Automatic Transcription of Continuous Speech for Indian Languages

A Thesis

Submitted by

G. Lakshmi Sarada

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS

December 2005

THESIS CERTIFICATE

This is to certify that the thesis entitled **Automatic Transcription of Continuous Speech for Indian Languages**, submitted by **G. Lakshmi Sarada**, to the Indian Institute of Technology, Madras, for the award of the degree of Master of Science (by Research), is a bona fide record of the research work carried out by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Date:

Chennai 600 036. (**Dr. Hema A. Murthy**)

Acknowledgements

I am grateful to my supervisor Dr. Hema A. Murthy for her guidance and support all through my stay as an MS student at the department of computer science. You did help me a lot in improving my writing skills. Thank you Ma'am!

Special thanks to my co-supervisor Dr. T. Nagarajan, whose door was always open to my discussions and arguments. I gratefully acknowledge his warm encouragement and patient guidance. Thank you sir!

I am grateful to the Head of the department Prof. Timothy A. Gonsalves, who has taught me Operating systems. I am also thankful to our former HoD Prof S. Raman for providing necessary facilities in the department. I would like to thank all the office staff of our department for their official support.

I am indebted to the members of the GTC committee: Dr. C. Chandra Sekhar and Dr. C. S. Ramalingam for reviewing my reports and for their useful remarks.

I would like to express my thanks to all members of DON Laboratory; too many to mention. Especially I would like to thank Sridhar, Deivapalan and Hemalatha for their support.

I have always cherished the hostel life in IIT with my friends Swati, Bhanu and Wenonah.

My special thanks go to my parents, brothers and sister. I reckon growing up in such a lively, stimulating and therefore inspiring family life has been of considerable influence on my conduct in life. The kind of love and support I received from my

parents is outright. To them I dedicate this thesis.

Needless to mention, I express my deepest gratitude to my husband Pramod for the constant support, understanding and love that I received from him. Without his support it would not be possible for me to finish this thesis.

I would like to thank everybody who has contributed to the realization of this thesis.

Abstract

Over the last decade, speech recognition systems have come to be increasingly used in automated systems with spoken language interfaces. Systems with spoken language interfaces make a significant contribution to the realization of local language interfaces in the Indian scenario. There has been a substantial progress in speech technology, with today's state-of-the-art systems being able to transcribe unrestricted broadcast news speech data with good accuracy. However, the number of speech recognition systems that support Indian languages is very small.

The conventional method of building a speech recognizer for any language requires a labeled and segmented speech corpus. However, obtaining such a corpus is a time consuming and expensive process. It also requires trained human annotators and a substantial effort in supervising them. For Indian languages, if manually annotated speech corpora are required, building such speech recognizers even for the official languages is a difficult task.

The focus of the work presented in this thesis is to automate the transcription task for Indian languages. Automatic transcription of continuous speech is the automatic conversion of arbitrary or unrestricted natural language sentences from its spoken form into its text form. Given that Indian languages are syllable centred, the focus of this work is to segment the continuous speech signal into syllable-like units and then perform an isolated style recognition of syllable-like units. In particular this thesis builds upon the technique used for automatic speech recognition presented in [1]. The syllable training technique used in [1] is considered as the baseline technique for this thesis. Though the syllable recognition results obtained using the baseline system are promising for speech recognition, there are many drawbacks in the segmentation and clustering process.

This thesis mainly outlines the issues in segmentation and clustering process. Sev-

eral refinements are then made to the baseline system and a significant improvement in the syllable recognition performance is obtained. A new feature extraction technique that uses features extracted from multiple frame sizes and frame rates during both training and testing is explored for the continuous speech recognition task. The syllable recognition performance is evaluated for two Indian languages namely, Tamil and Telugu. A syllable recognition performance of 48.7% and 45.36%, is obtained for Tamil and Telugu respectively.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the last decade, speech recognition systems have come to be increasingly used in automated systems with spoken language interfaces. There has been a substantial progress in speech technology, with today's state-of-the-art systems being able to transcribe unrestricted broadcast news speech data with good accuracy. However, in India that has 22 official and a number of unofficial languages, very few systems support speech recognition. The conventional method of building a speech recognizer for any language requires a labeled and segmented speech corpus. However, obtaining such a corpus is a time consuming and expensive process. It also requires trained human annotators and a substantial effort in supervising them. For Indian languages, if manually annotated speech corpora are required, building such speech recognizers even for the official languages is a difficult task.

The focus of this thesis is to automate the transcription task for Indian languages. Automatic transcription of continuous speech is the automatic conversion of arbitrary or unrestricted natural language sentences from its spoken form into its text form. Before embarking on the task of automatic transcription, we first review the subword based continuous speech recognition (CSR) system in Section 1.1. The issues involved in automatic transcription of continuous speech are then outlined in Section 1.2. Finally, the most commonly used techniques for the speech transcription task are briefly explained in Section 1.3. Section 1.4 gives an overview of the thesis.

1.1 Overview of continuous speech transcription

Continuous speech transcription aims at converting continuous speech data into text. The characteristics of the sound units in continuous speech are different from the characteristics of isolated utterances of the same speech unit. In continuous speech the characteristics of sub-word units are affected by the context and the speaking rate. Hence, the recognition of sub-word units excised from the continuous speech signal is a difficult task when compared to the recognition of isolated utterances of units.

In continuous speech transcription task, the sub-word unit can either be a phoneme or a syllable, depending on the language. The focus of this thesis is to develop an automatic continuous speech transcription tool for Indian languages. Since all Indian languages are syllable timed languages the sub-word unit considered in this thesis is a syllable.

A block diagram of the overall continuous speech transcription system based on sub-word speech units is shown in Figure 1.1. The first step in processing is spectral analysis to derive the feature vector, which is used to characterize the spectral properties of the speech input. In this thesis, we consider spectral vectors with 39 components consisting of 13 Cepstral components, 13 delta Cepstral components, 13 delta-delta Cepstral components. The second step in the speech recognizer is a combined word-level or sentence-level match. The way this is accomplished is as follows. Using the set of sub-word Hidden Markov Models (HMMs) and the word lexicon, a set of word models (HMMs) is created by concatenating each of the sub-word unit HMMs as specified in the word lexicon. The way in which the sentence-level match is done is via a finite state network (FSN) realization of the word grammar (the syntax of the system) and the semantics as expressed in a composite FSN language model. The implementation of the combined word-level match or sentence-level match is via any specified structure. In particular, most systems use structures similar to beam search to restrict the range

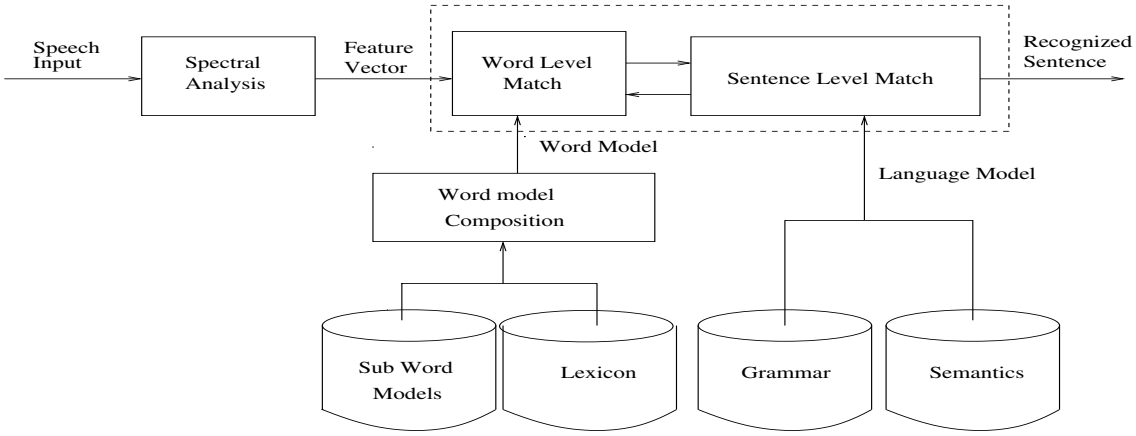

Figure 1.1: Overall block diagram of sub-word unit based continuous speech transcription system

of paths to find the best recognition sentence given an erroneous transcription. In this thesis, the focus is primarily on recognition of sub-word units. Since Indian languages are syllable centred, syllables correspond to the sub-word units.

Though researchers have tried several ways to automate speech transcription, without compromising accuracy of models trained from continuous speech data, there are several issues which make the automatic transcription of continuous speech a difficult task.

1.2 Issues in automatic transcription of continuous speech

There has been a substantial progress in speech technology, with today's state-of-art systems being able to transcribe unrestricted broadcast news. Aside from the substantial progress, there are many challenging issues in this area. Translation of a continuous speech signal into a sequence of words is a difficult task, as continuous speech does not have any natural pauses in between sub-word units. The conventional method of building a speech recognizer for any language requires a large amount of segmented and labeled speech corpus for training. However obtaining such a corpus

is a labour intensive and time consuming process. Another challenge is to reduce the cost, both in terms of human effort and financial needs, when it is required to adapt a recognition system to a new task or another language. In a country like India, that has 22 official and a number of unofficial languages, if manually annotated speech corpora are required, building such speech recognizers even for the official languages is a difficult task. The quality of automatic segmentation and labeling is potentially of great significance for continuous speech recognition systems, as word-error rate greatly depends on the accuracy of segmentation and sub-word unit recognition [2].

Allwood [3], has pointed out that the frequency of words and grammatical constructions in spoken and written language are vastly different. Therefore, speech cannot be viewed as an exact translation of written language into its spoken form. Further, speech is uttered in an environment of different sounds generally termed *noise* which consists of unwanted information in the speech signal. Channel variability caused by the prevalent noise and use of different microphones makes speech recognition a formidable task.

As speech recognition involves matching of speech signal with an already existing group of models, the lexicon needs to be kept small, in order to reduce the search space. This causes another problem called *out-of-vocabulary*, which means that the intended unit is not present in the lexicon. Automatic speech recognition (ASR) should also be capable of handling out-of-vocabulary issues in an efficient manner. Speaker variability, such as variation in speaking style, gender of the speaker and speaking rate, also affect speech recognition performance.

1.3 Overview of automatic speech transcription systems

Researchers have tried several ways to automate speech transcription, without compromising the accuracy of models trained from untranscribed data. Some of the commonly

used techniques for speech transcription are briefly explained here.

- **Rabiner** and **Rosenberg** [4], have proposed a new method called bootstrapping for speech recognition. The amount of training data, that is required to transcribe large amount of new data, is increased using this method. This newly transcribed data is then used to retrain the model parameters, thereby increasing the amount of training data.
- **Ljolje** [5], has used an automatic approach to segmentation and labeling of speech when only the orthographic transcription of speech is available. Orthographic transcription is nothing but a verbatim record of what is actually spoken.
- **Kemp** and **Waibel** [6], have proposed a method for unsupervised training of the speech recognizer for TV broadcasts. In this procedure, with the transcribed portion of the data, a bootstrap recognizer is built, which is used to generate transcripts of the untranscribed training material. To exclude the erroneous words from these transcripts, a measure of confidence is applied. As a last step, a new recognizer is trained on the remainder of the hypothesized words.
- **Jean-marc** [7], has proposed a hybrid approach that combines HMM and ANN for ASR, that initially uses manually segmented and labeled BREF [8] corpus.
- **Dilek** [9], has proposed a new method for reducing the transcription effort for training the ASR. The number of training examples to be labeled is reduced by automatically processing the unlabeled examples. The most *informative* (incorrectly recognized) ones are then selected with respect to a given cost function and added to the training set.

The basic idea behind all the above mentioned efforts is, to use an existing speech recognizer to transcribe large amount of untranscribed data, which can further be used to refine the trained models. For a new language, if no such speech recognizer is available,

few hours of data is manually transcribed and this is used to build a recognizer. This, in turn, is again used to increase the amount of training data by transcribing large quantities of untranscribed data. The two basic problems in the above mentioned techniques are (i) if there is a mismatch in the environment or language during transcription, the recognition performance is expected to be very poor, and (ii) if this newly transcribed data is going to be used for further refining the model parameters, convergence of the training process may be very slow and in few cases, it may be impossible.

The immediate alternative to this problem is, manually transcribing part of the new data, which is taken from a different environment, building models using this data and then using these models to transcribe the rest of the data.

A few methods for transcribing the untranscribed speech data and that uses minimal manually transcribed data considered from a different environment for training, are explained below.

- **Matthew** [10], has proposed a method in which for both segmentation and clustering problems, the symmetric Kullback-Leibler distance is taken as the solution. The symmetric Kullback-Leibler distance is an effective distance metric to facilitate the detection of long-term statistical differences in speech signals. In this, the system is able to detect changes in acoustic conditions and recognize previously observed conditions and this is used to further pool the data.
- In the experiments carried out by **Zavaliagkos** [11] at Bulletin Board Network (BBN) Technologies, completely unsupervised acoustic training data from a conversational speech corpus is combined with 3 hours of manually annotated data. It is shown that a lot of untranscribed data is needed to achieve comparable levels of performance with transcribed data.
- **Frank Wessel** [12], has proposed an approach in which a low-cost recognizer trained with one hour of manually transcribed speech is used to recognize 72

hours of unrestricted acoustic data. These transcriptions are then used to train an improved recognizer.

- **Lamel** [13], has shown that the acoustic models can be initialized using as little as 10 minutes of manually annotated data.
- **Gunawardana** [14], has proposed an unsupervised adaptation of acoustic models to a domain with mismatched acoustic conditions. Estimation of acoustic models from untranscribed acoustic data is done using Expectation Maximization (EM) algorithm.
- **Stavros** and **William** [15], have proposed an approach in which the acoustic model training is done using heterogeneous data sources. This approach is an application of model-based acoustic normalization technique that maps out-of-domain feature space onto the in-domain feature space.

In all the above mentioned techniques, some amount of manually transcribed data, which is considered from a different environment, is needed for transcribing the remaining untranscribed data. However, for transcribing even a small amount of data, trained human annotators are needed.

A few methods that do not require any manually annotated speech corpora for speech recognition are given below.

- Incremental maximum a posteriori estimation of HMMs that is proposed by **Yoshihoko** [16] improved the efficiency of training and also improved performance of a speech recognition system. This algorithm has randomly selected a subset of data from the training set, updated the model using maximum a posteriori estimation, and this process is iterated until convergence occurs. This training strategy is in sharp contrast with standard batch training where the model parameters are updated only after all the data in the training set are processed.

- **Lokendra Shastri** [17], has proposed a new method in which articulatory-acoustic phonetic features are extracted from each frame of the speech signal and classification of phone is done by special purpose neural networks. The output of these networks is processed by a Viterbi-like decoder to produce a sequence of phonetic segment labels along with boundary demarcations associated with each segment.

- **Meinedo** [18], has used acoustic model alignment and training procedures. The automatic segmentation procedure detects changes in the acoustic conditions and marks those time instants as segment boundaries. This work results in the segmentation of audio into homogeneous regions according to background conditions and speaker gender. A Hybrid HMM/MLP speech recognition system [19] [20] is used to build the acoustic models for the broadcast news task.

- **Nagarajan** [21] has proposed a novel approach for automatically segmenting and transcribing continuous speech signal without the use of any manually annotated speech corpora for Language Identification (LID) task. In this approach, the continuous speech signal is first automatically segmented into syllable-like units and similar syllable segments are grouped together using an unsupervised and incremental clustering technique. Separate models are generated for each cluster of syllable segments and labels are assigned to them.

Though the afore mentioned techniques do not require manual transcriptions, they are inferior to the speech recognition systems that use manually segmented and labeled speech corpus.

In this thesis, we build upon the technique used for automatic speech recognition presented by **Nagarajan** in [21]. The syllable training technique based on unsupervised incremental training (UIT) used in [21] is considered as the baseline technique for this thesis. Though the syllable recognition results obtained using the baseline system are comparable to the presently existing techniques, there are many draw backs in the

baseline system due to both segmentation and clustering errors. Since, in [21], the focus is primarily on LID, no effort has been made to validate the syllables obtained. This thesis mainly outlines the issues due to segmentation and clustering. Several modifications are then made to the baseline system and a significant improvement in the recognition performance is obtained. This technique is referred to as the modified unsupervised incremental training (MUIT). A feature extraction technique that uses features extracted from multiple frame sizes and multiple frame rates during both training and testing is explored to improve the syllable recognition performance. This technique is used along with the MUIT technique and a significant improvement in the syllable recognition performance is obtained over the baseline system. The syllable recognition performance is evaluated for two Indian languages namely, Tamil and Telugu. In the following Section a brief overview of the thesis is presented.

1.4 Organization of the thesis

In Chapter 2, several unsupervised training approaches for ASR are reviewed. The base-line Unsupervised and Incremental Training (UIT) technique for continuous speech transcription is then discussed in detail.

In Chapter 3, major issues in the existing baseline system are outlined. In order to overcome the issues in the baseline UIT system, several modifications are made to the segmentation and initial cluster selection algorithms. The resulting training technique is named as Modified Unsupervised and Incremental Training (MUIT) technique. The syllable recognition performance results that are obtained using MUIT are compared with the baseline UIT system.

In Chapter 4, a new feature extraction technique that uses Multiple Frame Sizes (MFS) and Multiple Frame Rates (MFR), is explored for automatic speech recognition. This technique is applied to manually segmented data. A significant improvement

in syllable recognition performance is obtained when MFS and MFR feature extraction technique is used both during training and testing for an isolated style syllable recognizer.

In Chapter 5, the MFS and MFR feature extraction technique is applied to the continuous speech transcription task where automatically segmented data is used. Features are extracted with multiple frame sizes and frame rates from an automatically segmented speech corpus and models are trained using the MUIT technique. The syllable recognition results obtained using MUIT system with Single Frame Size (SFS), is compared with the MUIT system with MFS and MFR features. These results are also compared with the baseline UIT system. Performance evaluation of the automatic speech transcription system that is built using the proposed techniques in this thesis, is done for the Indian languages Tamil and Telugu. The automatic transcription examples for both the languages Tamil and Telugu are given and they are compared with the corresponding manual transcriptions.

In Chapter 6, summary and main contributions of the work presented in this thesis are given. Future directions based on this work are outlined.

1.5 Major contributions of the thesis

The major contributions of the work presented in this thesis are listed below.

- Analysis of the drawbacks in the unsupervised and incremental training technique are analyzed.
- Several modifications such as duration analysis, silence normalization, pruning of final syllable clusters are made to the baseline system and a significant improvement in the syllable recognition performance is obtained for a CSR task.
- A feature extraction technique, that uses multiple frame sizes and frame rates

during both training and testing, is explored to improve the continuous speech recognition performance. This technique is used in the MUIT technique and a significant improvement in the syllable recognition performance is obtained over the baseline UIT system.

Chapter 2

Background to unsupervised training for automatic speech transcription

2.1 Introduction

Many areas in speech research such as speech synthesis, speech recognition and speech analysis require a large corpus of accurately transcribed speech. An accurate method of phonetic transcription could potentially facilitate development of automatic speech recognition systems for new speech corpora, both within and across languages, as well as increase robustness with respect to environment variability and variation in speaking style and pronunciation. In the past, in order to handle the environmental and speaker variability, large amounts of speech were recorded from different speakers and in different environments and were transcribed manually for training. But manually transcribing speech data is a labour intensive and time consuming process. To counter this problem, an unsupervised and incremental training procedure, that does not require any manually transcribed data, is explored in this thesis.

In Section 2.2 some of the relevant literature for ASR using supervised techniques is reviewed and their limitations discussed. Section 2.3 outlines few popular techniques for unsupervised training of speech models. The baseline unsupervised and incremental training technique which is studied in this thesis is reviewed in Section 2.4.

2.2 Supervised Training

In supervised training, the transcription corresponding to the unknown utterance is available. Several algorithms for supervised training of acoustic models have been proposed and refined over the last decade.

Subrata [22], has proposed a supervised training technique in which the models for training are built from one part of the training database. They are tested based on their effectiveness on an independent part of the same database. In **Lamel** [23], and **Nguyen** [24], a lightly supervised acoustic model training on data having less accurate transcripts is proposed. In this study, the automatic transcripts generated by a recognition system are filtered with closed-caption transcripts. Only a subset of this data is chosen for training.

In both these techniques, they require manually transcribed data to train the models. More over, the focus of these techniques is on word recognition rather than sub-word unit recognition. In these systems, the acoustic match is carried using word models. The complexity in finding the acoustic match increases with the length of speech utterance. The present day automatic speech transcription systems can not accept arbitrarily long speech utterances for speech recognition. Considering this issue, the continuous speech recognition can be modified as shown in Figure 2.1 that considers segmented sub-word units both during training and testing. In [25], it has been shown that syllable is a reliable candidate for segmentation. It is also shown that co-articulation is minimum at the syllable unit level compared to the phoneme units within the syllable.

In [25], considering syllable as the basic unit, it is shown that a simple isolated style recognition system gives a recognition performance of 65% (syllables recognized are the 1-best results), if segmentation is done accurately at the syllable boundaries. In this work, the recognition is performed on manually segmented data. In [26], an au-

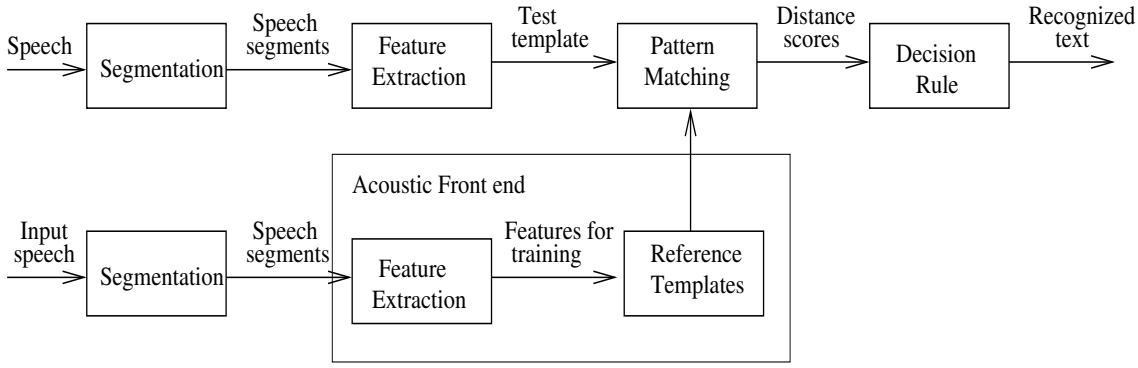

Figure 2.1: Model for Speech Recognition system that considers segmented sub-word units

tomatic segmentation procedure is proposed which automatically segments the speech signal into syllable-like units. Using this automatically segmented data, each segment has been tested against the syllable HMMs to find the HMM that gives maximum likelihood. In this approach, the syllable HMMs are trained using manually segmented and labeled speech corpus.

Though the supervised training techniques are superior in practice, they require annotated speech corpora for training the speech models. In a country like India that has 22 official and 5000 unofficial languages, building manually annotated speech corpora is a difficult and time consuming process. Hence, unsupervised model training techniques are required to overcome this limitation. In the following Section some of the unsupervised training techniques for building speech models, are described.

2.3 Unsupervised Training

The basic idea of unsupervised training is to train speech templates such that similar utterances are clustered together using untranscribed training data. In other words, these systems take only raw speech data without any manual transcription. We now review some of the unsupervised training methods..

- **Yoshihoko** [16], has proposed an approach for unsupervised training which uses Continuous Density Hidden Markov Model (CD-HMM) using a stochastic, incremental variant of the Expectation Maximization (EM) algorithm. This algorithm randomly selects a subset of data from the training set, updates the model using Maximum A Posteriori (MAP) estimation, and then iterates till a convergence criterion is satisfied.

- In [27], **Thomas** has proposed an approach which aims at training a speech recognizer with only a minimal amount (30 minutes) of transcriptions and a large portion (50 hours) of untranscribed data. A recognizer is bootstrapped on to the transcribed part of the data and initial transcripts are generated with it for the remainder (the untranscribed part). Using a lattice-based confidence measure, the recognition errors are partially detected and the remainder of the hypothesis is used for training.

- **Frank Wessel** [12], has proposed an unsupervised training technique for large vocabulary continuous speech recognition. A low-cost recognizer trained with only one hour of manually transcribed data is used to transcribe large amount of untranscribed data. They are further used in combination with confidence measures to train an improved recognizer.

- **Rukmini** [28], has proposed an unsupervised training technique to reduce transcription cost and time for developing a call routing application. An already existing transcription system trained with the Switchboard corpus is adapted to the call routing task using Maximum Likelihood Linear Regression (MLLR) adaptation and the final models are re-estimated using both Switchboard [29] and task dependent data.

- In [30], **Lori** has proposed an unsupervised training technique where the bootstrapped models are obtained using lightly supervised training and they are further used to transcribe untranscribed data using unsupervised training.

- **Dilek** [31], has proposed a novel approach that aims at reducing the amount of manually transcribed in-domain data required for building ASR models in spoken language dialog systems. This method is based on mining relevant text from various conversational systems and websites. It is claimed that the recognition performance is improved using unsupervised and active learning of the ASR models. The goal of active learning method is to select the examples which will have the largest improvement on the performance, thus reducing the amount of effort in human labeling.
- In [9], **Giuseppe** has proposed an active and unsupervised learning approach for automatic speech recognition. In this approach, the number of examples to be labeled are reduced by automatically processing the unlabeled examples and then selecting the most informative ones using a given cost function. Further, unsupervised training is used to transcribe the remaining untranscribed data by using the ASR output along with word confidence scores.

The main drawback of the above mentioned approaches is that they all require some amount of transcribed training data (in-domain or out-of-domain) to build the initial models for unsupervised training of the remaining data.

Although speech recognition performance using supervised training is superior to unsupervised training, it suffers from one important limitation. Models for all the speech utterances have to be trained using a large amount of segmented and labeled speech data. Using the above mentioned techniques, building large vocabulary continuous speech recognizer for languages, where transcriptions are not available, is an impossible task. Especially in India, building such speech recognizers, using these techniques, is a very difficult task even for the official languages. In the following Section, an unsupervised incremental training technique is explored to automate the continuous speech transcription task for Indian languages.

2.4 Unsupervised Incremental Training (UIT)

In [1], a novel approach is proposed for automatically segmenting and transcribing the speech signal without the use of any manually annotated speech corpora. This technique is used for Language Identification (LID) task. In this approach, the continuous speech signal is first automatically segmented into syllable-like units. Similar syllable segments are then grouped together using an unsupervised and incremental clustering technique. Separate models are generated for each cluster of syllable segments and labels assigned to them. It is observed that the unsupervised incremental training results in syllable clusters which are similar in CV or VC (C - consonant, V - vowel) and in some cases they correspond to different syllable examples of the same class. For LID task, each syllable identification is not important. It is sufficient if syllable or even polysyllable units with similarity in some of the constituent CV units are clustered together. Thus in the LID task, no effort is made to establish the identity of the units that are clustered together.

In the case of continuous speech transcription, syllable clustering greatly influences the transcription task. If each cluster consists of syllable segments corresponding to the same class, the syllable models that are generated will be statistically consistent, labeling will be easy and the final transcriptions will be accurate. The clustering process used in UIT results in some syllable clusters whose identity can be determined easily as syllable segments in each cluster correspond to the examples of the same class. But there are also a large number of clusters in which syllable segments corresponding to different classes of syllables are clustered together. The preliminary results obtained using the baseline system indicates that the UIT shows lot of promise for the continuous speech transcription task. Since the baseline UIT technique is modified in this thesis, we briefly review the UIT based syllable recognition system. This system consists of two phases: automatic segmentation followed by automatic labeling. The automatic segmentation technique is explained in Section 2.4.1. In the baseline technique, the

automatic labeling process comprises of two phases namely: (i) initial cluster selection (ii) incremental training. These techniques are explained in Section 2.4.2.

2.4.1 Automatic segmentation

Automatic segmentation of the speech signal is an essential tool for building large corpora for training speech recognition systems. Manual segmentation of a speech signal is both a time consuming and an error-prone task. The performance of segmentation algorithm is important for any speech recognition system. Segmentation is a process of decomposing the speech signal into a set of basic phonetic units. Further the model parameters of each phonetic unit are trained using different examples of the same phonetic unit. The basic phonetic unit can be a phoneme or a syllable, based on the language. Since Indian languages are syllable-timed languages, syllable is considered as the basic phonetic unit in this work. Several methods for automatic segmentation of speech have been proposed in the literature.

In [32] **Paul**, et al. (1975), has used a time-smoothed and frequency-weighted summation of the signal spectrum, for segmenting speech into syllable-like units. In this approach, the local minima in the loudness function gives the syllable boundaries.

Pfitzinger [33] uses a short-time energy based method to detect syllable nuclei. In this method, the speech signal is first band-pass filtered and then the short-term magnitude function (energy) is computed. Peaks of the resulting energy contour correspond to the syllable nuclei.

In [34], **Rudi** has proposed an automatic segmentation algorithm, which is similar to the Mermelstein's algorithm [33], in which intensity peaks are considered as syllable nuclei and intensity troughs as syllable boundaries. The Howitt algorithm [35] is implemented by using amplitude onset velocity as the key indicator of boundaries.

An automatic segmentation algorithm has been proposed in [26] for segmenting

the continuous speech signal into syllable-like units. In this approach, the minimum

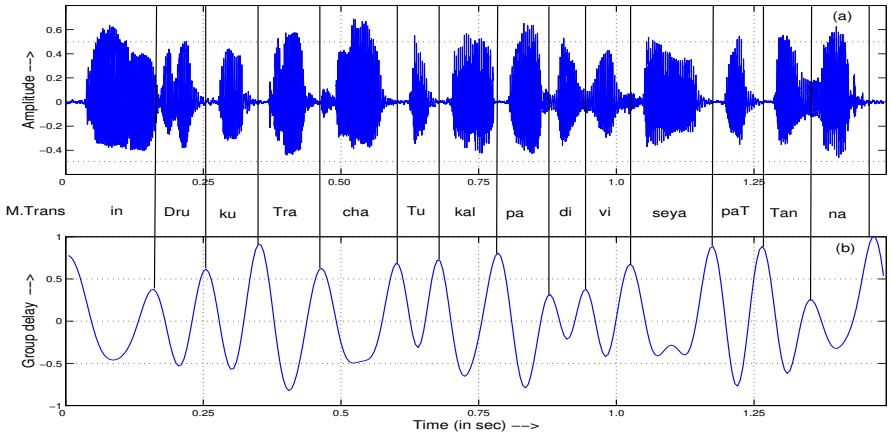

Figure 2.2: Segmentation of speech signal based on minimum phase group delay. (a) The speech signal (b) Group delay spectrum derived from the minimum phase speech signal (M.Trans - Manual transcription of the speech signal)

phase signal is derived from the short-term energy function. The group delay function of this signal is then computed. It is shown that the peaks in the group delay function correspond to the syllable boundaries. The same algorithm is used in this thesis to segment the speech signal into syllable-like units. An example of the segmentation of a speech signal is shown in Figure 2.2. Note that in Figure 2.2, location of peaks corresponds to the syllable boundaries. The syllable boundaries are shown by solid lines in the Figure. Figure 2.2.a shows the actual speech signal and Figure 2.2.b shows the group delay of the speech signal. In the Figure, M.Trans corresponds to the manual transcription of the speech signal.

2.4.2 Unsupervised and incremental clustering technique

The next stage in building a speech recognizer is to build acoustic models for the speech segments. This requires grouping of similar sounding units to get unique models for

each sound unit. Grouping of similar sounding units is performed using an unsupervised and incremental clustering algorithm that is explained in the following Sections.

2.4.2.1 Initial cluster selection

First the speech signals are automatically segmented into M syllable-like units. After segmenting the speech signal into syllable-like units, the first task is to select some unique syllable segments or groups of unique syllable segments for training. A subset of syllables are selected to initialize the HMMs. The flowchart for the initial cluster selection procedure is shown in Figure 2.3. We now outline the steps for initializing the HMMs.

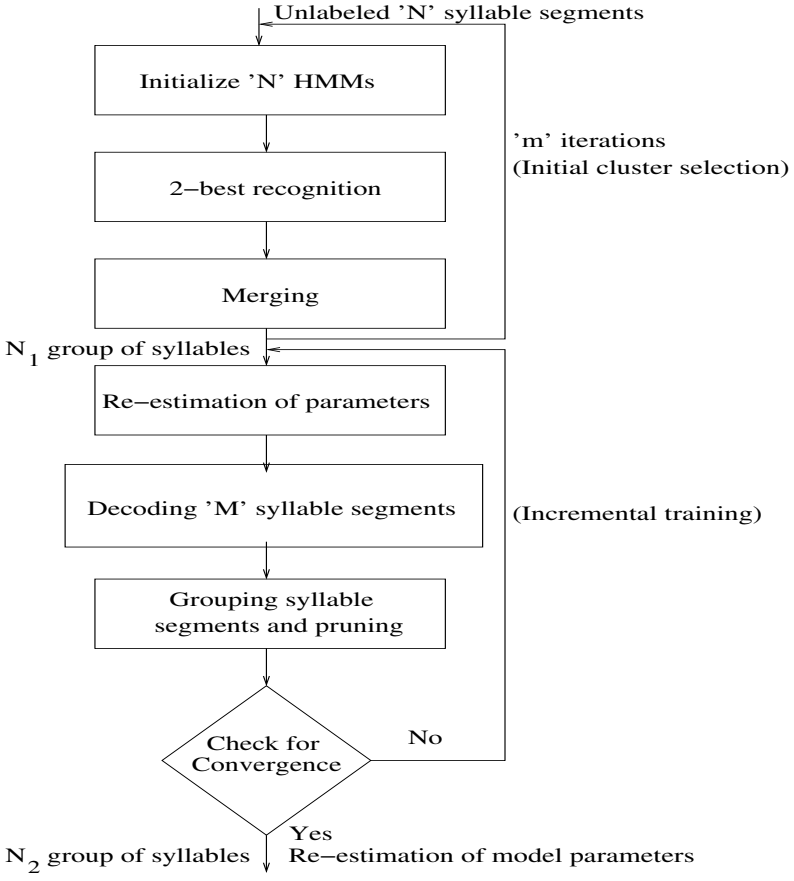

Figure 2.3: Flow chart of the unsupervised incremental training algorithm

1. Initially a subset of N syllable segments are chosen randomly from the M syllable segments to initialize N HMMs, where N < M. Each HMM corresponds to one of the N syllables. MFCCs are extracted from each of these N syllable segments with multiple frame sizes (MFS). The reason for using MFS feature extraction is, when a single example is considered and single frame size (SFS) features are used to build a HMM with more than one Gaussian mixture/state [1], the variance of these mixtures may become zero. There are several ways of increasing the variance and adjusting the mean value such as introducing some additive noise to the same example and considering it as a different example of the same class. In our case, instead of introducing some unknown noise, we consider features extracted from multiple frame sizes, thus increasing the number of examples for the corresponding class. In this approach, using a frame size, let us say F_1, MFCC features are extracted from a single example. The same example is considered again and features are once again extracted using a different frame size (say F_2). This process is repeated for several frame sizes (12,14,16,18, and 20ms) and the features extracted from these different frame sizes are given as input to the model initialization process. This procedure is carried out for all the N syllables and N HMMs are built. This procedure is illustrated below with an example. The spectrograms of a syllable /*kam*/ using different frame sizes of 12, 14, 16, and 18ms are shown in Figure 2.4. From Figure 2.4, a slight variation in the spectrograms for different frame sizes can be observed. This variation is sufficient to initialize the syllable models in the initial cluster selection procedure. This multiple frame size feature extraction procedure is repeated for several frame sizes and the features extracted from these different frame sizes are given to the model initialization process as separate examples.

2. After initializing N HMMs, all the N syllable segments are recognized against these N HMMs and 2-best recognition results are obtained, resulting in N pairs

[1] To capture the speaker variation of multiple speakers

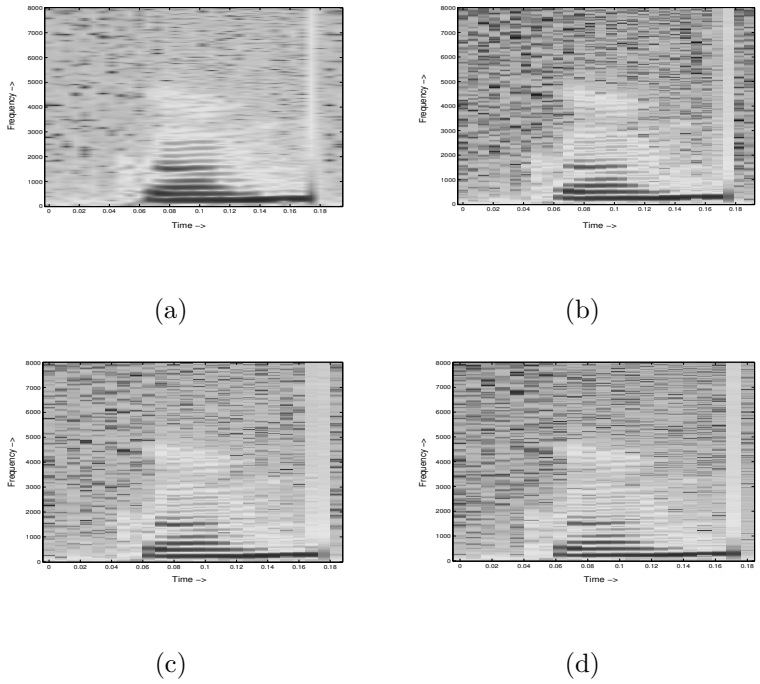

Figure 2.4: Spectrograms of a syllable segment using different frame sizes: (a) with 12ms frame size (b) with 14ms of frame size (c) with 16ms of frame size (d) with 18ms of frame size

of syllable segments.

3. The N pairs of syllable segments are then pruned. The pruning process is illustrated with an example in Figure 2.5.

   Let S_1, S_2, ..., S_N be the syllables obtained in step 1. If a syllable segment occurs more than once in the 2-best recognition results, the repeated syllable models are removed. From the Figure, the 2-best recognition result for the syllable S_i is $\{S_i, S_j\}$ and the 2-best recognition result for the syllable S_j is $\{S_j, S_k\}$. Since the syllable S_j occurs in the 2-best recognition results of both the syllable S_i and S_j, model for S_j is removed. The model for S_i is rebuilt using S_j as another example. The syllable S_k occurs in the 2^{nd}-best position in the 2-best result of the syllable S_k and also occurs in the 2-best results of the syllable S_j, model for S_k is therefore removed. This procedure is carried out using the 2-best recogni-

tion results of all the N syllable segments. This results in N_1 clusters, where N_1 < N.

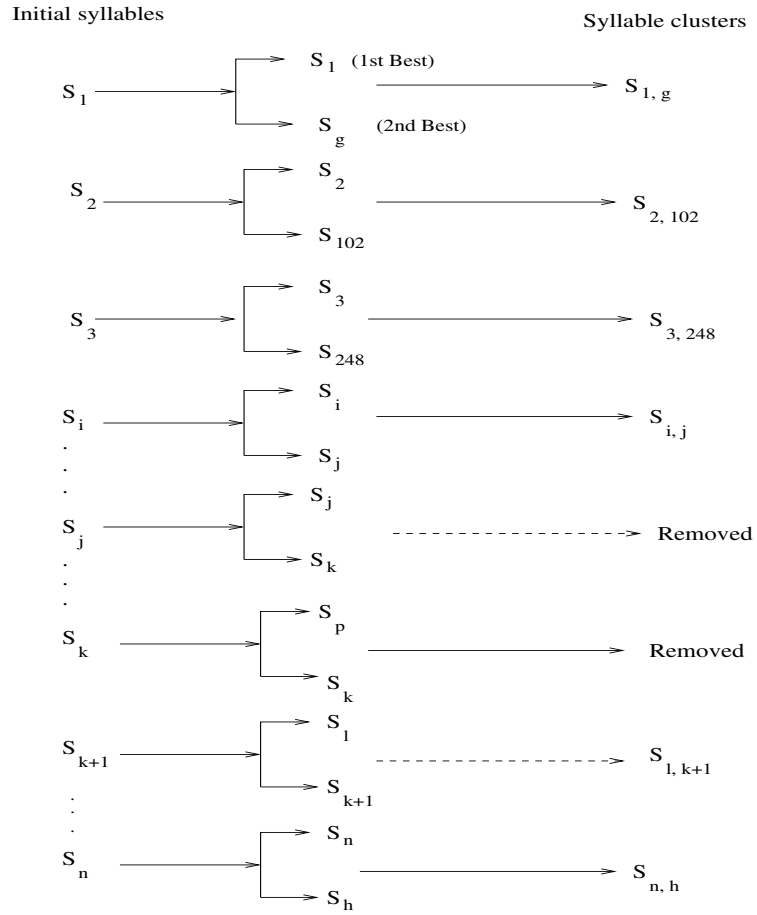

Figure 2.5: Pruning of 2-best recognition results obtained from the initial cluster selection procedure

4. Steps 2-3 are repeated m times (where m is a +ve integer), as shown in Figure 2.3, resulting in at least 2^m syllable segments in each cluster.

The N_1 models corresponding to the N_1 clusters are then subjected to the incremental training procedure that is described in the next Section.

2.4.2.2 Incremental training

After selecting the initial clusters, where the models are only initialized, the parameters of the models of each of the clusters are re-estimated. Since the *HMM* parameters are tuned before all the training data has been considered, this training procedure is referred to as *Incremental Training*. The training strategy is different from conventional batch training where the models are updated only after all the training data is processed. The flow chart for the unsupervised incremental training (UIT) algorithm is shown in Figure 2.3. The steps followed for the incremental training are given below.

1. After obtaining the N_1 initial clusters and their models, the model parameters of each cluster are re-estimated using Baum-Welch re-estimation algorithm [36].
2. The new models are used to decode all the syllable segments using Viterbi decoding. Clustering is now performed based on the decoded sequence.
3. If a particular cluster is found to have less than k (k = 6) syllable segments, that particular cluster is removed.
4. Steps 1-3 are repeated until the convergence criterion is satisfied resulting in N_2 clusters. The convergence criterion followed is explained below.

   In each iteration, as the model parameters are re-estimated, the number of syllable segments that migrate from one cluster to another is expected to reduce. Convergence is said to be met, if the number of syllable migrations between clusters reaches zero. When this convergence criterion is satisfied, incremental training procedure terminates.

The incremental training procedure gives N_2 syllable clusters and in turn N_2 syllable models. The syllable clusters thus obtained from the incremental training technique are labeled manually as S_1, S_2, ..., S_{N_2} and the corresponding HMMs are used for transcription task.

2.5 Speech corpus

The performance of the continuous speech recognition system using unsupervised and incremental training is evaluated on the DD news database [37]. The database for Indian languages consists of news bulletins of eight different languages. Each bulletin represents approximately 10-15 minutes of read speech. The average duration of each sentence in the news bulletin is observed to be about 2.5 seconds. Thus 20 bulletins spoken by 11 male and 9 female speakers for Telugu and 33 bulletins spoken by 10 male and 23 female speakers for Tamil were collected [38]. In this thesis, Tamil and Telugu news bulletins are considered. For majority of the experiments Tamil is used. Both Tamil and Telugu are used to evaluate the MUIT system discussed in Chapter 5.

2.6 Performance analysis of the UIT system

For both training and testing, Indian Tamil news bulletins [37] have been used. Four female speakers' data is considered for training the speech models. Using automatic segmentation approach [26], the four speakers' data is segmented into syllabic units, which gives M syllable segments (M is about 8000). These syllable segments are used for training the models using the UIT technique.

The data set considered for testing is divided into two categories namely (I) Speaker Dependent (SD) data and (II) Speaker Independent (SI) data. In speaker dependent transcription, new data of a speaker used in training is transcribed. In speaker independent transcription, data of speakers that do not appear in training is transcribed. The test speech utterances are segmented into syllabic units and the resultant syllable inventory for testing is about 4000. All 4000 syllables are tested against the models that are obtained using UIT. Speech recognition performance of the baseline system is shown in Table 2.1. The syllables that are recognized completely are named as *Complete syllable* in Table 2.1. In some cases, only the CV or VC part of the entire

CVC syllable has been recognized correctly and they are categorized as *CV/VC only* in Table 2.1. Similarly, the syllable segments that are recognized based on the similarity in consonant and vowel part are categorized as *Consonant only* and *Vowel only* respectively. From the Table, the average performance of the syllable recognizer is

Table 2.1: Syllable recognition performance of the baseline UIT system

Sound units	Baseline UIT system	
	Speaker dependent data	Speaker independent data
Complete syllable	41.9	34.9
CV/VC only	18.5	16.7
Vowel only	27.3	31.0
Consonant only	3.25	4.28

observed to be approximately 42% for speaker dependent data and approximately 35% for speaker independent data (refer Table 2.1).

The syllable recognition results obtained using the baseline UIT system are promising for the speech transcription task. Though the recognition results are promising, there are several drawbacks in the segmentation and clustering techniques. The performance of the recognition system is therefore not comparable to the present day techniques that use manually segmented labeled speech data [25] [16].

2.7 Summary

A novel approach for segmenting and transcribing continuous speech signal without using a manually annotated speech corpus is explored for speech recognition. Recognition performance of the baseline UIT system is tested for Indian Tamil news bulletins [37]

and only 35% syllable recognition is observed for speaker independent data. Though the syllable recognition performance is promising for automatic speech transcription task, there are few drawbacks in the segmentation and clustering algorithms. In the next Chapter, the issues in the baseline system due to segmentation and clustering techniques are explained. Several refinements are made to the baseline system to improve the syllable recognition performance.

Chapter 3

Modified unsupervised and incremental training

3.1 Introduction

In Chapter 2, an unsupervised and incremental training (UIT) technique is evaluated for the continuous speech transcription task. Although the results obtained using the UIT [1] are promising for transcription task, there are many draw backs in the baseline recognition system due to both segmentation and clustering errors. As a consequence of these problems, the performance of the recognition system is not good compared to the batch training [25] technique on manually segmented data. In this thesis, we take the UIT technique as a baseline technique and perform several refinements to it. The modified UIT system gives a significant improvement in the syllable recognition performance over the baseline UIT system.

We first address the major issues in the baseline speech recognition system due to segmentation and syllable clustering in Section 3.2. In Section 3.3, modifications made to the baseline UIT system, for improving speech recognition performance are discussed in detail.

3.2 Issues in the baseline UIT system

As discussed in Chapter 2, using the baseline UIT system, the continuous speech signal is first automatically segmented into syllable-like units and similar syllable segments are grouped together using the unsupervised and incremental clustering technique. Separate models are generated for each cluster of syllable segments. But the syllable recognition performance using the baseline technique is inferior to that of batch training [25] due to several drawbacks in the segmentation and clustering techniques. The major issues in the segmentation and clustering procedure, using the base-line UIT system, are now outlined.

1. When a speech signal with a large number of syllable units is automatically segmented, few segments may get merged because of strong co-articulation effect between the syllable units. The speech signal may therefore not be segmented exactly at the syllable boundaries, causing some segments to be poly-syllabic or consists of fragment of a syllable. The speech unit that is poly-syllabic is named as a merged syllable and the speech unit that has only a part of the syllable is named as a syllable fragment. To illustrate this, consider Figure 3.1, where the segment /*seya*/ should ideally be split into two segments, /*se*/ and /*ya*/, as shown by the dashed line. Owing to the strong co-articulation effect between these two syllables, they are merged into a single syllable. Clustering is poor throughout the incremental training procedure due to these merged syllables and syllable fragments. For example, the merged syllable /*seya*/ has two vowels (V) and two consonants (C). This is therefore clustered with another syllable based on the similarity in either V or C part of that syllable.

2. Presence of silence at the boundaries of syllables may result in poor clustering. It is noticed that while generating HMMs, a syllable that has a small silence portion at the beginning or end, results in a HMM with a separate state for silence.

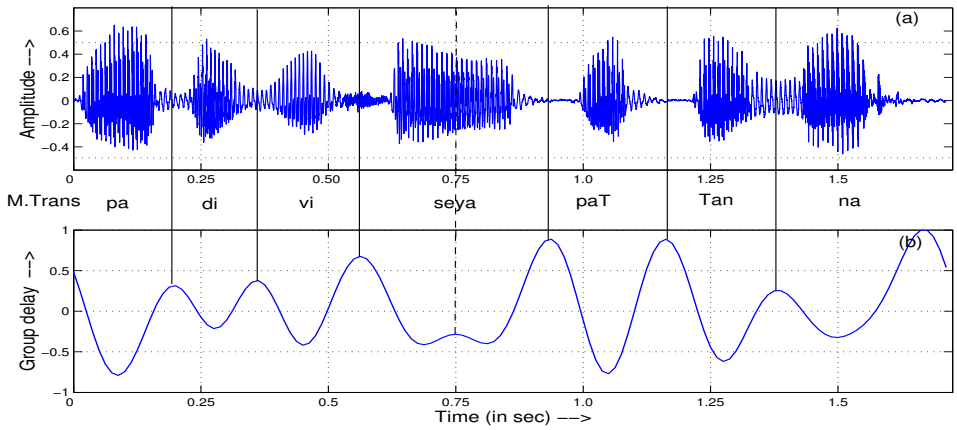

Figure 3.1: Segmentation of speech signal based on minimum phase group delay. (a) Actual speech signal (b) Group delay spectrum of the minimum phase speech signal.

Spectral characteristics of units corresponding to silence are almost the same for all syllables. The presence of silence thus dominates the syllable clustering process. As a result, syllables that have very similar silence characteristics may get clustered together. An example is illustrated where the syllable /*ni*/ (refer Figure 3.2), having a small silence segment at the beginning, is clustered with the syllable /*er*/ (refer Figure 3.3), that also has a small silence segment at the beginning. The hidden markov models for the syllables /*ni*/ and /*er*/ are shown in Figure 3.4 and Figure 3.5. The first state in the HMM of /*ni*/ (Figure 3.4) and HMM of /*er*/ (Figure 3.5) corresponds to silence. Based on the similarity in the silence portion of both the syllables, they are clustered together.

3. Similar syllable groups (e.g., /*kam*/ and /*van*/) are sometimes clustered together. Since the first *C* in this *CVC* is not identical, this cluster is assigned the label of *VC* rather than *CVC* that further leads to wrong transcription.

4. Labeling is easier if the syllables in each cluster are examples of the same class. However in many cases, clusters contain a number of syllable examples that are

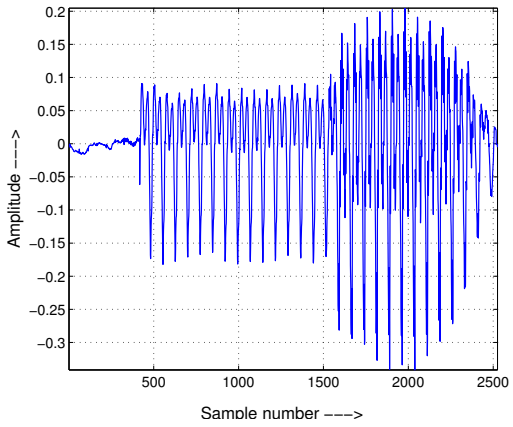

Figure 3.2: syllable /*ni*/ having small silence portion at the beginning

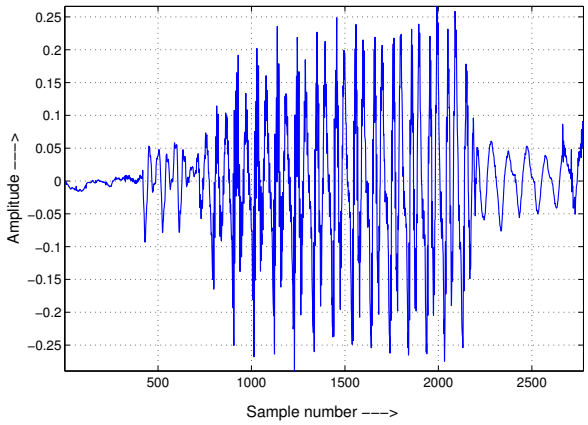

Figure 3.3: syllable /*er*/ having small silence portion at the beginning

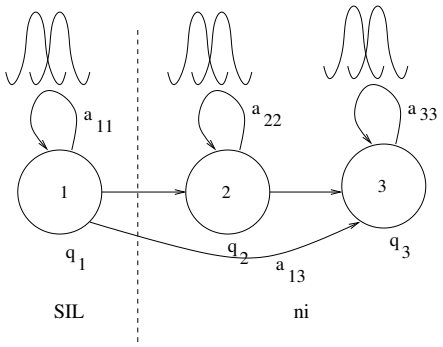

Figure 3.4: HMM of the syllable /*ni*/

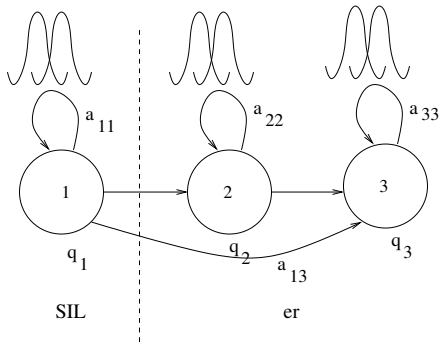

Figure 3.5: HMM of the syllable /*er*/

similar in either vowel or consonant part. In such cases, listening to all syllables and labeling manually is a difficult task. It is a tedious task, especially when the syllables in the clusters are entirely different. For example, the syllable cluster, consists of /*kri*/ and /*te*/, can not be given a proper label as both syllables are entirely different.

5. In the base-line system, only a subset of syllables are considered for the initial cluster selection procedure. As limited data is considered for training the syllable models initially, this can lead to inaccurate syllable clustering.

In the following Sections, the transcription system implemented using UIT is refined to alleviate some of these issues.

3.3 Modified unsupervised and incremental clustering

As a result of the issues outlined in Section 3.2, the syllable recognition performance is not as good as that of batch training [25]. In order to improve the performance of the syllable clustering algorithm, several refinements are made to the baseline clustering technique. In the following Sections, modifications made to the baseline UIT technique are explained.

3.3.1 Modifications to the syllable segments

In order to improve the performance of the syllable clustering algorithm, syllable segments that are obtained after the automatic segmentation process are pruned.

3.3.1.1 Duration analysis

The problem of syllable fragmentation and merging of syllables can be overcome by performing duration analysis on the syllable inventory. To illustrate this, duration analysis is carried out on four DD News bulletins [37], in which each bulletin corresponds to a different female speaker. The duration of each bulletin is about 15 minutes and the average duration of each sentence is observed to be about 2.5 seconds. The speech signals are first automatically segmented into S syllable segments. The duration of these syllable segments is then analyzed. The syllable duration distribution is shown in Figure 3.6. The duration analysis results show that the duration of approximately 95% of the syllables vary between 110 ms to 270 ms. The mean duration of syllable is observed to be 135 ms. If the syllable duration is either below 110 ms or above 270 ms, that particular syllable segment is removed. This pruning process ensures that most

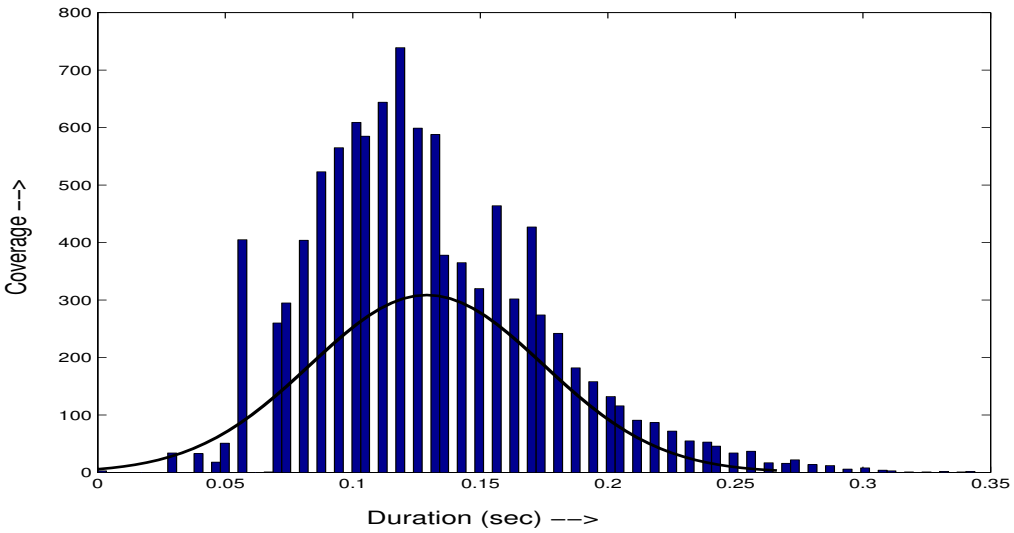

Figure 3.6: Duration analysis of the syllable segments

of the syllable fragments and merged syllables are removed, resulting in M number of syllables, where M < S.

3.3.1.2 Silence normalization

The drawback in syllable clustering due to the presence of a short duration silence segment at the boundaries of syllables can be overcome by prefixing and suffixing a silence segment of approximately 20 ms to each syllable segment. This procedure is named as the Silence Normalization (SN) procedure. This ensures that during the generation of a HMM for a syllable, a separate state is assigned to the silence portion at the boundaries of the syllable (refer Figure 3.7).

An experiment is illustrated to show the effect of the silence normalization procedure. For this experiment 10 syllable segments that consist of 10 examples in each class are considered. These syllable segments are prefixed and suffixed with a silence segment of about 20ms and they are manually grouped into clusters based on the similarity of syllable segments. Models with 5-states and 1-Gaussian mixture/state are

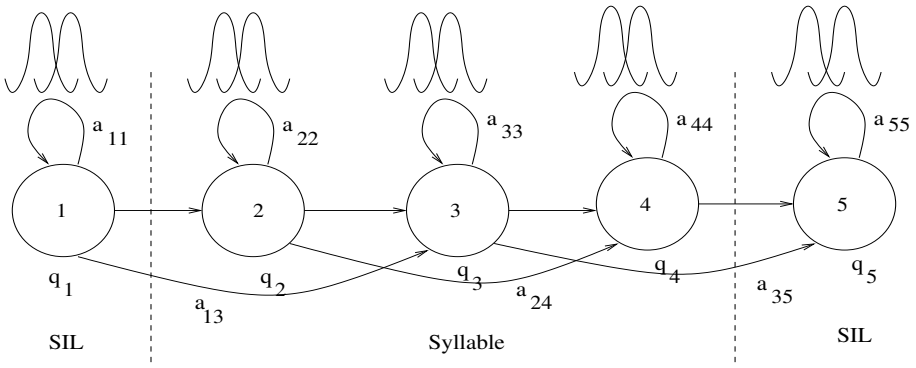

Figure 3.7: HMM of a typical syllable after silence normalization

initialized for these clusters.

To evaluate the syllable recognition performance with silence normalization technique, a set of syllables belonging to the syllable class /*ni*/ are considered. Here, Figure

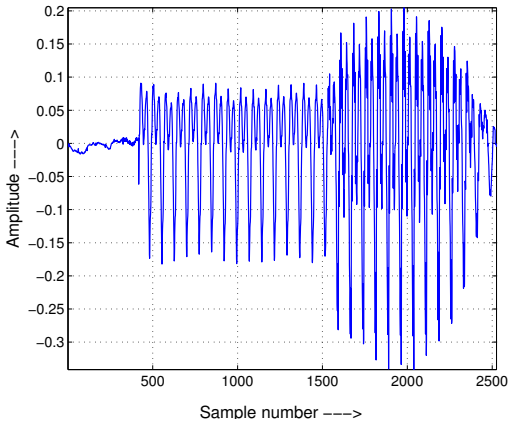

Figure 3.8: Syllable /*ni*/ before silence normalization

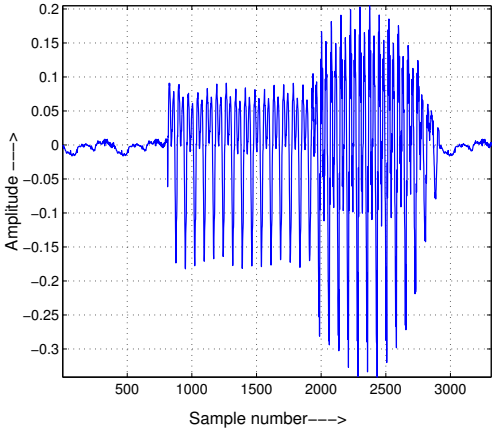

Figure 3.9: Syllable /*ni*/ after silence normalization

3.8 refers to the syllable /*ni*/ before silence normalization and the Figure 3.9 refers to the syllable /*ni*/ after silence normalization. This syllable set is silence normalized with a silence segment (duration of about 20ms) and tested against all the 10 models. Similarly, this syllable set is also tested against all the 10 models without silence

normalization.

Earlier the syllable /*ni*/ is wrongly clustered with a syllable /*er*/ which has a small silence segment at the beginning. Silence normalization ensures that all syllable models have a separate state in the HMM assigned for silence. As a result of the silence normalization procedure, the syllable /*ni*/ is properly clustered with another syllable from the same class.

Table 3.1: Isolated syllable recognition performance for the syllable *ni* before and after silence normalization

Syllable	Before silence normalization	After silence normalization
ni	26.58	48.10

The syllable recognition performance for the syllable /*ni*/ before and after silence normalization is shown in Table 3.1. From Table 3.1, it can be observed that the silence normalization procedure gives better syllable recognition performance compared to that without silence normalization. Silence normalization procedure can therefore be used to overcome the problem due to silence segments at the boundaries. HMM training can be carried out using the silence normalized syllable segments for the continuous speech recognition task. The training process of the base-line system is slightly modified. This is explained in Section 3.3.2 and Section 3.3.3.

3.3.2 Modifications to initial cluster selection

Automatic segmentation of the continuous speech signal gives S ($s_1, s_2, ..., s_S$) number of syllable segments. These syllable segments are subjected to duration analysis that results in M number of syllable segments, where M < S. Silence normalization is carried out on all the M syllable segments. After obtaining all M silence normalized syllable segments, the initial groups of syllables are carefully selected to ensure fast convergence

and unique syllable models. In the baseline UIT system, only a few syllable segments are considered at the initial cluster selection stage, as the focus there is primarily on LID. Therefore it is sufficient if a set of syllable units that are unique to each language are identified. In the automatic transcription task, every syllable in the language must be enumerated. This requires that a large number of examples corresponding to each syllable cluster are available after clustering.

As limited data is taken in the baseline system for selecting the initial clusters, only few clusters have syllable segments belonging to the same class where as the other clusters have syllables that are similar only in CV, VC or V part. To ensure each cluster, obtained after the initial cluster selection consists of many different examples of the same syllable, the *entire set* of syllables available in the training data is used in the initial cluster selection procedure. The flow chart for the modified initial cluster selection procedure is shown in Figure 3.10 and the procedure is explained here.

1. All the M silence normalized syllable segments are considered for the initial cluster selection process.
2. 39 dimensional MFCCs appended with velocity and acceleration parameters are extracted for each of these M syllable segments with multiple frame sizes (MFS) of 12, 14, 16, 18, and 20ms.
3. M HMMs (λ_1, λ_2, ..., λ_M) are initialized from these parameters. The M syllable segments are recognized against all the M HMMs and 2-best recognition results are obtained.
4. The pruning process adopted in the baseline initial cluster selection process (Section 2.4.2.1) may result in some syllable models being unnecessarily removed from further stages of training. To overcome this problem, a syllable model S_j (indicated in Section 2.4.2.1) is removed only if the 2-best results of S_i and S_j are $\{S_i, S_j\}$, and $\{S_j, S_i\}$, respectively. This pruning process ensures that the syllables that occur only once in the 2-best results are also retained. The number

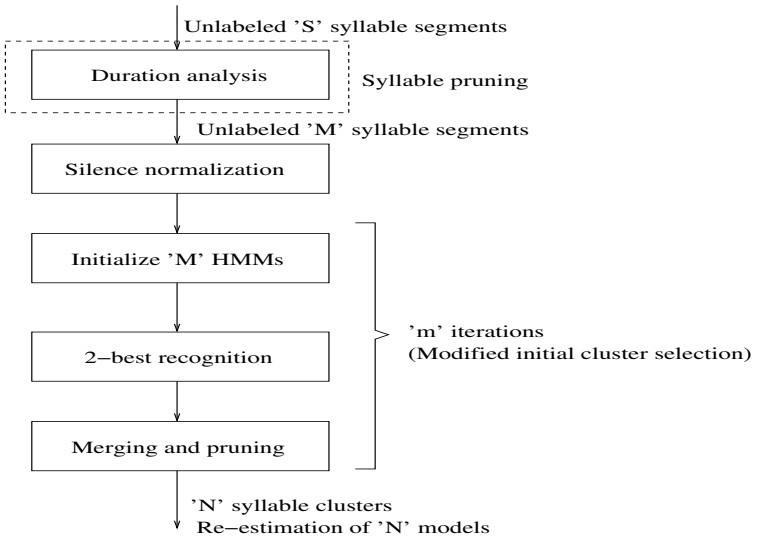

Figure 3.10: Flow chart of modified initial cluster selection procedure

of models are pruned to N, where N < M.

5. Steps 2-4 are repeated m times (where m is a +ve integer), as shown in Figure 3.10, resulting in at least 2^m syllable segments in each cluster.

The initial cluster selection procedure leads to N clusters $(c_1, c_2, ...c_N)$.

3.3.3 Modifications to incremental training

After selecting the N initial clusters $(c_1, c_2, ..., c_N)$, where models are only initialized, parameters of the models of each of the clusters are re-estimated incrementally using the following steps (refer 3.11).

1. Model parameters of the initial clusters $(c_1, c_2, ..., c_N)$ derived from the initial cluster selection procedure are re-estimated using Baum-Welch re-estimation algorithm.
2. These new models are used to decode all the syllable segments using Viterbi algorithm. Clustering is done based on the decoded sequence and n-best recognition

results are obtained for each syllable (where n is a variable).

3. If a particular cluster is found to have less than k ($k = 3$) number of syllable segments, that particular cluster is removed and the number of models is reduced to N_0, where $N_0 < N$.

4. Steps 1-3 are repeated for p iterations until the convergence criterion is satisfied. The convergence criterion followed is explained in step 5.

5. In each iteration, as the model parameters are re-estimated and the syllable segments are re-clustered, the number of syllable segments which migrate from one cluster to another is expected to reduce. The convergence is said to be met, if the number of syllable migrations between each cluster reaches zero and the incremental training procedure is terminated.

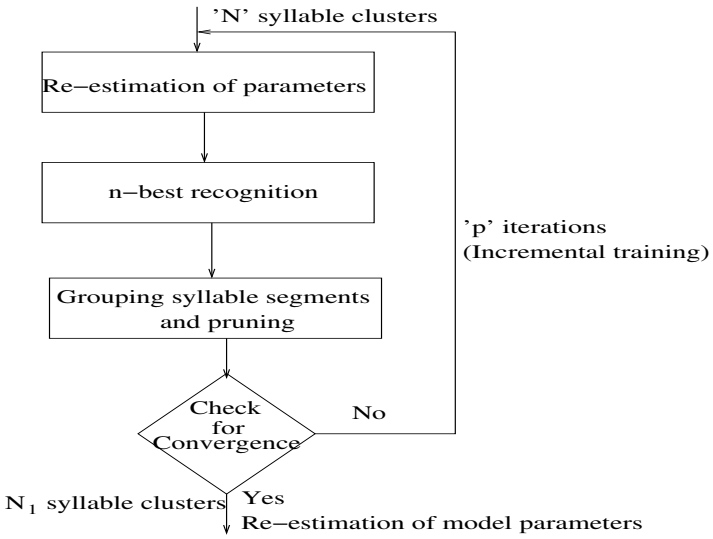

Figure 3.11: Flow chart of modified incremental training procedure

This entire technique is named as the modified unsupervised and incremental training (MUIT) technique. Once convergence is met, a set of S_1, S_2, ..., S_{N_1} HMMs are obtained, each corresponding to a syllable. These models are then labeled according to the syllable identity in the given language, by manually listening to segments in each cluster. Clusters for which a definite identity can not be given are removed.

3.4 Performance analysis of MUIT

The Indian television news bulletins of Tamil [37] are used to analyze the performance of MUIT algorithm for transcribing speech data not seen during training. Training is performed using four female speakers' news bulletins, each of about 15 minutes duration and for testing, two female speakers' news bulletins are used.

The speech signals are automatically segmented into syllable-like units that results in S (S=8804) syllables. Duration analysis and silence normalization with 20ms of silence is carried out on these syllable segments and M (M=8400) syllables are obtained. Using these syllable segments, the modified incremental training procedure is carried out. This results in N_1 number of HMMs, where $N_1 < M$. Each HMM is a 5-state and 2-mixture/state model. The total number of syllable clusters obtained after incremental training is 512. The final clusters are labeled manually using *iTRANS* method of labeling [39], and the corresponding models with labels assigned to them are used for the continuous speech recognition task.

The data set considered for testing is divided into two categories namely (I) Speaker Dependent (SD) data and (II) Speaker Independent (SI) data. In speaker dependent transcription, new data of a speaker that appear in training is transcribed. In speaker independent transcription, data of speaker that does not appear in training is transcribed. The performance of the syllable recognition system is computed using the formula given below:

$$Performance\ (\%) = \frac{Number\ of\ syllables\ correctly\ recognized}{Total\ number\ of\ syllables}$$

The recognition results for Tamil language are tabled in Table 3.2. As a syllable recognizer, the baseline UIT system gives a syllable recognition performance of 41.9% for the SD data and 34.9% for the SI data whereas the modified UIT gives a syllable recognition performance of 56.2% and 42.6% for SD and SI data sets, respectively. The syllables which are correctly recognized are named as *Complete syllable* in Table 3.2.

Table 3.2: Syllable recognition performance of the baseline UIT system and the modified UIT system. I - Speaker dependent data. II - Speaker independent data.

Sound units	Baseline UIT system		Modified UIT system	
	I	II	I	II
Complete syllable	41.9	34.9	56.2	42.6
CV/VC only	18.5	16.7	25.6	20.8

In some cases, only the *CV* or *VC* part of the entire *CVC* syllable has been recognized correctly and they are categorized as *CV/VC only* in Table 3.2. In the case of *Complete syllable*, a significant improvement in the syllable recognition performance of 15% and 8% is observed for SD and SI data respectively. Based on the similarity in *CV* or *VC* part, an improvement of 7.1% for SD data and 4.1% for SI data is obtained over the baseline system.

An example of the automatic transcription of a speech signal, that is considered from speaker independent data, is shown in Figure 3.12. Here, Figure 3.12 (a) shows

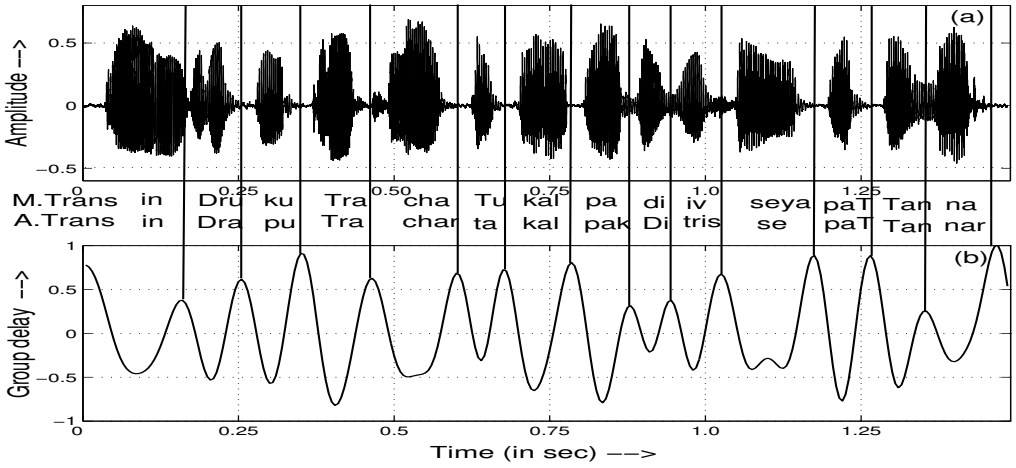

Figure 3.12: An example of automatic transcription. (a) Actual speech signal. (b) Minimum phase group delay of the speech signal. M.Trans - Manual Transcription. A.Trans - Automatic Transcription

the actual speech signal and Figure 3.12 (b) shows the group delay spectrum of the minimum phase speech signal. The manual transcription and the corresponding automatic transcription are given in the Figure. In the Figure 3.12, M.Trans corresponds to manual transcription and A.Trans corresponds to automatic transcription of the speech signal.

3.5 Syllable clustering - an analysis

The syllable clustering errors during training are analyzed. In order to analyze the syllable clustering results, we first present in Table 3.3, the categorization of speech sounds based on place and manner of articulation.

Table 3.3: Speech sounds based on place and manner of articulations

Place of articulation	Manner of articulation				Nasals	Semivowels	Fricatives
	Unvoiced		Voiced				
	Unaspirated	Aspirated	Unaspirated	Aspirated			
Velar	/ka/	/kha/	/ga/	/gha/	/kna/	–	/ha/
Palatal	/cha/	/chha/	/ja/	/jha/	/chna/	/ya/	–
Alveolar	/Ta/	/Tha/	/Da/	/Dha/	/Tna/	/ra/	/shha/
Dental	/ta/	/tha/	/da/	/dha/	/na/	/la/	/sa/
Bilabial	/pa/	/pha/	/ba/	/bha/	/ma/	/va/	–

During incremental clustering, it is observed that some of the syllables are clustered based on the manner and place of articulation of sound units. The most common errors observed in the syllable clustering algorithm are listed in Table 3.4. In Table 3.4, the first column corresponds to the categories of syllables that are confused based on place and manner of articulation. The second column corresponds to few examples of such confused syllables.

From the analysis made from the syllable clustering, a major confusion between alveolar and dental segments is observed. For *e.g.*, syllable */Ta/* is wrongly recog-

Table 3.4: List of confused syllables during incremental training

Categories of syllables confused	Examples of syllables confused
Alveolar and Dental	/*Ta*/, /*ta*/ /*Da*/, /*da*/ /*Ti*/, /*ti*/
Dental and Bilabial	/*da*/, /*ba*/ /*ta*/, /*pa*/ /*dha*/, /*bha*/
Unaspirated, voiced and Semivowels	/*ba*/, /*va*/ /*da*/, /*ra*/
Unaspirated, unvoiced and Fricatives	/*cha*/, /*sa*/ /*ta*/, /*ha*/

nized as /*ta*/ and vice versa. Major confusion between dental and bilabial segments is also observed. In few cases, unaspirated voiced segments such as /*ba*/ are wrongly recognized as semi vowels such as /*va*/. Similarly, wrong clustering between unaspirated, unvoiced segments and fricatives is also observed. There is hardly any confusion between different nasal sounds. Syllable segments, that do not exist in the obtained syllable inventory, are mostly recognized based on the similarity in CV/VC part of the syllable models.

Although, this analysis is not comprehensive, the results suggests that perhaps a hierarchical clustering algorithm can be used for classification of these syllables.

3.6 Summary

Major issues in the baseline system using unsupervised and incremental training are outlined. Several modifications are made to the baseline system in order to increase

the syllable recognition performance. The problems due to unusual duration syllables are overcome by performing duration analysis on the syllable inventory. Silence normalization technique is proposed in order to increase the performance of the syllable clustering algorithm. Significant improvement in the performance of the syllable recognition system is observed using the MUIT technique.

Chapter 4

Multiple frame size and multiple frame rate feature extraction for manually segmented data

4.1 Introduction

The conventional speech recognition systems use features that are extracted with a single frame size and single frame shift during training and recognition. But this may not address the issue of varying speaker characteristics, mainly speaking rate. In this Chapter, we propose a new approach to feature extraction that uses multiple frame size (MFS) and multiple frame rate (MFR) Cepstral features both during training and recognition. The main motivation for considering the MFS and MFR Cepstral features is explained below:

In UIT technique, discussed in Chapter 2, initially a single example is considered for each class during initial cluster selection procedure and features are extracted with multiple frame sizes to initialize HMMs (refer Section 2.4.2.1). It is observed that most of the syllable clusters obtained in the initial cluster selection procedure itself, consists of syllable segments that are different examples of the same syllable.

In the work presented in Chapter 2, the syllable inventory considered is large and hence the effect of MFS feature extraction alone could not be analyzed. In this Chapter, an attempt is made to analyze the effect of MFS features for speech recognition

on manually segmented data. Several techniques in the literature consider features extracted using different frame sizes or frame rates for speech recognition. But no effort is made to combine both multiple frame size and multiple frame rate based features for speech recognition. In this work, an effort is also made to combine the features extracted using MFS and MFR for an isolated style syllable recognizer. In order to analyze the effect of MFS and MFR feature extraction technique, a set of controlled experiments are performed on a limited number of syllables.

In this Chapter, we first briefly review some of the techniques that consider features extracted using different frame sizes and frame rates in Section 4.2. In Section 4.3, a set of controlled experiments are conducted to analyze the effect of MFS Cepstral features. In Section 4.4, MFS and MFR features are used both during training and recognition in an isolated speech recognizer. The syllable recognition results using MFS and MFR features are compared with that of single frame size (SFS) features.

4.2 Overview of feature extraction techniques using different frame sizes and frame rates

In most of the speech recognition systems, speech signals are first windowed into frames. Frames are typically 20-30ms in duration and frame step size is about 10ms. The conventional speech recognition systems use features that are extracted with single frame size and single frame rate. Though the conventional practice of using single frame size of about 20-30ms and frame shift of about 10ms is a good compromise between the requirements for time and frequency resolutions, this may not capture the speaker variability accurately.

4.2.1 Overview of feature extraction techniques using different frame sizes

In [40], **Vaseghi** has introduced multi-resolution Cepstral features for speech recognition. The multi-resolution Cepstral features idea is based on the hypothesis that for many speech sounds, the localized features of the time-frequency trajectory of speech can provide crucial clues for classification. In this technique, speech features are derived using three different time resolutions. These include sub-phonetic time windows (small duration of 5-10ms) and one phonetic time window. The main drawback of this approach is that, it requires phoneme boundaries.

Multi-resolution feature extraction technique is used to build acoustic models of a language identification system in [1]. In this work, it is claimed that multi-resolution feature extraction ensures a reasonable variance for each Gaussian mixture in the models.

In the literature, variable frame size (VFS) techniques have also been used in order to capture the speaking rate information in the context of ASR. In [41], an optimal frame size and frame rate is selected for a particular speaking rate. In other words, the frame rate depends on the average speaking rate of a test utterance. The analysis frame size does not change within the utterance, but changes across utterances.

A data dependent VFS technique [42], is proposed for modeling transition segments. In this method, the size of the analysis frame depends on the data that the frame contains. In these two approaches [41] [42], the analysis frame size depends on either speaking rate or measurement of non-stationarity of the speech signal.

4.2.2 Overview of feature extraction techniques using different frame rates

Changes in spectral characteristics are important clues for identifying speech sounds. Hence, ASR should be able to model the acoustic signals in regions with fast spectral

changes. Several techniques have been proposed [43] which vary the frame rate according to the value of a specific metric to better model the speech signal in regions with fast spectral changes. The metric is typically related to the level of spectral variations detected and computes more feature vectors in those regions in which spectral information changes significantly. A comparison between the measured spectral change and a pre-calculated threshold is done. Frames with a spectral change level under the selected threshold are discarded, otherwise they are considered. The classical approach for variable frame rate (VFR) used in [44] computes the Euclidean distance D(i, j) between the current frame i and the last retained frame j. This distance is compared with some threshold θ. Using this method, the current frame is considered only when D(i, j) $> \theta$.

In [45], **Qifeng** has proposed a variable frame rate (VFR) algorithm in which the frames are selected based on the energy weighted MFCC distance between every two adjacent *dense* frames. Though the above mentioned approaches are simple, they have some drawbacks: only two frames are considered in the decision making process. Sometimes, this may not represent the complete characteristics of a speech signal.

Philippe [46], has proposed a new frame rate analysis in which instead of calculating the distance between frames, the norm of the derivative parameters is used for deciding the frame rate.

In the above mentioned techniques, since acceptance or rejection of frames depends on the value of pre-calculated threshold, performance of the speech recognition system may degrade, if threshold is not calculated properly.

4.3 Significance of multiple frame size Cepstral features

In [40], it is mentioned that the multiple frame size (multi-resolution) Cepstral features can provide crucial clues for the classification of different sound units. In Chapter 2,

it is also observed that the MFS feature extraction technique gives better recognition results even though the number of training examples considered in each syllable class are less. Since the database considered in Chapter 3 consists of large number of unique syllables, the effect of MFS Cepstral features alone could not be analyzed. In this Section, an attempt is made to analyze the significance of MFS feature extraction.

4.3.1 Significance of MFS Cepstral features on number of training examples

Performance of the speech recognition system will be low if the number of examples considered for each class are very less. In the initial cluster selection procedure discussed in Chapter 3, initially a single example is considered for each class of syllable and models are initialized using MFS Cepstral features. Though a single example is considered for each class, the clusters obtained after the initial cluster selection procedure are observed to have examples of the same or similar (similarity in CV/VC) class of syllables. Hence, in this Section we first analyze the significance of MFS Cepstral features on number of training examples for each class of syllables. In order to study its significance on number of training examples, a set of experiments is conducted in which the number of training examples are incremented from experiment to experiment.

The data set considered for testing is divided into two categories:

(I) Speaker dependent data set

(II) Speaker independent data set

Two different experiments are conducted corresponding to the two different data sets given above. For these experiments, five female speakers' data from the Tamil news bulletins [37] is considered for training. First, each speaker's data is manually segmented into syllable segments and different syllable examples of the same syllable are grouped together manually. The syllables with CV units alone are considered

for these experiments. For analysis purpose of MFS feature extraction, 40 different CV units are considered for which enough number of training examples are available. Initially 2 syllable examples are considered for each class and models are initialized using 39 dimensional Cepstral features (13 $MFCC$ + 13 $velocity$ + 13 $acceleration$) using multiple frame sizes (12, 14, 16, 18 and 20ms). Here, each HMM is a 3-state and 1-Gaussian mixture/state model.

In all these experiments, the main focus is on comparing the performance of the proposed feature extraction techniques with single frame size feature extraction technique. Since the focus is only on relative performance, specific attention is not given to tuning the model parameters. Only acoustic likelihood based raw recognition results are analyzed in all the experiments. For the MFS experiment the frame shift is fixed to 10ms during training.

4.3.1.1 Significance of MFS Cepstral features on number of training examples: Speaker dependent data set

The models are tested against the data of two female speakers that are considered during training. For testing, cepstral features are extracted from the test speech utterances with a frame size of 20ms and a frame shift of 10ms. This experiment is repeated for each set of training examples (2-20, in steps of 2).

The performance of the recognition system using MFS features is compared with that of SFS Cepstral features (with 20ms frame size and 10ms frame shift). The recognition performance of both MFS and SFS models is shown in Figure 4.1. In Figure 4.1, the solid line corresponds to the recognition performance using MFS features where as the dotted line corresponds to that of SFS features. From Figure 4.1, it can be observed that when the number of training examples for each class is less, the recognition system using MFS Cepstral features shows better performance compared to SFS features. Even when the number of training examples are increased further,

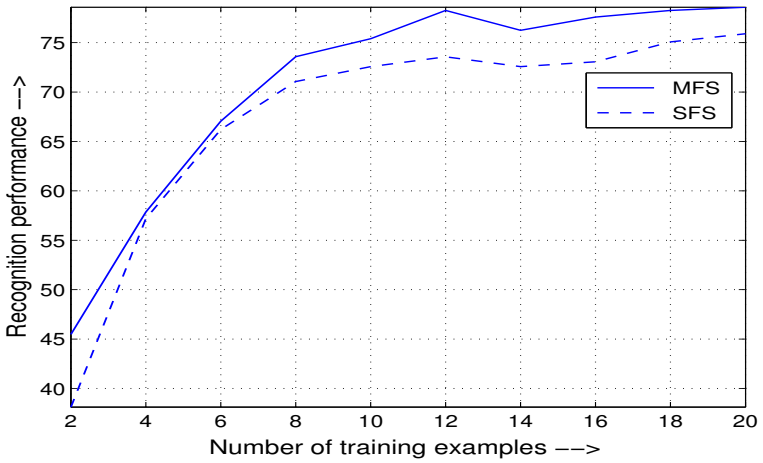

Figure 4.1: Syllable recognition performance using MFS and SFS Cepstral features for different number of training examples of speaker dependent data

for each class of syllables, MFS feature extraction technique shows better performance compared to that of SFS. Hence, MFS feature extraction technique can be used to improve the syllable recognition performance when speaker dependent data is used for recognition.

Now, in the next Section, performance of the speech recognition system is analyzed on a speaker independent data set.

4.3.1.2 Significance of MFS Cepstral features on number of training examples: Speaker independent data set

The models are tested against the data of two female speakers that are not considered during training. Once again during testing, cepstral features are extracted from the test speech utterances with a frame size of 20ms and a frame shift of 10ms. This experiment is repeated for each set of training examples (2-20, in steps of 2).

The performance of the recognition system using MFS features is compared with

that of SFS Cepstral features (with 20ms frame size and 10ms frame shift).

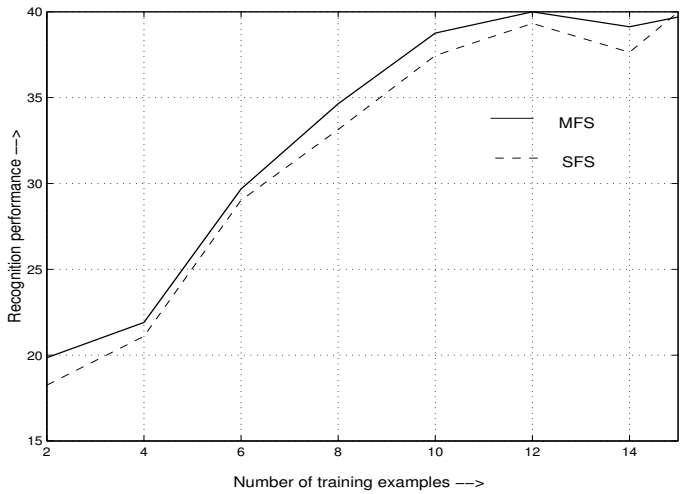

Figure 4.2: Syllable recognition performance using MFS and SFS Cepstral features for different number of training examples of speaker independent data

The recognition performance of both MFS and SFS models is shown in Figure 4.2.

From Figure 4.2, it can be observed that for each class the recognition system using MFS Cepstral features shows better performance compared to SFS features when the number of training examples considered is less. When the number of training examples is further increased (approximately to 15 examples in this case), both features shows almost the same performance. Hence, MFS feature extraction technique can be used when the number of training examples is less.

4.4 Multiple Frame Size (MFS) and Multiple Frame Rate (MFR) feature extraction for ASR

In the literature, cepstral features have been extracted either using different frame sizes or using different frame rates. No effort is made to combine both these features for speech recognition. In this Section, an attempt is made to combine both these features

during training and recognition.

The speech features extracted using MFS and MFR are combined for isolated style syllable recognition. Here the syllable models are generated with features extracted using multiple frame sizes and frame rates. Further, two different experiments are conducted:

1. MFS and MFR based feature extraction only during training
2. MFS and MFR based feature extraction both during training and testing

These techniques are explained in the following Sections.

4.4.1 MFS and MFR only during training

To illustrate the effect of MFS and MFR feature extraction, syllable models are generated with multiple frame sizes and frame rates. For a gender independent speech recognition system, data of five speakers is considered for training. The speech signals are manually segmented into syllable-like units and 40 different CV units, that have more than 50 training examples, are considered. For these syllable segments, features are extracted using multiple frame sizes[1] (12-20ms, in steps of 2ms) and multiple frame shifts[2] (10-18ms, in steps of 2ms) and models with 3-states and 1-Gaussian mixture/state are generated for each frame size and frame shift. For testing, two different female speakers and one male speaker not seen during training are considered. In this experiment, during testing, features are extracted with single frame size (SFS) and single frame rate.

[1]In the MFS task, the frame rate is constant (100 frames/sec).

[2]In the MFR task, the frame size is constant (20ms).

4.4.2 MFS and MFR during training and testing

The significance of MFS and MFR feature extraction on both training and testing is illustrated in this Section. In this technique, HMMs are generated with features extracted using multiple frame sizes and frame rates. During testing also, speech features are extracted using different frame sizes and frame rates. For each frame size and frame rate based features, a separate recognition experiment is conducted and 3-best results are considered for each of the syllable segment. The recognition results are combined using the following procedure. For the 3-best results, weights are assigned based on their n-best position (1.0 for the 1^{st} best, 0.8 for the 2^{nd} best and 0.6 for the 3^{rd} best). The value for each syllable model is computed based on its recognition position for different frame sizes and frame rates. An example of the recognition result of a syllable (say S_i) using MFS and MFR features is shown in Table 4.1.

Table 4.1: 3-best recognition results of a syllable using 3 different frame sizes and frame rates

	3-best recognition results		
Frame size or frame rate	1st best	2nd best	3rd best
F_1	$S_{25}(1.0)$	$S_{14}(0.8)$	$S_{35}(0.6)$
F_2	$S_{25}(1.0)$	$S_{35}(0.8)$	$S_{14}(0.6)$
F_3	$S_{14}(1.0)$	$S_{25}(0.8)$	$S_{35}(0.6)$

The entries in the first column corresponds to a specific frame size/frame rate. In the Table, S_{25}, S_{14}, and S_{35} represent the 25^{th}, 14^{th}, and 35^{th} syllable models respectively. For a given syllable S_i, the 3-best recognition results are obtained using a combination of frame size and frame rate. In the Table, F_1, F_2, and F_3 correspond to three different experiments. The 3-best results are ordered and weights are assigned as indicated in Table 4.1.

To determine the identity of the syllable S_i, the following is performed:

The 3-best recognition results from all the recognizers are combined by adding the weights associated with the n-best position of the syllable. For example, S_{25} gets a weight of 2.8 (1.0+1.0+0.8), while S_{14} and S_{35} get a weight of 2.4 and 2.0 respectively. The unknown syllable S_i is thus identified as S_{25}. This procedure is carried out for all the syllables during testing.

The block diagram of the speech recognition system using MFS and MFR Cepstral features both during training and testing is shown in Figure 4.3.

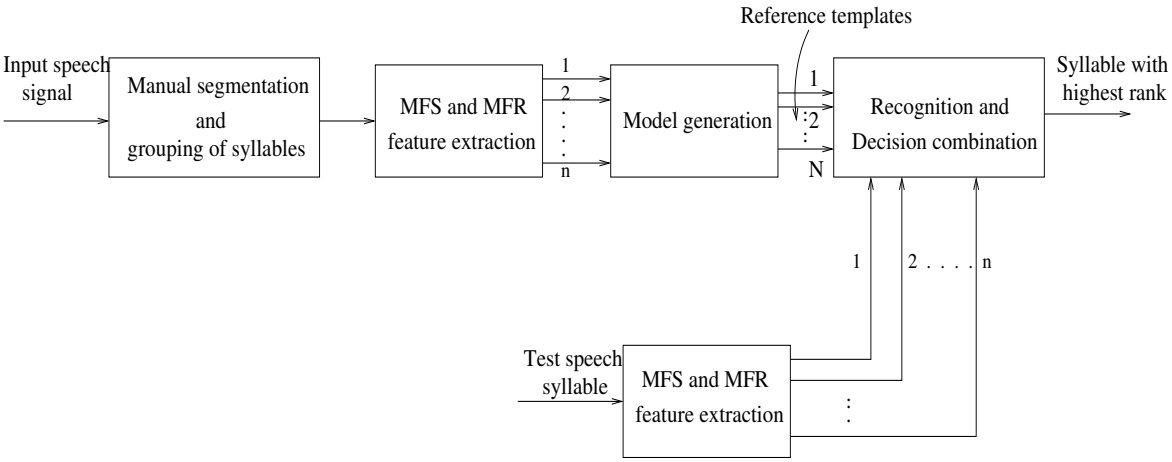

Figure 4.3: Isolated style syllable recognition system using MFS and MFR Cepstral features both during training and recognition

4.4.3 Performance analysis on manually segmented data

The final isolated style recognition results using MFS and MFR during both training and testing are compared with the system that uses MFS and MFR features only during training. These results are also compared with SFS feature based system. The isolated style syllable recognition results of speaker independent data (2-female speakers' DD news bulletins) are shown in Table 4.2.

Table 4.2: Performance analysis of speech recognition system using MFS MFR features and SFS features

Speaker	Syllable recognition performance (in %)		
	MFS and MFR only during training	SFS	MFS and MFR during training and testing
Speaker I (female)	46.7	49.1	53.3
Speaker II (female)	44.5	47.5	51.4

From Table 4.2, it can be observed that the speech recognition system using MFS and MFR Cepstral features both during training and recognition gives better performance compared to that of the remaining feature extraction techniques (SFS and MFS MFR).

4.5 Conclusions

In this Chapter, the significance of multiple frame size (MFS) feature extraction technique on number of training examples for each class of syllables, is first analyzed. It is observed that when the number of training examples are less in each class, the MFS Cepstral features give better syllable recognition compared to that of SFS features. Next, the significance of MFS Cepstral features on speaker independent data is analyzed. A new approach to feature extraction technique that uses features extracted from multiple frame sizes and multiple frame rates both during training and recognition is explored. A significant improvement in the syllable recognition performance is obtained using these features. In Chapter 5, we study the effect of MFS and MFR in the continuous speech recognition task.

Chapter 5

CSR using MUIT with MFS and MFR technique

5.1 Introduction

In CSR, the choice of speech features has a substantial influence on the separability of different classes of sound units. In Chapter 4, it has been shown that the MFS and MFR feature extraction technique can improve the recognition performance of an isolated style recognizer, if they are used both during training and recognition phases. In this Chapter, an attempt is made to make use of MFS and MFR Cepstral features for continuous speech recognition. We explore the significance of multiple frame size and multiple frame rate features both during training and recognition of a continuous speech recognizer that uses modified unsupervised and incremental training technique discussed in Chapter 3.

This Chapter is organized as follows. Section 5.2 explains the modified unsupervised incremental training technique using MFS and MFR Cepstral features. In Section 5.4, performance of this system is analyzed for two Indian languages, Tamil and Telugu. Section 5.5 gives few transcription examples for the language Tamil. Section 5.6 gives few transcription examples for the language Telugu.

5.2 MUIT using MFS and MFR Cepstral features

It is shown in the previous Chapter that the MFS and MFR feature extraction technique can improve the recognition performance of an isolated style syllable recognizer. In this Chapter, an attempt is made to make use of MFS and MFR cepstral features during modified incremental training and recognition also for the CSR.

First the continuous speech signal is automatically segmented into S syllable-like units as discussed in Section 2.4.1. The syllable segments are pruned using duration analysis and subjected to silence normalization technique. The resultant M syllable inventory is used for the training process using MFS and MFR cepstral features. The training procedure that consists of initial cluster selection followed by incremental training procedure with MFS and MFR cepstral features is discussed in the following Sections.

5.2.1 Initial cluster selection

The initial cluster selection process is the same as that of MUIT (Section 3.3.2). The only difference is that the cepstral features are extracted from the syllable segments using multiple frame sizes and frame rates. All M silence normalized syllable segments are considered and 39-dimensional cepstral feature vectors are extracted from these segments with multiple frame sizes (12,14,16,18 and 20ms) and frame rates (100, 83.3, 71.4, 62.5, 55.5 frames/sec)[1]. These features are considered to build the initial HMMs. The M syllable segments are then recognized against all M HMMs. During recognition also, speech features are extracted using different frame sizes and frame rates. For each frame size and frame rate based features, a separate recognition experiment is conducted and 3-best results are considered for each syllable segment. These recognition results are combined using the decision combination rule discussed in Section

[1]These frame rates correspond to a frame shift of 10, 12, 14, 16 and 18ms respectively.

4.4.2.

Thus the M HMMs are initialized and initial cluster selection procedure is carried out for m iterations (Here, m = 3). This initial cluster selection procedure results in N number of syllable clusters.

5.2.2 Incremental training using MFS and MFR features

After selecting the initial clusters where the models are only initialized, they are re-estimated and subjected to the incremental training procedure. In Section 3.3.2, MFS feature extraction is used only during the initial cluster selection procedure. Since MFS and MFR feature extraction technique has shown a moderate improvement in the performance over that of SFS for an isolated style recognizer, this technique is adopted during incremental training for CSR. The steps involved in the incremental training procedure are the same as in Section 3.3.3. The incremental training procedure is carried out until the convergence criterion is satisfied. This procedure leads to N_1 syllable clusters and thus N_1 syllable models. A label is manually assigned to each of the syllable model based on a listening test. These syllable models with labels are further used for the recognition task using MFS and MFR cepstral features. The recognition task using MFS and MFR cepstral features is discussed in the following Section.

5.3 MFS and MFR features for recognition

The test speech signals are first segmented into syllable-like units and they are silence normalized. During recognition, features are extracted using multiple frame sizes and frame rates for the test syllables. For each frame size and frame rate, a separate recognition experiment is conducted and 3-best results are taken for each of them. The recognition results are combined using the decision combination rule explained in

Section 4.4.2 in order to get the final syllable recognition output.

5.4 Performance analysis

Performance of the continuous speech recognition system using MUIT with MFS and MFR cepstral features is evaluated on the Indian language television news database [37]. The two Indian languages considered in this work are Tamil and Telugu. Duration of each news bulletin in the database is about 15 minutes. Section 5.4.1 analyzes the syllable recognition performance on Tamil News bulletins, whereas Section 5.4.2 analyzes the syllable recognition performance on Telugu News bulletins.

5.4.1 Performance analysis on Tamil data

The Tamil language news bulletins consisting of five female speakers are considered for training. The speech signals are automatically segmented into syllable-like units that result in S ($S = 8804$) syllables. Duration analysis and silence normalization with 20ms of silence is carried out on these syllable segments and M ($M = 8400$) syllables are obtained. Using these syllables, MFS (12-20ms, in steps of 2ms) and MFR (100, 83.3, 71.4, 62.5, 55.5 frames/sec)[2] features are extracted and the modified incremental training procedure is carried out. Here each model is a 5-state and 2-Gaussian mixture/state model. This procedure results in 621 syllable models. If a syllable model consists of examples which are entirely different, that particular model is discarded (*i.e.* once the syllable models are obtained, they are manually pruned). A syllable identity is assigned to each syllable model after listening to the syllable sounds in the corresponding cluster manually. For recognition, two female speakers, which are not seen during training, are considered. During recognition, features are extracted using different frame sizes and frame rates for the test syllables. For each frame size

[2]These frame rates correspond to a frame shift of 10, 12, 14, 16 and 18ms respectively.

and frame rate, a separate recognition experiment is conducted and 3-best results are combined to get the final recognition output. Syllable recognition results using MUIT with MFS and MFR feature extraction technique for the language Tamil are shown in Table 5.1. These results are compared with the MUIT technique with SFS cepstral features. The syllables which are correctly recognized are named as *Complete syllable*

Table 5.1: Syllable recognition performance of MUIT system using MFS, MFR features for training and testing and MUIT system using SFS features for the language Tamil

Syllables	Performance in %	
	MUIT with SFS	MUIT with MFS and MFR
Complete syllable	42.6	48.7
CV/VC only	20.8	19.2

in Table 5.1. If only the CV or VC part of the entire CVC is recognized correctly, they are categorized as CV/VC *only* in this Table. As a syllable recognizer, the CSR with MUIT gives a syllable recognition performance of 42.6% for test speakers not seen in training, whereas the MUIT with MFS and MFR features gives a syllable recognition performance of 48.7%.

In the next Section, performance of the syllable recognizer is analyzed for the language Telugu.

5.4.2 Performance analysis on Telugu data

For analysis purpose, the Telugu language news bulletins consisting of four male speakers are considered for training. The speech signals are automatically segmented into syllable-like units that results in S ($S = 13773$) syllables. Duration analysis and silence normalization with 20ms of silence is carried out on these syllable segments and M

($M = 11885$) syllables are obtained. Using these syllables, MFS (12-20ms, in steps of 2ms) and MFR (100, 83.3, 71.4, 62.5, 55.5 frames/sec)[3] features are extracted and the modified incremental training procedure is carried out. Here each model is a 5-state and 2-Gaussian mixture/state model. This procedure results in 485 syllable models. Once the syllable models are obtained, they are manually pruned. A syllable identity is assigned using *iTRANS* Telugu labels [47] to each syllable model after listening to the syllable sounds in the corresponding cluster manually.

For recognition, two male speakers, which are not seen during training, are considered. Using MFS and MFR cepstral features, syllable recognition is carried out. The syllable recognition results of MUIT using MFS and MFR feature extraction technique for the language Telugu are shown in Table 5.2. These results are compared with MUIT technique with SFS cepstral features. As a syllable recognizer, CSR with MUIT

Table 5.2: Syllable recognition performance of MUIT system using MFS, MFR features for training and testing and MUIT system using SFS features for the language Telugu

Syllables	Syllable recognition performance in %	
	MUIT with SFS	MUIT with MFS and MFR
Complete syllable	39.94	45.36
CV/VC only	14.05	14.06

gives a syllable recognition performance of 39.94% for test speakers not seen in training, whereas MUIT with MFS and MFR features gives a syllable recognition performance of 45.36%. Hence, MFS and MFR feature extraction technique can be used for continuous speech transcription task to improve the syllable recognition performance.

[3]These frame rates correspond to a frame shift of 10, 12, 14, 16 and 18ms respectively.

5.5 Transcription examples of Tamil News bulletins

The word level Transcription of few Tamil sentences is given below.

1. *mAnila mudalvarhaL kUttattai aDuttamAda mudal*
2. *kUTTam pradamar tiru aTalbihAriti vAjpEy*
3. *inDru kuTra chAtukal padivu sEya paTTana*
4. *nERRu naDai berra annaitu kaTci kUTTatil*
5. *idupaRRi muDiveDukkappaTadu*

The manual transcription and the corresponding automatic transcription of these sentences using MUIT with MFS and MFR feature extraction technique is given in Table 5.3. In Table 5.3, M.Trans corresponds to manual transcription and A.Trans corresponds to automatic transcription of the News bulletins. Here, each example is separated by a double line in Table 5.3.

If we observe the transcription examples of these sentences, silence is recognized accurately in almost all the sentences. In the 1^{st} example, the syllable */tai/* which is highlighted in the manual transcription is recognized as */Tai/* in its automatic transcription. The vowel part of both these syllables are identical and moreover the consonant part of these syllables belongs to the class of unvoiced and unaspirated sound units. In few cases, long vowels are recognized as short vowels and vice versa (*e.g.*, */mA/* as */mat/*, */kUT/* as */kuT/* and */a/* as */A/*). In almost all the cases, syllables with nasal consonants are correctly recognized. Since the syllables */pEy/* and */pai/* sound alike, the syllable */pEy/* is recognized as */pai/* in the 2^{nd} example of the automatic transcription.

Table 5.3: Transcription examples of Tamil News bulletins. SIL - Silence, M.Trans - Manual transcription, A.Trans - Automatic transcription (/∗/ indicates insertion and /@/ indicates deletion of syllables during segmentation)

M.Trans	**mA**	ni	la	mud	al	var	haL	kUT	Tat	**tai**	a	Dut	tam	mA	da	mu	dal
A.Trans	**mat**	ni	lam	mu	al	var	AL	kuT	Tap	**Tai**	A	da	tam	mA	da	mu	dal
M.Trans	SIL	kUT	Tam	pra	da	mar	ti	ru	**a**	Tal	bi	hA	ri	vA	j	**pEy**	
A.Trans	SIL	kUT	Tam	kra	da	ma	tir	@	**A**	Til	bi	hA	ri	vA	SIL	**pai**	
M.Trans	in	Dru	ku	Tra	chA	Tu	kal	pa	di	vu	sE	ya	paT	Tan	na		
A.Trans	in	Dra	pu	Tra	chA	tu	kal	pak	Di	va	sEy	@	paT	Tan	na		
M.Trans	SIL	nER	Ru	naD	ai	ber	ra	an	nai	tu	ka	Tci	**kUT**	Ta	til		
A.Trans	SIL	nER	@	naD	@	bo	ra	AN	nai	tu	Tac	Ti	**kuT**	Ta	til		
M.Trans	SIL	i	du	pa	RRi	mu	Di	ve	Duk	kap	paT	Ta	du				
A.Trans	SIL	i	du	A	kalai	mu	ri	vel	Duk	kap	paT	Tap	du				

To illustrate the readability of automatic transcription, we consider the word level manual and automatic transcriptions of a sentence. The word level manual transcription of 2^{nd} sentence is *kUTTam pradamar tiru aTalbihAriti vAjpEy*. The corresponding automatic transcription of this sentence at word level, is observed to be *kUTTam kradama tir ATil bihAri vA pai*. Here, we can observe that the automatic transcription of this sentence is readable and comparable to the corresponding manual transcription.

From the analysis made, automatic transcriptions of given Tamil sentences are observed to be readable using MUIT with MFS and MFR feature extraction technique.

5.6 Transcription examples of Telugu News bulletins

In this Section, few automatic transcription examples of Telugu News bulletins are given and they are analyzed. The word level transcriptions of few Telugu sentences is given below.

1. *kArgil lO pAkistAn chorabATudArula dADulu*
2. *E rOju vArtalloni mukhyAmsAlu*
3. *chainA vidEsAnga mantri tO*
4. *mAjI nyAya mUrti tO*

Corresponding manual and automatic transcriptions of these sentences using MUIT with MFS and MFR feature extraction technique is given in Table 5.4. Consider 1^{st} sentence in Table 5.4. The syllable */Du/* is recognized as */du/* because of identical vowel parts of these syllables. Moreover they belong to the same class of unaspirated voiced speech sounds. In 2^{nd} sentence syllable */muk/* is recognized as */mud/* because of identical CV part. In few cases, though the syllable recognition is correct in automatic transcription, it differs from word level manual transcription due to a shift in segment boundaries in the automatic segmentation procedure.

Table 5.4: Transcription examples of Telugu News bulletins. SIL - Silence, M.Trans - Manual transcription, A.Trans - Automatic transcription (/∗/ indicates insertion and /@/ indicates deletion of syllables during segmentation)

M.Trans:	kAr	gil	lO	pAk	is	-	tAn	cho	ra	bA	Tu	**dA**	ru	la	dA	**Du**	lu
A.Trans:	kAr	vI	lo	pAk	is	*	tan	@	ra	bA	Tu	**dAr**	@	gA	ga	**du**	lu
M.Trans	SIL	E	rO	ju	vAr	tal	lo	ni	**muk**	khyAm	sA	lu	SIL				
A.Trans	SIL	E	@	je	va	tal	lo	ni	**mud**	hud	sa	lu	SIL				
M.Trans	SIL	chai	nA	vi	dE	**sAn**	Nga	man	tri	tO							
A.Trans	SIL	chai	nA	@	dE	**tan**	Lam	man	tna	tO							
M.Trans	SIL	mA	jI	nyA	ya	mUr	ti	tO									
A.Trans	SIL	mA	je	nai	ya	mud	si	tO									

Consider the 1^{st} sentence in Table 5.4. As the automatic segmentation segments the poly-syllable /*dAru*/ into /*dAr*/ and discards fractional syllable /*u*/, the syllable /*dA*/ in manual transcription is recognized as /*dAr*/.

To illustrate the readability of Telugu sentences, let us consider the word level transcriptions of 1^{st} sentence in the given examples. The word level manual transcription of this sentecne is *kArgil lO pAkistAn chorabATudArula dADulu*. The corresponding word level automatic transcription obtained is *kArvI lo pAkistan rabATudArgA gadulu*. Here, we can observe that the automatic transcription of this sentence is readable.

From the analysis made, automatic transcription of Telugu sentences are also observed to be readable using MUIT with MFS and MFR feature extraction technique. Readability of these sentences can further be increased using a basic language model.

5.7 Summary and Conclusions

The work described in this Chapter has explored a new approach to speech feature extraction for a continuous speech recognizer. In this work, the Cepstral features are extracted using multiple frame sizes and multiple frame rates. Using these features the modified unsupervised and incremental training is carried out. During recognition also, Cepstral features with MFS and MFR are used. Experimental results on Tamil and Telugu News bulletins have shown that the features extracted using MFS and MFR give a significant improvement in syllable recognition performance over the base-line system. With the use of the modern computing facilities, increased number of frames to be processed during training may not be a serious issue.

Chapter 6

Summary and conclusions

6.1 Summary of the work

The conventional method of building a speech recognizer for any language requires a large segmented and labeled speech corpus. However obtaining such a corpus is a labour intensive and time consuming process. In order to adapt a recognition system to a new task or another language, it is required to reduce the cost, both in terms of human effort and financial resources. In India, there are 22 official, and a number of unofficial languages. If manually annotated speech corpora are required, building such speech recognizers even for the official languages is a difficult task. Thus, there is a need for a system that transcribes continuous speech signal without the benefit of manually annotated speech corpora.

In this work, an attempt is made to automate the continuous speech transcription task for Indian languages that does not require any manually segmented and labeled speech corpus. Several attempts are made to improve the syllable recognition performance of a continuous speech recognizer where the syllable models are trained using an unsupervised and incremental training (UIT) technique.

The following are the major contributions of the research work presented in this thesis.

- A continuous speech recognition system that uses unsupervised and incremental training (UIT) technique is considered as the baseline system for this thesis.

Though the syllable recognition results obtained using the UIT technique are promising for continuous speech transcription task, there are many issues in the baseline recognition system due to both syllable segmentation and clustering errors. In this thesis, an effort is made to analyze some of the major issues in the baseline UIT system.

- Several refinements such as duration analysis and silence normalization are made to the baseline system to improve the syllable recognition performance. The training technique after several modifications is named as MUIT technique.

  During incremental training process, improper clustering of syllables due to syllable fragments and merged syllables can be overcome by performing duration analysis on the syllable inventory. Unusual duration syllables are thus removed.

  During automatic segmentation of speech signals, some syllables may have silence at their boundaries. Since the spectral characteristics of silence are almost the same for all syllables, presence of silence at the syllable boundaries sometimes dominates the syllable clustering process. As a result, syllables that have very similar silence characteristics are clustered together. A new technique that normalizes all the syllables by prefixing and suffixing a small silence segment to the syllables is proposed. This technique is named as Silence Normalization (SN) technique. Using this technique, a significant improvement in the clustering performance is observed.

- A preliminary analysis on the syllable clustering is done based on the place and manner of articulation for the language Tamil.

- The conventional speech recognition systems use features that are extracted with a single frame size and single frame shift during training and recognition. But this may not address the issue of varying speaker characteristics in a particular speaking rate. In this work, a new feature extraction technique that uses multiple frame sizes and frame rates during both training and testing is explored

to improve the continuous speech recognition performance. This technique is used along with the MUIT technique and a significant improvement in the syllable recognition performance is obtained over the baseline system. The syllable recognition performance is evaluated for two Indian languages namely, Tamil and Telugu.

6.2 Key ideas presented in the thesis

- Major drawbacks in the unsupervised and incremental training technique are analyzed.
- Several modifications such as duration analysis, silence normalization and pruning of syllable clusters are made to the baseline system and a significant improvement in the syllable recognition performance is obtained for a CSR task. A preliminary analysis on the syllable clustering is also done based on manner and place of articulation.
- A feature extraction technique, that uses multiple frame sizes and frame rates during both training and testing, is explored to improve the continuous speech recognition performance. This technique is used along with the MUIT technique. Significant improvement in the syllable recognition performance is obtained over the baseline system for the two Indian languages Tamil and Telugu.

6.3 Criticism of the work

- No effort has been made to tune the HMM parameters such as number of states and mixtures.
- Increasing the computation time by using the features that are extracted from multiple frame sizes and multiple frame rates is not a serious issue for training the syllable models as most of the times it is done offline. Since, the MFS and

MFR feature extraction technique is used for recognition also, the computation time of the recognition process is increased. But for many practical tasks, the computation time involved in recognition should be less.

- The MFS and MFR feature extraction technique for continuous speech recognition is used only for a gender dependent system.

6.4 Future directions

- The proposed approaches to CSR have been applied only to the Tamil and Telugu language corpus which is of relatively restricted syllabic composition. In future it can be applied to other corpora for various Indian languages.
- From the syllable clustering analysis made in Chapter 3, perhaps a hierarchical classification of syllable segments can be done based on the place and manner of articulation. In the case of syllable, the basic sound unit can first be broadly classified based on manner of articulation and later they can be finely classified based on place of articulation.
- In the MFS and MFR feature extraction technique discussed in this thesis, multiple frame sizes and multiple frame rates are considered during syllable recognition. This increases the computation time during testing. Hence, a combination of variable frame size and frame rate can be tried to reduce the computation time during testing.

Bibliography

[1] **Nagarajan, T.**, *Implicit Systems for Spoken Language Identification*. PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, 2004.

[2] **Greenberg, S., Chang, S. and Hollenback, J.**, (2000), An introduction to the diagnostic evaluation of Switchboard corpus automatic speech recognition systems, *Proc. NIST Speech Transcription Workshop*, May.

[3] **Jens Allwood, E., A.**, (2001), Corpus-based research on spoken language, *Nordic Language Technology*, Museum Tusculanums forlag, pp. 59 – 68.

[4] **Rabiner, L. R., Rosenberg, A. E., Wilpon, J. G., and Zampini, T. M.** , (1982), A Bootstrapping training technique for obtaining demisyllabic reference patterns, *J. Acoust. Soc. Amer.*, vol. 71, pp. 1588 – 1595.

[5] **Ljolje, A. and Riley, M. D.**, (1991), Automatic segmentation and labeling of speech, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, Toronto, pp. 473–476.

[6] **Kemp, T. and Waibel, A.**, (1998), Unsupervised training of a speech recognizer using TV broadcasts, *Proc. of ICSLP 98*, vol. 5, Sydney, Australia, November 30 - December 4, pp. 2207–2210.

[7] **Jean-Marc and Christophe Ris**, (1999), Developement of a speech recognizer using a hybrid HMM/MLP System, *Europian Symposium on Artificial Neural Networks*, vol. 4, Bruges, Belgium, April 21-23, pp. 441–446.

[8] **Lamel L.F., Gauvain, J.L. and Eskenazi, M.**, (1991), BREF, a large vocabulary spoken corpus for French, *Proceedings of EUROSPEECH*, vol. 2, Gnes, Italy, 24-26 September, pp. 505 – 508.

[9] **Giuseppe Riccardi, Dilek Z. Hakkani-Tur**, (2003), Active and unsupervised learning for automatic speech recognition, *Proceedings of EUROSPEECH*, Geneva, Switzerland, pp. 1825 – 1828.

[10] **Matthew, A. S., Uday Jain, Bhiksha Raj, and Richard, M. S.**, (1997), Automatic segmentation, classification and clustering of broadcast news audio, *In Proc. of the DARPA speech recog. workshop*, Chantilly, VA, Feb., pp. 97–99.

[11] **Zavaliagkos, G. and Colthurst, T.** , (1998), Utilizing untranscribed training data to improve performance, *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, Feb., pp. 301–305.

[12] **Frank Wessel and Herman Ney**, (2001), Unsupervised training of acoustic models for large vocabulary continuous speech recognition, *IEEE Workshop on ASRU*, Madonna di Campiglio, Italy, 9-13 December, pp. 307–310.

[13] **Lamel, L., Jean-Luc Gauvain and Gilles Adda**, (2002), Unsupervised acoustic model training, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, May, pp. 877–880.

[14] **Asela Gunawardana and Alex Acero**, (2003), Adapting acoustic models to new domains and conditions using untranscribed data, *In* 8^{th} *European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 1-4, pp. 1633 – 1636.

[15] **Stavros, T. and William, B.**, (2005), Acoustic training from heterogeneous data sources: Experiments in Mandarin Conversational Telephone Speech Transcription, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, Philadelphia, PA, USA, March 18-23, pp. 461 – 464.

[16] **Yoshihoko Gotoh, Hochberg, M. M., Mashao, J. D. and Silverman, F. H.**, (1995), Incremental MAP Estimation of HMMs for Efficient Training and Improved Performance, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Detroit, MI, May, pp. 457–460.

[17] **Shuangyu Chang, Lokendra Shastri, and Steven Greenberg**, (2000), 'Automatic phonetic transcription of spontaneous speech, *Proceedings of Int. Conf. Spoken Language Processing*, vol. 4, Beijing, China, October 16-20, pp. 330–333.

[18] **Meinedo, H. and J. Neto**, (2003), Automatic speech annotation and transcription in a broadcast news task, *in Proc. of MSDR 2003*, Hong Kong, China, April 4-5, pp. 95–100.

[19] **Neto, J., Martins, C., Almeida, L.**, (1997), The developement of a speaker independent continuous speech recognizer for Portuguese, *Proceedings of EUROSPEECH*, Rhodes, Greece, pp. 1559–1562.

[20] **Neto, J., Martins, C., Almeida, L.**, (1998), A large vocabulary continuous speech recognition hybrid system for the Portuguese language, *Proc. of ICSLP 98*, Sydney, Australia, December.

[21] **Nagarajan, T. and Murthy, H. A.**, (2004), Non-bootstrap approach to segmentation and labeling of continuous speech, *National Conference on Communi-*

cation, IISc, Bangalore, January, pp. 508–512.

[22] **Subrata K. Das**, (1990), Supervised selection of prototypes for classification, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 697–700.

[23] **Lamel, L., Gauvian, J., L., and Adda, G.**, (2001), Investigating lightly supervised acoustic model training, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Salt Lake City, UT, May.

[24] **Nguyen, P., Rigazio, L., Jean-Claude, J.**, (2003), Large corpus experiments for broadcast news recognition, *Proceedings of EUROSPEECH*, Geneva, Switzerland, September 1-4, pp. 1837–1840.

[25] **Prasad V. K.**, *Segmentation and Recognition of Continuous Speech.* PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, 2002.

[26] **Prasad, V. K., Nagarajan, T., and Murthy, H. A.**, (2004), Automatic segmentation of continuous speech using minimum phase group delay functions, *Speech Communication*, vol. 42, pp. 429–446.

[27] **Kemp, T. and Waibel, A.**, (1999), Unsupervised training of a speech recognizer: Recent experiments, *Proceedings of EUROSPEECH*, vol. 6, Budapest, Hungary, September, pp. 2725–2728.

[28] **Rukmini Iyer, Herbert Gish and Dan McCarthy** , (2002), Unsupervised training techniques for natural language call routing, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4, Orlando, FL, May, pp. 3900–3903.

[29] *Linguistic data consortium: Switchboard corpus.* http://www.ldc.upenn.edu/: Linguistic data consortium, 1995.

[30] **Lamel, L., Jean-Luc, G. and Gilles, A.** , (2002), Lightly supervised and unsupervised acoustic model training, *Computer Speech and Language*, vol. 16, January, pp. 115 – 129.

[31] **Dilek, H., Gokhan, T., Mazin R., Giuseppe, R.** , (2004), Unsupervised and active learning in automatic speech recognition for call classification, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Montreal, Canada, May, pp. 429–432.

[32] **Paul Mermelstein** , (1975), Automatic segmentation of speech into syllabic units, *J. Acoust. Soc. Amer.*, vol. 58, October, pp. 880–883.

[33] **Pfitzinger, H.R., Burger, S., Heid, S.**, (1996), Syllable detection in read and spontaneous speech, *Proceedings of Int. Conf. Spoken Language Processing*, vol. 2, Philadelphia, pp. 1261–1264.

[34] **Villing, R., Joseph, T., Ward, T., and Costello, J.**, (2004), Automatic blind syllable segmentation for continuous speech, *ISSC 2004*, Belfast, June 30-July 2.

[35] **Howitt, A.**, (2000), Vowel landmark detection, *Proceedings of Int. Conf. Spoken Language Processing*, Beijing, China, October 16-20, pp. 628–631.

[36] **Rabiner, L. R. and B. H. Juang**, *Fundamentals of Speech Recognition*. New Jersy: Prentice Hall, 1993.

[37] *Database for Indian languages*. Speech and Vision Lab, IIT Madras, Chennai: India, 2001.

[38] **Nayeemulla Khan, A., Suryakanth, V.G., Rajendran, S.**, (2002), Speech database for Indian languages - A preliminary study, *in Proc. Int. Conf. Natural Language Processing*, IIT Bombay, Mumbai, December, pp. 295–301.

[39] **Avinash Chopde,** *ITRANS Indian Langauge Transliteration Package Version 5.2 Source*, http://www.aczoom.com/itrans/tamil/node5.html,

[40] **Vaseghi, S., Naomi Harte and Ben Milner**, (1997), Multi-resolution phonetic/segmental features and models for HMM-based speech recognition, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 1263–1266.

[41] **Okuda, K., Kawahara, T. and Nakamura, S.**, (2001), Lecture speech recognition considering the speaking rate variation, Technical Report *IEICE*, SP2001-103, Dec., pp. 13–18.

[42] **Samudravijaya, K.**, (2004), Variable frame size analysis for speech recognition, *Proceedings of Int. Conf. Natural Language Processing*, New Delhi, India, pp. 237–244.

[43] **Macias-Guarasa, J., Ordonez, J., Montero, J. M., Ferreiros, J., Cordoba, R., and Haro, L. F. D.**, (2003), Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition, *Proceedings of EUROSPEECH*, Geneva, Switzerland, September 1-4, pp. 1809–1812.

[44] **Ponting, K. M., and Peeling, S. M.**, (1991), The use of variable frame rate analysis in speech recognition, *Computer Speech and Language*, vol. 5, pp. 169–179.

[45] **Qifeng Zhu and Abeer Alwan**, (2000), On the use of variable frame rate analysis in speech recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 3, June, pp. 1783–1786.

[46] **Philippe Le Cerf and Dirk Van Compernolle**, (1994), A new variable frame rate analysis method for speech recognition, *Signal Processing Letters IEEE.*, vol. 1, Dec., pp. 185–187.

[47] **Avinash Chopde,** *ITRANS Indian Langauge Transliteration Package Version 5.2 Source*, http://www.aczoom.com/itrans/tlgutx/node3.html,

PUBLICATIONS

Journals:

1. G. L. Sarada, T. Nagarajan and Hema A. Murthy, " Automatic Transcription of Continuous Speech Without Bootstrapping", *IEEE Transactions on Speech and Audio Processing* (To be communicated).

Conferences:

1. G. L. Sarada, N. Hemalatha, T. Nagarajan, and Hema A. Murthy, "Automatic Transcription of Continuous Speech Using Unsupervised and Incremental Training", pp. 405-408, INTERSPEECH - 2004, October, Korea.
2. G. L. Sarada, T. Nagarajan, and Hema A. Murthy, "Multiple Frame Size and Multiple Frame Rate Feature Extraction for Continuous Speech Recognition", pp. 592-595, SPCOM-2004, December 11-14, Bangalore, India.
3. N. Hemalatha, G. L. Sarada, T. Nagarajan, Hema A. Murthy, "Connected Digit Recognition using Unsupervised and Incremental HMMs", NCC-2005, pp. 296-299, January, Kharagpur.
4. G. L. Sarada and Hema A. Murthy, " UIT based continuous speech recognition using MFS and MFR features", ICASSP-2006, May 14-19, Toulouse, France (communicated).

CURRICULUM VITAE

1 Name: G. Lakshmi Sarada

2 Permanent Address:

D/O G. G. K. Murthy
D.No: 26-39-37, 7^{th}lane
A. T. Agraharam
GUNTUR - 4
Andhra Pradesh, India

3 Educational Qualification:

MS (by Research) (Reg. Date 2003) Computer Science and Engineering	Indian Institute of Technology, Madras Chennai - 600 036 Tamil Nadu, India
B. Tech (2002) Electronics and Communication Engineering	Bapatla Engineering College Andhra Pradesh, India

General Test Committee members

1 Chairperson: **Prof. C. Siva Rama Murthy**

2 Guide: **Dr. Hema A. Murthy**

3 Members :

Dr. C. Chandra Sekhar
Dept. of Computer Science & Engg.

Dr. C. S. Ramlingam
Dept. of Electrical Engg.