Improving Training Data using Error Analysis of Urdu Speech Recognition System

Submitted by:

Saad Irtza 2009-MS-EE-109

Supervised by: Dr. Sarmad Hussain

Department of Electrical Engineering

University of Engineering and Technology Lahore

Improving Training Data using Error Analysis

Submitted to the faculty of the Electrical Engineering
Department of the University of Engineering and
Technology Lahore
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Electrical Engineering

Approved on -----------

Internal Examiner	External Examiner
Chairman Electrical Engineering Department	Dean Faculty of Electrical Engineering

Department of Electrical Engineering
University of Engineering and Technology Lahore

Declaration

I, Saad Irtza, declare that the work presented in this thesis is my own.

Signed: ____________________

Date: ____________________

Acknowledgments

I would like to express my sincere gratitude to my advisor Dr. Sarmad Hussain for the continuous support of my M.Sc study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my M.Sc study.

I would like to thank Miss Huda Sarfraz who was always willing to help and give her best suggestions.

I graciously thank Dr. Asim Loan for providing me the formats of synopsis and thesis. I am very thankful to Mr. Muhammad Iqbal and Mr. Muhammad Islam for arranging progress and final seminars of my thesis.

I would also like to thank my father and mother. They were always supporting me and encouraging me with their best wishes.

I would also like to thank NICT researchers to provide hands on training and APT to provide funds to attend it.

Dedicated to my family, especially…

to Dad for instilling the importance of higher education;

to Brother for encouragement;

and to Mom for love.

List of Figures

List of Tables

Contents

Abstract

Access to information is vital for development in today's age. However there are several barriers to this for the average Pakistani citizen and also for the visually impaired community in Pakistan. However, literacy rate in Pakistan is very low. According to UNICEF, literacy rate in Pakistan was 60 percent [1]. This leaves about half the population unable to access information that is available in textual form. This problem can be solved by creating an interface between illiterate people and technology so that they can use these facilities. An interface can be created by using automatic speech recognition (ASR). To achieve this goal, speaker independent automatic, continuous and spontaneous speech recognition system and integration to new technologies is required. This approach will bypass the barriers e.g. literacy, language and connectivity that Pakistani citizens face to access the online content. Moreover screen readers are a form of technology useful to people who are blind, visually impaired or illiterate. This technology often works in combination with other technologies, such as speech recognition system, text-to-speech system.

The current work has been done to investigate the issues in read and spontaneous speech recognition system developed in [3]. The word error rate of automatic speech recognition system that had been developed in [3] was 60%. The objective was to investigate the recognition issues. In this context, multiple experiments have been developed. Speech data has been cleaned by using error analysis techniques. Distribution of phonemes and their recognition results have been analyzed. Based on these results, possibility for developing minimally balanced corpus for speech recognition systems has been explored.

Chapter 1- Background and Introduction

The task of Automatic speech recognition (ASR) engine is to convert the speech signal into textual form [2]. This engine can be integrated with many modern technologies to play a vital role in creating a bridge between the Pakistani illiterate communities and online information. This system can evenly be helpful to our blind community and to those who are literate but don't have technical skills to operate information and communication technologies (ICT's). This can also be a challenging task to students to communicate with the robots with their speech rather electrical signals. It can be integrated e.g. 1) with computer commonly known as Human Computer interface 2) with mobile technology to access the information from online sources.

Through spoken dialog systems, a user can access the online information verbally over mobile channel. The information will be translated from any other language to the native language of the user and then converted in the form of speech. This technology will overcome all three barriers such as literacy, language and connectivity. It will serve as a simple and efficient information access interface. It can be equally beneficial for the visually impaired community.

Spoken dialog systems have been developed in a number of different languages for different domain e.g. weather, travel information, flight scheduling and customer support etc. No such system exists in Urdu language so the design of the dialog a system that has been developed in other languages can be used as guideline.

For example, Jupiter has been developed to provide weather forecast system for 500 cities over telephone channel. A user can access the weather information online available of several days. It also provides humidity, sunrise, precipitation, wind speed etc. The user can access this system by calling a toll-free number. Auto receptionist welcomes the user and indicates the free channel by a high tone. After that user can make any weather related query. When user stops making query, the will system play a low tone in order to indicate channel is busy. '*' key can be pressed to interrupt the system.

One of key component in spoken dialog systems is speech recognition engine. Speech recognizer in such systems plays the same role that mind has in human to human communication. A source-channel model is usually used to develop speech recognition systems. The listener's mind decodes the source word sequence W that is delivered by other person. It passes through a noisy communication channel that consists of the speaker's and speech information, also known as audio waveform. Finally, the human mind aims to decode the acoustic signal X into a word sequence ˆ W, which is the original word sequence W [16].

The signal processing module has been used to process the speech signal that extracts features for the decoder. It is used to remove the redundant information from speech signal. The decoder uses acoustic and language models to generate the word sequence for the input feature vectors [16].

Acoustic models represent the knowledge about phonetics, acoustics, environment and microphone variability and gender differences among speakers, etc. Language models represent a system's knowledge of original possible word. Many challenging tasks exist in speech recognition problem such as speaker characteristics, background noise interference, grammatical variation, nonnative accents. A good speech recognition system must contend with all of these problems. The acoustic uncertainties of the different accents and speaking styles of individual speakers are compounded by the lexical complexity and represented in the language model [16].

Chapter 2-Introduction to Speech Recognition

The ASR technology has been developed for many languages e.g. English, Japanese etc. It has also been developed for our local Urdu language but it's recognition accuracy is not good as described in [3].

There is some kind of variables involved in Automatic speech recognition system that affects the performance. These variables should be restricted at some level to improve the performance of ASR engine e.g. 1) accent of speakers 2) vocabulary size 3) gender and age 4) background noise level 5) continuous versus isolated words [3]. One way is to limit the effect of these variables to make gender dependent recognition module. ASR engine can be categorized in small, medium and large vocabulary systems. Usually small vocabulary ASR system are known as digit recognition systems which based counting e.g. aik (one), do (two), teen (three) etc. having vocabulary size in range of tens where as medium and large vocabulary ASR engines consists of vocabulary size of connected words or complete sentences in range of above 20,000. These sentences again can be categorized in read and spontaneous speech. The recording environment is also a key factor that affects the performance. A good environment is an echoing chamber but system in such kind of environment will not work in noisy environment and cannot be used in daily life routine. One way is to record the real noise from working environment and superimpose on noise free recording as it is difficult to record the data from working environments.

2.1 Speech Recognition Architecture

Speech recognition problem can be defined as [2]

"Given some acoustic observation 'O', what is the most likely sentence out of all the sentences in the language?"

In mathematical form it can be written as [2],

$$W' = arg_{W \in L}^{max} P(W|O) \quad \text{-------------} \ (1.1)$$

Where O set of individual observations and W is set of word:

$$O = o_1,\ o_2,\ o_3, \dots \dots, o_t$$

$$W = w_1,\ w_2,\ w_3, \dots \dots, w_t$$

Applying Bayes' rule on equation (1.1), we get a simpler version,

$$W' = \frac{arg_{W\in L}^{max}\ P(O|W).P(W)}{P(O)} \quad \text{--------------------- (1.2)}$$

In equation (1.2) P(O|W) is the observation likelihood which comes from the acoustic model and P(W) is the prior probability which comes from the language model. In the denominator, P(O) is prior probability of observation, it is constant and not easy to calculate. We can ignore it, as a constant taken out from whole calculation and equation (1.2) will be modified as [2],

$$W' = arg_{W\in L}^{max} P(O|W).P(W) \quad \text{--------------------- (1.3)}$$

Now, we can compute the observation likelihood by simply multiplying the prior probability and observation likelihood.

Speech Recognition task can be divided in two phases 1) Training 2) Decoding. In first phase we train the HMM's by giving input 1) recorded speech file 2) original transcription of speech files 3) dictionary file. First phase provides us a model that contain pattern of basic sound units and noise, known as acoustic model. In second phase we decode the HMM's by giving input 1) speech file 2) language model (probability of words) and it provides us with the transcription of the speech file.

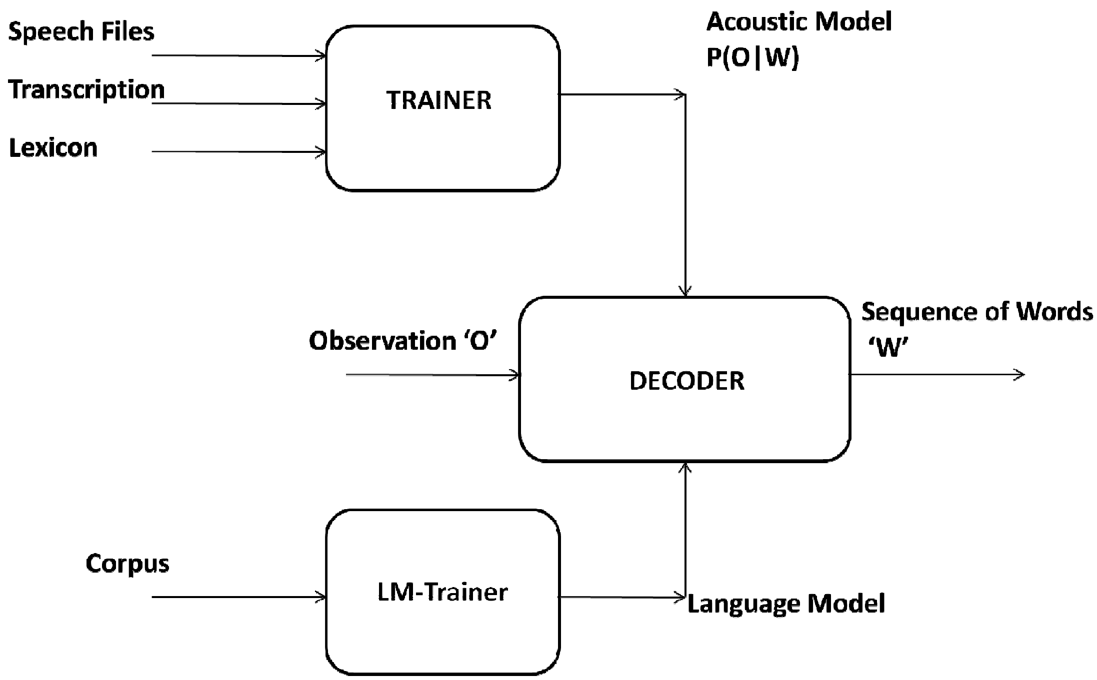


Figure 1: Block Diagram of speech recognition architecture [2]

Traditional SR software falls into one of three categories [4]. These categories are:

1- Template-based approaches
2- Knowledge-based approaches
3- Statistical-based approaches

Template-based approaches compare speech against a set of pre-recorded words [5]. A large number of traces are stored and incoming signal is compared with sequence of stored traces [5]. Knowledge-based approaches involve the hard-coding of known variations of speech into a system. Rules are defined from linguistic knowledge or from observation of speech spectrogram [4]. Both of these methods become impractical for a larger number of words. In Statistical-based approaches, (e.g. using Hidden Markov Models) variations in speech are modeled statistically using automatic learning procedures. This approach represents the current state of SR and is the most widely used technique today.

Block diagram of speech recognition is shown in figure1 [6]. Some useful features of speech are extracted by using either MFCC or LPC from speech waveform [6]. These feature vectors are scored against acoustic model and phoneme sequence is obtained.

HMM are statistical models and can be trained automatically and are simple and computationally feasible to use. In speech recognition, each basic unit (phoneme) is represented by a unique HMM. Each phoneme HMM can be represented by three state i.e. begin, middle and end state. HMM inputs a sequence of n-dimensional real-valued vectors and these vectors consist of cepstral thirty nine coefficients. A HMM for a sequence of words or phonemes is made by concatenating the individual trained HMMs for the separate words and phonemes. The Hidden Markov Model Toolkit (HTK) [7] is a portable toolkit for building and modeling of HMMs and used for speech recognition. It consists of a set of library tools like HMM training, testing and result analysis.

Implementation of Speech recognition system includes speech corpora development, training and tweaking of system for target language. Phonetic cover [6] and phonetic balance are two important terms in speech corpus development. In phonetic cover corpus contains all phones present in specific language and in phonetic balance these phones occur in same manner as in specific language [8], [9]. Phone based or context based are two types of phonetic cover [10]. Context based can either be diphone or triphone [11], [12]. Speech corpora can be developed for isolated words [13], continuous speech [[11], [12], [14]] and spontaneous speech [15].

2.2 Data Processing

In training phase, we are provided with the acoustic model which contains the basic pattern of sound units [2]. This process utilizes the transcription file and dictionary file to do the mapping of occurrences in speech files on the phones [16]. Dictionary file contains the mapping of words to phonemes that appear in transcription file. Original speech file has been segmented in the duration of 10ms using window function. This can be done by using following two functions [2]:

$$w[n] = \begin{cases} 1, & 0<L<N-1 \\ 0, & Elsewhere \end{cases} \quad \text{---------------------} \ (1.4)$$

$$w[n] = \begin{cases} 0.54-0.46\cos\left(\frac{2\pi n}{L}\right), & 0<L<N-1 \\ 0, & Elsewhere \end{cases} \quad \text{---------------------} \ (1.5)$$

Equation (1.4) is the rectangular function which can be used for segmenting the speech files. The drawback of using this function is that it generates noise which disturbs the each component equally (white noise) due to sharp discontinuity in time domain. Equation (1.5) shows the hamming function which overcomes the above problem and does the segmentation more efficiently. The segmented sound files contain both the speech information and as well speaker information. We are interested in speech information. Human ear is sensitive in the range of 20 to 1000 Hz and decrease after this range. Mel scale is used to intimate this effect. The mapping is done by using the following formula [2],

$$Mel(f) = 1127 * \ln(1+f/700) \quad \text{--------------------} \ (1.6)$$

The speech information lies in high frequency region, to separate this information cepstrum analysis has been performed. We are provided with 39-dimensioanl Mel Frequency Cepstral Coefficients (MFCC) [2]. These consist of 13 parameters to represent the phone value, 13 to capture the rate of change of these values (velocity) and 13 to capture the rate of rate of change of values (acceleration). In each these three sets of 13 MFCC's coefficients, one is energy coefficient and the rest of 12 are parameters representing phone, rate of change (delta) and rate of rate of change (delta-delta) of values respectively.

2.3 Training Phase

Hidden Markov Model (HMM) has been used to implement the acoustic model. HMM has been used with five states model. Start and end states are known as non-emitting while the middle ones are known as emitting states. The three emitting states contain the properties of phone. The first and last emitting states also depict the transition of current phone with the previous and next one respectively. To define a HMM, we need 1) set of states 2) transition probability matrix 3) set of observation 4) emission probabilities. MFCC's, calculated above, are used to model the emission probabilities and 39 dimensional multivariate Gaussian probabilities density functions as described in equation (1.7). Baum Welch algorithm is used to train these probabilities [2].

$$b_j(o_t) = \prod_{d=1}^{D} 1/\sqrt{2\pi\sigma_{jd}^2} \exp\left(-\frac{1}{2}\left[\left(\frac{o_{td}-u_{jd}}{\sigma_{jd}}\right)^\wedge 2\right]\right) \quad \text{--------------} \ (1.7)$$

Baum Welch algorithm consists of four major steps as described below [17],

1- Initialize the parameter φ (phi)

2- Determination of auxiliary function Q(φ, φ') based on φ.
3- Maximization of Q(φ, φ') function by re-estimation of φ.
4- Multiple iteration of step-2 by re-initializing $\varphi = \varphi'$ until it converges.

This process provides us the acoustic model i.e. P (O|W). From equation (1.3), P (W) still needs to be computed. This probability has been computed from Language Model. The Language Model can be based on unigram, bigram for small systems and trigram or 4-gram for large systems. Language Model can be constructed by using following equation

$$P(w_1^n) = \prod_1^n P(w_k \,|w_1^{k-1}) \qquad \text{-------------- (1.8)}$$

2.4 Decoding Phase

In decoding phase, we take the input of test observation sequence and find the best state sequence by using Viterbi Dynamic programming algorithm [2]. It takes observation sequence o[t], transition matrix a_{ij} and observation likelihood $b_i(o_t)$ as input and output path probability matrix$V_i(t)$. Being in one state at time t-1, it determines the probability of next state to reach at time t. It has following steps:

1- Initialize path probability matrix$V_i(t)$.
2- Calculate new maximum score by multiplyinga_{ij}, $b_i(o_t)$ and$V_{i-1}(t)$.
3- Find the best path probability matrix $\underset{1<i<N}{Max} V_i(t)$.
4- Now back trace through the maximum probability state.

Low priority paths have been pruned; this is done by using a threshold known as beam width. To evaluate the performance of decoding phase, Word error rate (WER) has been defined and calculated by using decoded string and the original one.

2.5 Overview of Toolkits

This section will look at some of the open source solutions available for speech recognition problems. The CMU Sphinx open source speech recognition toolkit (http://cmusphinx.sourceforge.net/) has been used in implementing the system [17]. Acoustic models built using SphinxTrain can be used by any of the decoders. Several tutorials are available, including tutorial projects, and training data is also available for English speech recognizers for use with Sphinx.

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating HMMs used primarily for speech recognition [7]. It consists of a set of library modules and tools for speech analysis, HMM training, testing and result analysis. Extensive documentation is available, including tutorials and training data for English. The toolkit is available in source form but there are some licensing restrictions.

Julius (http://julius.sourceforge.jp/en_index) is a high performance large vocabulary continuous speech recognition decoder using n-grams and context dependent HMMs, developed for Japanese speech recognition [18]. It uses standard formats for compatibility with other open source speech recognition toolkits such as those described in this section.

Speech recognition resources are also available through the Institute for Signal and Information Processing (ISIP) Internet Accessible Speech Recognition Technology Project (www.isip.piconepress.com/projects/speech).

CMU Sphinx open source toolkit has been used to implement the ASR system. It has been widely used previously in automatic learning and modeling HMM [17]. The following components are available in the toolkit.

1- PocketSphinx: lightweight recognizer library, focusing on speed and portability
2- SphinxBase: support library
3- Sphinx4: adjustable, modifiable recognizer
4- CMUclmtk: language model tools
5- SphinxTrain: acoustic model training tools
6- Sphinx3: decoder for speech recognition Research

Both toolkits are available but due to some licensing restrictions, CMU Sphinx open source speech recognition toolkit will be used.

The speech corpus for the training and testing will be developed as described in [3]. Speech data will be recorded in wav format at 16 kHz. Praat [19] will be used on the laptop to capture and manage the speech received over the microphone and will store in .wav format. The segmented speech files will be transcribed orthographically in Urdu script manually by a team of linguists. Each speech segment file name therefore will have a corresponding transcription string. The orthographic transcription will be converted into phonemic transcription using a transcription lexicon for use by the CMU Sphinx speech recognition toolkit. The general transcription rules have been based on [20]. In addition to the orthographic transcription of speech in segments, the Silence, Vocalization and Breath tags will be defined to represent non-speech areas in the segments.

This transcription files will then be converted to the format required by Sphinx using the Sphinx Files Compiler described in [21]. Following are some files required by SphinxTrain as input to build the acoustic models, A set of transcribed speech files, A dictionary file, containing transcriptions for all the words in the vocabulary, A filler dictionary file, containing entries for all non-speech sounds, e.g., vocalic pauses, throat clearing etc., A phone file, including all the phones used in the transcriptions. The transcription lexicon will be used with the Sphinx Files Compiler in order to generate phonemic transcriptions from Urdu orthography automatically. This transcription lexicon includes transcriptions of a base set of words. The speech transcriptions are also used for language model building using the SLM toolkit [22].

The Sphinxtrain will be used to integrate all created files and to train Speech recognition system. The Sphinx3 decoder will be used for testing and decoding the models

Chapter 3- Literature Review

3.1 Corpus development

There has been a lot of work done on development of speech corpora in different languages. These corpora have been used in many user applications such as ASR system development [23]. These corpora have been recorded from multiple speakers in different environments [24] and using different communications channel [25]. Greek speech corpus has been collected for development of dictation system [23]. This corpus has been recorded from 55 male and 70 female speakers in different environments. The recording sessions have been divided in three different environments. There are 180, 150 and 150 utterances in sound proof, quiet and office environment respectively. Transcription of this large corpus has been divided in two groups one for speech recognition group and second by linguists. Lowercase characters have been used to transcribe the corpus. Stress markers have been specified. External noise and articulation problem has also been marked with special characters. Speech recognition system has been trained on 46,020 utterances. The SRI's Decipher toolkit has been used to develop ASR system. Word error rate of this system has been found to be 21.01%. After analyzing the results text processing rules have been defined for the newspaper data because it may contains grammatically incorrect sentences.

Russian speech corpus, TeCoRus, has been collected over two telephony channel, narrowband and broadband [25]. One portion of this corpus consists of phonetically rich data to develop phone model and second portion consists of interview sessions and some spoken material. The speech data from first portion has been recorded from 6 speakers and consists of 3050 utterances. For second portion, 100 speakers have been selected to record the speech data. Chinese spontaneous speech corpus has been developed from university lectures and public meetings [26]. The speech data has been recorded in different noisy environments. The aim of this corpus selection was to capture phonetic variations and analyze the phoneme duration reduction, insertion and deletion. Six hours of speech data has been collected in noisy environment. Speech data has been transcribed at words, syllable and semi syllable level.

Initial contents of the corpora have been collected from many different resources such as books, internet, meetings etc to include all possible variation of phonemes [28] [29]. Many techniques have been used to collect corpora to have maximal phonetic coverage e.g. [27]. Speech corpus for Ethiopia language has been developed [27]. This corpus consists of read speech from newspaper and magazine articles. Phonetically rich sentences based on syllables have been integrated in the corpus. In first phase of computational method of phonetically rich large corpus consist of 100,000 sentences has been developed. In the second phase, sentences with highest phonetic score have been selected. In third phase, sentences with highest syllable balance score and having rare syllable has been selected. The performance of this corpus has been analyzed by developing an ASR system. This data has been divided in training and testing data. 20 hours of data from 100 speakers has been collected to develop speech corpus. The speech corpus has been cleaned and transcribed semi-automatically. English speech corpus has been collected over the mobile channel for American English corpus SALA-II [29]. The some portion of corpus has been selected from the Harvard and Timit corpus to increase phonetic richness. 4412 phonetically rich sentences have been short listed from these corpuses. The aim of this corpus was to train and develop speech recognition systems over mobile channel. 4000 speakers from different north, central and South American states have been selected to record this speech data in different environments.

Phonetically rich corpora have been developed in many languages e.g [24] [28] [30] [48]. Minimal Phonetically rich corpus has been collected from 560 speakers to develop speaker independent continuous speech recognition system [24]. The corpus has been collected from newspaper and website sources. Phonetically rich corpus has been selected from the larger set by using optimal text selection greedy algorithm. The aim of this corpus was to collect all the phonetic variations that occur in Tamil, Marathi and Telugu languages. This corpus consists of 155541, 303537 and 444292 sentences of Marathi, Telugu and Tamil languages. The speech data has been recorded over landline and cellular phone channel. The automatic speech recognition system has been developed on two speech corpora for Taiwanese language [30]. One corpus has been selected based on biphone phonetically rich data, second corpus based on triphone phonetically rich data. Performance of these corpora has been evaluated based on recognition results. Speech data has been recorded by a single male speaker over microphone. From the above experiments, it has been concluded that syllable recognition accuracy is better for biphone rich corpora. Hindi speech corpus has been collected from articles, magazines and online content available [28]. Phonetically rich sentences has been selected such that they are meaningful and do not contain any sensitive word. In first phase, 350,000 sentences have been selected. From the above corpus, 50,000 phonetically rich sentences have been short listed. This corpus has been used to develop ten speakers continuous speech recognition system. Urdu speech corpus has been developed on 82 speakers for speech recognition system. 45 hours of read and spontaneous speech data has been recorded from 40 female and 42 male. In this corpus, spontaneous speech data has been collected from designed questions set based on daily routines, hobbies, past experience and interests. For read speech data, 725 phonetically rich sentences and six paragraphs have been developed from 18 million Urdu words. Urdu native speakers mostly from university area have been recruited for recording purpose. Recording has been done in office room and lab environment. Three hours of data has been collected from each volunteer. This large data has been segmented into smaller portion, not more than 10 seconds duration. The linguists have transcribed the corpus in Urdu script using the rules defined in [20]. Different silence markers have been defined to represent non speech area at different locations in speech files.

Many greedy algorithms have been developed to collect corpora from different sources [31] [32] [49]. Greedy algorithms have been widely used to select corpora for speech synthesis purpose [32]. Turkish speech corpus has been developed using greedy algorithm on read speech data [31]. In first phase of greedy algorithm, all the sentences in the corpus has been assigned a cost based on occurrences of diphones. In second phase, sentences have been selected in multiple iterations based on maximum cost. Some special sentences, having unique diphones, have also been selected. Initially read data, from the internet, consisted of 11500 sentences has been selected. The greedy algorithm has been applied to this baseline corpus. Final corpus consists of 2500 sentences. Speech corpus for Irish language has been developed by using greedy algorithm [32]. This corpus has been used in development of text to speech system. In first phase, baseline source has been selected from source. In second phase, smaller corpus has been selected to have maximal unit coverage. In last phase, rare sentences have been selected. Phonetically balanced and distributed sentences have been selected by using defined iterative method for Thai language [49]. These sentences have been selected from ORCHID standard corpus. The aim of developing this greedy algorithm is to collect the phonetically balanced corpus to train large vocabulary speech recognition system. In first phase, phonetically balanced sentences has been selected and assigned an initial number. In second phase, this phonetically balanced sentences serve as initial set and phonetically distributed sentences have been selected using the method defined in [49]. The final results have compared with Japanese ATR and English TIMIT phonetically balanced corpus. The analysis of corpus shows 398 phonetically balanced and 802 phonetically distributed sentences have been selected in final set out of 27,634 sentences.

3.2 Speech Recognition Systems

There are two categories of ASR systems based on vocabulary size of corpus, spontaneous and isolated words. Spontaneous ASR systems have been developed on different corpora e.g. on English language Malach [38], NIST [50]. German speech recognition system has been developed by Sloboda using Janus 2 toolkit [34]. To improve the naturalness, system has the ability to add new pronunciation of words in database based on utterance frequency. An algorithm has been proposed to capture the pronunciation variation. It is not feasible to update dictionary for each pronunciation. The purpose of this algorithm is to optimize the dictionary based on statistical relevance. Spontaneous speech recognition system has been developed to evaluate the performance. Training and test data consists of 281160 and 110 words respectively. Word accuracy has been found to be 68.4%. Performance of spontaneous and dictation speech recognition systems has been compared [35]. WER of dictation is less as compares to spontaneous system due to inefficient language models for spontaneous systems. WER has been found to be 5% and 15% for dictation and spontaneous systems on broadcast news respectively. The reason of low accuracy has been found to be non fluent speech in spontaneous systems i.e. sentence breaks, repetition and hesitation. An algorithm has been proposed to modify the language models to solve the above issues using context manipulation technique. ASR system has been developed to evaluate the above technique. Training and test data consists of 310 and 2 hours of data respectively. Language model consists of 3 million words. WER has been reduced from 36.7% to 35.1%.

Repetition of words in spontaneous speech corpora is a common issue. An analysis has been performed on Fisher's English speech corpus to find the single and multiple word repetition [33]. Spontaneous speech recognition system has been developed using Fisher's corpus to address the disfluent repetition problem [33]. This problem has been addressed by defining repetition word error rate in spontaneous speech recognition system. This error rate has been determined by using different acoustic and language model. Acoustic prosodic classifier and multi word model techniques have been proposed to solve the above issue. Fisher corpus has been consisted of telephonic conversational data. It contains 17.8 million English words. Training and testing data consists of 220 and 2 hours of speech data from 20 speakers respectively. Absolute reduction of 2% has been achieved using above proposed solution. The analysis shows classifier approach is not very convincing. Using multi word approach 75.9% improvement has been achieved in repetition word error rate. Spontaneous English ASR system has been developed on NIST speech corpus using CMU sphinx3 toolkit [50]. Acoustic variation of phonemes has been modeled as different phone to capture acoustic variation in spontaneous speech. The training and test data consists of 2 and 0.5 hours of speech data. Gaussians, HMM likelihood and duration based phone splitting technique have been applied on AA and IY phonemes. WER has been reduced from 51.1% to 49.6% for AA and 49.3% for IY phoneme using Gaussians based splitting approach. While WER has been reduced to 49.8% and 49.6% in duration based splitting approach where as no improvement in HMM likelihood based splitting approach. Distributed speech recognition system over telecommunication channel has been analyzed with specified range of signal to noise ratio using HTK toolkit [36]. A database has been developed to analyze the performance of speech recognition algorithms. It has been consisted of connected Tidigits and recorded by American English speaker. This data has been cleaned by using low pass filters. Eight different kinds of noises has been selected from real word. These noises have been added superimposed on clean Tidigits with different signal to noise ratio. A number of experiment sets have been developed to compare performance of speech recognition system with clean and noisy training data. Vocabulary size consists of 8440 utterances from fifty two male and female speakers. The analysis shows performance of speech recognition system is worse which has noise from non stationary segments, The recognition results have been described with varied SNR [36].

Performance of many ASR systems have been evaluated and improved by using different methods e.g. improving SNR [38], improvement in language model and acoustic model [40] [41] [42]. English ASR has been developed on subset of Malach corpus [38]. Word error analysis has been performed on the above system to improve ASR performance. By improving signal to noise ratio and syllable rate absolute improvement of 1.1% has been achieved [38]. The role of acoustic and language model in unlimited vocabulary finish speech recognition system has been analyzed to improve the ASR system. Three acoustic models have been prepared one by using Maximum likelihood (ML), second by ML and three iteration of speaker adaptive training (SAT) [39] and third by ML, SAT and four iteration of minimum phone frame error criteria [40]. Error analysis has been performed on continuous speech recognition system Easy talk [41][42]. Two set of rules have been developed to identify the error type. Two methods have been defined to address the acoustic and syllable splitting error. Third method improves the Viterbi algorithm to improve the search process. Two speakers, Isolated words (0-9) Hindi (Swaranjali) speech recognition system has been developed [43]. Acoustic model has been trained from twenty utterance of a word for each speaker. Word accuracy for two speakers comes to be 84.49% and 84.27%.

There has been much work done in development of Hindi and Urdu ASR systems. Different methods, like HMM [3] [44], Artificial neural networks [45], Matlab [46], have been used to train and test the system. Speech recognition system on Hindi language has been developed in room environment for eight speakers on thirty isolated Hindi words. HTK toolkit has been used to train the acoustic word model. Overall word accuracy has been found to be 94.63% [44]. Urdu speech recognition system has been developed for 81 speakers. Acoustic model has been prepared on incremental basis in three stages by addition of two speaker's data. Three acoustic models have been tested on forty female, forty one male and eighty one combined speakers by using open source CMU sphinx toolkit. Word error rate has been found to be 60.2% [3]. Urdu Speech recognition system has been developed based on artificial neural network, pattern matching and acoustic modeling approaches [45]. Viterbi algorithm has been used for decoding the model. Single speaker isolated digit recognition system has been developed for Urdu language by using back propagated neural network approach using Matlab [46]. Multilayer neurons have been used in this architecture to train and recognize. Small vocabulary automatic speech recognition system has been developed for Urdu language by using sphinx4.

Acoustic model has been prepared from fifty two isolated spoken Urdu words and 5200 utterances of speech data from ten speakers. The average word error rate comes to be 5.33% [47]. Automatic speech recognition system has been developed for Urdu on single speaker medium vocabulary [3]. The acoustic model has been prepared form 800 utterances of read and spontaneous speech corpus combined in various ratios. Sphinx3 toolkit has been used to train and decode the model.

Chapter4- Methodology

There has been a lot of work done in speech recognition for other languages as described in section-4. Recently 81 speaker's large vocabulary continuous speech ASR system for Urdu [3] has been developed. Word error rate has been found to be 60.2% which seems to be very high for this system. The following experiments have been developed to analyze the recognition results and improve the accuracy.

Main objectives of this work will be to investigate if error analysis of recognition results can be used to improve new integration of collection of speech data

1- To develop an ASR system on single speaker large vocabulary continuous speech for Urdu (Experiment 1- Single speaker baseline)
2- To find the recognition Issues on above system (Experiment 2- Single speaker improved)
3- To develop an ASR system on ten speaker's large vocabulary continuous speech for Urdu (Experiment 3- Ten speaker baseline)
4- To find the recognition Issues on above system (Experiment 4- Ten speaker improved)
5- To replace one speaker of baseline Experiment-1 with revised Experiment-2 (Experiment 5- Ten speaker with one speaker cleaned data)
6- To find the criteria for minimal discriminative balanced corpus (Experiment 6- Minimal balanced corpus)

4.1 Experiment 1- Single speaker baseline

Single speaker ASR system on large vocabulary continuous speech on Urdu has been developed using the corpus developed and rules defined in [3]. Phonetically rich corpus has been used in training of ASR system. It consists of read and spontaneous speech. The speech files have been transcribed manually. Silence markers have been identified manually. This experiment has been developed on small scale to analyze the recognition issues. To analyze the recognition results, phoneme frequency confusion between different phonemes has been determined in training and testing speech data.

4.2 Experiment 2- Single speaker improved

Error analysis techniques have been performed on Experiment-1 to identify the recognition issues. Confusion matrix has been constructed to analyze the confusion between different phonemes. These issues have been addressed separately and modified the data set for ASR system.

4.3 Experiment 3- Ten speaker baseline

This experiment has been developed by increasing the number of speakers from one to ten. The acoustic model has been trained using same phonetically rich corpus recorded from ten speakers. The recognition issues have been analyzed on ten speaker's data.

4.4 Experiment 4- Ten speaker improved

In this experiment, based on recognition issues training data has been modified. Acoustic model has been developed on modified speech data. Revised ASR has been developed on modified data set.

4.5 Experiment 5- Ten speaker with one speaker cleaned data

This experiment has been developed by replacing the one speaker data from Experiment-4 with the revised data set of revised Experiment-2. The speaker's data has been replaced such that vocabulary size remains the same.

Training and test data has been described in the following table for each experiment.

Table 1: Training and testing data

Experiment	Number of training utterances	Number of test utterances	Read speech utterances	Spontaneous speech utterances
Baseline Experiment-1	620	45	351	269

Revised Experiment-1	671	60	400	269
Baseline Experiment-2	1946	119	873	1073
Revised Experiment-2	1946	119	873	1073
Experiment-3	1999	123	883	1116

4.6 Experiment 6-Minimal balanced corpus

In first phase of developing criteria for minimally balanced corpus, frequency and accuracy of each phoneme training data has been determined. In second phase, this training data has been divided in different ratio less than determined in phase-1 and phoneme accuracy has been analyzed. Phase-2 has been repeated by increasing amount of training data until saturation in phoneme accuracy achieved. In this way, training data for each phoneme has been determined. From the above results, phoneme training data has been analyzed to find the minimum amount of training data for phonemes to achieve maximum accuracy. Speech data has been updated using this minimal training data for phonemes and ASR system has been developed to compare the phoneme recognition accuracy.

This involves the development of speech corpora and ASR system training for Urdu. The aim is to find the recognition issues by analyzing the recognition results. From the above experiments recognition results, it can be analyzed is it a good way to increase more speaker's data in existing ASR system to make it speaker independent. From the above results, phoneme training data has been analyzed to find the minimum amount of training data for phonemes to achieve maximum accuracy. Speech data has been updated using this minimal training data for phonemes and ASR system has been developed to compare the phoneme recognition accuracy. It can be concluded from the above results weather it is a good way to develop minimal discriminative balanced corpus.

Chapter5- Experimental Results

5.1 Experiment 1- Single speaker baseline

In baseline Experiment-1, 56 minutes of data consisted of read and spontaneous speech has been used to develop this experiment as described in Table-1. Recognition results have been described in Table-2.

Table 2: Baseline Experiment-1 Recognition Results

No. of tied states	100
Beam width	1e-120
Language weight	23
Word error rate	18%

Error analysis technique has been developed to investigate the recognition issues. Confusion matrix has been created from the above results to find the phoneme accuracy and confusion with other phonemes. It has been summarized in the Table-3.

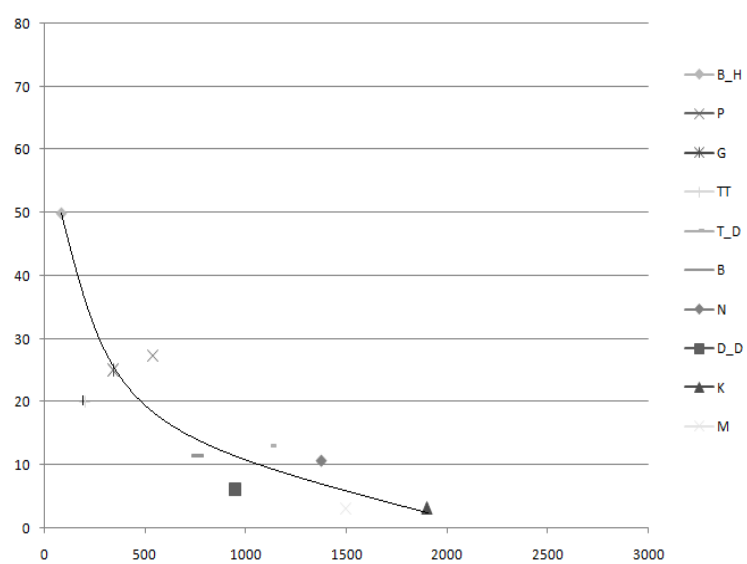
Table 3: Phoneme Confusion Summary

Phone	Confusion	Error Frequency	Phone	Confusion	Error Frequency
P	Sil	10	Z	R	1
TT	Sil	10	Z	Sil	9
T_D	Sil	6	F	Sil	2
T_D	D_D	5	SH	K	1
N	Sil	3	SH	H	3
K	Sil	2	S	Sil	7

K	P	5	H	Sil	2
K	B	4	T_SH	AA	8
M	Sil	1	D_ZZ	Z	2
V	R	3	D_ZZ	Sil	4
Z	D_D	2	R	Sil	6
J	Sil	6	AE	Sil	8
O	OON	2	U	AA	3
OO	O	8	U	Sil	4
OO	AE	1	I	II	7
AA	OO	2	I	Sil	5
AA	Sil	8	AA	Sil	7

Phoneme error rate has been calculated for each phoneme. Table-3 also shows the frequency of phoneme confusion with silence each other. Following graphs shows the phoneme error rate versus the amount of training data for each phoneme.

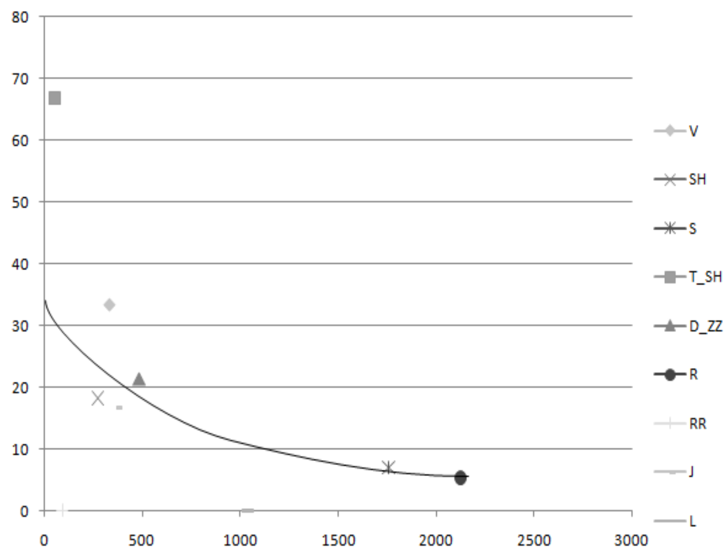
Percentage error rate



Amount of training data

Figure 2: Graph for Stops

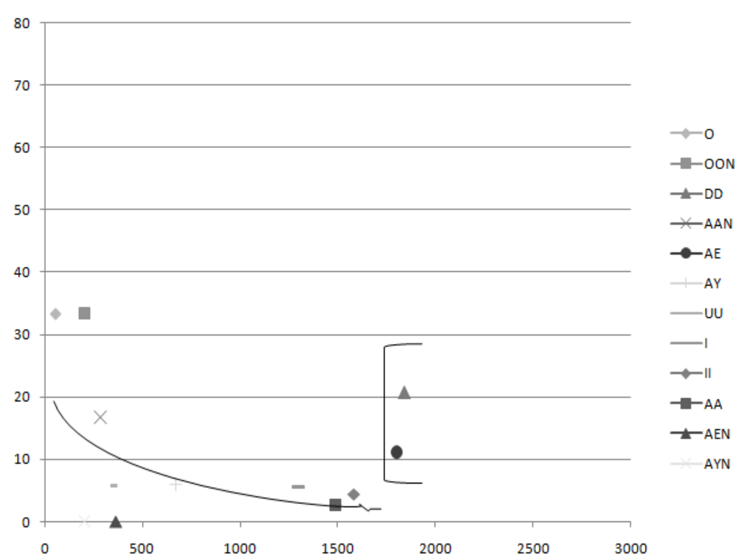
Percentage error rate



Amount of training data

Figure 3: Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure 4: Graph for Vowels

5.1.1 Experiment 1- Discussion

Word error rate has been described in Table-2. This error rate seems to be high on single speaker data. Decoded and original sentences have been compared. Phonemes that are mismatched with other ones are described in Table-3.

All the phonemes have been divided in three sections based on opening of vocal tract. Stops have been listed in first, Vowels in second and fricatives, affricates, Trills, Flap and Approximants in third category of phonemes. Stops phonemes have high order of confusion with silence e.g. phoneme P, TT and T_D. Vowels have less confusion with silence as compared to stops. Some fricatives also have confused with silence. There are fewer phonemes in Table-3 that have been confused with other category of phonemes e.g. Phoneme V from fricatives has been confused with trill R.

To analyze the distribution of phonemes in recorded speech data, training data of these phonemes have been plotted versus percentage error rate. Figure-2, 3, 4 shows this distribution. It can be seen from the graphs that phonemes that have the large y-axis and x-axis, large y-axis and small x-axis values indicate the high error region. Phonemes that have the small y-axis and x-axis, small y-axis and large x-axis values indicate the low error region. Phoneme distribution is not balanced. These issues are very common in developing ASR systems. Noise plays a major role in degrading the performance of such systems. Moreover, to have sufficient phonemes distribution in training data is a challenging task. Many greedy algorithms have been developed to have phonetically balanced data in corpus.

The following techniques have been proposed for the above problems. In high error region, for small x-axis values, one possibility is to increase the amount of training data. In high error region, for large x-axis values, one possibility is to carefully analyze the transcription of training and test data to remove the tagging error, if any. To increase the training data such that phoneme distribution will be balanced. Non-speech areas in speech files should be identified automatically. To add more data in language model using the perplexity rules.

There might be the possibility for phonemes, whose training data and accuracy is low, to increase the data. If the training data for some phonemes is sufficient then there might be the possibility that training data is not correctly transcribed or tagged. From Table-3 there is a lot of confusion between phonemes and silence region in speech data. Silence markers may be identified in speech data automatically to avoid these confusions.

5.2 Experiment 2- Single speaker improved

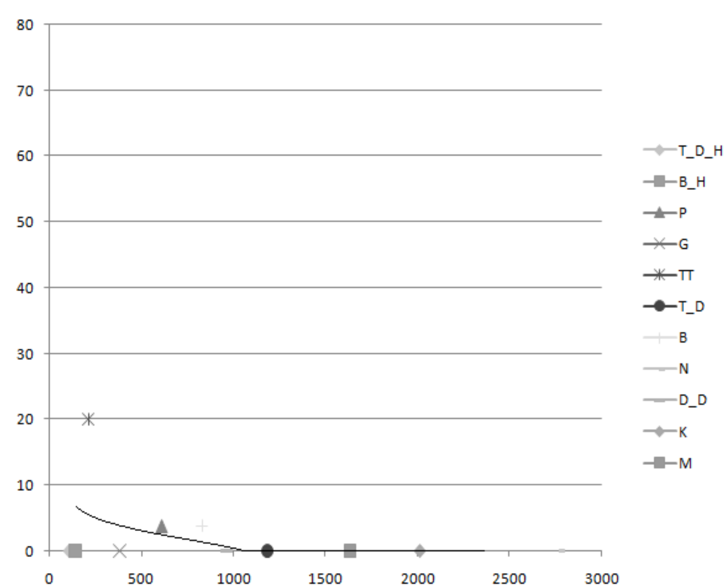
The above mentioned techniques have been applied to baseline Experiment-1. Followings recognition results have been obtained:

Table 4: Revised Experiment-1 Recognition Result

No. of tied states	100
Beam width	1e-120
Language weight	23
Word error rate	3.9%

Phoneme error rate has been calculated and plotted on above recognition results

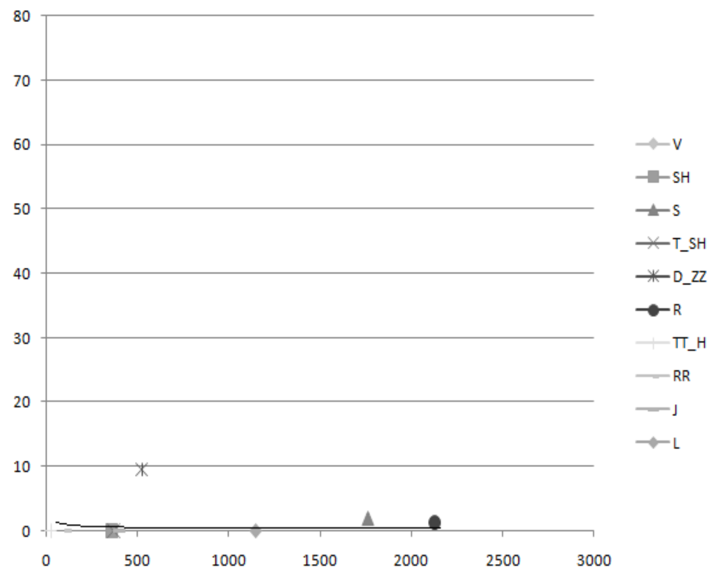
Percentage error rate



Amount of training data

Figure 5: Graph for Stops

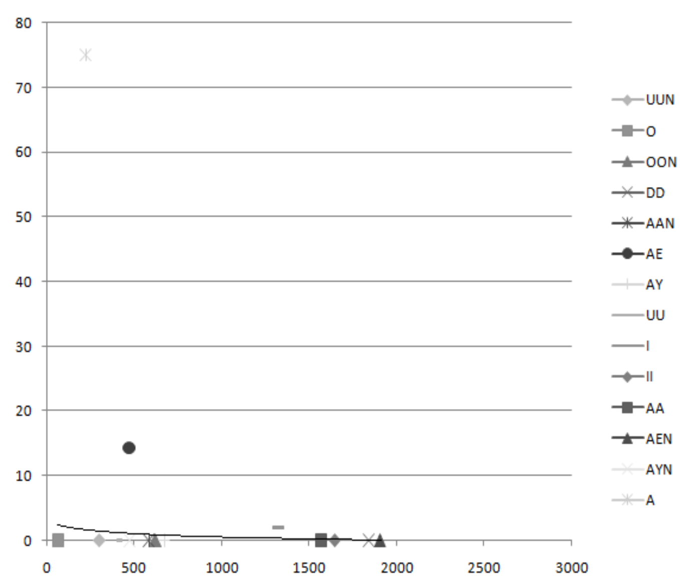
Percentage error rate



Amount of training data

Figure 6: Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure 7: Graph for Vowels

Following table shows the effect of carefully analyzing the corpus and problem with the phonemes in the transcription.

Table 5: Analysis of Transcription

Phoneme	Training Data	Previous Error rate (%)	Improved Error rate (%)	Percentage Improvement (%)
T_D	1127	13.04	6.52	50
DD	1842	20.69	6.89	66.67
AE	1804	11.11	6.67	39.69

Following table shows the effect of increasing training data to balance the distribution of phonemes on phoneme accuracy.

Table 6: Effect of increasing training data

Phoneme	Original training data	Increased training data	Improved accuracy from (%)-to(%)
B_H	82	142	50-0
P	540	608	27.3-3.7
G	342	415	25-0
SH	276	360	18.1-0
T_SH	55	515	66.3-0
D_ZZ	485	524	21.4-9.5

O	25	101	33.3-0
OON	203	621	33.3-0
AAN	285	585	16.6-0
AY	572	675	5.3-0
TT	290	974	20-20

5.2.1 Experiment 2- Discussion

Word error rate has been described in Table-4. The analysis techniques described in above section has been applied in this experiment. Training data has been analyzed based on confusions of phonemes as presented in Table-3. There has been a lot of confusion between different phonemes and silence. This problem has been mentioned in developing ASR systems for other languages by different authors. It has been usually improved by selecting non speech areas automatically. In this experiment, force alignment algorithm has been used to determine the silence region is speech files automatically. Moreover, phonemes transcription has also been analyzed for three selected phonemes whose error rate and training data is very high. Phonemes DD, AE and T_D have the tagging and pronunciation problem as described in Table-5. Table-5 shows the previous and improved error rate of these phonemes. Tagging error has been solved by updating the transcription files where as speech files have been updated for pronunciation error.

Speech and transcription files have been analyzed for phonemes whose training data and error rate is high. Speech files have been cleaned. Training data of phonemes has been increased randomly of those phonemes have less training data. The effect on accuracy has been presented in Table-6. The aim of increasing training data is to check weather increasing the training data will increase or decrease the phoneme error rate. Acoustic model has been trained after increasing training data. Table-6 shows the improvement in phoneme error rate. There has been some exception with some phonemes such as TT. Training data of this phoneme has been increased from 290 to 974 but error rate remains the same. Error rate of some phonemes has been decreased to zero as training data reaches a sufficient value. Error rate of some phonemes has been saturated on a non-zero value. There is no improvement on further increasing the training data of these phonemes.

The combined effect of all methods, that have been applied to improve accuracy, has been analyzed. Figure 5, 6 and 7 shows the improved graphs.

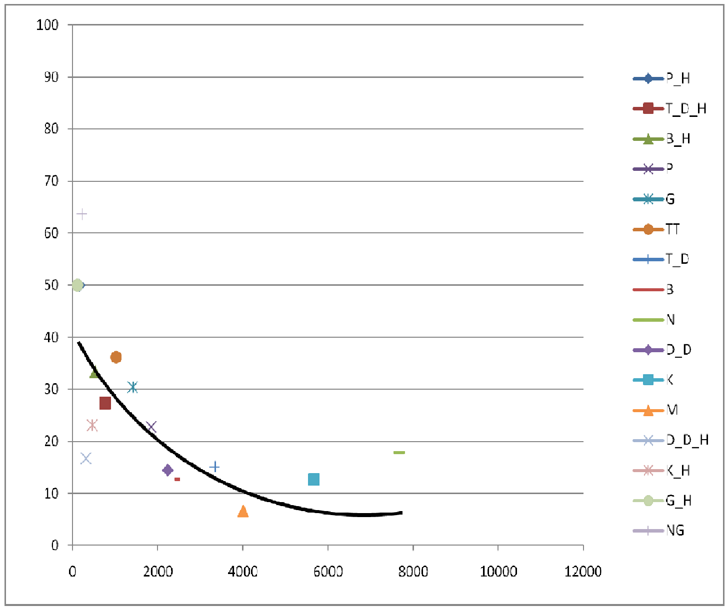
5.3 Experiment 3- Ten speaker baseline

The same concept of error analysis technique that has been applied on single speaker in baseline Experiment-1, extended to 10 speaker's data. The following table and graphs shows the recognition results and phoneme error rate.

Table 7: Baseline Experiment-2 Recognition Results

No. of tied states	500
Beam width	1e-120
Language weight	20
Word error rate	63.58%

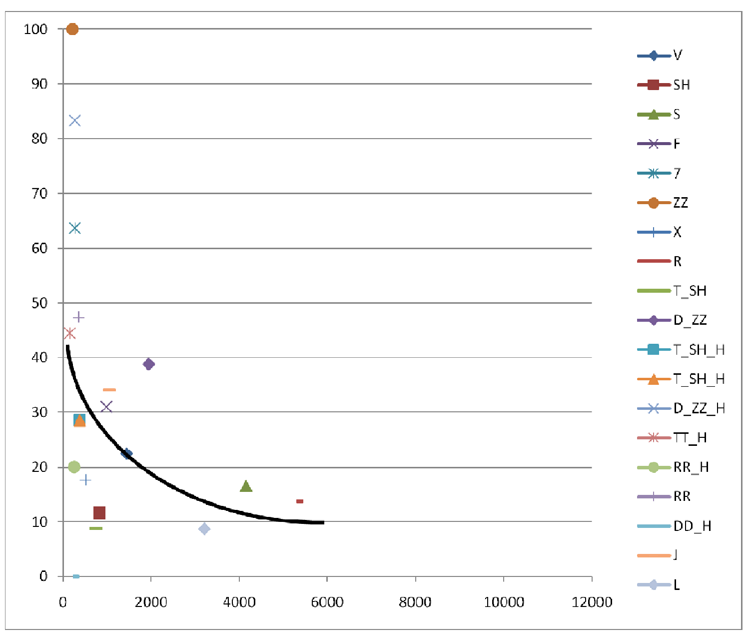
Percentage error rate



Amount of training data

Figure 8: Graph for Stops

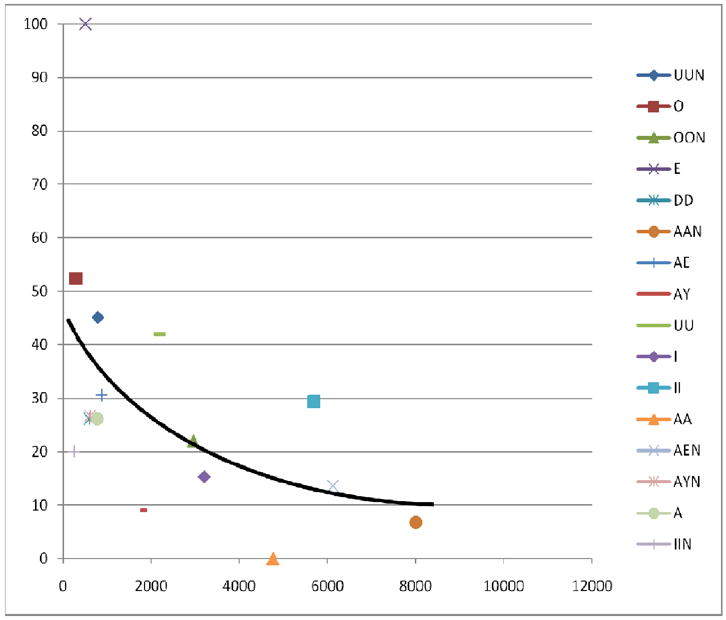
Percentage error rate



Amount of training data

Figure 9: Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure 10: Graph for Vowels

5.3.1 Experiment 3- Discussion

Word error rate has been described in Table-7. Word error rate seems to be very high for ten speakers. Phonemes in the above graphs again can be divided in two alarming regions i.e. low training data, high error rate (first category) and large training data, high error rate (second category). Training data of phonemes that fall in first category may be increased to analyze the effect on accuracy. Speech data and transcription of phonemes that fall in second category may be analyzed to find the recognition issues.

Analysis of the recognition results shows the same kind of problems that have been faced in Experiment-1. In Figure 8, as G_H phoneme has fewer amounts of training data so its error rate is nearly 50%. On the other hands, as M phoneme has sufficient amount of training data so its error rate is below 10%. Only two phonemes, L and T_SH, have below 10% training data in Figure 9. From Figure 10, it seems acoustic model has been trained well on AA phoneme as it shows 0% error rate.

A general trend has been observed from the above three figures that curves trend to be saturated with increase in training data. There are some outliers, e.g. N from stops and II from vowels. Phonemes of fricatives, trills, flap and approximants seems to have less training data as compared to the other two categories.

Same methodology has been used on ten speaker's ASR system as described in Experiment-1. Only testing data of speakers have been cleaned.

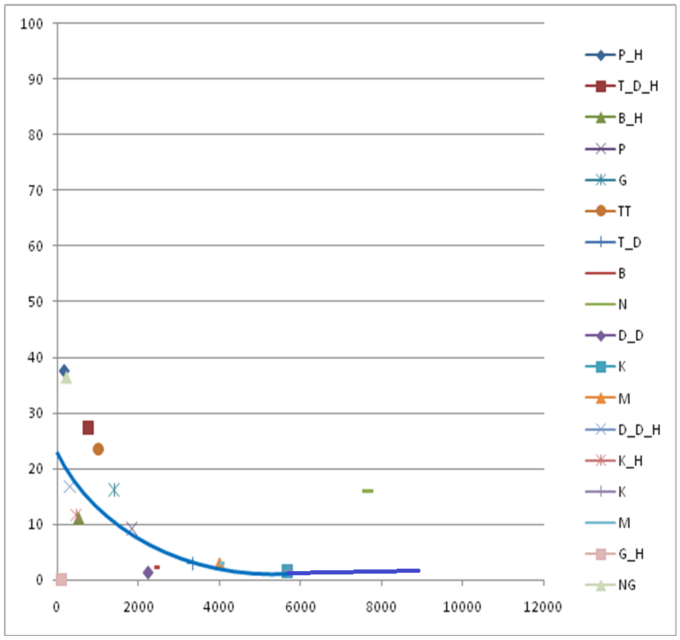
5.4 Experiment 4- Ten speaker improved

The techniques that have been proposed to improve the recognition results, applied to baseline Experiment-1. Following table and graphs show the improved error rate and phoneme recognition results.

Table 8: Revised Experiment-2 Recognition Results

No. of tied states	500
Beam width	1e-120
Language weight	20
Word error rate	25.88%

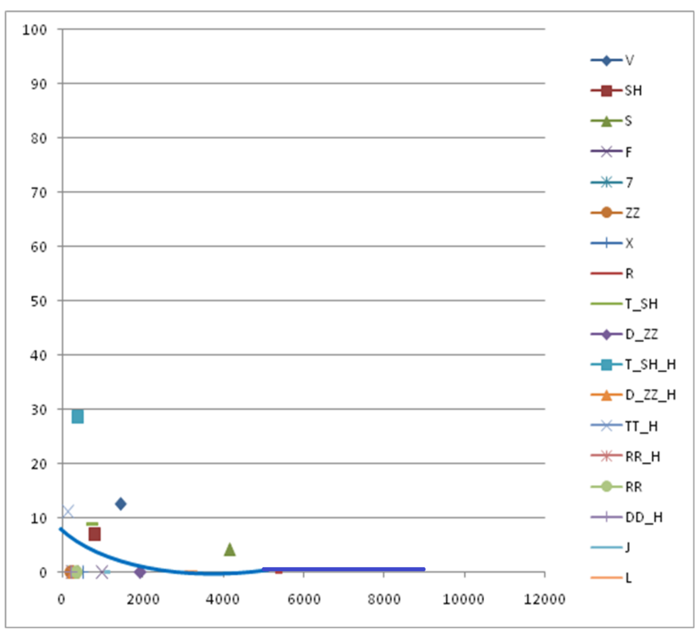
Percentage error rate



Amount of training data

Figure 11: Graph for Stops

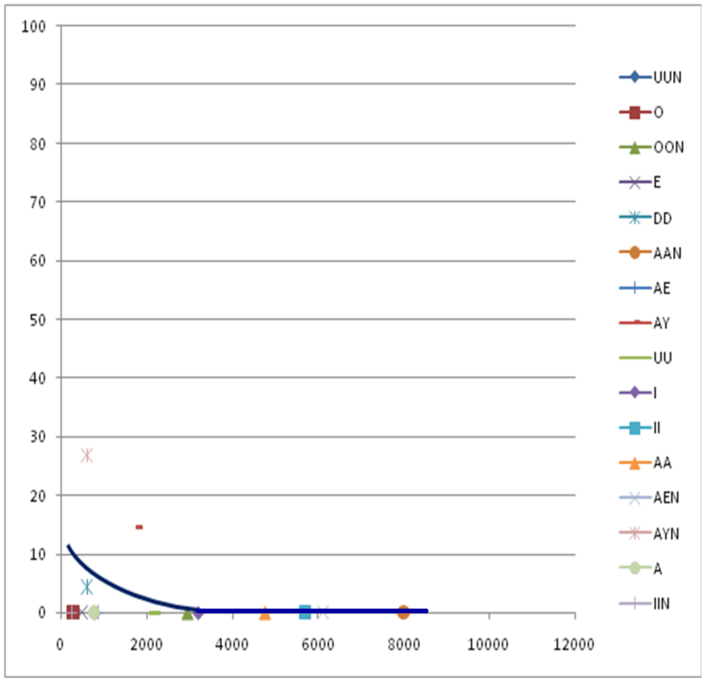
Percentage error rate



Amount of training data

Figure 12: Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure 13: Graph for Vowels

5.4.1 Experiment 4- Discussion

Word error rate has been described in Table-8. In this experiment, transcription of test speech data has been analyzed. Silence regions have been identified automatically by using force alignment algorithm. The above graphs show the improvement in phoneme error rate after applying the above techniques.

Figures 11, 12 and 13 show the improvement in phoneme error rate. Phoneme training data remains the same but error rate decreases due to cleaning of testing data. The outlier phoneme N, from Figure 8, has very less effect on error rate as shown in Figure 11. On the other hands error rate of outlier phoneme II, from Figure 10, has been dropped to 0%. From Figure 12 and 13, error rate of many phonemes have been decreased to 0%, e.g. F and ZZ in Figure 12, A and O in Figure 6. From Figure 4, error rate of phoneme G_H has been reduced to 0%. This phoneme has fewer amounts of training data and nearly 50% error rate in Figure 8.

During cleaning of test speech data, noise and pronunciation problems have been found. Pronunciation problems have been found largely in spontaneous speech data, e.g. "میں" (I) has been extended to longer duration of utterance and in transcription the extended period has been mapped to silence marker. Moreover, the phoneme error rate of 'TT' has been decreased from 18.75% to 0% while of 'N' from 17.82% to 12.7%. One reason found to be wrong pronunciation of phoneme 'N' e.g. the word 'ٹینجیبل'.

Such issues have been resolved. Moreover silence marker has been adjusted automatically using force alignment algorithm. Language model used for decoding has been prepared from 81 speaker corpus. From the Figures 11, 12 and 13, testing data equally critical to be cleaned as training data; otherwise the results achieved are not reflective of the system accuracy.

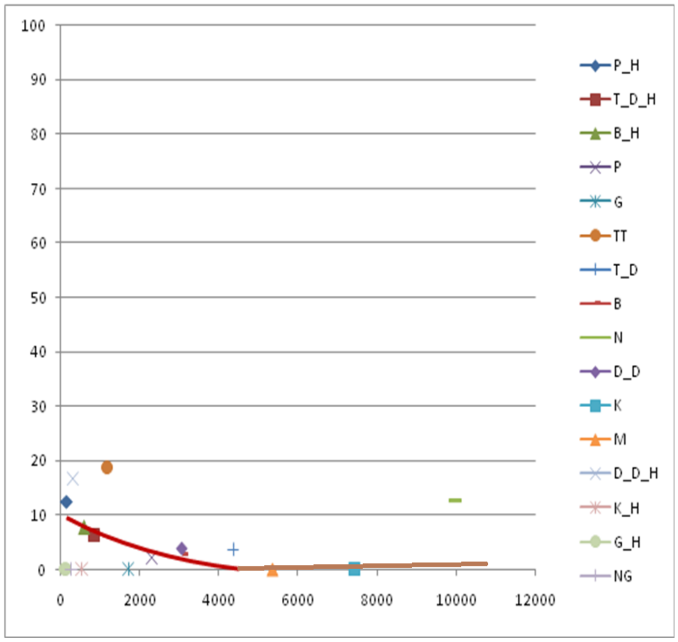
5.5 Experiment 5- Ten speaker with one speaker cleaned data

Single speaker data of revised Experiment-1 has been replaced with the equivalent amount of speaker's data of revised Experiment-2. Error analysis technique has been applied to check the phoneme error rate. Following table and graphs show the recognition result and phoneme error rate.

Table 9 : Experiment-3 Recognition Results

No. of tied states	500
Beam width	1e-120
Language weight	20
Word error rate	23.21%

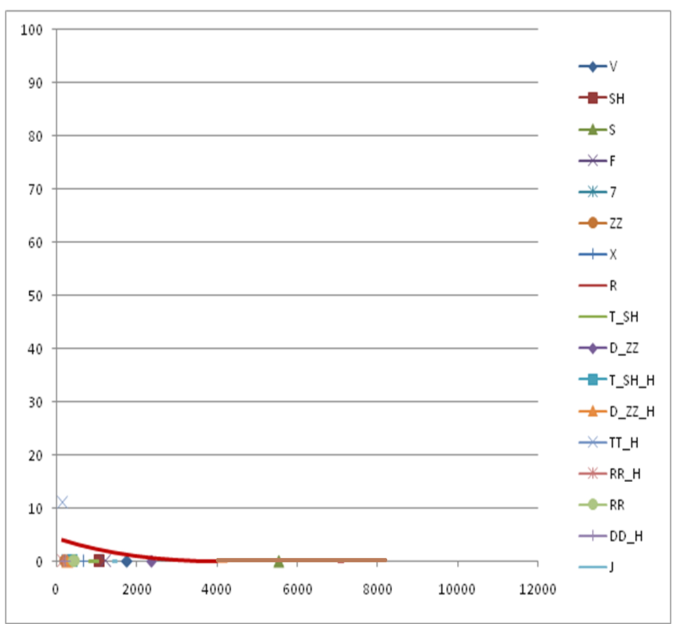
Percentage error rate



Amount of training data

Figure 14: Graph for Stops

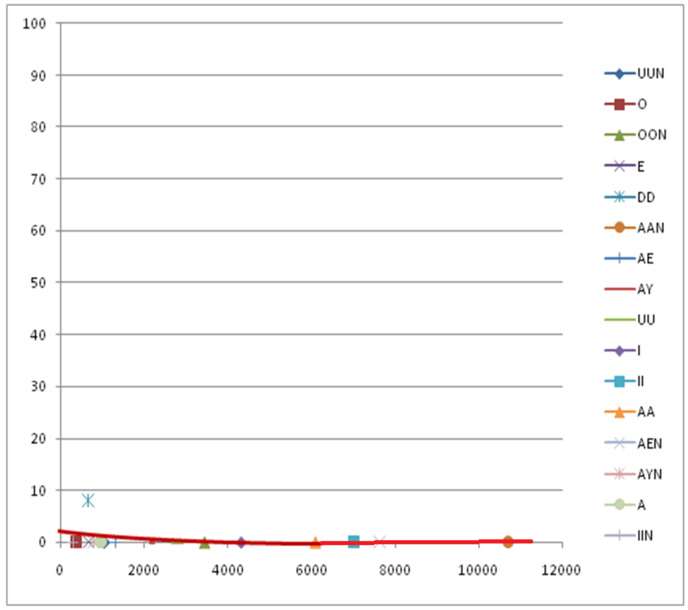
Percentage error rate



Amount of training data

Figure 15: Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure 16: Graph for Vowels

5.5.1 Experiment 5- Discussion

Word error rate has been described in Table-9. The above figures show the phoneme recognition results after integrating one speaker's data that has been collected using the error analysis technique developed in revised Experiment-1. The improvement in phoneme error rate, described in this experiment, is due to the cleaned data of one speaker that has been replaced from Experiment-2. It shows cleaned training data plays very important role in acoustic model preparation. By comparison of Experiment-3 and5, there is a significant difference in phoneme accuracy e.g. from Figure-8 TT phoneme has nearly 1200 training data and error rate is 38% where as from Figure-14 TT phoneme has slightly increased amount of training data but error rate has been reduced to 19.5%.

Figures 14, 15 and 16 show the improvement in phoneme error rate. Effect of addition of balanced training data can be analyzed from above figures. In Figure 14, error rate of outlier phoneme N has been reduced by addition of balanced speaker data. Training data of different phonemes have been increased, e.g. in Figure 14 training data of phoneme S has been increased from 4000 to 5500 and error rate reduced to 0%. Error rate of all phonemes in Figure 15 and 16 have been reduced to 0% except TT_H and DD respectively. Balanced training corpus seems to be critical for better performance of ASR system.

The improvement ratio in accuracy is not same for all phonemes e.g. from figure 8 and 14, error rate for K has been decreased from 8% to 0% where as for TT from 38% to 19.5%. There may be multiple possibilities for this behavior of phonemes such as to train HMM on TT phoneme is difficult as compared to K, training data for TT phoneme is more complex, variation in pronunciation of same phoneme and noise. In next section, this variation has been discussed through experiment.

Figure 14, 15 and 16 shows, the phonetically rich corpus has unequal amount of training data for different phonemes. It is unclear whether this amount of training data is optimal or not, as there is too much data for some phonemes and less data for other phonemes. E.g. AAN phoneme has 10600 where as D_ZZ has 935 amount of training data. Phoneme accuracy has been determined on different amount of training data and to reduce size of training corpus.

5.6 Experiment 6-Minimal balanced corpus

This experiment has been developed on speech data of Experiment-3. Following phonemes have been selected and their training data has been increased on incremental basis as shown in following Table-10.

Table 10 : Phonemes with default training data

Phoneme	Original training data	Accuracy (%)
AAN	10600	100
OON	3443	100
D_ZZ	935	100
T_SH	2364	100

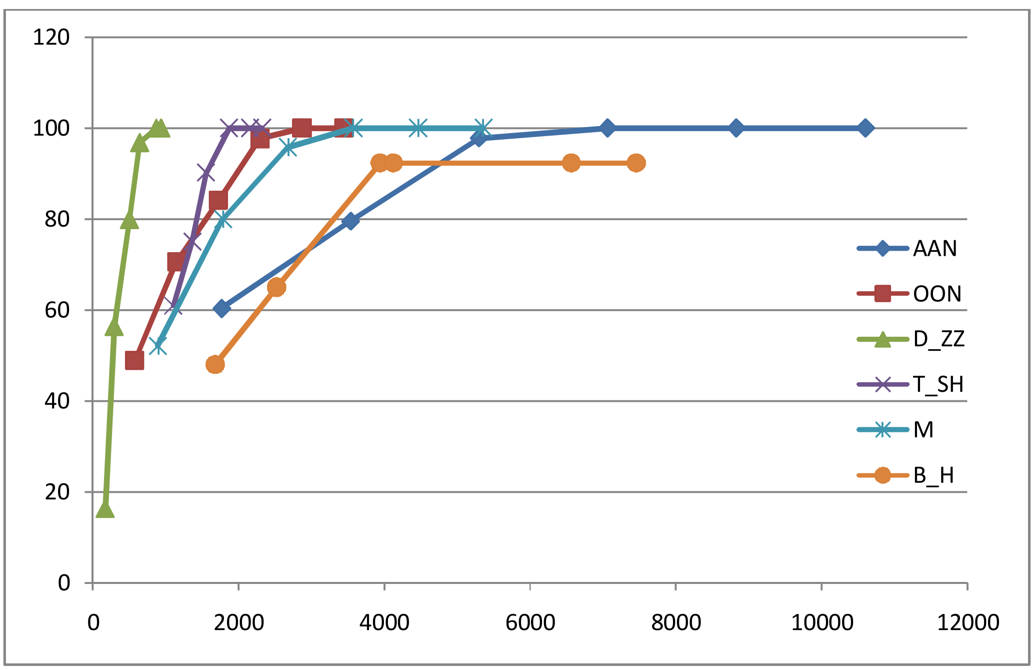
M	5354	100
B_H	7426	92.31

Table 11 : Phoneme accuracy with increamental training data

Phoneme	Original training data	Accuracy (%)
AAN	1766 3534 5298 7064 8830 10600	60.3 79.5 97.8 100 100 100
OON	574 1148 1722 2296 2870 3443	48.9 70.6 84.1 97.7 100 100
D_ZZ	167 289 500 640 870 935	16.3 56.4 79.9 96.8 100 100
T_SH	1100 1360 1552 1867 2151	61 75 90.2 100 100

	2318	100
M	892	52.1
	1784	79.9
	2676	95.8
	3568	100
	4460	100
	5354	100
B_H	1680	48
	2520	65
	3938	92.31
	4122	92.31
	6563	92.31
	7455	92.31

Percentage accuracy



Amount of training data

Figure 17 : Phoneme accuracy and training data

5.6.1 Experiment 6- Discussion

In Experiment-1, phonemes have been identified having low training data and high error rate. In Experiment-2, training data of these phonemes have been increased randomly and significant improvement in phoneme accuracy is achieved. This approach has been extended to ten speaker's ASR system. In Experiment-2, phonemes training data has been increased randomly to analyze the effect on accuracy. The aim in this experiment is to find the approximate value of training data of phonemes at which accuracy saturates. Greedy algorithms are very efficient in collecting phonetically rich and balanced training data. In such algorithms, same criteria have been developed to collect training data for all phonemes. From above experimental results, more efficient, cost effective and minimally balanced training corpus can be collected.

It can be seen from the Figures 14, 15 and 16 the amount of training data for each phoneme at which 100% accuracy can be achieved is different for each phoneme. E.g. from Figure 14, the amount of training data for AAN is 10600 and for OON is 3700. There might be a possibility that amount of training data has been saturated before 3700 for OON phoneme. This possibility has been explored and Table 10 shows the results.

In this experiment, training data has been increased incrementally to analyze the effect on phoneme accuracy. From Table-6, training data of phoneme AAN has been increased from 285 to 585 and error rate has been decreased from 16.6% to 0%. There might be a possibility that training data of this phoneme saturates below 585 and we can have same accuracy with less training data. It will reduce the recording and tagging effort. From the Figure 14, it can be seen that accuracy of AAN, OON, D_ZZ, T_SH, M and B_H phonemes saturates at amount of training data 7000, 2880, 850, 1800, 3500 and 3900 respectively. Table 10 and 11 shows the default and reduced amount of training data of these phonemes in phonetically rich corpus. Same accuracy for these phonemes has been achieved with minimal amount of training data.

Table-10 shows the phonemes training data that has been used to develop acoustic model in Experiment-3. Now this data has been divided in different parts and ASR system has been trained and tested. Training data of different phonemes and their accuracy has been shown in Table-11. It has been shown in Figure-17 that amount of training data at which phoneme accuracy saturates is different for every phoneme. The relationship between training data and saturation limits of different phonemes has been presented in Figure-17.

Each curve describing this relationship has different slope. The affricate D_ZZ has the best slope where as vowel AAN shows the worst slope. It means 100% accuracy can be achieved on minimum amount of training data for D_ZZ phoneme where as large training is required for AAN phoneme. From Table-11, the accuracy of phoneme B_H saturates at 92.31%. Further increase in training data of this phoneme has no effect on accuracy. There might be transcription or pronunciation error as same speech corpus has been used in Experiment-2 and 100% accuracy has been achieved for this phoneme.

It can be seen that accuracy of phonemes depends largely on training of ASR system. The slopes of saturation curves of phonemes largely depend on the speech corpora and pronunciation of phonemes in sentences. The saturation value is different for all these phonemes.

Chapter 6- Conclusion and Future Direction

The efficiency of spontaneous ASR system is not very convincing. Five set of experiments have been developed to analyze the performance of ASR systems. The aim is to find the recognition issues on small system and extend it to a larger one. In this regard, Experiment-1 has been developed one speaker. Recognition issues have been discussed. Generally, these problems are same as addressed by other researchers in developing ASR systems in other languages. Different reasons of confusions between phonemes have been discussed by performing phoneme error analysis. In Experiment-2, different techniques have been applied to Experiment-1 and improvement in recognition has been discussed. Training data of some phonemes has been increased whose original training data was low.

Experiment-3 has been developed by increasing number of speakers from one to ten. Phoneme analysis has been performed on this system. Same issues have been found as discussed in Experiment-1. Proposed techniques, for issues of Experiment-1, have been applied to Experiment-4. Performance of integration of one speaker data has been evaluated in Experiment-5. It also shows the performance of cleaned speaker data. In Experiment-6, criteria have been developed to collect minimally balanced corpus.

It can be concluded that the training data will improve the acoustic model if we collect speech data by taking in consideration the issues enlighten by error analysis of above systems. Moreover, error analysis of recognition results can be used to improve new integration of collection of speech data. Number of speakers can be increased after checking the combined performance. In this way speaker independent system with low word error rate can be developed. Criteria for developing minimal balanced corpus have been discussed. If we collect corpus by using this technique, less training data of phonemes is required to achieve maximum accuracy.

The aim of this thesis is to address and analyse the issues in developing large vocabulary speaker independent ASR system. The possibility of integrating new speaker's data in existing ASR system has been explored. In future, this method can be used to integrate multiple speakers' data in baseline ASR system. In second phase of this experiment, corpus development criteria have been discussed. Training data of five phonemes have been analysed. In future, this concept can be extended to all phonemes of Urdu. Phoneme analysis can be included in CMU sphinx toolkit to evaluate the performance. We have seen from above experiments that different training data is required for all phonemes so greedy algorithms can be modified to develop minimally balanced corpus. This method can be automated to develop minimally balanced corpus for Urdu.

Bibliography

[1] http://www.unicef.org/infobycountry/pakistan_pakistan_statistics

[2] J. Daniel & J. H. Martin, Speech and Language Processing: *An introduction to natural language processing computational linguistics, and speech recognition*, 2005.

[3] A. Raza, S. Hussain, H. Sarfraz, I. Ullah and Z. Sarfraz, "An ASR System for Spontaneous Urdu Speech", In the Proc. of Oriental COCOSDA, Kathmandu, Nepal. 24-25 November 2010.

[4] A.Samoulian,"Knowledge based approach to speech recognition", Department of Electrical and Computer Engineering University of Wollongong.

[5] L. Deng , H. Strik, "Structure-Based and Template-Based Automatic Speech Recognition, Comparing parametric and non-parametric approaches".

[6] Nirav S. Uchat, *Hidden Markov Model and Speech Recognition*.

[7] HTK, http://htk.eng.cam.ac.uk, accessed July 2010.

[8] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary Continuous speech recognition," ISCA, 2005. Ninth European Conference on Speech Communication and Technology.

[9] L. Villase.or-Pineda, M. Montes-y Gomez, D. Vaufreydaz, and J. F. Serignat, "Experiments on the construction of a phonetically balanced corpus from the web," Lecture notes in computer science, pp. 416–419, 2004.

[10] A. Li, F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani, and X. Chen, "Cass: A phonetically transcribed corpus of mandarin spontaneous speech," ISCA, 2000. Sixth International Conference on Spoken Language Processing.

[11] G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh, R. N. V. Sitaram, and S. P. Kishore, "Development of indian language speech databases for largevocabulary speechrecognition systems,"

[12] V. Chourasia, K. Samudravijaya, and M. Chandwani, "Phonetically rich hindi sentence corpus for creation of speech database," Proc. O-COCOSDA, p. 132–137, 2005.

[13] G. Raškinis, "Building medium-vocabulary isolatedword lithuanian hmm speech recognition system," Informatica, vol. 14, no. 1, pp. 75–84, 2003.

[14] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in

greek: Corpus and an automatic dictation system," ISCA, 2003. Eighth European Conference on Speech Communication and Technology.

[15] D. Binnenpoorte, C. Cucchiarini, H. Strik, and L. Boves, "Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling," p. 681–684, 2004. Proceedings of the International Conference on Language Resources and Evaluation (LREC).

[16] H. X Acero, A. and Hon, H-W. Spoken Language Processing: *A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[17] CMU Sphinx Open Source Toolkit for Speech Recognition Project by Carnegie MellonUniversity, http://cmusphinx.sourceforge.net/, accessed July 2010.

[18] Lee, E. Akinobu, T. Kawahara, K. Shikano, Julius an Open Source Real-Time Large Vocabulary Recognition, Nara Institute of Science and Technology, Japan Kyoto University, Japan, 2001.

[19] Praat: doing phonetics by computer, www.fon.hum.uva.nl/praat, accessed June 2010.

[20] S. Hussain, , "Letter to Sound Rules for Urdu Text to Speech Sytem", proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages, COLING 2004,Geneva, Switzerland, 2004.

[21] Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah and Zahid Sarfraz, "An ASR System for Spontaneous Urdu Speech, In the Proc. of Oriental COCOSDA, Kathmandu, Nepal. 24-25 November 2010".

[22] The CMU Statistical Language Modeling (SLM) toolkit. www.speech.cs.cmu.edu/ SLM_info.html, accessed July 2010.

[23] V. Digalakis, D. Oikonomidis, D.Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in greek: Corpus and and automatic dictation system, "ISCA, 2003, Eight European Conference on Speech Communication and Technology.

[24] G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh, R. N. V. Sitaram, and S. P. Kishore, "Development of Indian language speech database for large vocabulary speech recognition system".

[25] A. L. Ronzhin, R. M. Yusupov, I.V. Li, and A. B. Leontieva, "Survey of Russian speech recognition systems".

[26] A. Li, F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani, and X. X. Xhen, " Cass: A phonetically transcribed corpus of mandrain spontaneous speech, ISCA, 2000. Sixth International Conference on Spoken Language Processing.

[27] S. T Abate, W. Menzel, and B. Tafila, “An Amharic speech corpus for large vocabulary continuous speech recognition,” ISCA, 2005. Ninth European Conference on Speech.

[28] V. Chourasia, K. Samudravijaya, and M. Chandwant, “Phonetically rich Hindi sentence corpus for creation of speech database,” Proc. O-Cocosda,p. 132-137, 2005.

[29] P. A Heeman, “The American English sala-II data collection,” 2004. Proceedings LREC.

[30] G. Raskinis, “Building medium vocabulary isolated word Lithuanian HMM speech recognition system,” Informatica, vol. 14, no. 1, pp.75-84, 2003.

[31] Y. C. Yio, M. S Liang, Y.C. Chiang, and R. Y. Lyu, “Biphone rich versus triphone rich: a comparison of speech corpora in automatic speech recognition,” pp.194-197, 2005. Cellular Neural Networks and their applications, 2005 9th International workshop.

[32] A. C. Kelly, H. Berthelsen, N. Campbell, A. Chasaide, C. Gobl , “Corpus Design Techniques for Irish Speech Synthesis Phonetics and Speech Laboratory”, SLSCS, Trinity College Dublin, Ireland, 2006.

[33] V. Rangarajan and S. Narayanan, “Analysis of disfluent repetition in spontaneous speech recognition”, Proc. EUSIPCO 2006.

[34] T. Slobada and A. Waibel, “Dictionary learning for spontaneous speech recognition”, in spoken language, 1996. ICSLP 96. Proceedings., Fourth International Conference, 1996.

[35] J. Duchateau, T. Laureys, and P. Wambacq, “Adding robustness to language models for spontaneous speech recognition”, in COST278 and ISCA tutorial and research workshop (ITRW) on robustness issues in conversational interaction, ISCA, 2004.

[36] P. david, H. G. Hirsch, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, 2000.

[37] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajiˇc, D. Oard,M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, “Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Transactions on Speech and Audio Processing”, 2004.

[38] O. Siohan, B. Ramabhadran, G. Zweig, “Speech Recognition Error Analysis on the English MALACH Corpus”, 2004.

[39] T. Hirsimaki and M. Kurimo, “Analysing Recognition Errors in Unlimited-Vocabulary Speech Recognition”, 2009.

[40] C.K. Raut, K. Yu, and M.J.F. Gales, "Adaptive training using discriminative mapping transforms", ;in Proc. INTERSPEECH, 2008, pp.1697-1700.

[41] L. Chunhua, X.U. Mingxing, Z. Fang, “Acoustic Level Error Analysis in Continuous Speech Recognition”, Center of Speech Technology , State Key Laboratory of Intelligent

Technology and Systems, Department of Computer Sciences & Technology, Tsinghua University, Beijing, 2000.

[42] F. Zheng, Z. Song, M. X. Xu, J. Wu, Y. Huang, W. Wu, C. Bi., EasyTalk, “A Large-Vocabulary Speaker-Independent Chinese Dictation Machine”, EuroSpeech’99, Vol.2, pp.819-822, Budapest, Hungary, Sept.1999.

[43] T. Pruthi, Saksena, S and Das, P. K. Swaranjali, “Isolated Word Recognition for Hindi Language using VQ and HMM”, International Conference on Multimedia Processing and Systems (ICMPS), IIT Madras, 2000.

[44] K. Kuldeep, R. K. Aggarwal, “Hindi speech recognition system using htk”, Internation journal of computing and business ISSN(online) :2229-6166, volume 1,May 2011.

[45] M. U. Akram and M. Arif, “Design of an Urdu Speech Recognizer based upon acoustic phonetic modelling approach”, IEEE INMIC 2004, pp. 91-96, 24-26 December, 2004.

[46] S. M. Azam, Z. A. Mansoor, M. S. Mughal, S. Mohsin, “Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network”, IEEE Computer Graphics, Imaging and Visualisation CGIV, Bangkok, 14-17 August, 2007.

[47] J. Ashraf, N. Iqbal, N. S. Khattak, A. M. Zaidi, “Speaker Independent Urdu Speech Recognition Using HMM”, INFOS, IEEE, Cairo, 28-30 March, 2010.

[48] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen “Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System”, Oriental COCOSDA 2010 conference, Nov. 24-25, 2010, Katmandu, Nepal.

[49] W. Chai, P. Cotsomrong, S. Suebvisai, S. Kanokphara, Information Research and Development Unit National Electronics and Computer Technology Center, Phonetically Distributed Continuous Speech Corpus for Thai Language, COCOSDA, 2003.

[50] J. Nedel, R. Singh, and R. Stern, “Automatic Subword Unit Refinement for Spontaneous Speech Recognition via phone splitting,” in sixth international conference on spoken language processing, ISCA, 2000.