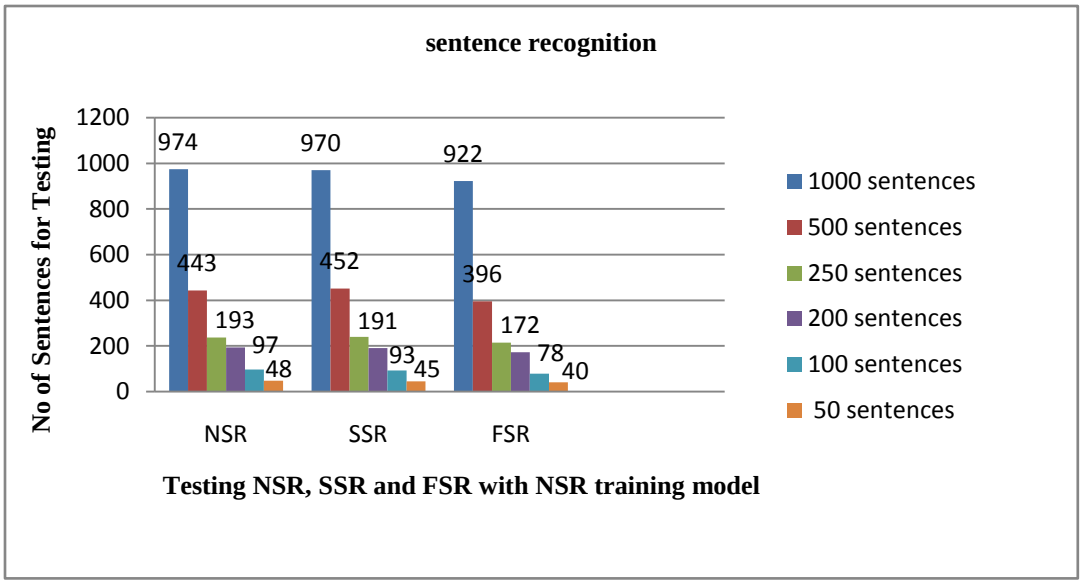

Figure 5.1: Sentence recognition for NSR training model

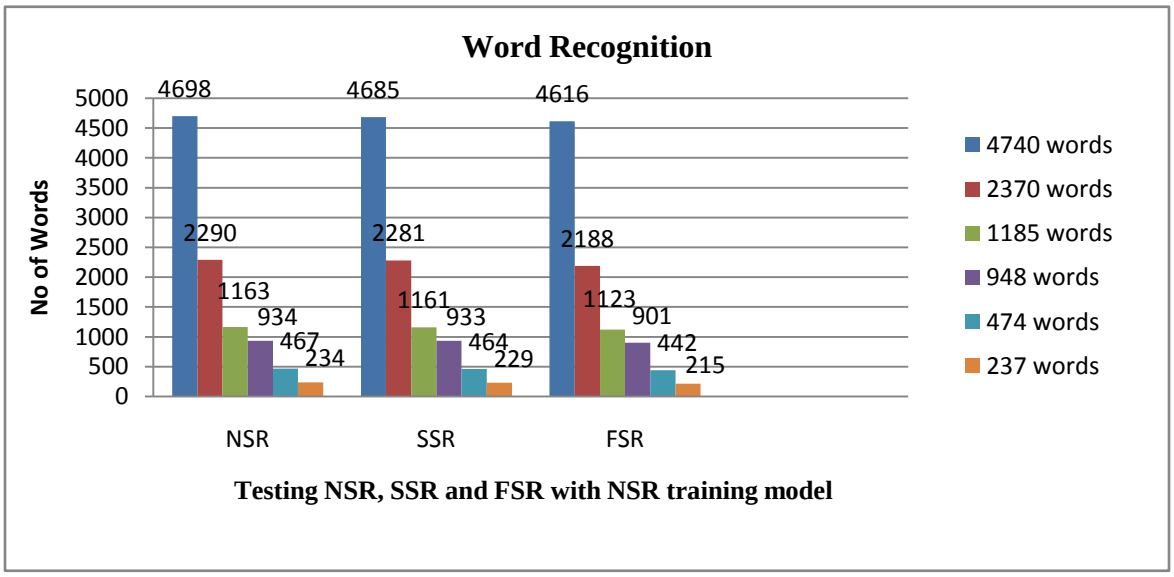

Figure 5.2: Word recognition for NSR training model

5.1.2. SSR Training

5.1.2.1. NSR is tested on SSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 SSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 NSR sentences are used for testing. 1000 SSR sentences consisting of 4740 words are used for training. 1000 NSR sentences consisting of 4740 words are used for testing. Here total errors are 92 in which 62 substitutions, 20 deletions and 10 insertions are noticed. Thus WER is 1.9% and WRR is 98.3%. Thus SRR is observed as 95.4%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 SSR sentences are used for training and voice of 1 speaker (1 female speaker) with 50 NSR sentences is used for testing. 950 SSR sentences consisting of 4503 words are used for training. 50 NSR consisting of 237 words are used for testing. 10 substitutions, 3 deletions and 5 insertion errors are observed here. Thus WER is 7.6% and WRR is 94.5%. Thus SRR is 88% in this case.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 SSR sentences used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 NSR sentences are used for testing. 900 SSR sentences consisting of 4266 words are used for training. 100 NSR sentences consisting of 474 words are used for testing. The total errors are 27 in which 19 substitutions, 3 deletions and 5 insertions are noticed. Thus WER is 5.7% and WRR is 95.4%. Thus SRR is 82% in this case.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 SSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 NSR sentences are used for testing. 800 SSR sentences consisting of 3792 words are used for training. 200 NSR sentences consisting of 948 words are used for testing. The total errors are 31 in which 20 substitutions, 4 deletions and 7 insertions are noticed. Thus WER is 3.3% and WRR is 97.5%. Here SRR is noticed as 90.5%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 SSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 NSR sentences are used for testing. 750 SSR sentences consisting of 3555 words are used for training. 250 NSR sentences consisting of 1185 words are used for testing. The total errors are 43 in which 24 substitutions, 10 deletion errors and 9 insertion errors are noticed. Thus WRR is 97.1% and WER is 3.6%. Thus SRR is noticed as 90.8%.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 SSR sentences are used for training and voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 NSR sentences are used for testing. 500 SSR sentences consisting of 2370 words used for training and 500 NSR sentences are used for testing. 87 substitutions, 26 deletions and 20 insertions are obtained in this case. The WER is 5.6% and WRR is 95.2%. The SRR is 86.2% in this case.

The above discussed results are tabulated in the following **Table 5.4.**

Table 5.4: Training with SSR and Testing with NSR

S.No	**Training Data in SSR**			**Test Data in NSR**			**Recognition Accuracy**		**Errors Obtained**			**Total errors & WER (%)**
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	954 95.4%	4658 98.3%	62 1.3%	20 0.4%	10 0.2%	92 1.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	44 88%	224 94.5%	10 4.2%	3 1.3%	5 2.1%	18 7.6%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	82 82%	452 95.4%	19 4%	3 0.6%	5 1.1%	27 5.7%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	181 90.5%	924 97.5%	20 2.1%	4 0.4%	7 0.7%	31 3.3%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	227 90.8%	1151 97.1%	24 2%	10 0.8%	9 0.8%	43 3.6%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	431 86.2%	2257 95.2%	87 3.7%	26 1.1%	20 0.8%	133 5.6%

5.1.2.2. SSR is tested with SSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 SSR sentences are used for training and same sentences are used for testing. 1000 SSR sentences consisting of 4740 words are used for training and testing. The total errors are 59 in which 38 substitutions, 8 deletions and 13 insertions are seen in this case. Thus WER is 1.2% and WRR is 99%. The SRR obtained in this case is 97.1%.

In the second case, voices of 19 speakers (10 Male speakers + 9 Female speakers) with 950 SSR sentences are used for training and voice of 1 speaker (1 Female speaker) with 50 SSR sentences is used for testing. 950 SSR sentences consisting of 4666 words are used for training. 50 SSR sentences consisting of 237 words are used for testing. Here 1 deletion and 1 insertion are obtained. But 8 substitutions are noticed. Thus WER is 4.2% and WRR is 96.2%. Thus SRR is 90%.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 SSR sentences used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) 100 SSR sentences are used for testing. 900 SSR sentences consisting of 3792 words are used for training. 100 SSR sentences consisting of 474 words are used for testing. In this case deletions are 0 (zero). But 10 substitutions and 2 insertions are noticed. Thus WER is 2.5% and WRR is 97.9%. In this case SRR is observed as 92%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 SSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 SSR sentences are used for testing. 800 SSR sentences consisting of 3792 words are used for training. 200 SSR sentences consisting of 948 words are used for testing. Here the SRR is 94.5%. Here also deletions are 0 (zero). The WER is 2.1% due to 15 substitutions and 5 insertions. Thus WRR is 98.4%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 SSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 SSR sentences are used for testing. 750 SSR sentences consisting of 3555 words are used for training. 250 SSR sentences consisting of 1185 words are used for testing. The total errors are 27 in which 24 substitutions and 3

insertions are noticed. In this case also deletions are 0 (zero). Thus WER is 2.3% and WRR is 98%. The SRR is 94% in this case.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 SSR sentences are used for training and voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 SSR sentences are used for testing. 500 SSR sentences consisting of 2370 words used for training and remaining 500 SSR sentences consisting of 2370 words are used for testing. Here 74 substitutions, 27 deletions and 19 insertions are noticed. The WER is 5.1% and WRR is 95.7%. Thus 86.6% of SRR is noticed in this case.

The above discussed results are tabulated in the following **Table 5.5.**

Table 5.5: Training with SSR and Testing with SSR

S.No	**Training Data in SSR**			**Test Data in SSR**			**Recognition Accuracy**		**Errors Obtained**			**Total errors & WER (%)**
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	971 97.1%	4694 99%	38 0.8%	8 0.2%	13 0.3%	59 1.2%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	45 90%	228 96.2%	8 3.4%	1 0.4%	1 0.4%	10 4.2%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	92 92%	464 97.9%	10 2.1%	0 0%	2 0.4%	12 2.5%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	189 94.5%	933 98.4%	15 1.6%	0 0%	3 0.5%	20 2.1%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	235 94%	1161 98%	24 2%	0 0%	3 0.3%	27 2.3%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	433 86.6%	2269 95.7%	74 3.1%	27 1.1%	19 0.8%	120 5.1%

5.1.2.3. FSR is tested on SSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 SSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 FSR sentences are used for testing. 1000 SSR sentences consisting of 4740 words are used for training. 1000 FSR sentences

consisting of 4740 words are used for testing. Here 139 substitutions, 41 deletions and 5 insertions are noticed. Thus WER is 3.9% and WRR is 96.2%. Here SRR is observed as 88.5%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 SSR sentences are used for training and voice of 1 speaker (1 Female speaker) with 50 FSR sentences is used for testing. 950 SSR sentences consisting of 4503 words are used for training. 50 FSR sentences consisting of 237 words are used for testing. The SRR in this case is 74%. The total errors are 28 in which 19 substitutions, 7 deletions and 2 insertions are noticed. Thus WER is 11.8% and WRR is 89%.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 SSR sentences used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with100 FSR sentences are used for testing. 900 SSR sentences consisting of 4266 words are used for training. 100 FSR sentences consisting of 474 words are used for testing. 34 substitutions, 10 deletions and 2 insertions are noticed. Thus WER is 9.7% and WRR is 90%. The SRR is observed as 72%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 SSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 FSR sentences are used for testing. 800 SSR sentences consisting of 3792 words are used for training. 200 FSR sentences consisting of 948 words are used for testing. 52 substitutions, 23 deletions and only 1insertion error are noticed in this case. Thus WER is 8% and WRR is 92.1%. In this case, SRR is noticed as 80%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 SSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 FSR sentences are used for testing. 750 SSR sentences consisting of 3555 words are used for training. 250 FSR sentences consisting of 1185 words are used for testing. 64 substitutions, 21 deletions and 1 insertion errors are noticed. Thus WER is 7.3% and WRR is 92.8%. The SRR obtained in this case is 78.8%.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 SSR sentences are used for training and voices of 10 speakers (5 Male speakers and 4 Female speakers) with 500 FSR sentences are used for testing. 500 SSR sentences consisting of 2370 words are used for training. 500 FSR sentences consisting of 2370 words are used for testing. The total errors are 270 in which 194 substitutions, 70 deletions and 6 insertions are noticed. Thus WER is 11.4% and WRR is 88.9%. The SRR seen in this case is 71.4%.

The above discussed results are tabulated in the following **Table 5.6.**

Table 5.6: Training with SSR and Testing with FSR

S.No	**Training Data in SSR**			**Test Data in FSR**			**Recognition Accuracy**		**Errors Obtained**			**Total errors & WER (%)**
	No. of Speakers (M&F)	No. of Sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	885 88.5%	4560 96.2%	139 2.9%	41 0.9%	5 0.1%	185 3.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	37 74%	211 89%	19 8%	7 3%	2 8%	28 11.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	72 72%	430 90.7%	34 7.2%	10 2.1%	2 0.4%	46 9.7%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	160 80%	873 92.1%	52 5.5%	23 2.4%	1 0.1%	76 8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	197 78.8%	1100 92.8%	64 5.4%	21 1.8%	1 0.1%	86 7.3%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	357 71.4%	2106 88.9%	194 8.2%	70 3%	6 0.3%	270 11.4%

The following observations are noticed from the above results mentioned in **Tables 5.4**, **5.5** and **5.6**:

(i) S>D>I in two cases and S>I>D in 4 cases when NSR is tested with SSR

(ii) S>D>I in one case and S>I>D in 5 cases when SSR is tested with SSR

(iii) S>D>I in all cases when FSR is tested with SSR

It has been observed that the word recognition accuracy is more when SSR is tested on SSR. Low recognition accuracy is observed when FSR is tested on the SSR due to mismatches between training and testing. WER is more when FSR is tested with SSR.

Thus SRR and WRR is more when SSR is tested with SSR. This is shown in **Figure 5.3** and **Figure 5.4.**

The following **Figure 5.3** shows sentence recognition accuracy for different speech rates.

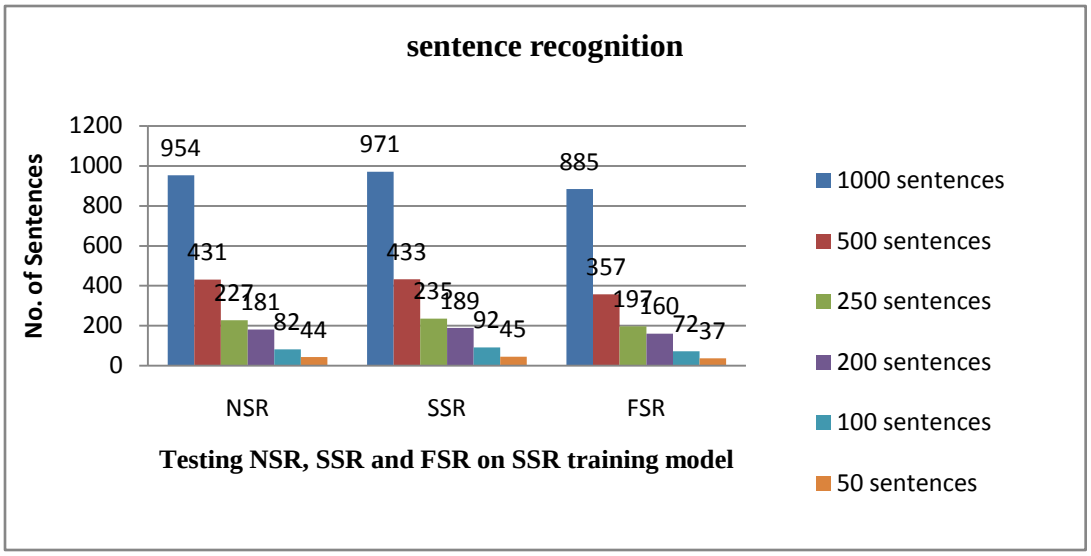

Figure 5.3: Sentence recognition for SSR training model

The following **Figure 5.4** shows word recognition accuracy for different speech rates.

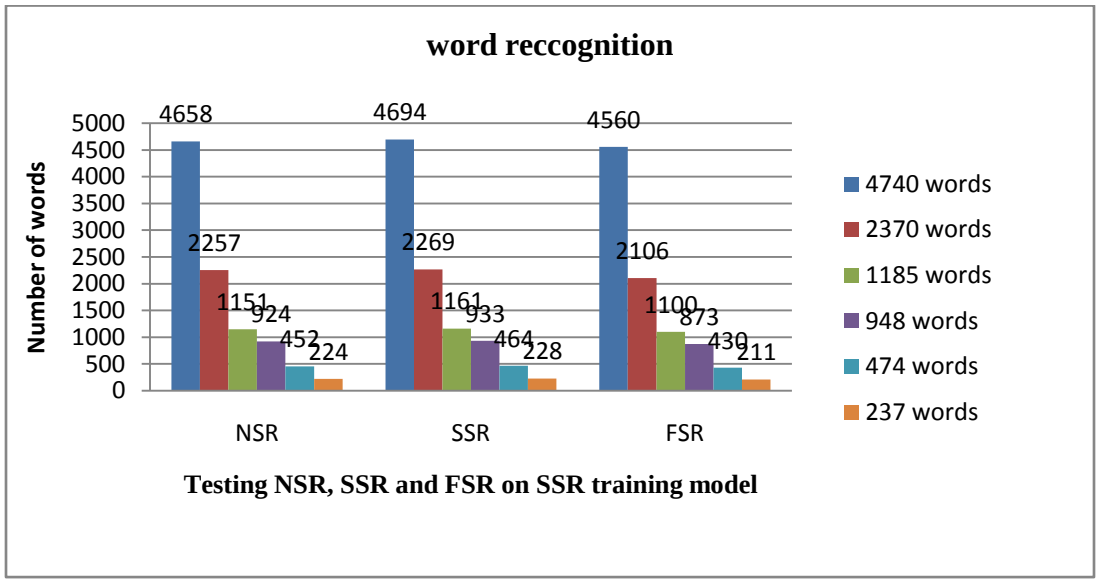

Figure 5.4: Word recognition for SSR training model

5.1.3. FSR Training

5.1.3.1. NSR is tested on FSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 FSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 NSR sentences are used for testing. 4740 words in 1000 FSR sentences are used in training. 4740 words in NSR are used for testing. 44 substitutions, 5 deletions and 11 insertions are obtained. Thus WER is 1.3%. The WRR in this case is 99%. Thus SRR is observed as 96.3%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with FSR sentences are used for training and voice of 1 speaker (1 Female speaker) with 50 NSR sentences is used for testing. 4503 words in 950 FSR sentences are used in training. 237 words in 50 NSR sentences are used in testing. Here deletions are 0 (zero). The total errors are 10 in which 7 substitutions and 3 insertions are noticed. The WER is 4.2% and WRR is 97%. Here SRR is seen as 90%.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 FSR sentences are used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 NSR sentences are used for testing. 900 FSR sentences consisting of 4266 words are used in training. 100 NSR sentences consisting of 474 words are used in testing. The total errors are 19 in which 12 substitutions, 2 deletions and 3 insertions are noticed. Thus SRR is 88% and WRR is 97%. The WER in this case is 4%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 FSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 NSR sentences are used for testing. 3792 words in 800 FSR sentences are used in training and 948 words in 200 NSR sentences are used in testing. 94.5% of SRR is observed in this case. The WRR is 98.7%. The total errors are 20 in which 15 substitutions and 5 insertions are noticed. Deletions are 0 (zero) in this case. The WER observed here is 2.1%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 FSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 NSR sentences are used for testing. 3555 words in 750 FSR sentences are used in training and 1185 words in 250 NSR sentences are used in testing. The total errors are 27 in which 22 substitutions, only 1 deletion and 4 insertions are noticed. The WER here is 2.3%. The WRR is 98.1% in this case. Thus SRR is seen as 93.4%.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 FSR sentences are used for training and voice of 10 speakers (6 Male speakers and 4 Female speakers) with 500 NSR sentences are used for testing. 2370 words in 500 FSR sentences are used in training and 2370 words in 500 NSR sentences are used in testing. 88.2% of SRR is observed in this case. The WRR in this case is 97%. The total errors are 94 in which 57 substitutions, 13 deletions and 24 insertions are noticed. Thus WER observed here is 4%.

The above discussed results are tabulated in the following **Table 5.7.**

Table 5.7: Training with FSR and Testing with NSR

S.No	**Training Data in FSR**			**Test Data in NSR**			**Recognition Accuracy**		**Errors Obtained**			**Total errors & WER (%)**
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	963 96.3%	4691 99%	44 0.9%	5 0.1%	11 0.2%	60 1.3%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	45 90%	230 97%	7 3%	0 0%	3 1.3%	10 4.2%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	88 88%	460 97%	12 2.5%	2 2%	5 5%	19 4%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	189 94.5%	933 98.4%	15 1.6%	0 0%	5 5%	20 2.1%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	234 93.6%	1162 98.1%	22 1.9%	1 0.1%	4 0.3%	27 2.3%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	441 88.2%	2300 97%	57 2.4%	13 0.5%	24 1%	94 4%

5.1.3.2. SSR is tested with FSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 FSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 SSR sentences are used for testing. 1000 FSR sentences consisting of 4740 words are used in training. 1000 SSR sentences consisting of 4740 words are used in testing. 39 substitutions, 12 deletions and 5 insertions are noticed in this case. Thus WER observed in this case is 1.2%. The SRR found in this case is 96.6%. The WRR is 98.9%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 FSR sentences are used for training and voice of 1 speaker (1 Female speaker) with 50 SSR sentences is used for testing. 950 FSR sentences consisting of 4503 words are used in training. 50 SSR sentences consisting of 237 words are used for testing. 88% of SRR is observed in this case. The WRR in this case is 95.8%. 8 substitutions, 2 deletions and 1 insertion errors are noticed in this case. Thus WER is 4.6%.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 FSR sentences used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 SSR sentences are used for testing. 900 FSR sentences consisting of 4266 words are used in training. 100 SSR sentences consisting of 474 words are used in testing. 84% of SRR is observed here. The WRR is 96%. 14 substitutions, 5 deletions and 3 insertions are noticed in this case. Thus WER is noticed to be 4.6%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 FSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 SSR sentences are used for testing. 800 FSR sentences consisting of 3792 words are used in training. 200 SSR sentences consisting of 948 words are used in testing. The SRR in this case is 91.5%. Here WRR observed as 97.7%. 15 substitutions, 7 deletions and 1 insertion errors are seen in this case. Thus WER is 2.4%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 FSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 SSR sentences are used for testing. 750 FSR sentences

consisting of 3555 words are used in training. 250 SSR sentences consisting of 1185 words are used in testing. The SRR is 91.2% and WRR is 97.3%. 24 substitutions, 8 deletions and 1 insertion errors are noticed in this case. Thus WER is noticed in this case is 2.8%.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 FSR sentences are used for training and voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 SSR sentences are used for testing. 500 FSR sentences consisting of 2370 words are used for training. 500 SSR sentences consisting of 2370 words are used for testing. The SRR is 87.6% and WRR is 96.2%. 73 substitutions, 17 deletions and 13 insertions are observed in this case. Thus WER observed here is 4.3%.

The above discussed results are tabulated in the following **Table 5.8.**

Table 5.8: Training with FSR and Testing with SSR

S.No	**Training Data in FSR**			**Test Data in SSR**			**Recognition Accuracy**		**Errors Obtained**			**Total errors & WER (%)**
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	966 96.6%	4689 98.9%	39 0.8%	12 0.3%	5 0.1%	56 1.2%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	44 88%	227 95.8%	8 3.4%	2 0.8%	1 0.4%	11 4.6%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	84 84%	455 96%	14 3%	5 1.1%	3 0.6%	22 4.6%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	183 91.5%	926 97.7%	15 1.6%	7 0.7%	1 0.1%	23 2.4%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	228 91.2%	1153 97.3%	24 2%	8 0.7%	1 0.1%	33 2.8%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	438 87.6%	2280 96.2%	73 3.1%	17 0.7%	13 0.5%	103 4.3%

5.1.3.3. FSR is tested on FSR training model

In the first case, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 FSR sentences are used for training and same sentences are used for testing. 1000 FSR sentences consisting of 4740 words are used for training and testing. The SRR

is 97.8%. The WRR is 99.2% in this case. 34 substitutions, 6 deletions and 10 insertions are noticed in this case. Thus WER in this case is 1.1%.

In the second case, voices of 19 speakers (10 Male speakers and 9 Female speakers) with 950 FSR sentences are used for training and voice of 1 speaker (1 Female speaker) with 50 FSR sentences is used for testing. 950 FSR sentences consisting of 4503 words are used for training. 50 FSR sentences consisting of 237 words are used for testing. Here the SRR is 92% and WRR is 97.5%. Only 1 insertion and 1 deletion errors occurred here. But 5 substitution errors are noticed. Thus observed WER is 3%.

In the third case, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 FSR sentences used for training and voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 FSR sentences are used for testing. 900 FSR sentences consisting of 4266 words are used in training. 100 FSR sentences consisting of 474 words are used in testing. 9 substitutions, 1 deletion and 3 insertions are noticed here. Thus SRR is seen as 91% and WRR is 97.9%. Thus WER is noticed in this case is 2.7%.

In the fourth case, voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 FSR sentences are used for training and voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 FSR sentences are used for testing. 800 FSR sentences consisting of 3792 words are used for training. 200 FSR sentences consisting of 948 words are used for testing. 12 substitutions, 1 deletion and 1 insertion errors are noticed in this case. Thus SRR is 95.5% and WRR is 98.6%. In this case WER is 1.7%.

In the fifth case, voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 FSR sentences are used for training and voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 SSR sentences are used for testing. 750 FSR sentences consisting of 3555 words are used for training. 250 FSR sentences consisting of 1185 words are used for testing. The SRR is 94.8% and WRR is 98.3%. The total errors are 25 in which 18 substitutions, 2 deletions and 5 insertions are noticed here. Thus WER in this case is seen as 2.1%.

In the sixth case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 FSR sentences are used for training and voices of 10 speakers (6 Male speakers

and 4 Female speakers) with 500 sentences FSR sentences are used for testing. 500 FSR sentences consisting of 2370 words are used for training and remaining 500 sentences consisting of 2370 words are used for testing. In this case SRR is 90.6% and WRR is 97.4%. The total errors are 83 in which 51 substitutions, 10 deletions and 22 insertions are noticed here. Thus WER observed as 3.5%.

The above discussed results are tabulated in the following **Table 5.9.**

Table 5.9: Training with FSR and Testing with FSR

S.No	**Training Data in FSR**			**Test Data in FSR**			**Recognition Accuracy**		**Errors Obtained**			**Total errors & WER (%)**
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	978 97.8%	4700 99.2%	34 0.7%	6 0.1%	10 0.2%	50 1.1%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	46 92%	231 97.5%	5 2.1%	1 0.4%	1 0.4%	9 3.7%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	91 91%	464 97.9%	9 1.9%	1 0.2%	3 0.6%	13 2.7%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	191 95.5%	935 98.6%	12 1.3%	1 0.1%	3 0.3%	16 1.7%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	237 94.8%	1165 98.3%	18 1.5%	2 0.2%	5 0.4%	25 2.1%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	453 90.6%	2309 97.4%	51 2.2%	10 0.4%	22 0.9%	83 3.5%

The following observations are noticed from the above results mentioned in **Tables 5.7**, **5.8** and **5.9**:

(i) S>I>D in all cases when NSR is tested with FSR

(ii) S>D>I in all cases when SSR is tested with FSR and

(iii) S>I>D in all cases when FSR is tested with FSR.

Thus WER is more SSR is tested with FSR. It has been observed that the WRR is more when FSR is tested with FSR. Low recognition accuracy is observed when SSR is tested with FSR due to mismatches between training and testing. This is shown in the following **Figure 5.5** and **Figure 5.6**

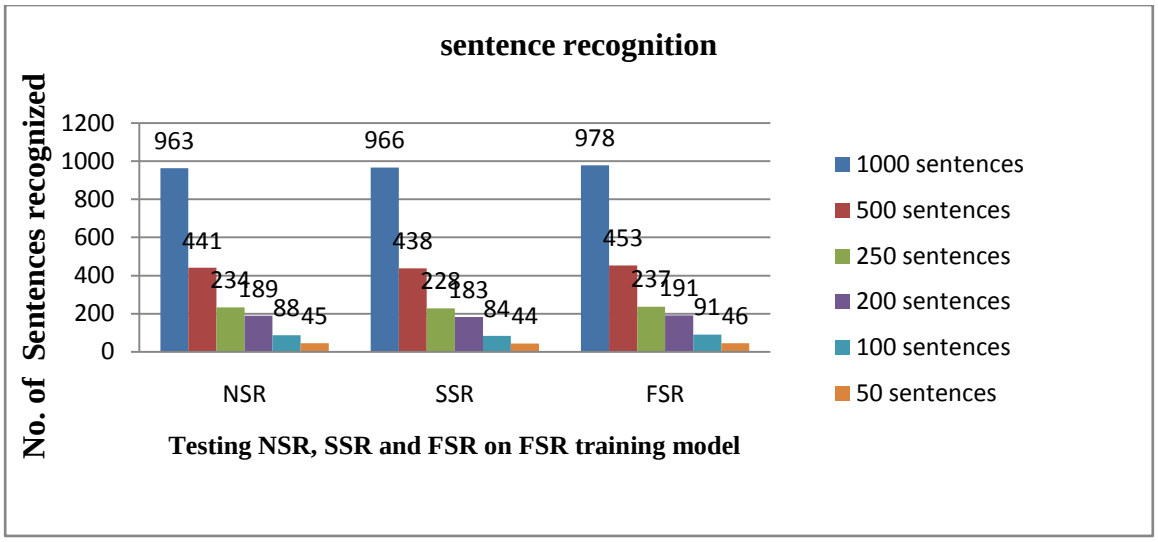

Figure 5.5: Sentence recognition for FSR training model

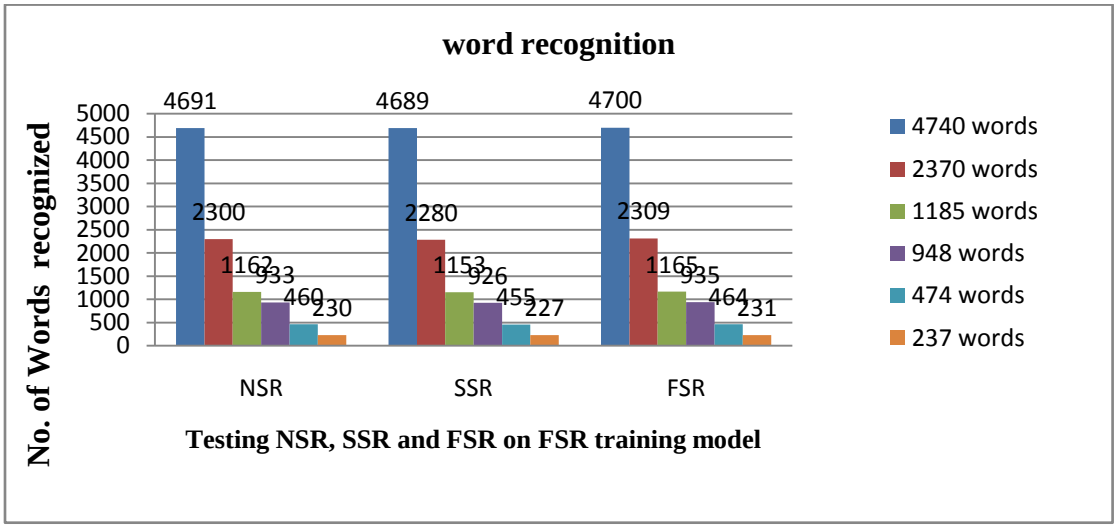

Figure 5.6: Word recognition for FSR training model

5.2. TYPES OF ERRORS

Errors in the SIASR system output can be classified into three types: substitution errors, insertion errors and deletion errors **[58]**.

5.2.1. Substitution errors

Substitution errors increase the error rate of SIASR. The misrecognition of a word in the place of original word causes substitution error which affects the accuracy of SIASR system. Phonetically similar words are confused by the SIASR system. It is interesting to note that both words in the confused pair are valid according to the vocabulary. The present work focuses on minimizing the confusion pairs so that substitution errors will be reduced. For this purpose, dictionary is modified during testing. Pronunciation Dictionary Modification Method (PDMM) is used to modify the pronunciation dictionary. This modified dictionary reduces substitution errors. PDMM is discussed in next **Chapter 6**. The following are some of the errors that occurred during experiments performed in present work.

(i). REF: yashwanthpur ekspres KAACHIGOODAKU eppudu vasthundhi
HYP: yashwanthpur ekspres KAAKINAADAKU eppudu vasthundhi

(ii) REF: eypi ekspres taimings EMITI
HYP: eypi ekspres taimings PEYREMITI

(iii) REF: gowthami ekspres kaakinaadaku eppudu VELUTHUNDHI
HYP: gowthami ekspres kaakinaadaku eppudu CHEYRUTHUNDHI

In the first sample, KAAKINAADAKU is substituted in the place of KAACHIGOODAKU. The phonemes present at the beginning and end are similar in both words. The above two words are confused by the SIASR system, thus the substitution error taken place in this sample. Similarly in the second sample, PEYREMITI is recognized which is the substitution of the original word EMITI. Here the phonemes at the end of the words are same in this situation. Hence the confusion occurs. In the third sample, CHEYRUTHUNDI is substituted in the place of

VELUTHUNDHI due to the confusion of the similar phones present at the end in both words.

5.2.2. Insertion errors

The accuracy of SIASR system degrades due to the insertion of one or more words in the place of single original word. This error type comes under the insertion error which drastically increases the error rate.

(i) REF: thelangaana ekspres sirpur nundi ekkadiki veluthundhi

HYP: thelangaana ETU ekspres sirpur nundi ekkadiki veluthundhi

(ii) REF: anownsement ela cheyaali

HYP: CHEYAALI anownsement ela CHEYAALI cheyaali

(iii) REF: delli velle train eppudu vundhi

HYP: NUNDI delli velle train eppudu vundhi

In the first sample, ETU is the extra word which inserted in the result. In the second sample, CHEYAALI is the word inserted more than once which drastically reduces the accuracy rate. In the third sample, NUNDI is the extra word inserted in this case. Excluding this word, remaining words are correctly recognized. WER is increased because of the extra word, though required accuracy obtained.

5.2.3. Deletion errors

Error rate will drastically increase when some words are deleted during recognition. Low recognition accuracy is observed due to the deletion of words in hypothesis file compared to reference file.

(i) REF: SREE kalahasthiki e train veluthundhi

HYP: kalahasthiki e train veluthundhindhi

(ii) REF: varangalku velle TRAIN PEYREMITI

HYP: varangalku velle

(iii) REF: buuking aafis ekkada VUNDHI

HYP: buking aafis ekkada

In the first sample, SREE word is deleted. In the second sample, two words TRAIN and PEYREMITI are deleted. In the third sample, VUNDHI word is deleted.

5.3. ERROR ANALYSIS

From the experimental results discussed in **section 5.1**, it is observed that the WER is more, if the size of training data is reduced and the size of test data increased. The performance of SIASR system suffers a lot due to mismatch between training and testing. This mismatch can be due to the variation of speech rate. The distribution of acoustic properties in speech and linguistic interpretation of an utterance is different at different rates of speech **[4]**. The recognition accuracy affect when different speech rates are used in training and testing. More substitution and insertion errors occurred when SSR is tested with NSR and FSR training models. Similarly more errors occurred when NSR and FSR are tested with SSR training models. More deletion errors occurred when FSR is tested with NSR and SSR training models. These errors are also more when NSR and SSR are tested with FSR training models.

A number of substitution errors occurred in all cases explained in **section 5.1**. The main reason for substitution errors is due to the presence of phonetically similar words in the vocabulary. It is observed that the word is confused because of the similar phones present either in the beginning or at the end of the word. The following are some of the confusion pairs obtained frequently in **section 5.1**.

1. EMITI PEYREMITI
2. CHEYRUTHUNDHI VELUTHUNDHI
3. KAACHIGOODAKU KAAKINAADAKU
4. EKKADA EKKADIKI
5. ENNI ANNI

The phonetic transcriptions for the first confused pair are given below:

EMITI AH M IH T IY

PEYREMITI P EY R AH M IH T IY

The phonetic transcriptions for the second confused pair are given below:

CHEYRUTHUNDHI CH EY R AH TH AH N D HH IY

VELUTHUNDHI V EH L AH TH AH N D HH IY

The phonetic transcriptions for the third confused pair are given below:

KAACHIGOODAKU K AA CH AH G UW D AH K UW

KAAKINAADAKU K AA K AH N AA D AH K UW

The phonetic transcriptions for the fifth confused pair are given below:

EKKADA EH K AH D AH

EKKADIKI EH K AE D AH K IY

The phonetic transcriptions for the sixth confused pair are given below:

ENNI IH N IY

ANNI AE N IY

WER will drastically reduce when the number of substitution errors is decreased. This motivates to modify the pronunciation dictionary. Dictionary is modified by applying PDMM. Substitution errors will be reduced if the modified dictionary is used in all cases of training and testing performed in this chapter. The procedure of PDMM and the experimental results are shown in **Chapter 6.** FBSM is applied on the decoder output which reduces error word relating to insertion or deletion or substitution. The procedure and results are explained in **Chapter 7.**

CHAPTER-6
PRONUNCIATION DICTIONARY MODIFICATION METHOD

This chapter describes the importance of pronunciation dictionary (lexicon) in SIASR system. The Pronunciation Dictionary Modification Method (PDMM) is explained in this chapter. The PDMM is used to modify the pronunciation dictionary. From the experimental results, it has been observed that the errors are reduced with modified dictionary**.**

Pronunciation dictionary plays a vital role in the SIASR system. Transcription is essential for the development of dictionary. Multiple pronunciations of same words should be dealt carefully in writing the transcription. Accurate transcription eliminates recognition errors **[59]**. Care should be taken to develop the dictionary during training phase. Lexicon is nothing but the utterance and its corresponding phonetic transcription. Dictionary maker tool (DictMaker) is used to develop pronunciation dictionary. The word list, the phoneme set and the grapheme set were initialized to the DictMaker without any pronunciation information. The DictMaker provides categories to sort the phoneme set for training. The DictMaker chooses a word from the wordlist and predicts its pronunciation **[60]**.

Continuous speech data is used to learn stochastic lexicons along with pronunciation mixture models **[61]**. Accent and speech rate also influence the pronunciation dictionary. Pronunciation dictionary should be adapted according to the speech rate **[62]**. This is also one of the causes to occur more confusion pairs which degrade the performance of the SIASR system. Confusions also occur when more frequent words are there in the vocabulary **[63]**.

The words will be confused if the phonetic transcription is similar in the words. Frequency of occurrence of confusion pair is also examined to know the number of times the words are confused in decoding phase. If frequency of occurrence of confusion pair is

1, it means that the word is confused for one time. If the frequency of occurrence is 2, it means that the word is confused for two times. Similarly if the frequency of occurrence is n, then the word is confused for n times during decoding.

In order to reduce the number of confusion pairs, proposed Pronunciation Dictionary Modification Method (PDMM) is applied. This method also reduces the frequency of occurrence of confusion pairs. Thus the performance is increased by reducing substitution errors. In order to reduce the number of confusion pairs, it is desirable to change the phonemes of confused words in the dictionary. Reducing the confusion pairs leads to reduction of substitution errors. Minimizing the number of confusion pairs increases the performance of SIASR system. This method is used to minimize the different types of errors which are discussed in the previous **Section 5.2**.

The proposed PDMM updates the pronunciation dictionary (lexicon) by taking the confusion pairs from the decoder output and pronunciation dictionary as the inputs. For each confusion pair, update the phonetic transcription of a word as the next possible pronunciation in the pronunciation dictionary by the phonetic transcription of a confused word when the Levenshtein distance is less than the length of $3/4^{th}$ of the longest word in the confusion pair. Algorithmically, the PDMM is outlined as below:

Algorithm PDMM

//Input**:** confusion pairs are obtained from the decoder output, pronunciation dictionary

//Output: modified pronunciation dictionary

Step 1: Start

Step 2: Read C confusion pairs and pronunciation dictionary

// C is the number of confusion pairs obtained from the decoder output

Step 3: Calculate Levenshtein distance for each confusion pair (a, b)

//Calculate Levenshtein distance for the confusion pair (a, b) where a and b

//are the confused words.

LevenshteinDistance(a, len_a, b, len_b)

// len_a is length of the word a i.e. word a is holding m characters and

// len_b is the length of the word b i.e. word b is holding n characters

3.1. for i= 1 to len_a

3.1.1. d[i,0]:= i // d[i,j] is holding levenshtein distance

// end for

3.2. for j = 1 to len_b

3.2.1. d[0, j] := j

3.3. for i= 1 to len_a

for j = 1 to len_b

if a[i] = b[j] then
d[i, j] := d[i-1, j-1] *// no operation required*
else
d[i, j] := minimum (d[i-1, j] + 1, *// a deletion*
d[i, j-1] + 1, *// an insertion*
d[i-1, j-1] + 1 *// a substitution*
)

return $d[len_a, len_b]$ //returned levenshtein distance

Step 4: if $d[len_a, len_b] \leq 3/4^{th}$ (max (len_a, len_b))

Update the phonetic transcription of word a as the next possible pronunciation in the pronunciation dictionary by the phonetic transcription of word b

//end if

Step 5: Repeat step 3 to step 4 for C confusion pairs

Step 6: Stop

Thus the modified lexicon obtained from the above procedure is used for decoding. The following misrecognized sentences are taken to illustrate PDMM.

1. yashwanthpur ekspres KAAKINAADAKU eppudu vasthundhi
2. THIRUPATHI velle train eppudu vundhi
3. eypi ekspres taimings PEYREMITI

The following are confusions obtained for the above misrecognized sentences:

Confusion Pairs

THIRUPATHIKI	THIRUPATHI
KAACHIGOODAKU	KAAKINAADAKU
EMITI	PEYREMITI

The phonetic transcription appears in the pronunciation dictionary for confused words as follows:

EMITI	AH M IH T IY
KAACHIGOODAKU	K AA CH IH G UW D AH K UW
KAAKINAADAKU	K AA K IH N AA D AH K UW
PEYREMITI	P EY R AH M IH T IY
THIRUPATHI	TH AH R UW P AH TH IY
THIRUPATHIKI	TH IH R AH P AE TH AH K IY

The Levenshtein distance for the confusion pair (THIRUPATHIKI THIRUPATHI) is shown below:

		T	H	I	R	U	P	A	T	H	I
	0	1	2	3	4	5	6	7	8	9	10
T	1	0	1	2	3	4	5	6	7	8	9
H	2	1	0	1	2	3	4	5	6	7	8
I	3	2	1	0	1	2	3	4	5	6	7
R	4	3	2	1	0	1	2	3	4	5	6
U	5	4	3	2	1	0	1	2	3	4	5
P	6	5	4	3	2	1	0	1	2	3	4
A	7	6	5	4	3	2	1	0	1	2	3
T	8	7	6	5	4	3	2	1	0	1	2
H	9	8	7	6	5	4	3	2	1	0	1
I	10	9	8	7	6	5	4	3	2	1	0
K	11	10	9	8	7	6	5	4	3	2	1
I	12	11	10	9	8	7	6	5	4	3	2

From the above, the Levenshtein distance for the confusion pair (THIRUPATHIKI THIRUPATHI) is 2. Lexicon will be updated if the Levenshtein distance≤ 3/4th[Max$_{\textbf{length}}$ (THIRUPATHIKI, THIRUPATHI)]. Here 2<=3/4*12. Thus the condition satisfied here and hence the phonetic transcription of THIRUPATHI is added as an alternative pronunciation for the word THIRUPATHIKI.

The Levenshtein distance for the confusion pair (KAACHIGOODAKU KAAKINAADAKU) is shown below:

		K	A	A	K	I	N	A	A	D	A	K	U
	0	1	2	3	4	5	6	6	8	9	10	11	12
K	1	0	1	2	3	4	5	6	7	8	9	10	11
A	2	1	0	1	2	3	4	5	6	7	8	9	10
A	3	2	1	0	1	2	3	4	5	6	7	8	9
C	4	3	2	1	1	2	3	4	5	6	7	8	9
H	5	4	3	2	2	2	3	4	5	6	7	8	9
I	6	5	4	3	3	2	3	4	5	6	7	8	9
G	7	6	5	4	4	3	3	4	5	6	7	8	9
O	8	7	6	5	5	4	4	4	5	6	7	8	9
O	9	8	7	6	6	5	5	5	5	6	7	8	9
D	10	9	8	7	7	6	6	6	6	5	6	7	8
A	11	10	9	8	8	7	7	6	6	6	5	6	7
K	12	11	10	9	8	8	8	7	7	7	6	5	6
U	13	12	11	10	9	9	9	8	8	8	7	6	5

From the above, the Levenshtein distance for the confusion pair (KAACHIGOODAKU KAAKINAADAKU) is 5. The lexicon is updated when Levenshtein distance≤ 3/4th[Max$_{\textbf{length}}$ (KAACHIGOODAKU KAAKINAADAKU). Here 5<=3/4*13. Thus the condition satisfied here and hence the phonetic transcription of KAAKINAADAKU is added as an alternative pronunciation for the word KAACHIGOODAKU.

The Levenshtein distance for the confusion pair (EMITI PEYREMITI) is shown below:

		P	E	Y	R	E	M	I	T	I
	0	1	2	3	4	5	6	7	8	9
E	1	1	1	2	3	4	5	6	7	8
M	2	2	2	2	3	4	4	5	6	7
I	3	3	3	3	3	4	5	4	5	6
T	4	4	4	4	4	4	5	5	4	5
I	5	5	5	5	5	5	5	5	5	4

Thus the Levenshtein distance for the (EMITI PEYREMITI) is 4. The lexicon is updated when Levenshtein distance≤ $3/4^{th}$[$Max_{\textbf{length}}$ (EMITI PEYREMITI). Here 4<=3/4*9. Thus the condition satisfied here and hence the phonetic transcription of PEYREMITI is added as an alternative pronunciation for the word EMITI.

Thus the modified pronunciation dictionary after PDMM is appeared as follows:

EMITI AH M IH T IY

EMITI (1) P EY R AH M IH T IY

KAACHIGOODAKU K AA CH AH G UW D AH K UW

KAACHIGOODAKU (1) K AA K AH N AA D AH K UW

KAAKINAADAKU K AA K AH N AA D AH K UW

PEYREMITI P EY R AH M IH T IY

THIRUPATHI TH AH R UW P AH TH IY

THIRUPATHIKI TH IH R AH P AE TH AH K IY

THIRUPATHIKI (1) TH AH R UW P AH TH IY

The following are the correctly recognized sentences after giving the modified dictionary (lexicon) to the decoder.

1. yashwanthpur ekspres KAACHIGOODAKU eppudu vasthundhi
2. THIRUPATHIKI velle train eppudu vundhi
3. eypi ekspres taimings EMITI

During recognition the Sphinx-3.6 decoder combines both Acoustic model scores (AM scores) and language model probabilities into a single score scores in order to compare various hypotheses. The scores reported by the decoder are log-likelihood in this peculiar log-base. The default base is 1.0003 and can be changed using the –log base configuration argument **[57]**. To get optimum recognition accuracy, it is required to exponentiate the language model probability using a language weight (LM weight) before combining the result with acoustic likelihood. LM weight is the multiplicative factor applied to LM log- probabilities. The LM weight parameter is typically ranges from 6 to 13. Word insertion penalty parameter is another multiplicative factor for the computation of language model probability. Word insertion penalty is the fixed value when it transits from the end of one word to the start of the next word.

The following **Table 6.1** illustrates the AM scores and LM scores for the misrecognized sentences which are taken from the experimental results

Table 6.1: AM scores and LM scores for misrecognized sentences

SNo.	Incorrectly Recognized Sentences	AM score	LM score
1	yashwanthpur ekspres KAAKINAADAKU eppudu vasthundhi	12170575	-1532484
2	THIRUPATHI velle train eppudu vundhi	6277459	-1441114
3	eypi ekspres taimings PEYREMITI	20153054	-1349383

The following **Table 6.2** illustrates the AM scores and LM scores after decoding with modified dictionary using PDMM. In this table, it is seen that AM scores are same after testing with modified dictionary. But the LM scores are changed after testing with modified dictionary using PDMM.

Table 6.2: AM scores and LM scores for recognized sentences after PDMM

SNo.	Correctly Recognized Sentences	AM score	LM score
1	yashwanthpur ekspres KAACHIGOODAKU eppudu vasthundhi	12170575	-928854
2	THIRUPATHIKI velle train eppudu vundhi	6277459	-982026
3	eypi ekspres taimings EMITI	20153054	-846956

6.1. EXPERIMENTAL RESULTS AT DIFFERENT RATES OF SPEECH IN TRAINING AND TESTING WITH MODIFIED PRONUNCIATION DICTIONARY USING PDMM

Pronunciation Dictionary Modification Method (PDMM) is used to modify the dictionary (lexicon). The modified dictionary is used in the decoder of SIASR system. This modified dictionary is used for the same training and test data sets (eighteen experiments) mentioned in **Section 5.1** in **Chapter 5.** The SRR, WRR and WER are calculated after decoding with modified dictionary. These results are tabulated in this section. It has been observed that the errors are reduced with modified dictionary**.** These results are compared with the SRR, WRR and WER in **Section 5.1** in **Chapter 5.**

6.1.1. NSR Training

6.1.1.1. NSR is tested with NSR training model after PDMM

The following **Table 6.3** shows improved SRR, WRR, SER, DER, IER and WER when NSR is tested with NSR training model with modified pronunciation dictionary using PDMM.

Table 6.3: Testing NSR with NSR after PDMM

S.No	Training Data in NSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	978 97.8%	4702 99.2%	27 0.5%	11 0.2%	8 0.2%	46 1%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	49 98%	235 99.2%	0 0%	2 2%	0 0%	2 0.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	98 98%	469 98.9%	0 0%	5 1.1%	0 0%	5 1.1%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	194 97%	938 98.9%	0 0%	10 1.1%	0 0%	10 1.1%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	243 97.2%	1178 99.4%	0 0%	7 0.6%	2 0.2%	9 0.8%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	467 93.4%	2343 98.9%	13 0.5%	14 0.6%	22 0.9%	49 2.1%

The above results are compared with results mentioned in **Table 5.1** as follows:

In the first case, the SRR is increased from 97.4% to 97.8%. The WRR is increased by 0.1%. Here SER is reduced by 0.2%.The numbers of error words are reduced from 52 to 46. Thus the WER reduced by 0.1%.

In the second case, the SRR is increased from 96% to 98%. Thus the improvement of SRR is 2%. The WRR is improved from 98.7% to 99.2%. The substitution errors become zero in this case. The number of error words reduced from 3 to 2 words. Thus the WER is reduced from 1.3% to 0.8%.

In the third case, the SRR is increased from 97% to 98%. Thus the SRR is improved by 1%. The WRR is enhanced by 0.4%. The substitutions are reduced to zero in this case. The total number error words reduced from 7 to 5 words. Thus WER is reduced from 1.5% to 1.1%.

In the fourth case, the SRR is improved by 0.5%. The WRR is enhanced by 0.4%. The substitutions errors become zero in this case. Thus the number of errors reduced from 14 to 10 words. As the errors are reduced, WER is reduced by 0.4%.

In the fifth case, the SRR improved by 1.2%. Thus WRR is raised by 0.9%. Here the substitution errors reduced from 15 to 0(zero). The total errors reduced from 22 to 9 errors. Thus WER is reduced from 1.9% to 0.8%.

In the sixth case, the SRR is raised by 2.4%. The WRR increases from 96.6% to 98.9%. Thus 2.3% of improvement is observed. The substitution errors reduced from 65 to 13. Thus the WER is reduced by 2% in this case.

6.1.1.2. SSR is tested with NSR training model after PDMM

The following **Table 6.4** below shows SRR, WRR, SER, DER, IER and WER when SSR is tested with NSR training model with modified pronunciation dictionary using PDMM. These results are compared with results tabulated in **Table 5.2**.

Table 6.4: Training with NSR and testing SSR after PDMM

S.No	Training Data in NSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	976 97.6%	4691 98.9%	41 0.9%	8 0.2%	15 0.3%	64 1.4%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	46 92%	230 97%	7 3%	0 0%	3 1.3%	10 4.2%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	94 94%	465 98.1%	8 1.7%	1 0.2%	2 0.4%	11 2.3%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	192 96%	935 98.6%	11 1.2%	2 0.2%	4 0.4%	17 1.8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	241 96.4%	1164 98.2%	14 1.2%	7 0.6%	0 0%	21 1.8%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	456 91.2%	2294 96.8%	61 2.6%	15 0.6%	18 0.8%	94 4%

In the first case, the SRR is increased from 97% to 97.6%. The WRR is improved from 98.8% to 99%. Here the substitution errors are reduced from 47 to 41 words and thus WER is reduced from 1.5% to 1.4%.

In the second case, the reduction of 1 substitution error enhanced SRR by 2%. The WRR is improved from 96.6% to 97%. Thus WER is reduced from 4.6% to 4.2%.

In the third case, the reduction of 1 substitution increased SRR by 1%. The WRR is improved from 97.9% to 98.1%. The SER is reduced from 1.9% to 1.7%. The WER reduced by 0.2%.

In the fourth case, the SRR is improved by 0.5%. The WRR is increased from 98.4% to 98.6%. The SER is reduced from 1.4% to 1.2%. The WER is reduced by 0.2% in this case.

In the fifth case, The SRR is increased from 94.8% to 96.4%. Thus the improvement of SRR is 0.9%. The WRR is improved from 97.8% to 98.2%. The substitution errors are reduced from 17 to 14. Thus WER is reduced from 2.2% to 1.8%.

In the sixth case, the SRR is raised by 2.6%. The WRR is increased from 96.2% to 96.8%. The substitution errors are reduced from 72 to 61 words. Thus WER is reduced from 4.4% to 4%.

6.1.1.3. FSR is tested with NSR training model after PDMM

The following **Table 6.5** shows the SRR, WRR, SER, DER, IER and WER when FSR is tested with NSR training model with modified pronunciation dictionary using PDMM.

Table 6.5: Testing FSR with NSR after PDMM

S.No	Training Data in NSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	925 92.5%	4621 97.5%	91 1.9%	28 0.6%	11 0.2%	130 2.7%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	41 82%	216 91.1%	13 5.5%	8 3.4%	1 0.4%	22 9.3%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	79 79%	444 93.7%	23 4.9%	7 1.5%	1 0.2%	31 6.5%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	173 86.5%	902 95.1%	32 3.4%	14 1.5%	2 0.2%	48 5.1%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	216 86.4%	1126 95%	37 3.1%	22 1.9%	2 0.2%	61 5.1%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	397 79.4%	2191 92.4%	141 5.9%	38 1.6%	3 0.1%	182 7.7%

The above results are compared with the results mentioned in **Table 5.3** as follows:

In the first case, the SRR is increased from 92.2% to 92.5%. The WRR is improved from 97.4% to 97.5%. The number of error words is reduced from 135 words to 130 words. Thus the reduction of WER is 0.1%.

In the second case, 2% improvement in SRR is observed. The WRR is improved from 90.7% to 91.1%. The error words are reduced from 23 to 22 words. Thus WER is reduced by 0.4%.

In the third case, the SRR is raised by 1%. The WRR is improved by 0.5%. The substitution errors are reduced from 25 to 23 words. Thus reduction of WER is 0.5%.

In the fourth case, the SRR is increased by 0.5%. The WRR is increased by 0.1%. The error words are reduced from 49 to 48 words. Thus the WER is reduced by 0.1%.

In the fifth case, the SRR is improved from 85.6% to 86.4%. Thus the SRR is enhanced by 0.8%. The WRR is increased by 0.2%. The total number of error words is reduced from 64 to 61 words. Thus the WER is reduced by 0.3%.

In the sixth case, 0.2% improvement in SRR is observed. The WRR is raised by 0.1%. The error words are reduced from 185 to 182 words. Thus the reduction of WER is 0.1%.

From the results tabulated in **Tables 6.3**, **6.4** and **6.5**, the observations are as follows:

(i) S<D<I in 1 case; S≤I<D in 4 cases and S>D>I in one case when NSR is tested with NSR after PDMM

(ii) S>I>D in 5 cases; S>I>D in 1 case when SSR is tested with NSR

(iii) S>D>I in all cases when FSR is tested with NSR

It is found that substitutions are reduced rapidly when NSR is tested with NSR with modified dictionary and in some cases substitution errors become 0 (zero). Reduction of substitution errors are observed when SSR and FSR are tested with NSR.

As the drastic reduction of substitution errors, the reduction of WER is noticed in all cases. Thus considerable improvement in WRR and SRR is noticed in all cases. This improvement is shown in **Figure 6.1** and **Figure 6.2**.

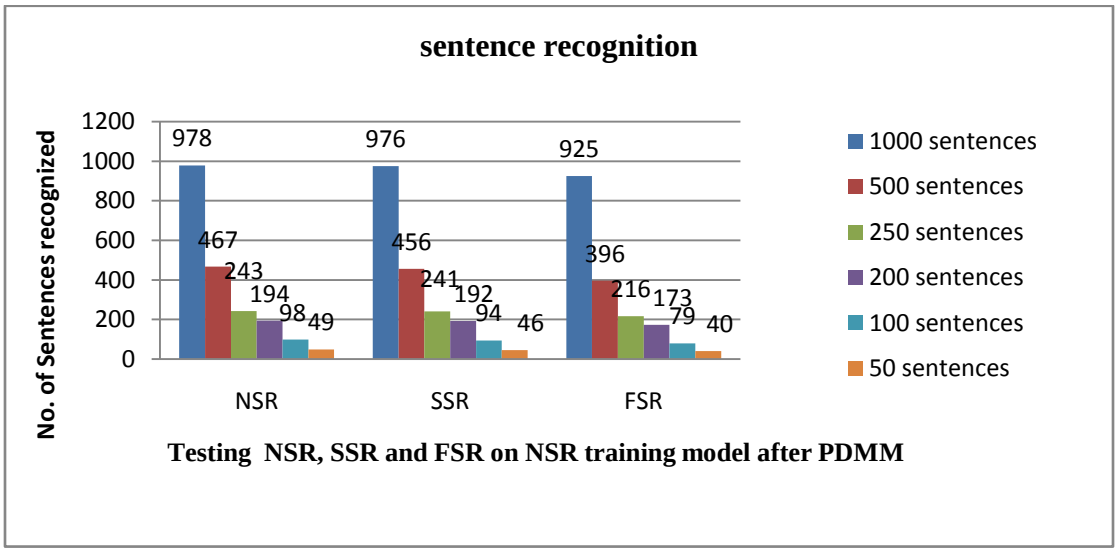

Figure 6.1: Sentence recognition with modified dictionary using PDMM (NSR training)

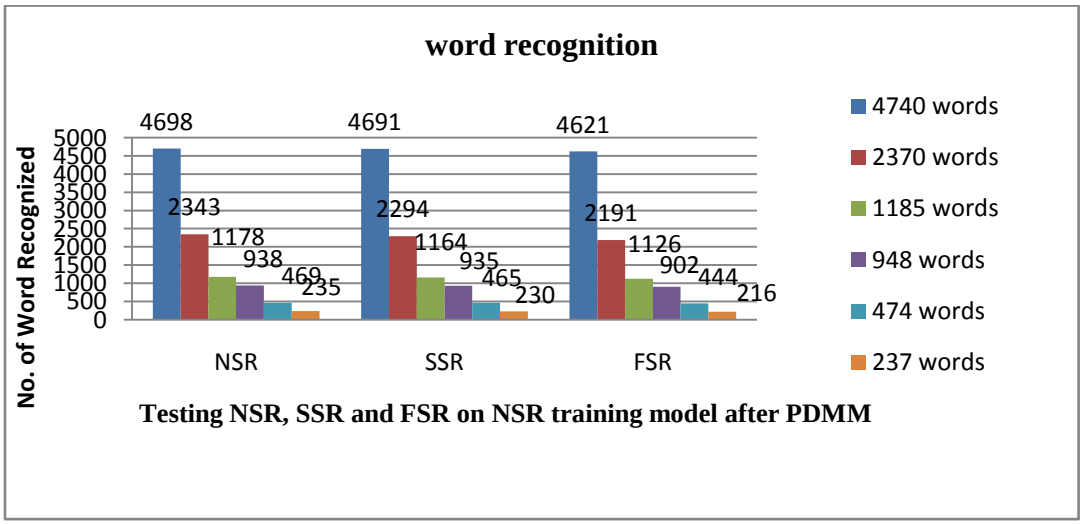

Figure 6.2: Word recognition with modified dictionary using PDMM (NSR training)

6.1.2. SSR Training

6.1.2.1. NSR is tested with SSR training model after PDMM

The **Table 6.6** shows SRR, WRR, SER, DER, IER and WER when NSR is tested with SSR training model with modified lexicon using PDMM.

Table 6.6: Testing NSR with SSR after PDMM

S.No	Training Data in SSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of Words	No. of speakers (M&F)	No. of sents	No. of words	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	958 95.8%	4664 98.4%	56 1.2%	20 0.4%	10 0.2%	86 1.8%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	45 90%	225 94.9	9 3.7%	3 1.3%	5 2.1%	17 7.2%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	84 84%	454 95.8%	17 3.6%	3 0.6%	5 1.1%	25 5.3%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	183 91.5%	926 97.7%	18 1.9%	4 0.4%	7 0.7%	29 3.1%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	229 91.6%	1153 97.3%	22 1.9%	10 0.8%	9 0.8%	41 3.5%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	435 87%	2263 95.5%	81 3.4%	26 1.1%	20 0.8%	127 5.4%

The present results are compared with the results mentioned in **Table 5.4** as follows:

In the first case, the SRR is raised by 0.4%. The WRR is improved by 0.1%. The substitution errors are reduced from 62 to 56 words. Thus the total error words are reduced from 92 to 86 words. As the errors are reduced, WER is reduced by 0.1%.

In the second case, 2% improvement in SRR is observed with the reduction of one substitution error. Thus error words are reduced from 18 to 17 words. Hence the WER is reduced by 0.4%. The improvement of WRR is 0.4%.

In the third case, SRR is increased by 2% with the reduction of 2 substitutions. The total errors are reduced from 27 to 25 errors. Thus the WER is reduced by 0.4%. The WRR is raised by 0.4%.

In the fourth case, 1% of SRR improvement is noticed. 0.2% improvement in WRR is observed. The SER is reduced from 2.1% to 1.9%. The numbers of error words are reduced from 31 to 29 words. Thus the WER is reduced by 0.2%.

In the fifth case, the SRR is improved by 0.8%. The WRR is increased by 0.2%. The SER is reduced by 0.1%. The number of error words is decreased from 43 to 41 words. Thus WER is reduced by 0.1%.

In the sixth case, the SRR is improved from 86.2% to 87%. Thus 0.8% improvement is noticed in SRR. The WRR is improved by 0.3%. The substitution errors are reduced from 87 to 81. Thus the reduction of WER is 0.2%.

6.1.2.2. SSR is tested with SSR training model after PDMM

The **Table 6.7** shows SRR, WRR, SER, DER, IER and WER when SSR tested with SSR training model with modified pronunciation dictionary using PDMM.

Table 6.7: Testing SSR with SSR after PDMM

S.No	Training Data in SSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	974 97.4%	4697 99.1%	35 0.7%	8 0.2%	13 0.3%	56 1.1%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	46 92%	229 96.6%	7 2.9%	1 0.4%	1 0.4%	9 3.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	95 95%	467 98.5%	7 1.5%	0 0%	2 0.4%	9 1.9%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	191 95.5%	935 98.6%	13 1.4%	0 0%	5 0.5%	18 1.9%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	238 95.2%	1164 98.2%	21 1.8%	0 0%	3 0.3%	24 2%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	436 87.2%	2273 95.9%	70 3%	27 1.1%	19 0.8%	116 4.9%

The present results are compared with the results mentioned in **Table 5.5** as follows**:**

In the first case, 0.3% improvement in SRR is observed. The WRR is increased from 99% to 99.1%. The number of error words is reduced from 59 to 56 words. Thus the reduction of WER is 0.1%.

In the second case, the SRR is increased from 90% to 92%. Thus 2% of SRR is improved with the reduction of 1 substitution error. It has been observed that the WER is reduced from 4.2% to 3.8%. Thus WRR is raised by 0.4%.

In the third case, the SRR is raised from 92% to 95%. Thus 3% of improvement in SRR is noticed. The WRR is improved by 0.4%. The substitution errors are reduced from 10 to 7 words. Thus WER is reduced by 0.6%.

In the fourth case, 1% improvement in SRR is found by the reduction of 2 substitution errors. The number of error words is reduced from 20 to 18 words. Thus the WER is reduced by 0.2%.

In the fifth case, the SRR is increased from 94% to 95.2%. Thus 1.2% improvement in SRR is observed. The WRR is improved by 0.2%. The substitution errors are reduced from 24 to 21 words. Thus WER is reduced by 0.2%.

In the sixth case, the SRR is raised by 0.6%. The error words are reduced from 120 to 116 words. As 4 substitutions are reduced, WER is reduced by 0.2%.

6.1.2.3. FSR is tested with SSR training model after PDMM

The following **Table 6.8** shows SRR, WRR, SER, DER, IER and WER when FSR is tested with SSR training model with modified pronunciation dictionary using PDMM. These results are compared with the results mentioned in **Table 5.6.**

Table 6.8: Testing FSR with SSR after PDMM

S.No	Training Data in SSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	892 89.2%	4569 96.4%	132 2.8%	39 0.8%	5 0.1%	176 3.7%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	39 78%	213 89.8%	17 7.2%	7 3%	2 8%	26 11%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	73 73%	433 91.4%	31 6.5%	10 2.1%	2 0.4%	43 9.1%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	161 80.5%	875 92.3%	50 5.3%	23 2.4%	1 0.1%	74 7.8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	199 79.6%	1102 93%	62 5.2%	21 1.8%	1 0.1%	84 7%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	359 71.8%	2110 89%	190 8%	70 3%	6 0.3%	266 11.2%

In the first case, the SRR is enhanced by 0.7%. Thus WRR is improved by 0.2%. The error words are reduced from 185 to 176 words. The reduction of WER is reduced by 0.2%.

In the second case, the SRR is increased by 4%. The WRR is improved by 0.8%. The number of error words is reduced from 28 to 26 words. Thus the WER is reduced by 0.8%.

In the third case, 1% of improvement in SRR is observed. The WRR is increased by 0.7%. The SER is reduced from 7.2% to 6.5%. Thus WER is reduced by 0.7%.

In the fourth case, the SRR is enhanced by 0.5%. The improvement of WRR is 0.2%. The number of error words reduced from 76 to 74 words. The WER is decreased by 0.2%.

In the fifth case, the SRR is enhanced by 0.8%. The improvement of WRR is 0.2%. The number of error words decreases from 86 to 84 words. The WER is reduced by 0.3%.

In the sixth case, the SRR is increased by 0.4%. The WRR is improved by 0.2%. The substitution errors are reduced from 194 to 190 words. Thus WER is reduced by 0.2%.

From the **Figure 6.3** and **Figure 6.4,** it is observed that SRR and WRR is more when SSR is tested with SSR. Thus recognition accuracy is less when FSR is tested with NSR.

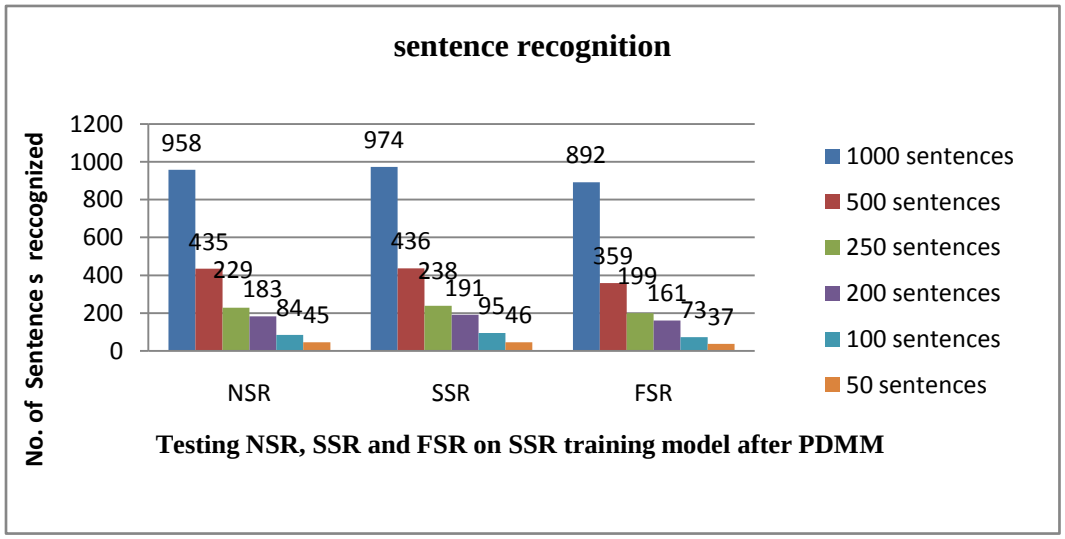

Figure 6.3: Sentence recognition with modified dictionary using PDMM (SSR training)

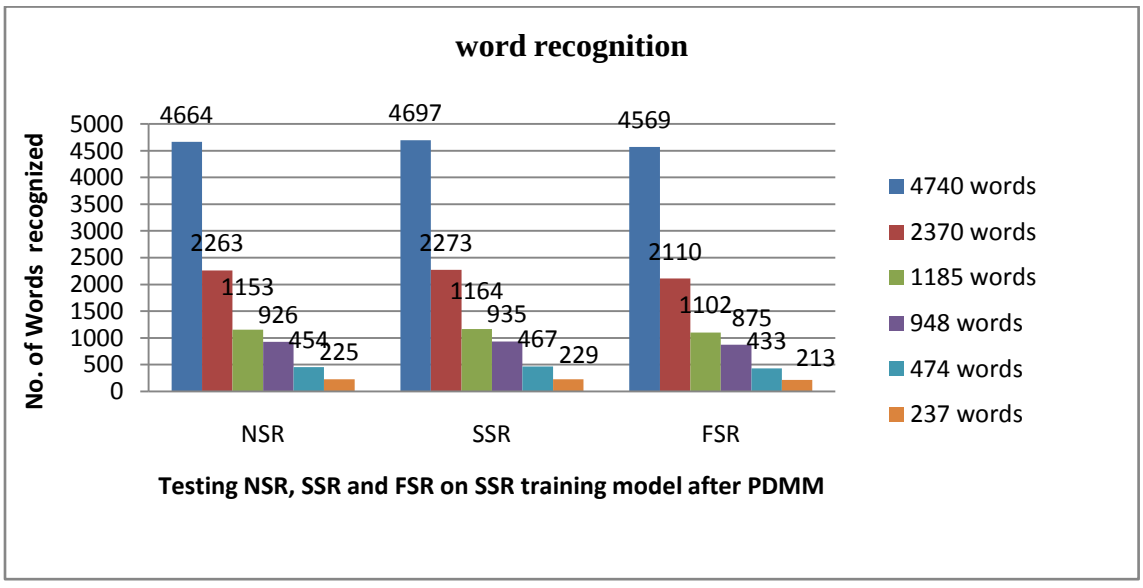

Figure 6.4: Word recognition with modified dictionary using PDMM (SSR training)

The following observations are noticed from the above results mentioned in **Tables 6.6**, **6.7** and **6.8**:

(i) S>D>I in all cases when NSR is tested with SSR with modified dictionary

(ii) S>I>D in 5 cases; S>D>I in 1 case when SSR is tested with SSR with modified dictionary

(iii) S>D>I in all cases when FSR is tested with SSR

6.1.3. FSR Training

6.1.3.1. NSR is tested with FSR training model after PDMM

The below **Table 6.9** shows accuracy when NSR is tested with FSR after PDMM. The above results are compared with the results mentioned in **Table 5.7** as follows:

Table 6.9: Testing NSR with FSR after PDMM

S.No	Training Data in FSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	978 97.8%	4702 99.2%	32 0.7%	6 0.1%	10 0.2%	48 1%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	47 94%	232 97.9%	4 1.7%	1 0.4%	1 0.4%	6 2.5%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	90 90%	462 97.5%	10 2.1%	2 0.4%	5 1.1%	17 3.6%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	194 97%	936 98.7%	12 1.3%	0 0%	5 5%	17 1.8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	238 95.2%	1166 98.4%	18 1.5%	1 0.1%	4 0.3%	23 1.9%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	448 89.6%	2306 97.3%	51 2.2%	13 0.5%	22 0.9%	86 3.6%

In the first case, the SRR is improved by 1.5%. The WRR is improved by 0.2%. The substitution errors are reduced from 44 to 32 words. The number of error words reduced from 60 to 48 words. Thus the WER is reduced by 0.3%.

In the second case, the improvement in SRR is 4%. The WRR is improved by 0.9%. The number of error words is reduced from 10 to 6 words. The reduction of WER is 1.7%.

In the third case, 2% of improvement in SRR is observed. The WRR is improved by 0.5%. The number of error words decreases from 19 to 17 words. Thus WER is reduced by 0.4%.

In the fourth case, the SRR is enhanced by 2.5%. The WRR is improved by 0.3%. The number of substitution errors is reduced from 15 to 12 words. The reduction of WER in this case is 0.3%.

In the fifth case, the SRR improved by 1.6%. The WRR improvement is 0.3%. 4 substitution errors are reduced in this case. Thus WER is reduced by 0.4%.

In the sixth case, the SRR is increased by 1.4%. The error words are decreased from 94 to 86 words. Thus WER is reduced from 4% to 3.6%.

6.1.3.2. SSR is tested with FSR training model after PDMM

The following **Table 6.10** shows the SRR, WRR, SER, DER, IER and WER when SSR tested with FSR training model with modified pronunciation dictionary using PDMM.

Table 6.10: Testing SSR with FSR after PDMM

S.No	Training Data in FSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents	No. of words	Corr-sents &SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	970 97%	4693 99%	35 0.7%	12 0.3%	5 0.1%	52 1.1%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	46 92%	229 96.6%	6 2.5%	2 0.8%	1 0.4%	9 3.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	86 86%	457 96.4%	12 2.5%	5 1.1%	3 0.6%	20 4.2%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	185 92.5%	928 97.9%	13 1.4%	7 0.7%	1 0.1%	21 2.2%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	229 91.6%	1154 97.4%	23 1.9%	8 0.7%	1 0.1%	32 2.7%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	443 88.6%	2285 96.4%	68 2.9%	17 0.7%	13 0.5%	98 4.1%

These results are compared with the results mentioned in **Table 5.8.**

In the first case, the improvement in SRR is 0.4%. The WRR is improved by 0.2%. The number of error words is decreased from 56 to 52 words. 1% reduction in SER is observed in this case. Thus the WER is reduced by 0.1%.

In the second case, the SRR is enhanced by 4%. The WRR increases from 95.8% to 96.6%. The SER is reduced from 3.4% to 2.5%. Thus number of error words is reduced from 11 to 9 words. Thus the reduction of WER is 0.8%.

In the third case, 2% of improvement in SRR is observed. The WRR is enhanced by 0.4%. The error words are reduced from 22 to 20 words. Thus the reduction of WER is 0.4%.

In the fourth case, the SRR is increased by 1%. The number of error words is reduced from 23 to 21 words. Here 2 substitution words are reduced. Thus the reduction of WER is 0.2%.

In the fifth case, 0.4% of improvement in SRR is observed. The WRR is increased by 0.1%. The error words are decreased from 33 to 32 words. Thus the reduction of WER is 0.1%.

In the sixth case, the improvement in SRR is 1%. The WRR is raised by 0.2%. The error words are reduced from 103 to 98 words. Thus the WER is reduced by 0.3%.

6.1.3.3. FSR is tested with FSR training model after PDMM

The following **Table 6.11** shows SRR, WRR, SER, DER, IER and WER when FSR is tested with FSR training model with modified pronunciation dictionary using PDMM.

Table 6.11: Testing FSR with FSR after PDMM

S. No	Training Data in FSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	980 98%	4706 99.3%	29 0.6%	5 0.1%	11 0.2%	45 0.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	48 96%	233 98.3%	4 1.7%	0 0%	3 1.3%	7 3%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	95 95%	468 98.7%	5 1.1%	1 0.2%	3 0.6%	9 1.9%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	194 97%	940 99.2%	7 0.7%	1 0.1%	3 0.3%	11 1.2%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	241 96.4%	1169 98.6%	14 1.2%	2 0.2%	5 0.4%	21 1.8%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	459 91.8%	2315 97.7%	45 1.9%	10 0.4%	22 0.9%	77 3.2%

The above results are compared with the results mentioned in **Table 5.9** as follows:

In the first case, the SRR is enhanced by 0.2%. The WRR is raised by 0.1%. The error words are reduced from 50 to 45 words. Here substitution errors are reduced from 34 to 29. Thus the reduction of WER is 0.2%.

In the second case, 4% of improvement in SRR is observed. The WRR is increased by 0.8%. The error words are reduced from 9 words to 7 words. The WER reduces from 3.7% to 3%.

In the third case, the SRR is enhanced by 4%. The WRR is improved from 97.9% to 98.7%. The number of error words reduces from 13 to 9 words. 4 substitution errors are reduced in this case. Thus the WER is reduced by 0.8%.

In the fourth case, the SRR is increased by 1.5%. The WRR increases from 98.6% to 99.2%. The SER is reduced from 1.3% to 0.7%. Thus the error words are reduced from 16 to 11 words. Hence the reduction of WER is 0.5%.

In the fifth case, the SRR improvement is 1.6%. The SER is reduced from 1.5% to 1.2%. The number of error words is reduced from 25 to 21 words. Thus the reduction of WER is 0.3%.

In the sixth case, the SRR improvement is 1.2%. The SER reduces from 2.2% to 1.9%. The error words are reduced from 83 to 77 words. Thus WER reduces by 0.3%.

The following observations are noticed from the above results mentioned in **Tables 6.9**, **6.10** and **6.11**:

(i) S>I>D in all cases when NSR is tested with FSR

(ii) S>D>I in all cases when SSR is tested with SSR

(iii) S>I>D in all cases when FSR is tested with FSR

From the **Figure 6.5** and **Figure 6.6**, it is noticed that the accuracy is increased when the dictionary is modified through PDMM. The reduction of confusion pairs which reduces substitution errors is noticed from the experimental results. Thus the accuracy automatically improved to considerable extent. But deletion and insertion errors also affect recognition accuracy. This type of errors also should be reduced.

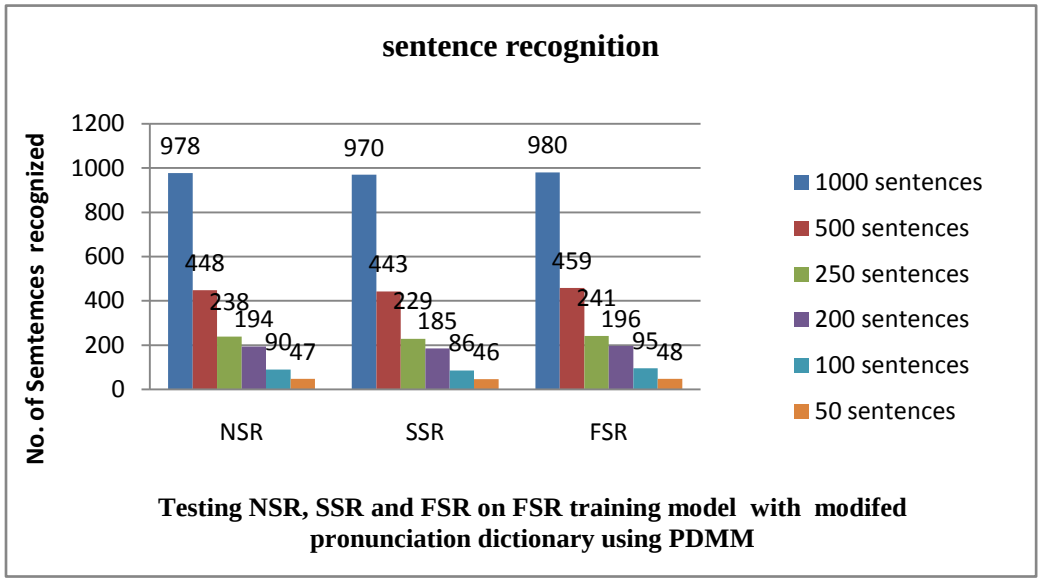

Figure 6.5: Sentence recognition with modified dictionary using PDMM (FSR training)

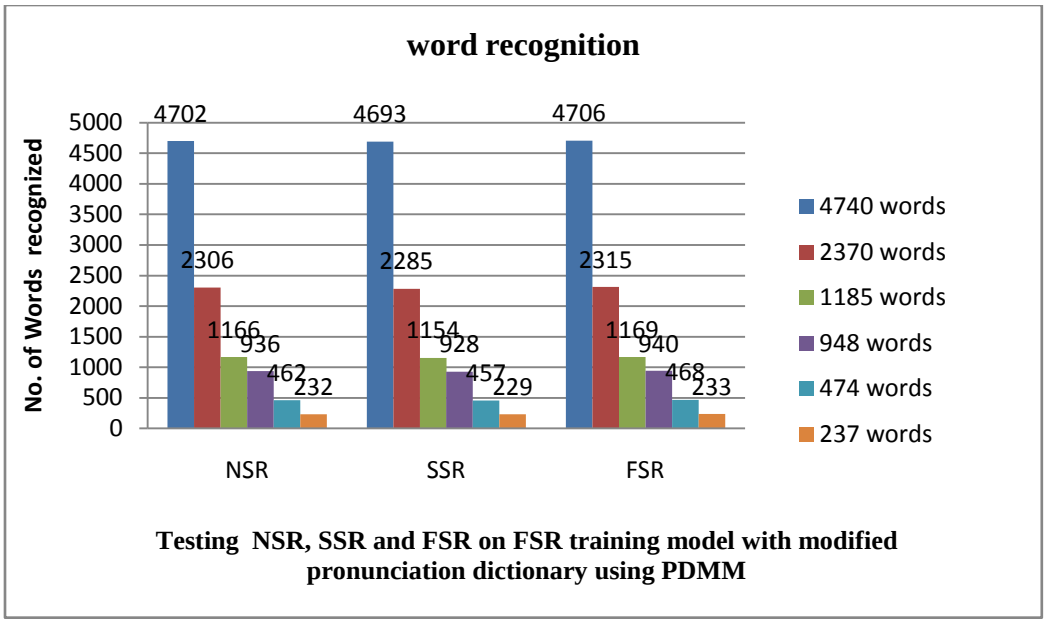

Figure 6.6: Word recognition with modified dictionary using PDMM (FSR training)

6.2. REDUCTION OF CONFUSION PAIRS WITH MODIFIED DICTIONARY USING PDMM

The main aim of PDMM is to minimize the confused pairs. The confusion between the words is responsible for occurring substitution errors. The confusion pairs discussed in **Section 5.3** in **Chapter 5** are reduced by modified dictionary using PDMM**.** In this section, the reduction of confusion pairs in all cases is shown below:

(A). <u>Training with NSR and testing with NSR, SSR and FSR</u>

The confused pairs obtained when NSR, SSR and FSR are tested with NSR in all cases are shown in **Figure 6.7**. Less number of substitution errors is observed when NSR is tested with NSR. Here more substitutions are observed when FSR and SSR are tested with NSR training model. This is because of the variation in speech rates used in training and testing.

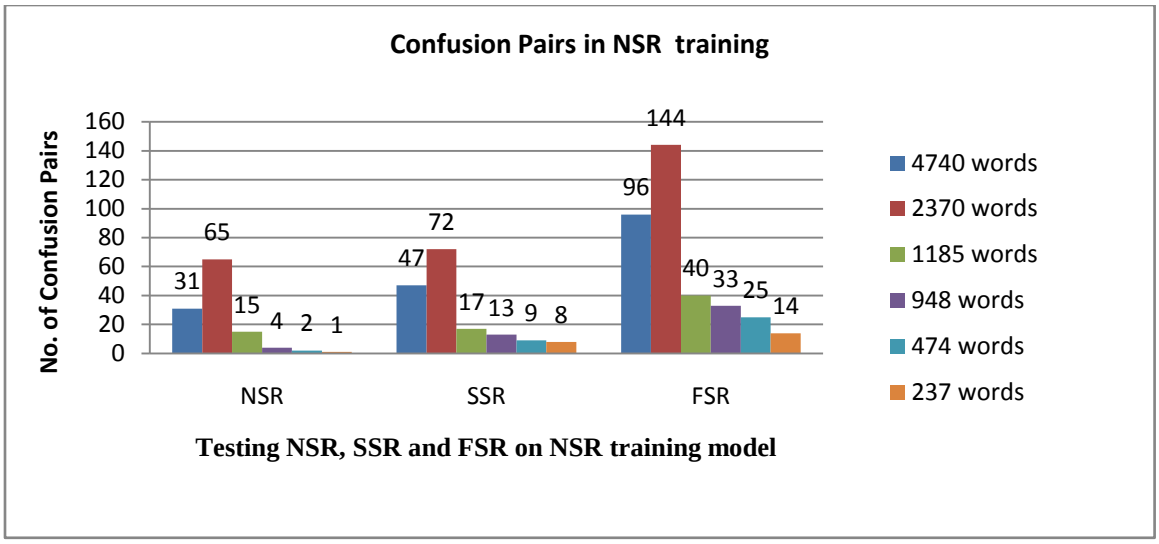

Figure 6.7: Confusions pairs obtained when NSR, SSR and FSR are tested with NSR

The number of confusion pairs drastically reduces with modified dictionary using PDMM. This is shown in the **Figure 6.8**. Confusions reduce more when NSR is tested with NSR training model. Reduction of confused pairs is low when SSR and FSR are tested with NSR.

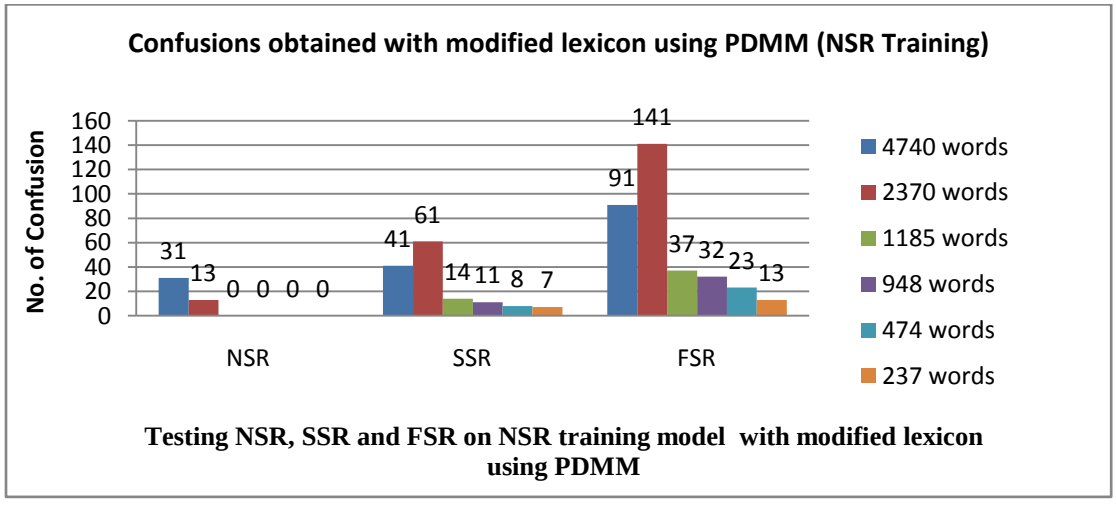

Figure 6.8: Confusions pairs obtained when NSR, SSR and FSR are tested with NSR training model using modified dictionary

B). <u>Training with SSR and testing NSR, SSR and FSR</u>

The confused pairs obtained when NSR, SSR and FSR are tested with SSR training model are shown in **Figure 6.9**. More confusion pairs are observed when NSR and FSR are tested with SSR training model. Confusions are less when SSR is tested with SSR.

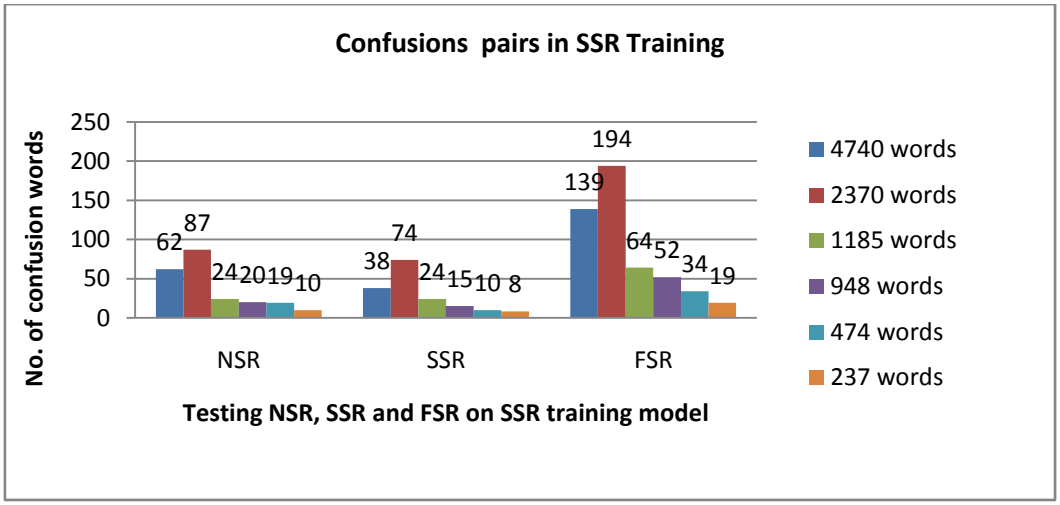

Figure 6.9: Confusions pairs obtained when NSR, SSR and FSR are tested with SSR training model

More confusion pairs are reduced after modifying the lexicon using PDMM. This is presented in the **Figure 6.10**. More substitution errors are reduced when SSR is tested with SSR training model. It is noticed that the reduction of confused pairs is low when FSR and NSR is tested with SSR**.**

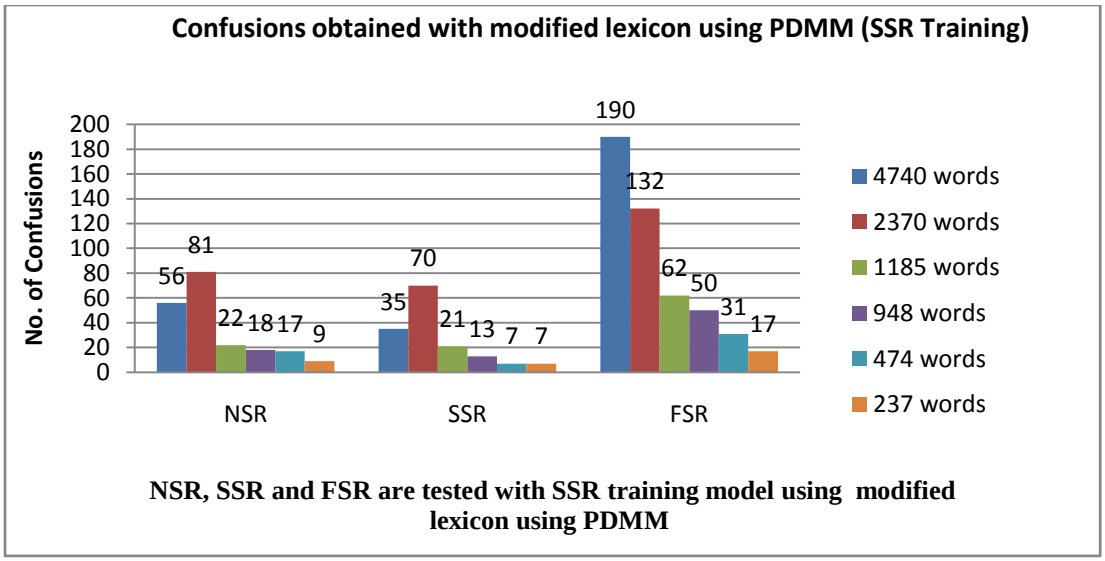

Figure 6.10: Confusions obtained when NSR, SSR and FSR are tested with SSR training model using modified pronunciation dictionary

(C). Training with FSR and testing with NSR, SSR and FSR

The confused pairs obtained when NSR, SSR and FSR are tested with FSR training model is shown in **Figure 6.11**. Confusion pairs are less when FSR is tested with FSR. More confusion pairs are observed when NSR and SSR are tested with FSR training model. This is because of the variation in speech rate between training and testing models.

Confusions are drastically reduced with modified dictionary using PDMM which is shown in the **Figure 6.12**. More confusion pairs are reduced when FSR is tested with FSR. Reduction of confused pairs is low in the case of NSR and SSR is tested with FSR.

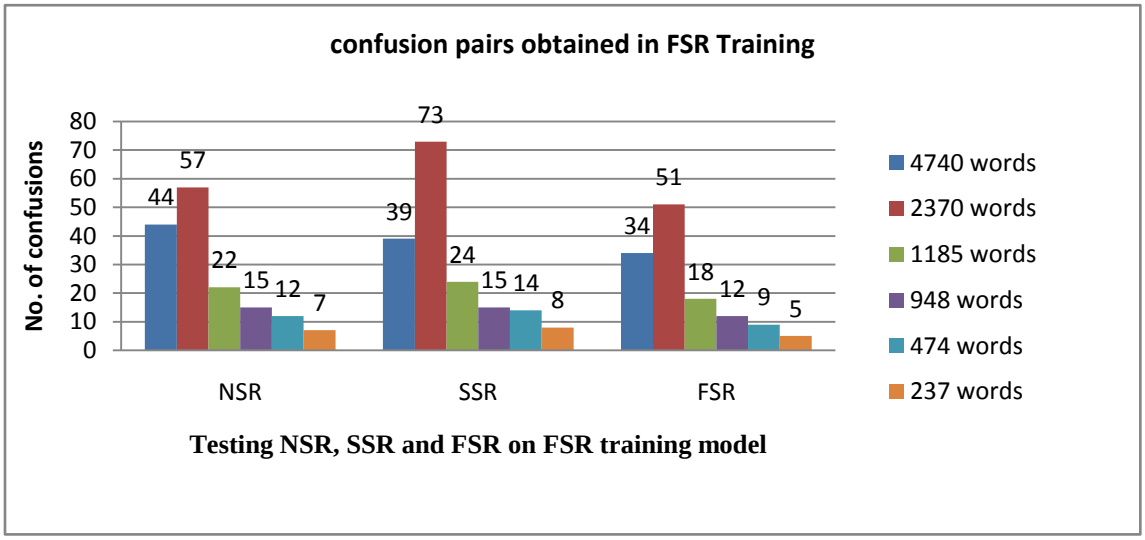

Figure 6.11: Confusions obtained when NSR, SSR and FSR are tested with FSR training model

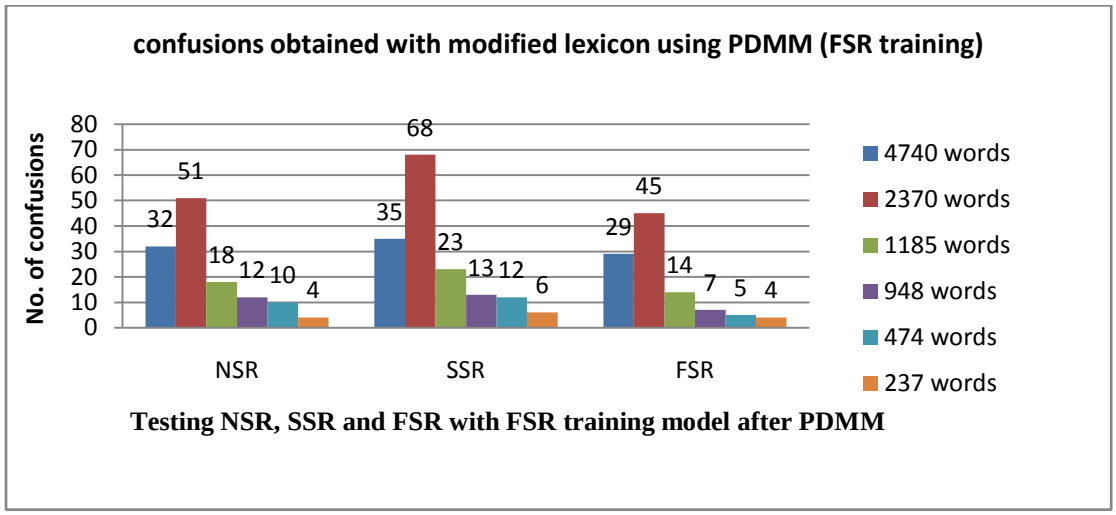

Figure 6.12: Confusions pairs obtained when NSR, SSR and FSR are tested with FSR training model using modified dictionary

CHAPTER-7
FORWARD AND BACKWARD SEARCH METHOD

The performance of the SIASR system depends on the quality of the speech recognition result. Forward and Backward Search Method (FBSM) is applied on the decoder results to reduce errors. FBSM is explained in this chapter.

It is important to correct the errors in the results of speech recognition to improve the performance **[64]**. Specific words in a large corpus tend to co-occur frequently with certain other context words, and misrecognitions of those specific words will also tend to co-occur with the same context words **[65]**. Post-editing (i.e. post-processing) implies that detecting and correcting errors are done after the input wave has been transformed into text **[66]**.

The proposed Forward and Backward Search Method (FBSM) will applied on the decoder output of the SIASR system to reduce errors. This method improves SRR as well as WRR. If the decoder output contains even one substituted error or one insertion error or one deleted error, it will affect the performance of SIASR system to a great extent. The FBSM corrects the substitution error by placing correct word, deleting the inserted error word and inserting correct word in the place of deleted word. Thus the performance of SIASR will be improved a lot.

To apply FBSM on the decoder output, it is necessary to get the sentences with errors. Success factor is calculated on all sentences for this purpose. Success factor is calculated as given below:

$$\text{success factor} = \frac{F+B}{N} \qquad (7.1)$$

where

F = number of consecutive words in hypothesis file matched with reference file using forward search

B = number of consecutive words in hypothesis file matched with reference file using backward search

L_{hyp} = length of the hypothesis file (number of words in hypothesis file)

L_{ref} = length of the reference file (number of words in reference file)

$N = \max(L_{\mathbf{hyp}}, L_{\mathbf{ref}})$

$n = \min(L_{\mathbf{hyp}}, L_{\mathbf{ref}})$

$F = \sum_{i=1}^{n} f_i$ where $f_i = 1$ if $hyp_{\mathbf{i}} = ref_{\mathbf{i}}$

$= 0$ stop forward search

$B = \sum_{i=1}^{n} b_i$ where $bi = 1$ if $hyp_{\mathbf{len-i+1}} = ref_{\mathbf{len-i+1}}$

$= 0$ stop backward search

f_i and b_i are the values of forward and backward search at the index i

If the success factor is 2, then the sentence in both reference and hypothesis is same. The sentences from the decoder output are selected if their success factor lies between 0.5 and 1. The following are observed from the above:

1. If $L_{hyp} > L_{ref}$, a particular sentence in a decoder result have insertion errors.
2. If $L_{hyp} < L_{ref}$, a particular sentence is recognized with deleted words.
3. If $L_{hyp} = L_{ref}$, a particular sentence is recognized with substitution errors.

Then modification factor needs to be applied on each selected sentences to get sentences with one error (substitution, insertion and deletion). The modification factor is given below:

$$\text{modification factor} = (N - (F + B)) \qquad (7.2)$$

If the modification factor is 1, the following actions are performed on the sentences with one error:

1. If the sentence having one substitution error word at i^{th} position, correct word is replaced in the i^{th} position.

2. If the sentence having one insertion error at i^{th} position, the error word is deleted.
3. If the sentence having one deleted word at i^{th} position, correct word is inserted at i^{th} position.

The following samples are taken for illustrating FBSM:

Sample-1:

SIASR Output: **EY TIRUMAL EKSPRES ENNI GANTALU AALASYANGAA VASTUNDHI**

Here $f_1 = 0$, stop the forward search. Thus the value of F is 0. Here $b_1= 1$, $b_2= b_3= b_5=b_6=1$ and $b_7=0$, stop the backward search. Thus the value of B is 6 ($b_1+b_2+ b_3+ b_5+b_6$).

The success factor = (0+6)/7 = 0.85, the above sentence is selected from the decoder output. The modification factor = (7-6) = 1. Also observed the condition $L_{hyp} > L_{ref}$ which means there is only one insertion error as 'EY'. Hence the sentence is modified by deleting the error word 'EY'. After FBSM, the result will be as follows:

FBSM output: **TIRUMALA EKSPRES ENNI GANTALU AALASYANGAA VASTUNDHI**

Sample-2:

SIASR Output: **PADHMAAVATHI EKSPRES VASTUNDHI**

Here $f_1 = 1$, $f_2 = 1$ and $f_3 = 0$ then stop the forward search, thus value of F is 2 ($f_1 + f_2=2$). Here $b_1= 1$, $b_2= 0$ then stop the backward search. Thus the value of B is 1.

The success factor = (2+1)/4 = 0.75, the above sentence is selected from the decoder output. The modification factor = (4-3) =1. The condition is also observed as $L_{hyp} < L_{ref}$ which means there is only one deletion error as 'EPPUDU'. Hence the sentence is modified by inserting correct word 'EPPUDU'. After FBSM, the result will be as follows:

FBSM output: **PADHMAAVATHI EKSPRES EPPUDU VASTUNDHI**

Sample-3:

SIASR Output: **YASHWANTHPUR EKSPRES KAAKINAADAKU EPPUDU VASTHUNDHI**

Here $f_1 = 1$, $f_2 = 1$ and $f_3 = 0$ then stop the forward search, thus value of F is 2 ($f_1 + f_2 = 2$). Here $b_1 = 1$, $b_2 = 1$, $b_3 = 0$ then stop the backward search. Thus the value of B is 2 ($b_1 + b_2$).

The success factor = (2+2)/5 = 0.8, the above sentence is selected from the decoder output. The modification factor = (5-4) =1. The condition is also observed as $L_{hyp} = L_{ref}$ which means there is only one substitution error as '**KAAKINAADAKU**'. Hence the sentence is modified by substituting correct word '**KAACHIGOODAKU**'. After FBSM, the result will be as follows:

FBSM output: **YASHWANTHPUR EKSPRES KAACHIGOODAKU EPPUDU VASTHUNDHI**

Sample-4:

SIASR Output: **THIRUMALA EKSPRES EKKADA ELA VELUTHUNDHI**

Here $f_1 = 1$, $f_2 = 1$ and $f_3 = 0$ then stop the forward search, thus value of F is 2 ($f_1 + f_2 = 2$). Here $b_1 = 1$, $b_2 = 0$, then stop the backward search. Thus the value of B is 1. The success factor = 2+1/5= 0.6; this is selected from decoder output. The modification factor is (5-(2+1)) = 2, It denotes more than one error is occurred. So FBSM cannot be applied.

FBSM output: **THIRUMALA EKSPRES EPPUDU VELUTHUNDHI**

7.1. EXPERIMENTAL RESULTS AFTER APPLYING FBSM ON SIASR OUTPUT

Forward and Backward Search Method (FBSM) is applied on the decoder results of eighteen experiments to improve the performance of SIASR system. FBSM is used to reduce substitution errors or deletion errors or insertion errors. SRR, WRR, SER, DER, IER and WER are calculated for each experiment after applying FBSM and these results are tabulated in this section**.** These tabulated results are compared with the results obtained with the modified pronunciation dictionary using PDMM in **Section 6.2** in **Chapter 6**. Considerable improvement is observed in every experiment.

7.1.1. NSR Training

7.1.1.1. FBSM applied on the decoder output of NSR test data

The following **Table 7.1** shows the results after applying FBSM on the decoder output obtained when NSR is tested with NSR training model.

Table 7.1: **FBSM applied on the decoder output of NSR test data (NSR training)**

S.No	Training Data in NSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	985 98.5%	4705 99.3%	25 0.5%	10 0.2%	8 0.2%	43 0.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	49 98%	235 99.2%	0 0%	2 2%	0 0%	2 0.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	99 99%	470 99.2%	0 0%	4 0.8%	0 0%	4 0.8%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	196 98%	941 99.2%	0 0%	7 0.7%	0 0%	7 0.7%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	247 98.8%	1181 99.7%	0 0%	4 0.3%	1 0.1%	5 0.4%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	490 98.0%	2359 99.5%	5 0.2%	6 0.3%	15 0.6%	26 1.1%

The above results are compared with the results mentioned in **Table 6.3** for each test case as follows:

In the first case, the SRR is increased from 97.8% to 98.5%. Here 2 substitution errors and 1 deletion error are reduced. Thus WRR is increased by 0.1%. The WER is reduced by 0.1%.

In the second case, the SRR, WRR remains constant.

In the third case, the SRR is increased by 1%. 1 deletion error is reduced here. Thus the WER is reduced from 1.1% to 0.8%. Thus WRR is improved by 0.3%.

In the fourth case, the SRR is increased by 1%. The WRR is raised by 0.3%. 3 deleted words are reduced. Thus the WER is reduced by 0.4%.

In the fifth case, the SRR improved by 1.6%. The WRR is improved by 0.3%. Deletion errors are reduced from 7 to 4 errors. Insertion errors are decreased from 2 to 1 error. Thus 4 error words are reduced in total. Thus WER is reduced by 0.4%.

In the sixth case, the SRR is increased by 4.6%. The WRR is improved by 0.6%. The number of substitution errors is reduced from 13 to 5 errors. 8 deleted errors are reduced here.7 insertion errors are decreased in this case. Thus the WER is reduced by 1%.

The following **Figure 7.1** shows improvement in sentence recognition before and after applying PDMM and FBSM. SRR improvement is seen in every case with modified dictionary using PDMM. Significant improvement in SRR is noticed in every case when FBSM is applied on the decoder output.

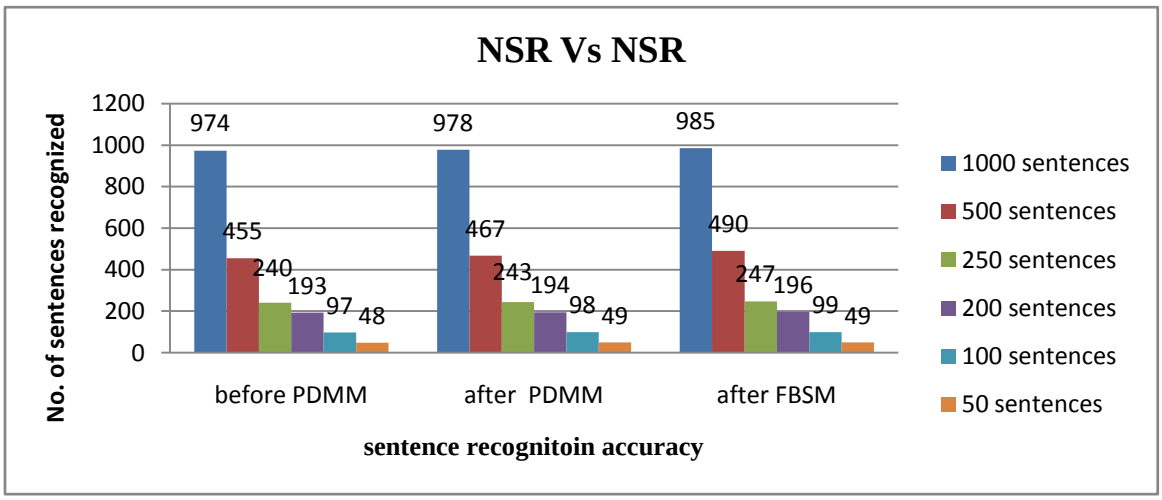

Figure 7.1: Accuracy improvement (before, after PDMM, after FBSM) when NSR is tested with NSR

7.1.1.2. FBSM applied on decoded output of SSR test data

The following **Table 7.2** shows results after applying FBSM on the decoder output obtained when SSR is tested with NSR training model.

Table 7.2: FBSM applied on the decoder output of SSR test data (NSR training)

S.No	Training Data in NSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	982 98.2%	4693 99.0%	39 0.8%	8 0.2%	11 0.2%	58 1.2%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	47 94%	231 97.5%	6 2.5%	0 0.0%	3 1.3%	9 3.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	98 98%	469 98.9%	5 1.1%	0 0.0%	2 0.4%	7 1.5%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	195 97.5%	938 98.9%	9 0.9%	1 0.1%	4 0.4%	14 1.5%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	242 96.8%	1165 98.3%	14 1.2%	6 0.5%	0 0.0%	20 1.7%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	471 94.2%	2308 97.4%	49 2.1%	13 0.5%	17 0.7%	79 3.3%

The above results are compared with results mentioned in **Table 6.4** for each test case as follows:

In the first case, the SRR is increased from 97.6% to 98.2%. The WRR is improved from 98.9% to 99%. 2 substitution errors and 4 insertion errors are reduced in this case. Thus 6 error words are reduced. The WER is reduced from 1.4% to 1.2%.

In the second case, 2% improvement is observed in SRR. Thus WRR is improved from 97% to 97.5%. Only one substitution error is reduced in this case. WER is reduced by 0.4%.

In the third case, the SRR is increased by 4%. The substitution errors are reduced from 8 to 5 errors and deletion errors are reduced by 1. As the substitution errors and deletion errors are reduced, WER is reduced by 0.8%. Thus WRR is increased from 98.1% to 98.9%.

In the fourth case, The SRR is improved by 1.5%. The WRR is increased from 98.6% to 98.9%. The substitution errors are decreased from 11 to 9 and only 1 deletion error is decreased in this case. As the substitution errors and deletion errors are reduced, WER is reduced by 0.3%.

In the fifth case, the SRR increased by 0.4%. The WRR is improved from 98.2% to 98.3%. Here only 1 deletion error reduced with this method. Hence the WER is reduced by 0.1%.

In the sixth case, the SRR is increased by 3%. The WRR is increased from 96.2% to 96.8%. 12 substitution errors, 2 deletions and 1 insertion error are reduced in this case. Hence the total errors are reduced from 94 to 79. Thus the WER is reduced by 0.7%.

The following F**igure 7.2** shows SRR when SSR is tested with NSR training model. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

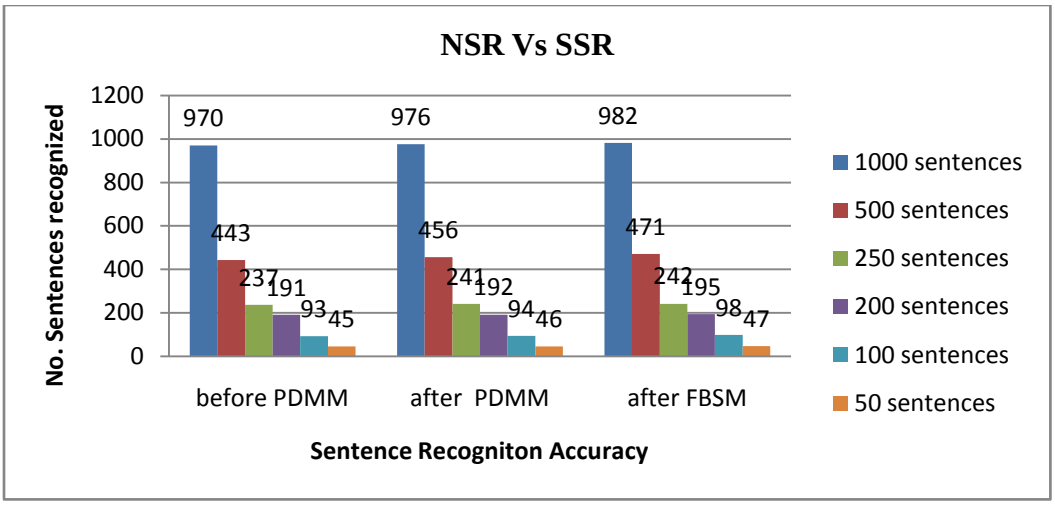

Figure 7.2: Accuracy improvement (before, after PDMM, after FBSM) when SSR is tested with NSR

7.1.1.3. FBSM applied on FSR decoder output

The following **Table 7.3** shows results after applying FBSM on the decoder output obtained when SSR is testes with NSR.

Table 7.3: FBSM applied on the decoder output of FSR test data (NSR training)

S.No	Training Data in NSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	958 95.8%	4654 98.2%	64 1.4%	22 0.5%	11 0.2%	97 2.0%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	44 88%	220 92.8%	9 3.8%	8 3.4%	1 0.4%	18 7.6%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	87 87%	452 95.4%	17 3.6%	5 1.1%	1 0.2%	23 4.9%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	185 92.5%	914 96.4%	22 2.3%	12 1.3%	2 0.2%	36 3.8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	233 93.2%	1143 96.5%	25 2.1%	17 1.4%	2 0.2%	44 3.7%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	445 89.0%	2240 94.5%	96 4.1%	34 1.4%	3 0.1%	133 5.6%

The above results are compared with results mentioned in **Table 6.5** for each test case as follows:

In the first case, the SRR is increased from 92.5% to 95.8%. Thus SRR is enhanced by 3.3%. The substitution errors are decreased from 91 to 64 errors. The deletion errors are reduced from 28 to 22 errors. Thus the total errors are reduced from 130 to 97 errors. With this reduction, the WER is reduced by 0.7%. Thus the WRR is enhanced by 0.7%.

In the second case, the SRR is improved from 82% to 88%. Thus SRR is raised by 6%. 4 substitution errors are decreased in this case. The WER is reduced from 9.3% to 7.6%. Thus the WRR is improved from 91.1% to 92.8% in this case.

In the third case, the SRR is increased from 79% to 87%. Thus 8% improvement in SRR is noticed. 6 substitution errors and 2 deletion errors are reduced here. As the substitution and deletion errors are reduced, WER is reduced by 1.7%. Thus WRR is enhanced by 1.7%.

In the fourth case, the SRR is improved by 6%. 10 substitution errors and 2 deletion errors are reduced in this case. With this reduction of error words, WER is reduced by 1.3%. Thus WRR is enhanced by 1.3%.

In the fifth case, the SRR is raised by 6.8%. 12 substitutions errors and 5 deletion errors are decreased in this case. Thus 17 error words are reduced here. The WER is reduced by 1.5%. The WRR is increased by 1.5%.

In the sixth case, the SRR is enhanced from 79.4% to 89%. Thus 9.6% of SRR improvement is found in this case. 45 substitution errors and 4 deletion errors are reduced in this case. Thus 49 error words are reduced. The WER is reduced by 2.1%.

The following **Figure 7.6** shows SRR when FSR is tested with NSR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

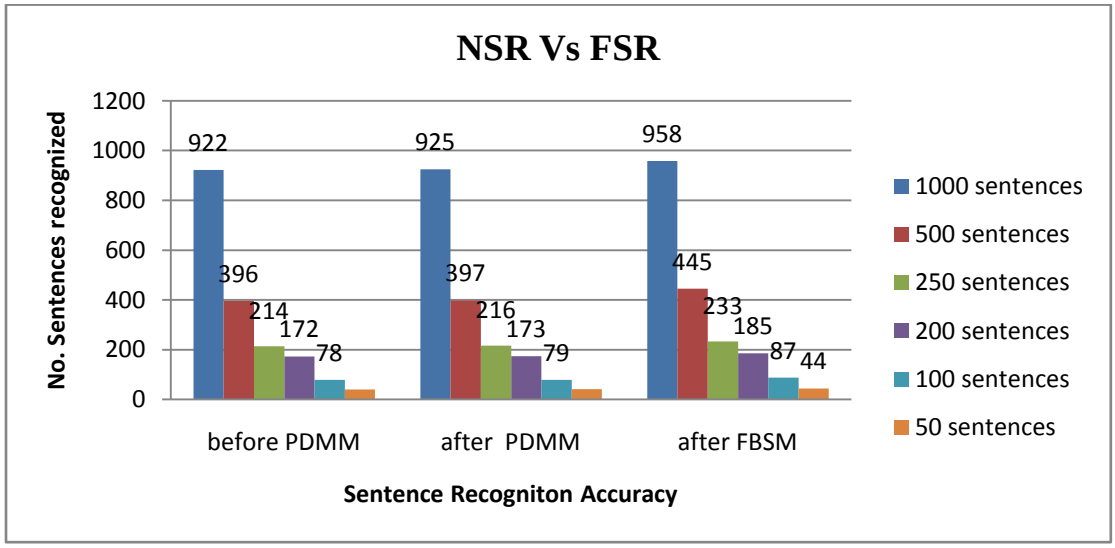

Figure 7.3: Accuracy improvement (before, after PDMM, after FBSM) when FSR is tested with NSR

The following observations are noticed from the above results mentioned in **Tables 7.1**, **7.2** and **7.3:**

(i) S>D>I in 1 case; S<I<D in 4 cases; S<D<I in 1 case when FBSM applied on the decoder output obtained when NSR is tested with NSR

(ii) S>I>D in 2 cases; S>I>D in 2 cases; S>I>D in 1 case and S>D>I in 1 case when FBSM applied on the decoder output obtained when SSR is tested with NSR

(iii) S>D>I in all cases when FBSM is applied on the decoder output obtained when FSR is tested with NSR.

7.1.2. SSR Training

7.1.2.1. FBSM applied on NSR decoder output

The following **Table 7.4** shows results after applying FBSM on the decoder output obtained when NSR is testes with SSR.

Table 7.4: FBSM applied on the decoder output of NSR test data (SSR training)

S.No	Training Data in SSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	976 97.6%	4678 98.7%	44 0.9%	18 0.4%	6 0.1%	68 1.4%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	46 92.0%	226 95.4%	8 3.4%	3 1.3%	5 2.1%	16 6.8%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	94 94.0%	464 97.9%	8 1.7%	2 0.4%	5 1.1%	15 3.2%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	195 97.5%	937 98.8%	8 0.8%	3 0.3%	6 0.6%	17 1.8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	240 96.0%	1163 98.1%	13 1.1%	9 0.8%	8 0.7%	30 2.5%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	475 95.0%	2300 97.0%	46 1.9%	24 1.0%	17 0.7%	87 3.7%

The above results are compared with the results mentioned in **Table 6.6** for each test case as follows:

In the first case, the SRR is improved from 95.8% to 97.6%. Thus the improvement in SRR is 1.8%. 12 substitution errors and 2 deletion errors are reduced in this case. As the substitution and deletion errors are, the WRR is improved by 0.3%. The WER is reduced by 0.4%.

In the second case, the SRR increased by 2%. The SRR improvement is noticed with the reduction of 1 substitution error. The improvement of WRR is 0.5%. The WER is reduced by 0.5%.

In the third case, the SRR increases from 84% to 94%. Thus the improvement in SRR is 10%. 9 substitution errors and 1 deletion error are reduced in this case. So the total errors are reduced by 10 error words. Thus the WER is reduced from 5.3% to 3.2%. The WRR is raised by 2.1%.

In the fourth case, the SRR is increased by 6%. Here 10 substitution errors, 1 deletion error and 1 insertion error are reduced. Hence the WER is reduced by 1.3%. Similarly the WRR is improved by 1.1%.

In the fifth case, the improvement in SRR is 5.6%. Here 9 substitution errors, 1 deletion error and 1 insertion error are reduced. Thus the total errors reduced to 41 to 30. Thus WER is reduced by 1%. The WRR improvement is 0.8%.

In the sixth case, the SRR increased from 87% to 95%. Hence the improvement in SRR is 8%. 35 substitutions, 2 deletions and 3 insertions are reduced in this case. Thus 40 error words are decreased. As the total errors reduced, WER is reduced by 1.7%. The WRR is improved by 1.5%.

The following **Figure 7.4** shows SRR when NSR is tested with SSR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

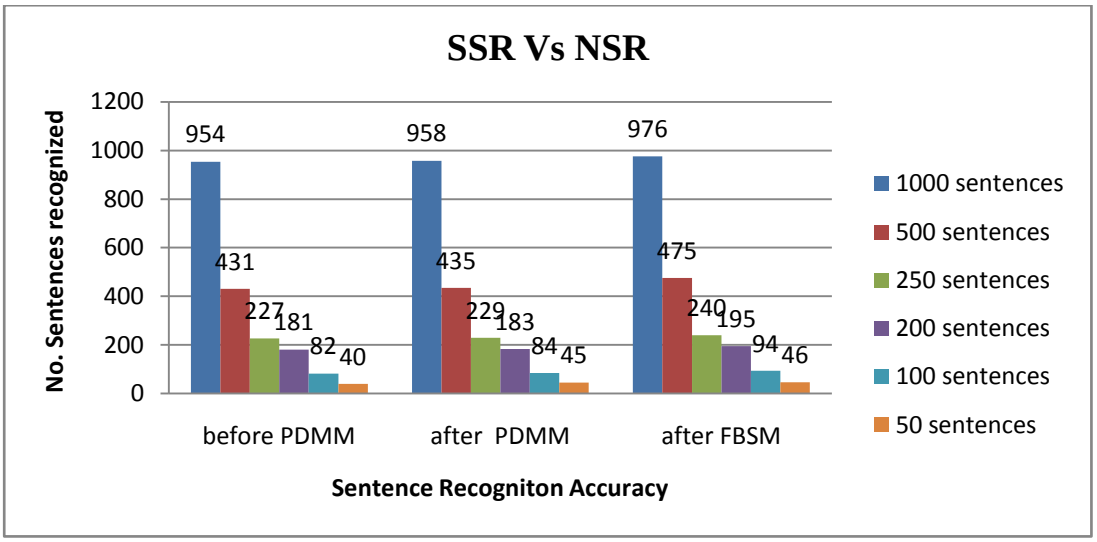

Figure 7.4: Accuracy improvement (before, after PDMM & after FBSM) when NSR is tested with SSR

7.1.2.2. FBSM applied on SSR decoder output

The following **Table 7.5** shows results after applying FBSM on the decoder output obtained when SSR is testes with SSR.

Table 7.5: FBSM applied on the decoder output of SSR test data (SSR training)

S.No	Training Data in SSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	986 98.6%	4706 99.3%	30 0.6%	4 0.1%	10 0.2%	44 0.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	47 94%	230 97%	6 2.5%	1 0.4%	1 0.4%	8 3.3%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	98 98%	469 98.9%	5 1.1%	0 0.0%	1 0.2%	6 1.3%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	197 98.5%	939 99.1%	9 0.9%	0 0.0%	3 0.3%	12 1.3%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	244 97.6%	1169 98.6%	16 1.4%	0 0.0%	1 0.1%	17 1.4%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	474 94.8%	2303 97.2%	42 1.8%	25 1.1%	9 0.4%	76 3.2%

The above results are compared with the results mentioned in **Table 6.7** for each test case as follows:

In the first case, the SRR increased by 1.2%. The substitution errors are decreased from 35 to 30 errors. The deletion errors are decreased from 8 to 4 words. 3 insertion errors are decreased in this case. As the errors are decreased, WER is reduced by 0.2%. Thus WRR is increased from 99.1% to 99.3%.

In the second case, the reduction of 1 substitution errors increased SRR to 2%. The WRR is raised by 0.4%. It has been observed that the WER is reduced by 0.5%.

In the third case, 2 substitution errors and 1 insertion error are decreased. With this reduction, SRR is enhanced by 3% and WRR is raised by 0.4%. Here the reduction of WER is noticed as 0.6%.

In the fourth case, 4 substitution errors and 2 insertion errors are decreased. Thus the SRR is raised by 3%. The WRR is increased from 98.6% to 99.1%. Thus WER is reduced by 0.6%.

In the fifth case, 5 substitution errors and 2 insertion errors are decreased. With this reduction of errors, SRR is enhanced by 2.4%. Thus WRR is improved by 0.4%. WER is reduced by 0.6%.

In the sixth case, 28 substitutions, 2 deletions, 10 insertions are reduced. With this reduction, SRR is improved from 87.2% to 94.8%. Thus SRR is raised by 7.6%. The WRR is increased from 95.9% to 97.2%. Thus the WRR is raised by 1.3%. The WER is reduced by 1.7%.

The following **Figure 7.5** shows SRR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

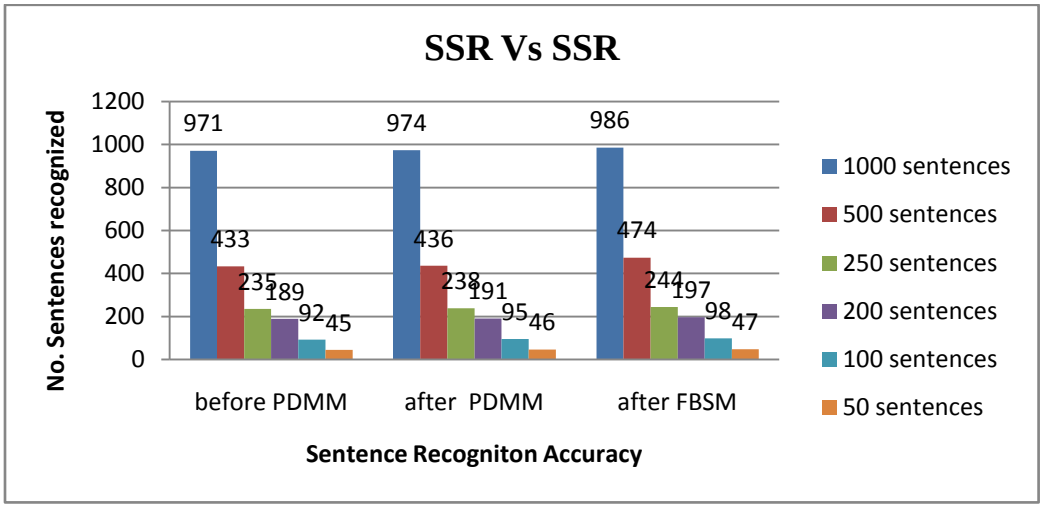

Figure 7.5: Accuracy improvement (before, after PDMM, after FBSM) when SSR is tested with SSR

7.1.2.3. FBSM applied on FSR decoder output

The following **Table 7.6** shows results after applying FBSM on the decoder output obtained when SSR is testes with SSR training model.

Table 7.6: FBSM applied on the decoder output of SSR test data (SSR training)

S.No	Training Data in SSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	950 95.0%	4627 97.6%	82 1.7%	31 0.7%	5 0.1%	118 2.5%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	41 82.0%	215 90.7%	16 6.8%	6 2..5%	2 0.8%	24 10.1%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	88 88.0%	448 94.5%	19 4.0%	7 1.5%	2 0.4%	28 5.9%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	180 90.0%	895 94.4%	33 3.5%	20 2.1%	1 0.1%	54 5.7%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	230 92.0%	1134 95.7%	35 3.0%	16 1.4%	1 0.1%	52 4.4%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	427 85.4%	2180 92.0%	128 5.4%	62 2.6%	5 0.2%	195 8.2%

The above results are compared with the results mentioned in **Table 6.8** for each test case as follows:

In the first case, 50 substitution errors, 7 deletion errors are reduced. With this reduction, the SRR is increased from 89.2% to 95%. Thus SRR is enhanced by 5.8%. The WRR is improved by 1.2% and WER is reduced by 1.2%.

In the second case, the reduction of 1 substitution and 1 deletion error improved SRR by 4%. Thus SRR is increased from 78% to 82%. The WRR is improved by 0.9%. The WER is reduced by 0.9%.

In the third case, the reduction of 12 substitutions and 3 deletions improved SRR by 15%. The WRR is increased by 0.7%. The total number of errors reduced from 43 to 28 errors. The WER is decreased from 9.1% to 5.9%. Thus the WER is reduced by 3.2%.

In the fourth case, the reduction of 17 substitutions and 3 deletions improved SRR by 9.5%. Thus the SRR is increased from 80.5% to 90%. The total errors are decreased by 20. The WRR is improved by 2.1% and WER is reduced by 2.1%.

In the fifth case, the reduction of 27 substitutions and 5 deletion errors improved SRR from 79.6% to 92%. Thus SRR is improved by 12.4%. Thus the improvement of WRR is 2.7% and WER is reduced by 2.6%.

In the sixth case, the reduction of 38 substitutions, 8 deletions and 1 insertion errors improved SRR from 71.8% to 85.4%. Thus the SRR is improved by 13.6%. The WRR is improved by 3% and WER is reduced by 3%.

The following observations are noticed from the above **Table 7.4, Table 7.5** and **Table 7.6:**

(i) S>D>I in 3 cases; s>I>D in 3 cases when FBSM is applied on the decoder output obtained when NSR is tested with SSR.

(ii) S>I>D in 5 cases; S>D>I in 1 cases when FBSM is applied on the decoder output obtained when SSR is tested with SSR.

(iii) S>D>I in all cases when FBSM is applied on the decoder output obtained when FSR is tested with SSR.

The following **Figure 7.6** shows SRR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

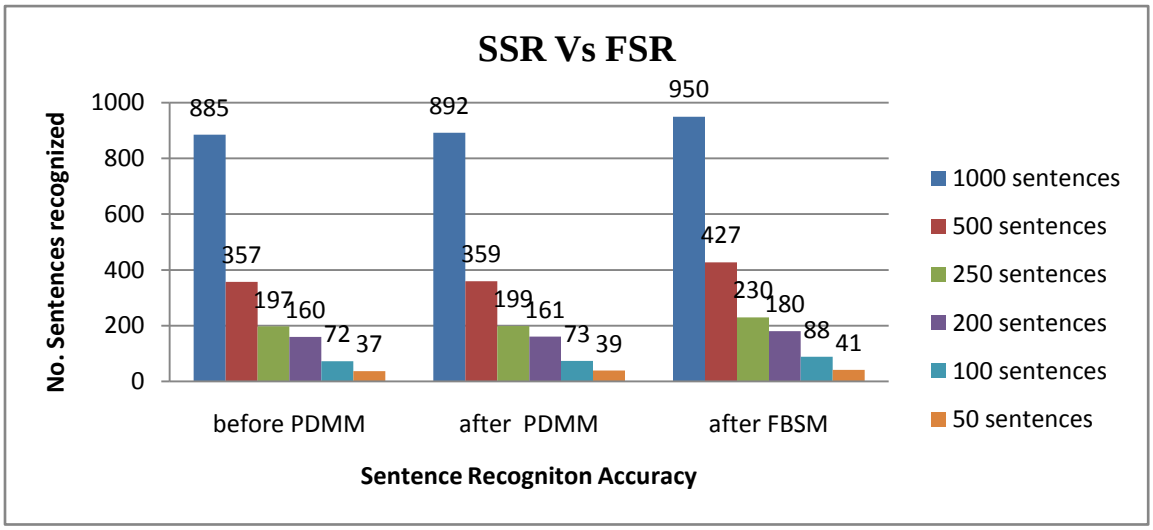

Figure 7.6: Accuracy improvement (before, after PDMM, after FBSM) when FSR tested with SSR

7.1.3. FSR Training

7.1.3.1. FBSM applied on NSR decoder output

The following **Table 7.7** shows results after applying FBSM on the decoder output obtained when NSR is testes with FSR training model.

Table 7.7: FBSM applied on the decoder output of NSR test data (FSR training)

S.No	Training Data in FSR			Test Data in NSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	981 98.1%	4706 99.3%	29 0.6%	5 0.1%	8 0.2%	42 0.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	48 96%	233 98.3%	3 1.3%	1 0.4%	1 0.4%	5 2.1%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	98 98.0%	469 98.9%	5 1.1%	0 0%	4 0.8%	9 1.9%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	197 98.5%	939 99.1%	9 0.9%	0 0%	5 0.5%	14 1.5%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	246 98.4%	1173 99.0%	10 0.8%	2 0.2%	4 0.3%	16 1.4%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	480 96%	2329 98.3%	32 1.4%	9 0.4%	13 0.5%	54 2.3%

The above results are compared with results mentioned in **Table 6.9** for each test case as follows:

In the first case, 3 substitutions, 1 deletion and 2 insertion errors are decreased. As errors are reduced, SRR improved from 97.8% to 98.1%. Thus the improvement in SRR is 0.3%. The WRR is improved by 0.1% and WER is reduced by 0.1%.

In the second case, the reduction of 1 substitution error improved SRR by 2%. Thus the WRR is improved by 0.4%. The reduction of WER is 0.4% is seen in this case.

In the third case, the reduction of 5 substitutions, 2 deletions and 1 insertion error improved SRR from 90% to 98%. Thus SRR is raised by 8%. Thus the WRR is improved by 1.4% and WER is reduced by 1.5%.

In the fourth case, 3 substitutions reduced which improved SRR by 1.5%. Thus WRR is improved by 0.4% and WER is reduced by 0.3%.

In the fifth case, 8 substitutions and 1 deletion errors are reduced here. Thus total errors are reduced by 9. With this reduction, SRR is raised by 0.5%. The WRR improvement is 0.6% and WER is reduced by 0.5%.

In the sixth case, 19 substitutions, 4 deletions and 9 insertion errors are reduced here. Thus 32 errors are decreased in this case. As the errors are reduced, SRR is increased from 89.6% to 96%. Thus SRR is improved by 6.4%. The WRR improvement is 1%. The WER is reduced from 3.6% to 2.3%.

The following **Figure 7.7** shows SRR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

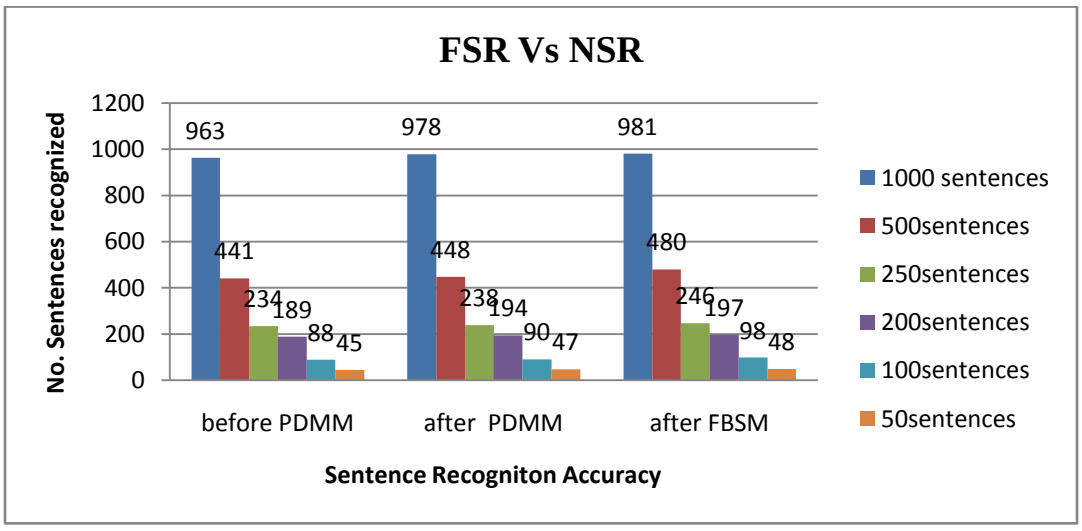

Figure 7.7: Accuracy improvement (before, after PDMM, after FBSM) when NSR is tested with FSR

7.1.3.2. FBSM is applied on SSR decoder output

The following **Table 7.8** shows results after applying FBSM on the decoder output obtained when NSR is testes with FSR.

Table 7.8: FBSM applied on the decoder output of SSR test data (FSR training)

S.No	Training Data in FSR			Test Data in SSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	983 98.3%	4704 99.2%	30 0.6%	6 0.1%	9 0.2%	45 0.9%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	47 94%	230 97%	5 2.1%	2 0.8%	1 0.4%	8 3.3%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	93 93%	463 97.7%	8 1.7%	3 0.6%	1 0.2%	12 2.5%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	194 97%	937 98.8%	7 0.7%	4 0.4%	1 0.1%	12 1.3%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	240 96%	1165 98.3%	15 1.3%	5 0.4%	1 0.1%	21 1.8%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	471 94.2%	2311 97.5%	49 2.1%	10 0.4%	10 0.4%	69 2.9%

The above results are compared with the results mentioned in **Table 6.10** for each test case as follows**:**

In the first case, 5 substitutions, 6 deletions and 4 insertions are reduced. Thus the WRR is raised by 0.2%. The SRR is increased from 97% to 98.3%. Thus SRR is enhanced by 1.3%. Thus the WER is reduced by 0.2%.

In the second case, the reduction of 1 substitution error raised SRR by 2%. Thus the SRR is increased from 92% to 94%. The WRR is raised by 0.4%. The reduction of WER is 0.5%.

In the third case, 4 substitutions, 2 deletions and 2 insertion errors are reduced. As the errors are reduced, SRR is enhanced by 7%. Thus SRR is increased from 86% to 93%. Thus the reduction of WER is 1.7%.

In the fourth case, the SRR is raised by 4.5%. Thus the SRR is increased from 92.5% to 97%. The WRR is raised by 0.9%. 5 substitutions and 3 deletions are decreased in this case. The WER is reduced by 0.9%.

In the fifth case, the reduction of 8 substitutions and 3 deletions improved SRR by 4.4%. Thus the SRR is increased from 91.6% to 96%. The WRR is increased by 0.9%. The WER is reduced by 0.9%.

In the sixth case, the SRR is raised from 88.6% to 94.2%. 19 substitutions, 7 deletions and 3 insertion errors are decreased. Thus SRR is improved by 5.6%. The WRR is increased from 96.4% to 97.5%. Thus WRR is raised by 1.1%. The WER is reduced by 1.2%.

The following **Figure 7.8** shows SRR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. Significant improvement is noticed in SRR when FBSM is applied on decoder output.

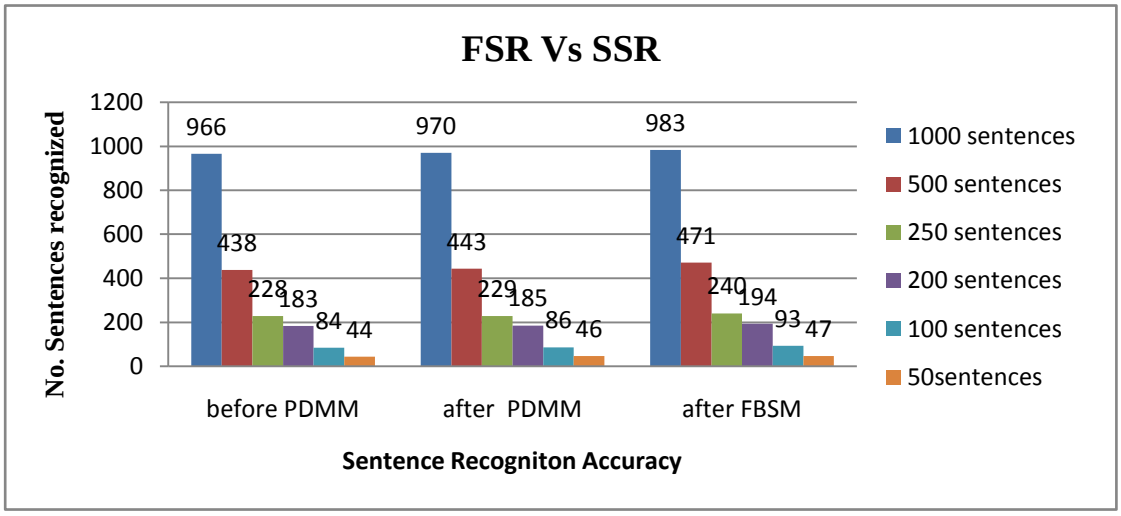

Figure 7.8: Accuracy improvement (before, after PDMM, after FBSM) when SSR is tested with FSR

7.1.3.3. FBSM is applied on FSR decoder output

The following **Table 7.9** shows results after applying FBSM on the decoder output obtained when NSR is testes with FSR training model

Table 7.9: FBSM applied on the decoder output of SSR test data (FSR training)

S. No	Training Data in FSR			Test Data in FSR			Recognition Accuracy		Errors Obtained			Total errors & WER (%)
	No. of Speakers (M&F)	No. of sents	No. of words	No. of speakers (M&F)	No. of sents (M)	No. of words (N)	Corr-sents & SRR (%)	C & WRR (%)	S & SER (%)	D & DER (%)	I & IER (%)	
1	20 (10M +10F)	1000	4740	20 (10M+10F)	1000	4740	984 98.4%	4707 99.3%	25 0.5%	8 0.2%	5 0.1%	38 0.8%
2	19 (10M +9F)	950	4503	1 (1F)	50	237	49 98%	234 98.7%	3 1.3%	0 0%	3 1.3%	6 2.5%
3	18 (9M +9F)	900	4266	2 (1M +1F)	100	474	98 98%	470 99.2%	4 0.8%	0 0.0%	2 0.4%	6 1.3%
4	16 (8M +8F)	800	3792	4 (2M +2F)	200	948	197 98.5%	942 99.4%	5 0.5%	1 0.1%	2 0.2%	8 0.8%
5	15 (7M +8F)	750	3555	5 (3M +2F)	250	1185	245 98.0%	1173 99%	12 1.0%	0 0.0%	4 0.3%	16 1.4%
6	10 (4M +6F)	500	2370	10 (6M +4F)	500	2370	483 96.6%	2334 98.5%	31 1.3%	5 0.2%	17 0.7%	53 2.2%

The above results are compared with the results mentioned in **Table 6.11** for each test case as follows:

In the first case, 4 substitutions, 3 deletions and 6 insertion errors are decreased. As the errors are decreased, SRR is increased by 0.4%. The WER reduced by 0.1%.

In the second case, the reduction of 1 substitution error enhanced SRR by 2%. Thus SRR is increased from 96% to 98%. The WRR is increased by 0.4% and WER is reduced by 0.5%.

In the third case, 1 substitution, 1 deletion and 1 insertion error are reduced. As the errors are reduced, SRR is increased from 95% to 98%. Thus SRR is raised by 3%. The WRR is raised from 98.7% to 99.2%. The WER is reduced by 0.6%.

In the fourth case, the SRR improvement is 1.5%. Thus the SRR is increased from 97% to 98.5%. The WRR is raised by 0.2%. 2 substitution errors and 1 insertion error are reduced in this case. Thus WER is reduced by 0.4%.

In the fifth case, 2 substitutions, 2 deletion errors and 1 insertion errors are reduced. As the errors are decreased, SRR is raised from 96.4% to 98%. Thus SRR is raised by 1.6%. The WRR is enhanced by 0.4%. Here WER is reduced by 0.4%.

In the sixth case, the reduction of 14 substitutions, 5 deletions and 5 insertions improved SRR by 4.8%. Thus SRR is increased from 91.8% to 96.6%. The WRR is raised by 0.8%. The total errors are reduced from 77 to 53 words. Thus WER is reduced by 1%.

Form all the above results, the following observations are noticed:

(i) S>I≥D in all cases when FBSM is applied on the decoder output obtained when NSR is tested with FSR

(ii) S>I>D in 2 cases; S>D>I in 4 cases when FBSM is applied on the decoder output obtained when SSR is tested with FSR

(iii) S>I>D in 5 cases; S>D>I in 1 case when FBSM is applied on the decoder output obtained when FSR is tested with FSR.

The following **Figure 7.9** shows SRR. Considerable improvement in SRR is observed with the modified dictionary using PDMM. SRR is improved after applying FBSM on the decoder output.

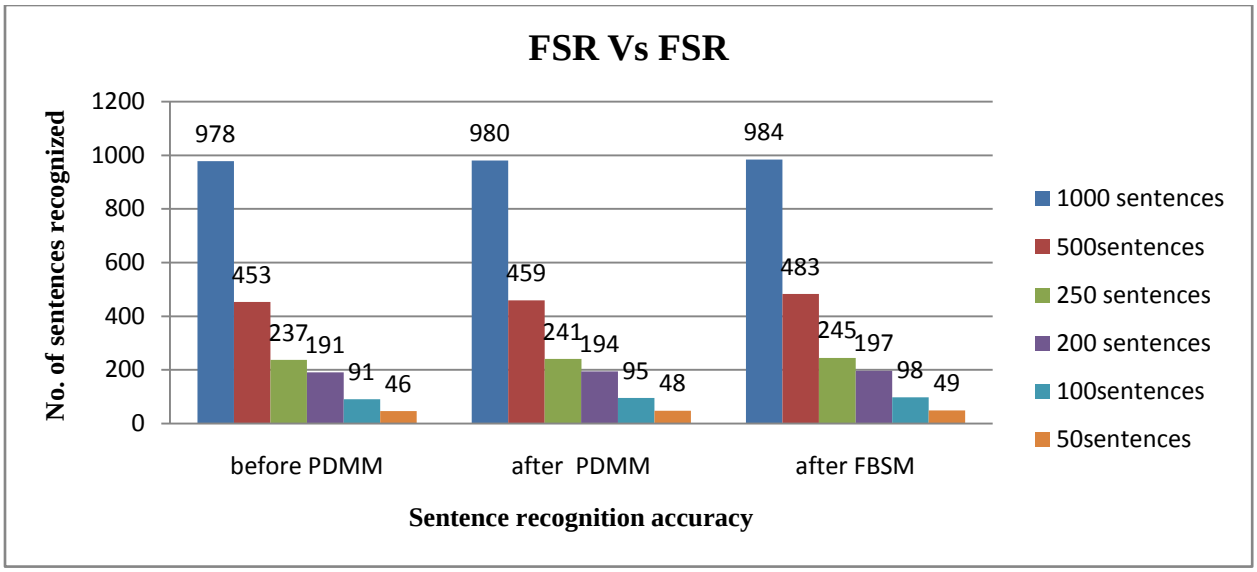

Figure 7.9: Accuracy improvement (before, after PDMM, after FBSM) when FSR is tested with FSR

7.2. COMPARISON OF PRESENT SIASR SYSTEM WITH SOME OF THE EXISTING SIASR SYSTEMS

The aim of this section is to compare the present SIASR system with some of the existing SIASR systems. The results of the present research work in some cases are compared with the results already provided in some existing SIASR systems. The following are the comparisons:

In the present work voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 sentences in NSR are used for training and the same sentences are used for testing. 1000 NSR sentences consisting of 4740 words are used for training and same sentences are used for testing. 99.1% of WRR is observed in this case. Also it is observed for the FSR. Voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 NSR sentences are used for training and voices of 20 speakers (10 male speakers and 10 female speakers) with FSR sentences are used for testing. 1000 NSR sentences consisting of 4740 words are used in training. 1000 FSR sentences consisting of 4740 words are used for testing. 99.2% of SRR is observed in the case of FSR training

and testing. This work is compared with work done by Ki-Seung Lee **[17].** In their work, 10-digit mobile phone numbers pronounced in Korean by 16 speakers (8 speakers for training and 8 speakers for testing). Two sets are developed for training and testing. Each set consisting of 1600 utterances in which 19200 words are present. Recognition accuracy was verified for normal and fast separately as HMM-N and HMM-F, later the incoming utterances are modified according to the class developed. Overall error rate is reduced by 17.8%.

One of the cases in present SIASR system is taken for comparing purpose. In this case, voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 sentences in NSR are used for training and the remaining voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 sentences in NSR are used for testing. 2370 words in 500 NSR sentences are used for training and the remaining 2370 words are used for testing. Here WER is 3.9%. The WER is reduced to 2.1% with modified dictionary using PDMM. Further the WER is reduced to 1.1% using FBSM. These results are compared with the work done by Gopala Krishna and et al **[67].** They developed SIASR for Telugu, Tamil and Marathi languages has developed with the CIIL corpus**.** The corpora consist of speech spoken by landline, cell phone with different age groups from 18 years to above 60 years. The WER is 30.3% for Telugu, 27.9% for Tamil and 27.3% for Marathi. They applied forced alignment in which inserting silences in the training transcription and retraining the models. The new transcriptions have silence markers at appropriate places, the intermediate tie states and hence the overall models are improved. Thus WER is reduced to 28% for Telugu, 20.2% for Tamil and 23.2%.

Another case from the present Telugu SIASR system is taken for comparing the results with other SIASR system. Voices of 15 speakers (7 Male speakers and 8 Female speakers) with 750 sentences in NSR are used for training and remaining voices of 5 speakers (3 Male speakers and 2 Female speakers) with 250 sentences in NSR are used for testing. 750 NSR sentences consisting of 3555 words are used for training. 250 NSR sentences consisting of 1185 words are used for testing. The WRR is 98.1% and WER is 1.9%. These results are compared with the work done by Gaurab and et al. **[68].** They used Hindi language for SIASR system. 30 speakers (12 females and 18 males) recorded

relating to geometry to children. 43 sentences are collected from each speaker. Totally 1806 utterances spoken by 30 speakers are prepared for training. 8 speakers (4 males and 4 females) are used for testing. The sentence recognition accuracy is 68.56 and word recognition accuracy is 88.81% for Hindi language.

One more case in present SIASR system is taken for comparison purpose. Telugu language is used for the corpus collection. Voices of 10 speakers (4 Male speakers and 6 Female speakers) with 500 sentences in NSR are used for training and the remaining voices of 10 speakers (6 Male speakers and 4 Female speakers) with 500 sentences in NSR are used for testing. 2370 words in 500 NSR sentences are used for training and the remaining 2370 words are used for testing. The WRR is 96.8% and WER is 3.9%. The percentage of substitution errors is 0.7%, deletion errors percentage is 0.2% and insertion errors is 1.1%. B. Das and et al. has done work on the Bengali language. They developed SIASR for Bengali language with young (30 male + 20 female) group belongs 20 to 40 years of age **[69]**. 10 young speakers are used for testing. Training models are developed for both young. Testing is performed on young. The accuracy of young with young is accuracy is 85.3% with 12.1% of substitution errors, 2.6% of deletions and 0.7% of Insertion errors. There are other combinations such as young with old and old with young and old with old are not considered for the present comparison.

WiQuas Ghai and Navdeep **[70]** Singh has done three phases of experiments. 9 speakers (5 male + 4 female) recorded 100 sentences in Punjabi language are used for training. In the first phase of their work, 86 sentences are recognized correctly out of 102 sentences. 665 words out of 698 words are correctly recognized. This phase is compared with one of the cases in present work. In present work, voices of 18 speakers (9 Male speakers and 9 Female speakers) with 900 sentences in NSR are used for training and remaining voices of 2 speakers (1 Male speaker and 1 Female speaker) with 100 sentences in NSR are used for testing. 900 NSR sentences consisting of 4266 words are used for training. 100 NSR sentences consisting of 474 words are used for testing. 97 sentences recognized correctly out of 100 sentences. 467 words out of 474 words recognized correctly in this case.

In the second phase of their work, 90 sentences are used for testing. 74 sentences are correctly recognized out of 90 sentences which are different from training set. 522 words out of 558 words in 90 sentences are recognized correctly. This phase is compared with the other case of present work. This phase is compared with one of the phase in present work. Voices of 16 speakers (8 Male speakers and 8 Female speakers) with 800 sentences in NSR are used for training and remaining voices of 4 speakers (2 Male speakers and 2 Female speakers) with 200 NSR sentences are used for testing. 800 NSR sentences consisting of 3792 words are used for training. And 200 NSR sentences consisting of 948 words are used for testing. 193 sentences are recognized correctly. 934 words are recognized correctly.

In the third phase of their work, 40 sentences are used for testing. 32 sentences are correctly recognized. 224 words out of 240 words are correctly recognized. This is compared with present work. In the present work, voices of 20 speakers (10 male speakers and 10 female speakers) with 1000 sentences in NSR are used for training and the same sentences are used for testing. 1000 NSR sentences consisting of 4740 words are used for training and same sentences are used for testing. 48 sentences are recognized correctly. 234 words are recognized correctly.

Matthew A. Stegler and Richard M.Stern **[71]** has done work relating to the speech rate. In their work**,** Wall Street Journal (WSJ) corpus containing 20000 words are used for training and 100 utterances of different speech rates are used for testing. Pronunciation dictionaries are modified by inserting the space between the words in the transcription due deletion errors which occur more in fast speech due to the compressed nature of fast speech. But in the present work, dictionary is modified with PDMM. In the present work, voices of 20 speakers (10 Males + 10 Females) with 1000 NSR sentences are used for training. 1000 NSR sentences consisting of 4740 words are used for training. Voices of 20 speakers with 1000 FSR sentences are used for testing. 1000 FSR sentences consisting of 4740 words are used for testing. 97.4% of accuracy is observed with 96 substitution errors, 28 deletion errors, and 11 insertion errors are obtained. The WRR is increased to 97.5% with modified dictionary using PDMM. Here 91 substitution errors,

28 deletion errors and 11 insertion errors are obtained with modified dictionary. Thus substitution errors are reduced from 96 to 91.

Hiroaki Nanjo and Tatsuya Kawahara **[72]** Corpus of Spontaneous Japanese (CSJ) consists of variety of oral presentations at technical conferences and informal monologue talks on given topics. Speech rate variation will be more in this type of data. 224 presentations among 612 presentations and talks of distinct speakers are used for determining 35.5%. 19158 words are used for training and 47896 words are used for testing in normal conditions. The WER obtained in this case is 35.8%. For fast speech, shortening was done on the frame length and shift for spectral analysis. Another way for the fast speech, state-skipping transitions in phone models was taken place. Syllable duration modeling was performed for the fast speech. Insertion penalty was used to suppress the insertion errors in slow speech. Here 2517 utterances are used for the testing which is taken from training set after segmenting into required rate. By applying above mentioned techniques, the WER reduces from 35.8% to 37.1% for fast speech. For the slow speech, the WER reduces from 35.8% to 36.3%. This work is compared with the results obtained in one the present work. In the present work, Time stretching is used to convert the normal speech into slow speech and time compressing is used to convert the normal speech into fast speech. Voice of 20 speakers (10 females + 10 males) of 4740 words is used for training and testing in different rates of speech (Normal, Slow and Fast). 2.8% of WER is obtained when fast speech rate is tested with normal speech rate. 1.5% of WER is obtained when normal speech rate is tested with slow. WER is reduced from 2.8% to 1.1% when fast speech rate is tested with fast speech rate.

In the present work, voices of 15 speakers (7 Male + 8 Female) with 750 NSR sentences are used for training and voices of 5 speakers with 250 FSR sentences (3 Male + 2 Female) are used for testing. 750 NSR sentences consisting of 3555 words are used for training. 250 FSR sentences consisting of 1185 words are used for testing. The recognition accuracy is increased from 94.8% to 98.1% with modified dictionary. The substitution errors are reduced from 40 to 15 words, deletion errors are reduced from 22 to 7 words and insertion errors are reduced from 2 to 0 words. This work is compared with Peng Gang **[73]** work on speaking rate. Speaking rate affects mainly on word

duration. Incorporating word duration using gamma function was used for word modeling. 6000 sentences of Mandarin digits from 1 to 7 uttered by 44 male speakers and 31 female speakers are used for training. 1360 sentences of Mandarin digits from 1 to 7 uttered by 11 male speakers and 6 female speakers. Extension of the basic Viterbi algorithm is used to incorporate word duration models. During decoding, if the syllable of the word recognized with less duration than its actual duration, then penalty was added to the state sequence according to the gamma function, so that correct word will be recognized. The word recognition increases form 97.51% to 98.31%. Deletion errors are reduced from 54 to 28 words, Insertion errors are reduced from 42 to 34 words and substitution errors are reduced from 40 to 30 words.

8. CONCLUSION AND FUTURE WORK

The present research work compared recognition accuracies when training and testing are done at different rates of speech namely NSR, SSR and FSR. It is observed that the WRR is more when same speech rate is involved in training and testing. WER is more when different rates of speech are involved in training and testing. Substitution errors, deletion errors and insertion errors occurred during testing (decoding).

Analysis of errors is made from the experimental results. Existence of more phonetically similar words gets confused. These confusion leads to occur substitution errors. These confusions are reduced by modifying the pronunciation dictionary using PDMM. In PDMM, Levenshtein distance method is applied on the confusion pairs and it is compared with the threshold value. Here the threshold is taken as the ¾ of the length of longest word in the confusion pair. If it satisfies the assumed length, the pronunciation dictionary is modified with the phonetic transcription of confused word. Thus the substitution errors are reduced with the modified dictionary using PDMM. Significant improvement is observed in recognition accuracy with the modified lexicon. This work can be enhanced to develop the lexicon for a particular language at different speech rates.

SRR affects more due to deletion and insertion errors apart from the substitution errors. An error word related to deletion or insertion or substitution in a sentence makes a greater affect on the performance of SIASR system. FBSM is applied on the decoder output to improve the performance of SIASR system. In this method, the success factor is used to select sentences with errors from decoder output and modification factor is used to notice the position of the error. This FBSM successfully reduces error word related to substitution or insertion or deletion. Considerable improvement is observed in SRR. Further this method can be extended by representing knowledge base and applying grammar rules on the decoder output for context independent SIASR system.

References

1. M. Benzeguiba, R.De Mori, O.Deroo, S.Dupont, T.Erbes, D.Jouvet, L.Fissore, P.Laface, A.Mertins, C.Ris, R.Rose, V.Tyagi, C.Wellekens, "Automatic Speech Recognition and Intrinsic Speech Variation", In Proceedings of International Conference on Acoustics, Speech, and Signal Processing - ICASSP, Vol. 5, pp. 1021-1024, 2006.
2. Florian Müller, "Invariant Features and Enhanced Speaker Normalization for Automatic Speech Recognition", Ph.D thesis, University of Lübeck, July 2012.
3. Daniel Elenius and Mats Blomberg, "Comparing speech recognition for adults and children", In the proceeding of FONETIK 2004, Dept. of Linguistics, Stockholm University, XVIIth Swedish Phonetics Conference pp. 156-159, 2004.
4. Noraini Seman and Kamaruzaman Jusoff, "Acoustic Pronunciation Variations Modeling for Standard Malay Speech Recognition", Journal of Computer and Information Science, pp.112- 120, 2008.
5. M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", In the Proceedings of International Journal of Computer Science and Information Security, (IJCSIS), Vol. 6, No. 3, pp: 181-205, 2009.
6. John Kirriemuir, Speech Recognition Technologies, Technical report, TSW 03-03, JISC, 2003.
7. Nivja H. de Jong and Ton Wempe, "Automatic measurement of speech rate in spoken Dutch", ACLC Working Papers, PP.51-60, 2007.
8. Koreman, Jacques, "Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech", Journal of the Acoustic Society of America, 119(1), pp.582-596, 2006.
9. T.Pfau, R.Falthauser and G.Ruske, "A Combination of speaker normalization and speech rate normalization for automatic speech recognition", In Proceedings of ICSLP, Vol 4, pp.362-365, 2000.
10. Eric K.Ringger and James F.Allen, "Robust Error Correction of Continuous Speech Recognition", In Proceedings of the ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, 1997.

11. Llya Oparin, “Language Models for Automatic Speech Recognition of Inflectional Language”, Ph.D thesis, University of West Bohemia, 2008.

12. A.Zgank and Z.Kacic, “Predicting the Acoustic Confusability between Words for a Speech Recognition System using Levenshtein Distance”, IEEE, pp. 81-84, 2012.

13. Matthew A.Seigler, “Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition”, CMU, 1995.

14. Bartosz Ziotko, Suresh Manandhar, Richard C. Wilson, “Analysis of Phonetic Similarities in Wrong Recognitions of the Polish Language”, International Conference on Audio, Language and Image Processing - ICALIP , 2008

15. Steven Greenberg, Shuangyu Chang and Joy Hollenback, “An Introduction to the Diagnostic Evaluation of Switchboard-Corpus Automatic Speech Recognition Systems”, In Proceedings of NIST Speech Transcription Workshop, 2000.

16 M.Richardson, M. Hwang, A. Acero, and X.D. Huang, “Improvements on Speech Recogniton for Fast talkers”, In proceeding of: Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, pp: 411-414, 1999.

17 Ki-Seung Lee, “Robust Recognition of Fast Speech”, IEICE Transactions, Vol E89-uD, No: 8, pp: 2456-2459, 2006.

18 An Overlap-Add Technique based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech”, In Proceedings of International Conference on Acoustic Speech and Signal Processing, pp.554-557, 1993.

19 Mike Demol, Werner Verhelst, Kris Struyve, Piet Verhoeve, “Efficient Non-Uniform Time-Scale of Speech with WSOLA”, Symposium on Computer Assisted Learning, In ICALL, 2004.

20 Morgan N, Fosler E and Mirghafori N, “Speech Recognition using on-line Estimation of Speaking Rate”, In proceeding of EUROSPEECH, pp. 2079-2082, 1997.

21 C. J.van Heerden and E. Barnard, “Speech rate normalization used to improve speaker verification”, Journal of SAIEE, pp. 129-135, 2007.

22 Yukari Hirata, “Effect of speaking rate on the vowel length distinction in Japanese”, Journal of Phonetics, pp. 565-589, 2004.

23 Yukari Hirata and Kimiko Tsukada, “The Effects of Speaking Rates and Vowel Length on Formant Movements in Japanese”, proceedings of the 2003 Texas Linguistics Society Conference, pp: 73-85, 2003.

24 V.R. Rao Gadde, Modeling Word Durations for better Speech Recognition”, In Proceedings of NIST Speech Transcription Workshop, 2000.

25 V.R. Rao Gadde, “Modeling Word Durations”, In Proceedings of International Conference on Spoken Language Processing, Vol 1, pp. 601-604, 2000.

26 Janse, E., “Word perception in fast speech: Artificially time-compressed vs naturally produced fast speech”, Speech Communication 42(2), pp. 155-173, 2004.

27 Mirghafori, N., Fosler, E., Morgan N., “Fast speakers in large vocabulary continuous speech recognition: analysis & antinodes”, In Proceedings of Eurospeech, pp.491-494, 1995.

28 E.Fosler-Lussier, S.Greenberg and N.Morg, “Incorporating Contextual Phonetics into Automatic Speech Recognition”, International Congress of Phonetic Sciences, pp. 1611-1614, 1999.

29 Francois Pellegrino, J.Farinos and J L Rousas, “Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech”, In Proceedings of International Conference on Speech Prosody, 2004.

30 Jing Zheng, Horacia Franco and Andreas Stolcke, "Rate-of-Speech Modeling for Large Vocabulary Conversational Speech Recognition”, In the proceedings of ASR-2000, Automatic Speech Recognition: Challenges for the new Millenium, pp: 145-159, 2000.

31 Keith Vertanen, “Speech and Speech Recognition during Dictation Corrections”, INTERSPEECH, ICSLP, pp. 1890-1893, 2006.

32 Veri Ferdiansyah and Ayu Purwarianti, “Indonesian Automatic Speech Recognition System Using English-Based Acoustic Model”, American Journal of Signal Processing, pp. 60-63, 2012.

33 Marelie Davel and Olga Martirosian, “Pronunciation Dictionary Development in Resource-Scarce Environments”, In Proceedings of INTERSPEECH, pp. 2851-2854, 2009.

34 Teemu Hirisimaki, Mathias Creutz, Vesa Siivola, Mikko Kurima, Sami Virpioja and Janne Pylkkonen, “Unlimited vocabulary speech recognition with morph language models applied to Finnish”, Computer Speech and Language, Vol 20(4), pp. 515-541, 2006.

35 Zheng Chen, Mingjing Li, Kai-Fu Lee, “Discriminative Training on Language Model”, International Conference on Spoken Language Processing, Vol 1, pp. 493-496, 2000.

36 Jun Ogata and Masataka Goto, “Speech Repair: Quick Error Correction just by using selection operation for Speech Input Interfaces”, INTERSPEECH, pp. 133-136, 2005.

37 K.Georgila, A. Tsopanoglou, N.Fakotakis and G.Kokkinakis, “Improved Large Vocabulary Speech Recognition Using Lexical Rules”, In Proceeding of PCHI’01, 2001.

38 Roeland Ordelman, Arjan Van Hessen, Franciska de Jong, “Lexicon Optimization for Dutch Speech Recognition in Spoken Document Retrival”, In the proceeding of Eurospeech 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, pp. 1085-1088, 2001.

39 Arjan Van Leeuwen, “Improving the Language Model for Automatic Speech Recognition in the Dutch Court of Law”, In the proceedings of 8th Twente Student Confernece on IT conference, 2008.

40 Olga Martirosian and Marelie Davel, “Error analysis of a public domain pronunciation dictionary”, In the proceedings of PRASA, 2007.

41 Ming-Yi Tsai, Fu-chiang Chou and Lin-Shan Lee, “Pronunciation Variation Analysis with respect to Various Linguistic Levels and Contextual Conditions for Mandarin Chinese”, In the proceeding of Eurospeech’01, Aalborg , Denmark, pp. 1445-1448, 2001.

42 Irina Kipyatkova,, “Modeling of Pronunciation, Language and NonVerbal Units at Convesational Russian Speech Recognition”, International Journal of Computer

Science and Applications, Technomathematics Research Foundation, Vol. 10, No. 1, pp. 11 – 30, 2013

43 Tetyana Lyudovyk, Valeriy Pylypenko, "Code-Switching Speech Recognition for Closely Related Languages", SLTU-2014, St. Petersburg, Russia, 14-16 May 2014.

44 Pertti Vayrynen, Johannes Peltola and Tapio Seppanen, "Enhancing Phoneme Recognizer performance with a Simple Rule-based Language model", In Proceedings of STeP- Finnish Artificial Intelligence Days, pp. 171-178, 2000.

45 Jan Anguita, Stephane Peillon, Javier Hernando and Alexandre Bramoulle, "Word Confusability Prediction in Automatic Speech Recognition", INTERSPEECH, 2004.

46 Arup Sarma and David D.Palmer, "Context-based Speech Recognition Error Detection and Correction", In proceedings of HLT-NAACL, pp. 85-88, 2004.

47 Kimberly Voll, Stella Atkins and Bruce Forster, "Improving the Utility of Speech through Recognition Error Detection" Journal of Digital Imaging, pp. 371-377, 2008.

48 Lawrence Rabiner and Biling-Hwang Juang, "Fundamentals of Speech Recognition", published by Prentice-Hall International, 1993.

49 Michael Seltzer, "SPHINX III Signal Processing Front End Specification", CMU Speech Group, 31 August 1999.

50 http://sistemic.udea.edu.co/wp-content/uploads/2013/10/introSR.pdf

51 http://www.speech.cs.cmu.edu/cgi-bin/cmudict#phones

52 http://speech.tenet.res.in/wiki/uploads/9/9e/Tutorial2_sphinxtrain.pdf

53 Chiori Hori, "A Study on Statistical Methods for Automatic Speech Summerization", Ph.D thesis, Tokyo Institute of Technology, 2002.

54 Janne Pylkkönen, "Towards efficient and Robust Automatic Speech Recognition: Decoding Techniques and Discriminative Training", Aalto University Publication series Doctoral dissertations, 2013.

55 http://www.fon.hum.uva.nl/praat/

56 http://research.cs.tamu.edu/prism/lectures/sp/l19.pdf

57 http://www-2.cs.cmu.edu/~robust/Tutorial

58 Yongmei Shi, "An Investigation of Linguistic Information for Speech Recognition Error Detection", Ph.D thesis, University of Maryland, Baltimore, 2008.

59 Bo-June(paul) Hsu and James Glass, "Language Model Parameter Estimation using User Transcriptions", ICASSP, pp. 4805-4808, 2009.

60 Mc NKosi, MJD Manamela and Gasela, "Creating a Pronunciation Dictionary for Automatic Speech Recognition – a Morphological approach", In the proceedings of SATNAC, Network Services, 2011.

61 Ibrahim Badr, Ian McGraw, James Glass, "Pronunciation learning from continuous speech", INTERSPEECH, pp. 549-552, 2011.

62 Stefan Benus, Milos Cernak, Milan Rusko, Marian Trnka, Sacia Darjaa, "Adapting Slovak ASR for native Germans speaking Slovak", In the proceedings of EMNLP, pp. 60-64, 2011.

63 Panagiota Karansonou, Francois Yvon, Lori Lamel, "Measuring the confusability of pronunciations in speech recognition", In the proceedings of 9th international workshop on finite state methods and Natural Language Processong , Association for Computational Linguists, pp.107-115, 2011.

64 Santoshi Kaki, Eiichiro Susmita and Hitoshi Iida, "A Method for Correcting Errors in Speech Recognition Using the Statistical Features of Character Co-occurance", Proceedings of 17th International Conference on Computational Linguistics, COLING-ACL, pp. 653-657, 1998.

65 A rup Sarma and David D. Palmer**, "**Context-based speech recognition error detection and correction", In the proceedings of HLT-NAAL-Short Papers, pp. 85-88, 2004.

66 Youssef Bassil and Mohammad Alwani, "Post-Editing Error Correction Algorithm For Speech Recognition using Bing Spelling Suggestion", International Journal of Advanced Computer Science and Applications, (IJACSA) Vol. 3, No.2, 2012.

67 Gopala Krishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R.N.V. Sitaram, S P Kishore "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems",

In the Proceedings of International Conference on Speech and Computer (SPECOM), 2005.

68 Gaurab, Devanesamoni Shakina Deiv, Gopal Krishna Sharma, Mahua Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", Journal of Signal and Information Processing, 394-401, 2012,

69 B. Das, S. Mandal, P. Mitra, "Bengali Speech Corpus for Continuous Automatic Speech Recognition System", The Oriental COCOSDA 2011 International conference on Microelectronics and Information Systems, Research Center, National Chiao Tung University, Hsinchu, Taiwan, Oct 26-28, pp.51-55, 2011.

70 WiQuas Ghai and Navdeep Singh, "Continuous Speech Recognition for Punjabi Language", International Journal of Computer Application, Vol 72-No 14, pp. 23-28, May 2013.

71 Matthew A. Stegler and Richard M.Stern, "On the Effects of Speech Rate in Large Vocabulary Speech Recognition system", In the proceeding of International Conference on Acoustics, Speech and Signal Processing Volume:1 pages 612-615, ICASSP-95, 1995.

72 Hiroaki Nanjo and Tatsuya Kawahara, "Speaking-Rate Dependent Decoding and Adaptation for Spontaneous Lecture Speech Recognition", IEEE, pp: 725-728, 2002.

73 Peng Gang, Zhang Bo and Wang Willian S-Y, "Duration Modeling in Mandarin Connected Digit Recognition", International Symposium on Chinese Spoken Language processing (ISCSLP 2000) Oct 13-15, 2000.

Publications:

1. N.Usha Rani and P.N.Girija, "Reduction of Confusion Pairs on Difference Rates of Speech in Telugu Language", In the Proceedings of 15th International Conference on Advanced Computing Technologies (ICACT-2013), 21-22 September, 2013. **(Scopus Indexing)**

2. N.Usha Rani and P.N.Girija, "Error Analysis to Improve the Speech Recognition Accuracy on Telugu Language", Journal of Sadhana, Indian Academy of Sciences Vol. 37, Part 6, pp. 747–761, December 2012**. (Scopus Indexing)**

3. N.Usha Rani and P.N.Girija, "Error Analysis and Improving the Speech Recognition Accuracy on Telugu Language", In the Proceedings of Third International Conference on Advances in Communication Network and Computing CNC-2012 –, LNICST pp. 301–308, 2012**. (Scopus Indexing).**

4. N. Usha Rani and P.N.Girija, "Analyzing and Correction of Errors to Improve the Speech Recognition Accuracy for Telugu Language", CiiT International Journal of Artificial Intelligent Systems and Machine Learning", June 2011.

5. N.Usha Rani and P.N.Girija, "Statistical Modification Method for Speech Recognition System Results", In the proceedings of International Conference on Advances in Mathematical & Computational Methods, Vol 1, pp.200-203, jan5-7**,** 2011.

6. N.Usha Rani and P.N.Girija, "Effect of Speech Rate on the Speech Recognition Accuracy: A Review", CiiT International Journal of Artificial intelligent Systems and Machine Learning, Vol 2, pp. 251-258, Oct 2010.

7. N.Usha Rani and P.N.Girija, "Modification of Pronunciation Dictionary to Improve the Speech Recognition Accuracy for Telugu Railway Enquiry System", In the proceedings of National Conference on Computing & Communication Technologies, pp. 158-162, 6th and 7th October, 2010**.**