Connected Digit Recognition over Long Distance Telephone Lines using the SPHINX-II System

Uday Jain

Submitted to the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for the
Degree of Master of Science at

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

May 1996

Abstract 4
Introduction 5
1.1 Large vocabulary vs. small vocabulary tasks 5
1.2 Overview 6
The SPHINX II System 8
2.1 Overview of SPHINX II 8
2.2 Hidden Markov Models (HMMs) 8
Recognition Unit 9
2.3 The SPHINX-II Trainer 10
Acoustical Feature Extraction 10
Lexical Feature Extraction 10
Context-Independent System Training 11
Segmentation and Senonic Clustering 11
Context-Dependent System Training 12
2.4 Decoder. 12
2.5 The Training Procedure 13
Bootstrapped Training 13
Acoustic Feature Extraction 13
Lexical Feature Extraction 13
Segmentation 13
Creation of CI-DHMMs 13
Creation of the Senonic Decision Trees and Mapping Table 14
Creation of CD-SCHMMs 14
Fine Tuning the CD-SCHMMs 14
Training from Scratch 15
Acoustic Feature Extraction 15
Lexical Feature Extraction 15
Creation of CI-SCHMMs 15
Segmentation and the Creation of Senonic Decision Trees and Mapping Table 15
Creation of CD-SCHMMs 16
Fine Tuning the CD-SCHMMs 16
2.6 Summary 16
Speech Corpora 17
3.1 MALL Databases 17
3.2 Other Databases 18
Reduced Bandwidth AN4 19
Filtered WSJ0+WSJ1 19
Macrophone 20
3.3 Summary 20
Performance of Existing Systems 21
4.1 Reduced Bandwidth AN4 21
4.2 Filtered WSJ 1PD models. 22
4.3 Macrophone models 23
4.4 Summary 24
Bootstrapped Training 25
5.1 Filtered WSJ models. 25
5.2 Macrophone models 26
5.3 Summary 27
Data-Driven Training 28
6.1 MALL 88 29
Context-Independent Semi-Continuous HMMs 29
Context-Dependent Semi-Continuous HMMs 30
6.2 MALL91. 32

Context-Independent Semi-Continuous HMMs 32
Context-Dependent Semi-Continuous HMMs 32
Context-Dependent Semi-Continuous HMMs 34
6.3 Summary 34
Word-Based Systems 36
7.1 System Description 36
7.2 Training 37
The MALL88 Database 38
Context-Independent Semi Continuous HMMs 38
Context-Dependent Semi-Continuous HMMs 39
Further Processing 40
The MALL91 Database 40
Context-Independent Semi-Continuous HMMs 40
Context Dependent Semi-Continuous HMMs 41
Further Processing 42
7.3 Summary 42
Towards a Word-based System using an Approximation to Continuous HMMs 43
8.1 Training using Multiple Gaussian Set (MGS) 44
8.2 Gender-dependent training 45
8.3 Summary 46
Environmental Adaptation 47
9.1 CDCN 47
9.2 Cross Environment Normalization. 47
MALL88 48
MALL91 49
9.3 Test set Normalization 50
9.4 Summary 51
Conclusion 52
10.1 Training models closer to the training data 53
10.2 Making the system completely digit oriented 53
10.3 Increased models size and reduced parameter sharing 54
10.4 Normalization 54
10.5 Future Work 55
Power variance training 55
Silence removal 55
References 56

Acknowledgments

I would like to thank the following people whose help and guidance made this work possible. Dr. Stern, my advisor who guided my research by encouraging me to question the methods used rather than follow them blindly. He showed me that the path to the answer has treasures more valuable than the final result. Dr. Raj Reddy for sponsoring my stay at Carnegie Mellon University. Dr. Pedro Moreno for his interest and guidance which proved invaluable. Matthew Siegler for patiently guiding me through the SPHINX II trainer. Bhiksha Raj, Evandro Gouvea and Vipul Parikh for the numerous discussions we had. Eric Thayer, Ravi Mosur and Bob Weide for the support they provide by maintaining the SPHINX II system.

Abstract

This report documents the performance of the SPHINX II system on two connected digit databases. The two databases, MALL88 and MALL91 were collected on long distance telephone lines at two sites, and they have the identical vocabulary of the ten digits + “OH”.

We describe the performance of the system using existing models developed for large vocabulary speech recognition. To further improve recognition accuracy two training procedures (bootstrapped training and data-driven training) are described and compared. It was found that further improvements can be obtained by untying parameters trained and increasing the model size. This is possible due to the small vocabulary size. To facilitate the untying of the parameters a new phone set was defined and trained, which separated distributions that were shared in previous training procedures. The new phones also ensure that training data is not shared between words. Results for all three training procedures are presented. Application of the procedures described in this report reduced the word error rate from 18 percent (using systems that had been trained for large vocabulary recognition) to 1.6 percent.

Finally the use of channel and test-set normalization was also explored. It was found that the CDCN algorithm (which compensates for unknown additive noise and unknown linear filtering) did not provide further improvements to recognition accuracy for these data.

Chapter 1
Introduction

Connected digit recognition, the recognition of the digits ZERO (and/or OH) through NINE, is an important task domain for continuous speech recognition. It shows up in a variety of applications such as speaker identification, telephone banking, form or database entry, remote or hands-free credit card transactions. In many of these applications the telephone is a very convenient interface between the user and the machine providing the services. Above all the telephone provides for remote access to all users accessing the system, because it has become truly ubiquitous. The limited vocabulary of the digit task, with its reduced acoustic confusability, also make it an ideal testbed for new recognition systems.

1.1 Large vocabulary vs. small vocabulary tasks

Traditional speech recognition systems have been very task dependent. Their performance degrades when they are tested on a domain that is different from the one with which they were trained. This drop in performance is due to the fact that systems are fine tuned to the task they have been built for, which reduces their adaptability to different tasks. In this work we explore the adaptability of the SPHINX II system across the dimension of vocabulary size.

Systems built to work on large-vocabulary tasks with dictionary sizes in the thousands of words are optimized as follows:

- Phone models. Phone units are easily trained and shared across large vocabularies.
- Large dictionaries. The large vocabularies increase the acoustic variability that has to be modelled which results in increased confusability between hypotheses.
- Shared parameters across models. To successfully model the greater variability of the data, the amount of training data required also increases. Model parameters have to be shared due to the reality of limited training data.
- Language models. Additional gain in performance is obtained from modelling the language structure.

On the other hand, systems that have been built to work on small vocabulary tasks are optimized differently:

- Word models. Words are modelled rather than the phones. This is advantageous for continuous speech because coarticulation gets better modelled.
- Small dictionaries. The smaller vocabulary results in a reduced amount of acoustic confusability.
- Independent parameters. Because the training data are being used to model fewer parameters (a digit system requires fewer acoustic classes than a large-vocabulary system), parameter sharing is no longer necessary.
- No language models. A language structure is not always available or learned for small vocabulary tasks. This means that the entire model differentiation power must come from the acoustic dimension.

In this body of work we start with a large-vocabulary configuration of the SPHINX II system, and try to adapt the system to a digit recognition task. We are working with the additional complexity that the digits were collected over the telephone network. The telephone network reduces the bandwidth of the signal available for analysis, adds noise, and adds spectral coloration to the speech signals.

1.2 Overview

We start by exploring the performance level that is obtained by models trained for large-vocabulary tasks when used to recognize a digit database test set. We test models trained on (1) bandlimited speech, (2) speech passed through an average telephone channel, and (3) real telephone speech. By so doing we observe the change in recognition accuracy as the models approach the test set across the dimension of degradation caused by limited bandwidth and other channel effects.

To further improve performance we realize that training using the digit data is required. Here we explore the usual training paradigm involving the initialization of the training data using existing models. This is the method routinely used when training models for a large vocabulary task. This method is then compared to training where no initial models are used and the training is completely driven by domain-specific data. We observe that the second procedure provides better models for the small vocabulary task.

We would still like to obtain a final system that is optimized for the digit task using SPHINX II. In an attempt to increase the model size from phone models to word models and to reduce the amount of parameter sharing we explore the possibility of word based phones. We define a completely new phone set for the digit recognition task and compare the results when models are trained using this new phone set. An improvement in performance is observed over recognition using the usual phone models.

Finally we explore the possibilities for channel and test set normalization to reduce the acoustic differences between the training and the testing speech data. Since stereo data (simultaneously recorded using noisy and clean speech) are not available, we use the CDCN algorithm in our investigation. This algorithm does not require stereo data and attempts to correct for degradations due to linear channel and additive noise.

Through this work we show that it is possible to obtain acceptable results on a digit recognition task using a system that was built for large vocabulary speech recognition.

Chapter 2

The SPHINX II System

The speech recognition system used in this project is the SPHINX II [8]. system. This chapter provides an overview of various training methods for SPHINX II, discussing each of the subsections that make up the training system. The decoder used will also be discussed to a limited degree.

2.1 Overview of SPHINX II

SPHINX-II is a large vocabulary, speaker independent, Semi-Continuous Hidden Markov Model (SCHMM) based continuous speech recognition system. The original SPHINX [10] system, developed at CMU in 1988, was one of the first systems to demonstrate the feasibility of accurate, speaker-independent, large vocabulary continuous speech recognition. Since then the system has gone through significant improvements and changes, such as the use of semi continuous HMMs and the use of senonic clustering [9], that have considerably enhanced its performance.

2.2 Hidden Markov Models (HMMs)

Hidden Markov Modelling [7],[14] is the most widely accepted and successful technology used for contemporary speech recognition. It is a statistical method that characterizes the spectral and temporal properties of speech. HMMs can model speech on many levels, including phones, syllables and words. Depending on the availability of training data and the size of the task, the model size can be selected to ensure that the parameters of the models will be estimated with minimum error. This selection of the model size is usually carried out in an *ad hoc* manner.

Speech recognition systems are adversely affected by any changes in the training and testing environment. The performance of HMMs also degrades as acoustic differences between the training and testing data increase. To alleviate this problem researchers try to either train models with training data that is very similar to the testing data, or to increase the robustness of their systems by using normalization algorithms.

2.2.1 Recognition Unit

SPHINX II models phone-level acoustic phenomena. It uses a phone set of 50 Context Independent (CI) phones along with additional phones for noise and silence segments. Since SPHINX II was designed to recognize continuous speech, it is necessary to model co-articulation of phones in different contexts. Hence instead of only modelling phones in isolation, it models phones in their left and right context. This modelling unit is called a triphone. Since the number of possible triphones can potentially exceed the number of triphones that a training corpus can suitably train, SPHINX II allows for the sharing of acoustic states between triphones that share the same central phone. The shared acoustic state is known as a senone [9].

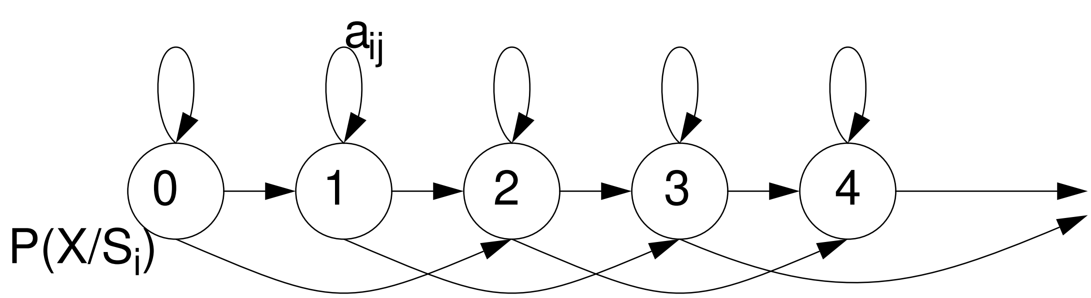


Figure 1: The 5 state left to right Bakis model used to model a triphone in SPHINX II.

Each triphone in the system is modelled as a 5 state left to right Bakis HMM [3] as shown in figure 1. The parameters that must be trained for this model are the transition probabilities (within and between states), the output distributions, and the initial probabilities for the states. The output distributions are formed by a mixture of 4 Gaussian distributions. The parameters to be learned for the output distributions are the mixture weights and the tied means and variances. As SPHINX II is a semi-continuous system the tying is across all states of all phones. However SPHINX II allows the tying to be within phone classes rather than among the entire phone set, which results in a greater pool of distributions to form the output distributions from. This increases the resolution of the models to the various acoustic events in a multi-speaker multi-environment task.

2.3 The SPHINX-II Trainer

We now describe the SPHINX II trainer. While we divide the system into individual subsections for easier understanding, it should be remembered that information from each subsection interacts at multiple levels.

2.3.1 Acoustical Feature Extraction

SPHINX II works on speech that has been parametrized using Mel-scale Frequency Cepstral Coefficients (MFCCs) [8]. The cepstral coefficients are static features that carry the short time frequency information of the speech signal. Dynamic features are obtained by taking the first and second order differences of the cepstral coefficients. The following are the steps followed in the SPHINX II front end to obtain acoustic features for telephone bandwidth speech:

- Input speech is digitized at a sampling rate of 8 kHz
- A pre-emphasis filter is used to suppress the effects of glottal pulses and radiation impedance while enhancing the spectral properties of the vocal tract.
- A Hamming window of 25.6 ms duration is applied at the rate of 100 windows/sec
- The power spectrum of each windowed signal is computed using a 512-point DFT
- 40 Mel Frequency Spectral Coefficients are derived based on mel-frequency bandpass filters with 13 linear bands for 200 Hz to 1 kHz and 18 logarithmic bands for 1 kHz to 3.5 kHz
- For each frame, 12 MFCCs are computed using the cosine transform of the logarithm of the modified spectrum that was obtained from the above filter outputs
- The dynamic features are obtained by taking the first- and second-order differences. Normalized cepstral power, difference power, and second-order difference power form the fourth feature stream

These streams are then individually clustered into 256 groups using a hierarchical clustering algorithm [11]. The data are then vector quantized [6] using these clusters. Diagonal variance matrices are calculated for each cluster.

2.3.2 Lexical Feature Extraction

The basic lexical feature is the phone set that is to be used in the final system. SPHINX II is built around a phone set of 50 unique context-independent phones, 10 noise phones (*e.g.* lipsmacks, exhale, inhale and other such extraneous acoustic phenomena), and 3 types of silence phones. The next level of acoustic information is the set of words to be recognized. These, along with their pronunciations are defined in the dictionary using the above phone set. SPHINX II allows multiple pronunciations of the same word.

Once the phone set and the dictionary are ready, a list of all possible triphones is created using the dictionary. If these triphones appear in the training set with a preset frequency, they will be trained. The rest are classified as unseen triphones. All unseen (untrained) triphones are decoded using the senonic mapping table. Unseen triphones are modelled as a sequences of trained states derived from trained triphones. The mapping table consists of a list of all the trained states. An initial mapping table is constructed using the triphone list and senonic decision trees from the selected initial models.

2.3.3 Context-Independent System Training

The acoustic data clusters serve as initial models for the training of semi-continuous models. These models start from flat output distributions, uniform initial states, and uniform state transition probabilities. This training requires the vector-quantized data, the transcriptions, and the segmentation information. Segmentation information is generated during forced recognition of the training data. This procedure will be described in the next section.

Baum-Welch training [4] results in Context-Independent Discrete HMMs (CI-DHMMs). These models are not necessarily optimized. They are obtained to serve as initial models for semi-continuous training and as seeds for the initial discrete context-dependent models from which the senonic clustering trees are based.

2.3.4 Segmentation and Senonic Clustering

Segmentation involves forced recognition of the entire training corpus using existing models. The segmentation allows us to classify automatically the frames of training data into distinct acoustic classes, which in our case are the CI phones and the triphones defined in the triphone list. Since the transcriptions are available, it is assumed that the segmentation output will result in clean clusters for the models to be trained. However, phoneme-level segmentation of continuous speech is not an easy task due to the high degree of co-articulation that blurs phoneme boundaries. Hence it is best to always use the best possible models for the segmentation task.

The clustering process also generates revised transcripts that include alternate pronunciations for words with multiple pronunciations in the dictionary. The transcriptions are used to compile an accurate list of all the triphones that can be trained from the training corpus. The triphone list, the CI-DHMMs, and the classified training data are used to generate unclustered Context-Dependent DHMMs.

The next step is the creation of senonic clustering trees. A senone is the shared output distribution associated with a cluster of similar states. The output distributions of the unclustered CD-DHMMs are used to decide which states should be clustered together based on an entropy measure. Senonic decision trees are generated using linguistic questions. Each leaf on these trees is a senone that can be trained. This allows the modelling of unseen triphones that might appear in test data.

2.3.5 Context-Dependent System Training

Context-dependent modelling is achieved by modelling phones in left and right context, called triphones. Baum-Welch training is performed using the CI-DHMMs and the variance information as initial models. The resulting Context-Dependent (CD) models are Semi-Continuous (SCHMMs) and speaker independent.

Further fine tuning to groups of speakers is possible by separating the training data into the desired groups and training individual models for each group. Usually this just involves separating the data into males and females for gender-dependent models. However further classifications into finer sets of speaker types is possible. The only constraints are that more training data and a more complex decoding scheme are needed.

2.4 Decoder

The SPHINX II decoder nominally performs 3 passes on test utterances [2]. In this task only the forward pass was used on the test utterances, as the backward pass did not improve the results in any appreciable way. The third pass involves a rescoring of the lattices generated based on acoustic and language models. This pass was not performed as no language model was used in this task. The digit task did not require a language model because we assumed every word to be equally likely.

When working with different sets of models (such as male models and females models), the decoder generates a pre-selected number of the most likely transcription hypotheses for each set of models. The most likely hypothesis from the combined set of hypotheses for the particular utterance is selected based on a reevaluation of the acoustic score. Working with multiple models obviously increases the run time, as multiple acoustic models have to be used for decoding.

2.5 The Training Procedure

We now provide a detailed description of the two training procedures, “bootstrapped training” and “data-driven training” that are evaluated in this work. The difference between these procedures lies in the use of existing models to bootstrap the training procedure.

2.5.1 Bootstrapped Training

This procedure relies on existing models to bootstrap the training of a new set of models. The models selected have been previously trained on the same data or data from a similar domain. Along with the models the corresponding senonic decision trees are also required.

2.5.1.1 Acoustic Feature Extraction

The acoustic feature extraction procedure described above is followed to obtain cepstra for the entire training database. The cepstral vectors obtained are vector quantized.

2.5.1.2 Lexical Feature Extraction

The lexical feature extraction also proceeds as described above. The phone set is dictated by the bootstrapping models being used. The senonic decision trees are used along with the triphone table for the training set to generate a mapping table for the training set. This mapping table is required by the context-dependent bootstrapping models.

2.5.1.3 Segmentation

Segmentation of the training set is carried out using the bootstrapping models, associated mapping table, triphone list, and dictionary for the training set.

2.5.1.4 Creation of CI-DHMMs

Training vectors are classified into phones by the segmentation process. These classified vectors are now used to train CI-DHMMs. The process usually uses existing discrete models as initial models, but models with uniform output distributions also can be used as initial models. The CI-DHMMs, along with the variance information derived during VQ, serve as initial models in the training of the CD-SCHMMs. They are also required in the senone clustering process which will be described in the next section.

2.5.1.5 Creation of the Senonic Decision Trees and Mapping Table

The CI-DHMMs along with the triphone list and the classified frames provide us with rough discrete estimates for the CD phones. The output distributions of these CD models are compared using an entropy measure and similar ones are clustered (they share the output distributions) into senones. The senonic decision trees and a mapping table for the models being trained are created.

2.5.1.6 Creation of CD-SCHMMs

Using the CI-DHMMs as initial models, and the mapping table containing senone clustering information, the CD-SCHMMs are trained using the Baum-Welch algorithm. As the CD-SCHMMs are the final models that will be used for the recognition process, care is taken to insure that they are optimally trained. This is usually done by ensuring that the output probabilities keep increasing and by monitoring their performance on a development set (to ensure that they are not over trained).

2.5.1.7 Fine Tuning the CD-SCHMMs

Further improvement in performance of the models is obtained by focusing the models on specific classes of speakers. This allows the models to be sharper, as they can be used to model distinct subsets of speakers rather than the entire set. This technique is very useful in improving the performance of speaker-independent systems as it provides us with a way to approach the performance of a speaker-dependent system in the limit. A common classification of the speakers is based on the gender (or pitch information) of the training speakers.

However it should be noted that training gender-dependent models increases the demand on training data as we are splitting the training data into two sets. The existing training data are now used to estimate twice the number of parameters. Another drawback with multiple models is that a specific set of models must be selected for testing at run time. A computationally expensive solution is to run the test utterances on all the final models and select *a posteriori* the hypothesis with the best score.

2.5.2 Training from Scratch

This method differs from bootstrapped training in that no existing models are used to bootstrap the training procedure. This might be because of a lack of previous models for the task at hand or similar tasks. The number of parameters that needs to be estimated increases with the size of the task. Without existing models to provide good initial estimates for large tasks, this procedure becomes computationaly expensive and can result in suboptimal models.

2.5.2.1 Acoustic Feature Extraction

The acoustic feature extraction is carried out as described above.

2.5.2.2 Lexical Feature Extraction

The lexical feature extraction is also as described above but with one difference. Because no bootstrapping models are used, a mapping table for the training data cannot be generated. This does not pose a problem as we shall show that an initial mapping table is not required when models are being trained from scratch.

2.5.2.3 Creation of CI-SCHMMs

Differences in the two procedures emerge with the creation of the CI-SCHMMs. For each of the phonemes that must be trained, uniform HMMs with flat output distributions and equally-likely transition probabilities are constructed. Using the transcriptions, maximum likelihood training is used to train the CI-DHMMs. The difference between the bootstrapped process and this process is that these models are trained using the transcriptions in a maximum likelihood framework rather than using preclassified training vectors. The estimated means and variances for the clusters, along with the previously-trained DHMMs, are used for further Baum-Welch iterations to obtain CI-SCHMMs. Care must be taken that the models are optimally trained as these models will be used to create the context-dependent models.

2.5.2.4 Segmentation and the Creation of Senonic Decision Trees and Mapping Table

The segmentation and the creation of the senonic decision trees and the mapping table is carried on as before, but using the CI-SCHMMs for segmentation rather than existing models. Since CI models do not require parameter sharing, and there are no unseen acoustic classes in the training data, no initial mapping table is required.

2.5.2.5 Creation of CD-SCHMMs

Once the senonic trees and mapping table are available, Baum-Welch training proceeds with the CI-SCHMMs used as initial models. In larger tasks, as we have noted, the eventual models are sometimes sub-optimal. In this case the models are used to resegment the data so that the senones can be reclustered and the models are further iterated with this new clustering information. This process is continued until the models achieve optimal performance or an *ad hoc* criteria to stop the process is reached.

2.5.2.6 Fine Tuning the CD-SCHMMs

The training set can be clustered into speaker clusters using the same procedure as in bootstrapped training, once the generic CD-SCHMMs have been created.

2.6 Summary

We have provided a brief overview of the SPHINX II system. The two training procedure we describe are the training procedures compared in this work. The two procedures differ on the technique used to initialize the training data. The optimum way to initialize training data would be by hand segmenting and labelling all the training utterances into the acoustic classes being trained. This is too expensive given the amount of training data used to generate models. Hence we attempt to use automatic procedures to initialize the training data.

In bootstrapped training, existing models are used to initialize the training data. It is the usual procedure used to train models on the SPHINX II system. This method presupposes the existence of models that are from a domain close to the domain for which models are being trained. This ensures that the training data will be initialized without the errors caused by domain mismatches.

Training from scratch is usually carried out when existing models are not suitable to initialize the given training data. This usually occurs when the new task is significantly different from the tasks for which models had been created.

Chapter 3
Speech Corpora

This body of work was performed on two corpora of connected digits that were collected on long distance telephone lines. This chapter describes the similarities and the differences of the two databases. Their individual characteristics have led to differences in the training processes employed to obtain models.

Existing speech corpora were also used for comparison experiments and to initialize the training of the digit recognition systems. These speech databases and their features that make them useful to this body of work are also described.

3.1 MALL Databases

The two major databases used in this work are known as the MALL88 and the MALL91 corpora. These data were provided by AT&T Bell Labs. MALL88 was collected in the year 1988 while MALL91 was collected three years later. In each case data were collected over long distance telephone lines from two sites, Long Island and Boston. Within a given year of data collection, data from the two sites are similar enough to be combined for training and testing. The speech was recorded on electret and carbon button microphones, and training and a test set were collected from each site. As the corpora were collected at different times, the recording equipment used differs in the two cases. This difference contributes to differences in performance between the two databases.

The MALL88 and MALL91 data consist of utterances of connected digits from male and female speakers. The corpora are hence ideally suited for speaker independent continuous speech systems. The vocabulary is limited to the ten digits with ‘0’ being uttered either as “ZERO” or as “OH”, resulting in a total vocabulary of eleven words. The utterances are of variable length. Even though within a corpus the maximum utterance length (in number of words) can be ascertained from the training set, such information was not used in this work.

As the speaker is equally likely to speak any digit no language model was provided and none was used. Each word in the vocabulary is assumed to be equiprobable. This allows us to concentrate purely on the problem of acoustic modelling.

	Train utt	Test utt	M/F in Tr. Set	M/F in Tr. Set
MALL88	9039	4584	26/62	11/35
Long Island Set	5109	2083	26/24	11/10
Boston Set	3930	2501	0/38	0/25
MALL91	2585	2688	117/125	123/126
Long Island Set	1306	1331	61/60	59/66
Boston Set	1279	1357	56/65	64/60

Table 1: Comparison of the MALL88 and MALL91 databases across the two sites, the training and test sets and the speaker distributions.

Table 1 shows the distributions of utterances across recording sites, speaker gender, and training-set vs. testing-set size, for the two MALL databases. The MALL88 data is sampled at 8kHz and stored in PCM linear format wav files. We see in the above table that this database is female heavy as no male speech was recorded at the Boston site. Each speaker recorded 124 utterances, most of which are included in the corpus. A few were left out due to errors in the recording process. The utterance length ranges from a single digit to 7 digits. In contrast the MALL91 utterances are longer, ranging from 14 to 16 digits with some 10 digit utterances. The longer utterances have the additional characteristic of being spoken faster and hence there is a greater degree of co-articulation in this database. Each speaker recorded 12 utterances, most of which are included in the corpus. A few were left out due to errors in the recording process. The MALL91 data are also sampled at 8 kHz but mu-law encoded. The mu-law speech samples were coded in gray code and stored in .wav files. Programs to decode the speech were provided by Bell Labs.

3.2 Other Databases

The MALL databases are different from the ones usually used with the SPHINX II system. To measure the performance of SPHINX II on the MALL data existing models were initially used.

These models had been trained on the data described below. Additionally models trained on these data were also used to initialize the training for the MALL databases. The following are some telephone-bandwidth databases for which SPHINX II models have been created.

3.2.1 Reduced Bandwidth AN4

AN4 [1] is a 106 word alphanumeric database. Its training set is composed of 74 speakers (53 males and 21 females). There are about 14 utterances recorded by each speaker. It was collected over two channels in an office environment. The “clean” channel was recorded using a Sennheiser HMD224 close talking microphone while the “noisy” channel was recorded using a Crown PZM6FS omnidirectional desk-top microphone.

Since this database was originally sampled at 16 kHz, it contains more acoustic data than a telephone channel. In this version of the database the clean channel was downsampled and passed through a low-pass filter with a 3600-kHz cutoff frequency. This was done to limit the bandwidth to that of telephone speech.

3.2.2 Filtered WSJ0+WSJ1

The Filtered WSJ0+WSJ1 [12] database was created for the 1994 ARPA Evaluation Hub2 evaluation. This evaluation was aimed at measuring the recognition performance in an unlimited-vocabulary speech recognition task where the speech was recorded on long distance telephone networks. The database was created by passing WSJ0 and WSJ1 clean speech through a filter with a frequency response that was designed to resemble that of an average telephone channel. The power spectrum of the average telephone channel used is shown in Fig. 1, [12].

The Wall Street Journal (WSJ) [13] databases were collected to enable research in general-purpose, large English vocabulary and high perplexity speech recognition tasks. These data were collected from 1987 to 1993.

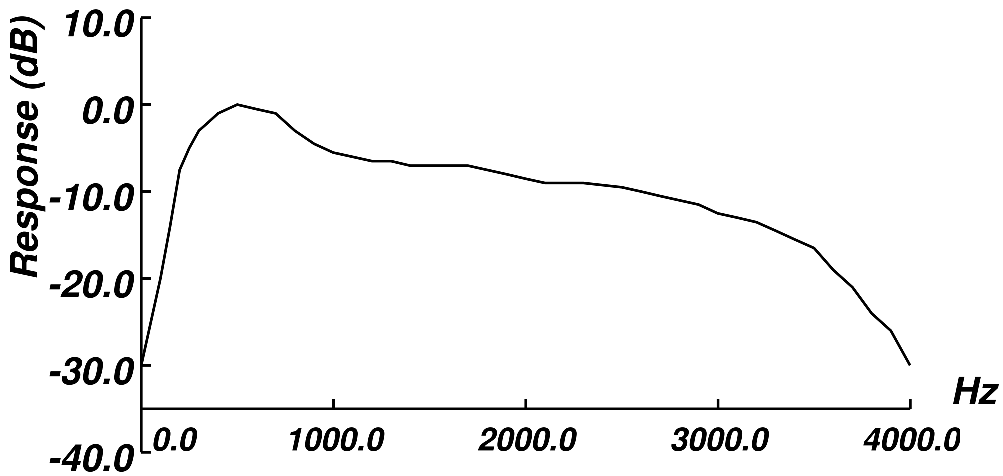


Figure 1: Equalization filter applied to WSJ0 and WSJ1 utterances to approximate the power spectrum of telephone speech.

3.2.3 Macrophone

Macrophone is a large telephone speech database (200,000 utterances recorded by about 5000 American speakers) that was sponsored by the Linguistic Data Consortium (LDC) [5]. The utterance were collected in 8-bit mu-law format directly from a T1 line. It contains varied styles of speech data (digits, alphabets, dates, places, single word commands and longer utterances) adaptable to different tasks.

3.3 Summary

We have provided descriptions of the various databases that were used in this work. The AN4, Filtered WSJ0+WSJ1 and the MACROPHONE databases were only used in the initial part of this work. Most of the development work solely depended on the MALL databases.

Chapter 4

Performance of Existing Systems

Small vocabulary continuous speech recognition on the telephone network is a task significantly different from the large vocabulary tasks for which the SPHINX II system is normally used. Hence to calibrate our performance the MALL88 database was tested on some of the existing models and decoding systems at our disposal.

We test models trained on bandlimited speech, speech passed through an average telephone channel and real telephone speech. The improvement in performance observed as the models approach the test set across the dimension of degradation caused by channel effects show how the system depends on the acoustic differences between the training and the test data being a minimum.

4.1 Reduced Bandwidth AN4

Models trained on the downsampled and lowpass filtered version of the AN4 database [1] were used. While the models were trained on speech with the same bandwidth as telephone speech, the training speech lacks other degradations found in telephone speech, such as attenuation of higher frequencies and noise in the telephone channel. This is a source of mismatch between the Reduced Bandwidth AN4 and the MALL88 data. Because the MALL88 vocabulary is a subset of the AN4 task, the test was first run using the AN4 dictionary. The mismatch between models and test data compounded with the larger dictionary caused a large number of recognition errors. The test set was then run with the subset of the AN4 dictionary that contained only the vocabulary of the MALL88 database. This time acoustic confusions involving words not in the MALL vocabulary did not occur and the results were a purer measure of the acoustic mismatch between the models and the test environment.

Models used	WER
AN4 reduced band models and AN4 dictionary	34%
AN4 reduced band models and Digits dictionary	18%

Table 2: Results with AN4 Reduced Band models. A comparison of the acoustic confusability by the system dictionary size.

We see that removing extraneous words from the dictionary caused a relative drop in error rate of 47%. From this point on the digit dictionary was used in all recognition systems. To further improve performance, models that better model the degradations in the telephone environment and that are more robust to them are required.

4.2 Filtered WSJ 1PD models

The filtered WSJ 1PD models were trained on the filtered WSJ0 and WSJ1databases, as described above. The 1PD indicates that the output probability distributions are modelled by Gaussians that are derived from a single set of 256 Gaussian distributions. In this way the models for all the phones are tied to one another.

Filtered WSJ System	WER on entire test set	WER on part of test set matched to models used
Perfect sex classification of test utt. (male and female models used)	12.9	
Male models only	14.5	7.2
Female models only	15.9	14.9

Table 3: Performance of the filtered WSJ sex-dependent models. Perfect classification of the test utterances is compared to running solely on either male or female models. The performance of the male models on male speech and female models on female speech is also provided for comparison.

The male models perform better than the female models. After perfect sex classification of the test utterances (using knowledge of the identity of the speaker), performance is better than using only male or only female models. This is because these models do not perform well in cross conditions. The WER rate after perfect sex classification however has fallen another 28% compared to results obtained using the reduced bandwidth AN4 models. This improvement is primarily due to the fact that the filter used is a better model of the frequency response of telephone channels than the down-sampling used in the AN4 database. Furthermore, the filtered WSJ models were better trained as a greater amount of continuous speech data was used.

4.3 Macrophone models

The Macrophone [5] models, like the filtered WSJ models, were trained for the 1994 ARPA Hub 2 evaluation. However these models were trained on real telephone speech. The WSJ and TIMIT utterances that are part of the Macrophone database made up the training corpus. The rest of Macrophone corpus was not used because it was not suitable for training models for unlimited vocabulary continuous speech. These models were also 1PD models.

Macrophone System	WER on entire test set	WER on part of test set matched to models used
Perfect sex classification of test utt. (male and female models used)	8.8	
Male models only	8.8	4.1
Female models only	12.8	10.4

Table 4: Performance of the sex dependent Macrophone models

The Macrophone models also produce the difference in performance between the male and female models that was observed in the filtered WSJ models. With the use of Macrophone models the WER has fallen another 32%. This further improvement is due to the fact that the Macrophone models do more than simply model the shape of the telephone channel. Other sources of degradations have also been modelled and hence there is a better match in the Macrophone models and the test speech from the MALL88 database.

4.4 Summary

We have seen that as the mismatch in the models and dictionary of the recognition system was reduced, performance on the MALL88 database improved. The best WER achieved so far was obtained with models trained on a subset of the Macrophone corpus and using a dictionary consisting of the digits. The fact that the best result were obtained by models trained on real telephone speech also show us that bandlimiting and linear filtering are only approximation to the effects the telephone network has on speech.

Further improvement in this task can be achieved if the models are adapted to the MALL domain. This can be achieved by using the MALL training data. Since enough data exist for training a unique set of models, models trained on the MALL88 database should provide us with better results.

Chapter 5

Bootstrapped Training

As stated in the previous chapter, further improvements in acoustic modelling of the MALL environment would reduce the acoustic mismatch between the models and the test environment and further decrease the WER. The next step in the process was to train models specific to the MALL88 and MALL91 databases. Since the MALL databases have their own training sets there are enough data to train models specific to the task rather than just adapting existing models.

Adapting existing models requires training on existing models using training speech from the new domain. Training new models requires the generation of a new senone mapping table and senone clustering trees that are unique to the individual databases.

The training was carried using existing models to bootstrap the training process, hence this training paradigm is called bootstrapped training. The models best suited to this task were the existing models trained on telephone environments, namely the filtered WSJ and the Macrophone models that were tested in the previous chapter. The male models were used to use to bootstrap the training process as they perform much better than the female models.

The MALL test sets were recognized using the final models. The models were tested both in “matched conditions” (training and testing data from the same database) and “cross conditions” testing (training data from one database and test data from the other database). This was done to assess the performance and the robustness/adaptability of the models that had been generated. We also wanted to see that if there was much difference in performance using models generated by the two databases.

5.1 Filtered WSJ models

We first used the filtered WSJ models as bootstrap models during the training procedure described above. Generic CD-SCHMMs were obtained, that were not fine tuned to any particular speaker set.

Five iterations were required for the MALL88 data and six iterations for the MALL91 data. This is because the variability in the MALL91 data is greater than that in the MALL88 database.

	MALL88 generic models	MALL91 generic models	MALL88 gender dep. models	MALL91 gender dep. models
MALL88 test set	8.2	12.2	6.2	11.0
MALL91 test set	9.7	6.3	7.8	4.7

Table 5: Filtered WSJ used to bootstrap the MALL models. Results of the generic and gender dependent MALL88 and MALL91 models.

We see from the results in Table 5 that the models do not perform well in cross conditions. This suggests that the acoustic mismatch in the two databases is greater than the models' ability to generalize. The absolute error rates, however, have dropped further.

There were enough training data to train gender-dependent models, which produced improved recognition accuracy. Male and female models were individually selected on a sentence-by-sentence basis according to the hypothesis transcription that had the best score. Further improvements were observed in absolute error rate.

5.2 Macrophone models

The Macrophone models were also used to bootstrap the training process for the MALL88 and the MALL91 databases. The results obtained using the resulting generic and gender dependent models trained are shown in Table 6.

	MALL88 generic models	MALL91 generic models	MALL88 gender dep. models	MALL91 gender dep. models
MALL88 test set	8.1	11.9	6.0	10.8
MALL91 test set	9.6	6.1	7.7	4.5

Table 6: Macrophone models used to bootstrap the MALL models. Results of the generic and gender dependent MALL88 and MALL91 models.

The performance of these models has improved over the performance achieved by the Macrophone models because of the use of MALL data for additional training. The performance however is not an improvement on the models that were bootstrapped using the filtered WSJ data. This suggests that the parameters of the bootstrapped models have settled in a local maxima.

5.3 Summary

Task-specific training has improved the system's performance on the connected digit task even further. However the WER is still a lot worse than is achieved by digit recognition systems, for clean speech. In this case the data were obtained from telephone networks so there is a greater degree of acoustic mismatch between training and testing environments. Telephone environment notwithstanding, the classification power of the models must be improved further. Improvements can be obtained by improving the inputs to the training process.

Thus far the inputs to training have been the initial models and the mapping table used in the training and even further back the models used for the segmentation of the data. Further improvements presumably can be obtained by using better models for segmentation. The initial models will have to be closer to the optimum models and the clusters created for the mapping table will have to be cleaner. These points will be tackled in the following chapters.

Chapter 6

Data-Driven Training

We saw in the previous chapter that using task specific training data improves the accuracy of the recognition system. However to further improve performance we have to improve the performance of the segmenting models and start the training process with better initial models. Even though the existing telephone models from the WSJ tasks provide good results, they are not a perfect environmental match. They were trained for an unlimited vocabulary task while the vocabulary of the MALL databases is limited to the digits.

This mismatch between training and testing conditions concerns more than just the size of the dictionaries in both the systems: the filtered WSJ and Macrophone models have to be very robust and acoustically adaptable test data that are not seen during the training phase (the unseen triphone problem). On the other hand, due to the limited vocabulary of the MALL database, there is a negligible amount of unseen test data. Since we know that the test data will be very similar to the training data we want the models to model the training data perfectly rather than also provide acoustic coverage for unseen test data.

Another difference, though indirect, is that the filtered WSJ and the Macrophone models rely on language models to provide discriminating information in the case of acoustic confusions. Since the systems built for digit tasks do not have language models and hence acoustic confusion will have to be further minimized in these systems. Hence it was decided that attempts should be made to disconnect the filtered WSJ/Macrophone models from the training process as far as possible. There are two ways that this can be achieved.

The first method of reducing the dependence on cross-domain models is to use the models created in the current training run to resegment the data and carry out a complete retraining. Since the entire training process uses Maximum Likelihood estimation it is assumed that the models achieved at the end of retraining will perform better than the previous models. While these models will perform better on training data, it is assumed that the train and test sets are acoustically similar so that performance on the test set will improve as well. Several iterations of segmentation and training will produce optimally trained models that bear very little resemblance to the cross-domain models used initially. Unfortunately, this process is extremely time consuming even for a simple digits task. In addition, there is no lower bound on the rate of convergence of the models. Furthermore, an *ad hoc* criteria must be used to stop the process which might also be sub-optimal. For these reasons this method was not used.

Another way of reducing the dependence on cross-domain models is to create models for the new task from scratch, and using these models to segment the data. We call this method the data-driven approach as the training is completely driven by the training data with no influence from previous models. Furthermore when there is acoustic confusability in the training set, decisions will be based on what has been learned from the same database, acoustic properties of another training database will not corrupt the current training process. We observed that models obtained by this procedure outperformed the models obtained using the training procedures previously discussed.

A similar approach was used for the MALL91 database with minor differences which were due to the differences in the two databases. The training procedure used for both MALL88 and MALL91 will be described in detail in the rest of the chapter.

6.1 MALL 88

The procedure described above to train models from scratch was used to obtain generic CD SCHMMs. These were further fine tuned to separate male and female models.

6.1.1 Context-Independent Semi-Continuous HMMs

Once the acoustic and lexical features were extracted as described above, five iterations of maximum likelihood (Baum-Welch) training were run to train discrete models. These are the initial models for training CI-SCHMMs. To train optimally the CI-SCHMMs the output probabilities were closely monitored. Baum-Welch was stopped when the current iteration failed to increase the output probabilities by more than 5%. It was also seen that at this stage further training did not improve the decoding performance of the models. Five iterations were required to train these models to acceptable level.

The performance of these models (after the fifth iteration) was as follows. They were tested on test data from the training database (MALL88) as well as on test data from MALL91.

MALL88 models	MALL88 test set	MALL91 test set
Filtered WSJ	6.2	11.0
Macrophone	6.0	10.8
CI-SCHMMs	3.7	4.1

Table 7: The performance of the MALL88 Context-Independent models trained from scratch. Comparison points with the filtered WSJ and Macrophone gender dependent models is also provided.

Table 7 shows that these models, on the two test sets, have already dramatically outperformed the models that were obtained after a complete training dependent on the filtered WSJ or the Macrophone models. This is an encouraging result as this training procedure is not complete yet and the context dependence has yet to be trained into the models. We now clearly see the advantage of using only training data and models from the environment at hand. The cross-domain nature of the filtered WSJ and the Macrophone models caused mismatches in the phone cluster generation used to train the initial models that the training process could not recover from, causing the final models to settle at a sub-optimal performance level. The improved performance of these models will help minimize this effect.

6.1.2 Context-Dependent Semi-Continuous HMMs

As the best possible models are used for the segmentation task, the CI-SCHMMs that were trained from scratch were the models used for segmentation. This insured that in this entire training process, of the MALL88 models, only information derived from the MALL88 training database was used.

Once the mapping table was created Baum-Welch (maximum likelihood) training was performed using all the available training data from MALL88. The progress of these models was also closely monitored. The models were monitored for increasing output probabilities, and tested at every iteration to make sure that over training had not occurred. The models converged to the final generic models after four iterations of the BW algorithm. They were tested on the MALL88 and the MALL91 test sets.

Similar training was carried out to produce male and female models. The male models required four iterations and the female models required three iterations of further Baum Welch training. This difference is attributed to the fact that the MALL88 training set has three times more female speech data than male data. This causes the generic models to be closer to the final female models than the male models, so it took less iterations to train the female models.

MALL88 models	MALL88 test set	MALL91 test set
CI-SCHMMs	3.7	4.1
CD-SCHMMs:		
generic	2.2	2.5
gender dependent	1.9	2.6

Table 8: The performance of the MALL88 Context-Dependent models trained from scratch. Comparison point with the Context-Independent models is also provided.

We see that there has been further improvement in the performance of these models. The generic CD-SCHMMs have improved by 40.5% on the CI models in the matched condition. Further improvement is expected as we move from context-independent models to context-dependent models in a continuous speech recognition task because of the improved modelling of co-articulation effects of continuous speech. We observed a 73% improvement of the generic CD-SCHMMs on the generic models and 64.6% improvement using the gender-dependent models trained by bootstrapping with the filtered WSJ system. Similar improvements are observed when the results are compared to the models bootstrapped from Macrophone models, 73% for the generic models and 63% for the gender-dependent models.

It should also be noted that the cross-domain performance of these models has improved as well. The performance on the MALL91 test set has improved 74.2% relative to (filtered WSJ-bootstrapped generic models). Hence we see that only using the MALL88 data to train these models does not have an adverse effect on the cross domain adaptability of these models.

With gender-dependent training, a further decrease of 13.5% in the word error rate is observed for the MALL88 test set. The error rate on the MALL91 test set, however, has increased

4%. As before, we have seen that further fine tuning of the generic MALL88 models has improved the performance on the MALL88 test set but reduced the robustness of the models as applicable to the MALL91 test set.

6.2 MALL91

A similar training process was used for the MALL91 database. However certain differences required special treatment in the training procedure in this case. The main difference was that the speech in this database has a greater degree of co-articulation. The digit strings are longer and this caused the utterances to be spoken at a faster rate which caused the greater co-articulation. This required additional iterations of Baum-Welch and segmentation in the training process.

6.2.1 Context-Independent Semi-Continuous HMMs

The training of initial models for the MALL91 set followed the same lines as that for the MALL88 set. In this case, however, the performance of the CI-SCHMMs was not better than that obtained in the bootstrapped training.

MALL91 models	MALL88 test set	MALL91 test set
Filtered WSJ	7.8	4.7
Macrophone	7.7	4.5
CI-SCHMMs	8.4	7.8

Table 9: The performance of the MALL91 Context Independent models trained from scratch. Comparison points with the filtered WSJ and Macrophone gender dependent models is also provided.

We assume that as this training progresses the performance of the models will improve beyond the level of the previously-trained models. By training context dependent models we will be able to better model the coarticulation effects in this data.

6.2.2 Context-Dependent Semi-Continuous HMMs

We stated earlier that segmentation should be carried out with the models that provide the best recognition result. The CI-SCHMMs are clearly not the models with the best result but they were used to produce segmentation information as we wanted to create the mapping table with information derived from the MALL91 database only. With the mapping table created the training proceeded as before with iterations of the Baum-Welch algorithm. The following are the results obtained from the models that were obtained after the fifth iteration of training:

MALL91 models	MALL88 test set	MALL91 test set
Filtered WSJ	7.8	4.7
Macrophone	7.7	4.5
CD-SCHMMS	4.7	4.3

Table 10: Word Error Rate of first pass Context-Dependent generic MALL91 models tested on the MALL88 and the MALL91 test sets. Comparison points with filtered WSJ and Macrophone bootstrapped gender-dependent model is also provided.

The CD-SCHMM models have moved away from the local maxima around the bootstrapped models. This is seen in the improved performance in the matched and cross conditions. Additionally the result obtained is a 44.8% further reduction in word error rate in the matched conditions when compared to the performance of the CI-SCHMMs. This is comparable to the 40.5% error rate reduction that was obtained in the similar case for the MALL88 models. However to achieve the level of performance that we have with the MALL88 models further training for this database is required.

For databases where the greater degree of acoustic confusability causes the first pass models to be sub-optimally trained the accepted procedure is to resegment the training data with the CD-SCHMMs and rerun the BW algorithm. This causes the models to move closer to the optimum performance level. This resegment-train cycle is continued until the performance of the models does not improve any further. At this point it is assumed that the resulting models are optimum. A resegmentation-training cycle on the MALL91 data improved the performance on the MALL91 test set to 3.5% Word Error Rate.

The training procedure now calls for another resegmentation-train iteration. This process is very time consuming and with this database we have seen that the convergence of the models is slow. A reason for the slow progress is that the senone clusters do not improve drastically with every resegmentation. At this point we assumed that since the mapping table had been generated after two iterations of segmentation it reflects the optimum senone clusters. Therefore the latest mapping table was used with the MALL88 generic models as initial models resulting in the optimum models without having to invest in the expensive resegmentation process iteratively. After 4 iterations the performance of the models stabilized at a WER of 1.9% on the MALL91 test set.

6.2.3 Context-Dependent Semi-Continuous HMMs

Further training of male and female models was also performed. The results obtained using these models are presented in Table 11.

MALL91 models	MALL88 test set	MALL91 test set
CI-SCHMMs	8.4	7.8
CD-SCHMMs:		
generic	2.4	1.9
gender dependent	2.4	1.8

Table 11: The performance of the MALL91 Context-Dependent models trained from scratch. Comparison point with the Context-Independent models is also provided.

Compared to the Context-Independent models these models have reduced the word error rate by 75% on the MALL91 test and 71% on the MALL88 test set. These models have also improved performance by 62% on the filtered WSJ and 60% on the Macrophone bootstrapped models. Fine tuning the models to the speaker gender further improved the performance of these models. The performance of the MALL91 models are now at par with the performance of the MALL88 models

6.3 Summary

Training from scratch has resulted in models that perform at a much better level than the models that were trained from existing systems. The process was cycle intensive and care had to be taken at each stage that the system was not over-trained. Every initial model used had to be as optimum as possible in order to obtain optimum models at later stages. In cases where the co-articulation or other acoustic features (such as noise) make training harder, more of CPU-intensive iterations of generating the mapping table are required. This process, though expensive, does improve the overall performance of the final models.

At this point the best models that we have for both databases are the gender-dependent, context-dependent, semi-continuous HMMs. The output probabilities in these models are all tied, having been derived from the same set of 256 Gaussian distributions. One way to increase the differentiability of the models would be to use multiple sets of Gaussians for the models. This is usually done by grouping certain phonemes according to their acoustic characteristics with each group of phonemes sharing its distributions with a different set. This approach will be used to train the next set of word-based models described in the following chapter.

Chapter 7

Word-Based Systems

The main difference between the MALL databases and the ones with which SPHINX II is usually used is the vocabulary size. SPHINX II was developed to perform well on large-vocabulary tasks. In such tasks normally there are not enough training data to train models perfectly for every word in the database. SPHINX II overcomes this problem by training context-dependent phones. In continuous speech the articulation of a single phone differs according to the word it is present in. SPHINX II reduces this dependency to the left and right phones in the current context. Though triphone modelling reduces the amount of training data required, there are still not enough data to train individually every triphone. For this reason senones that share the same central phone are tied to share the same output distribution. This tying is part of the senone mapping table.

In the digit task the vocabulary is small enough that word models may be used. If there are enough training data, word models perform better than phone models because they do a better job of modelling intra-word context dependencies. However, due to the continuous nature of the data, inter-word context dependency must be incorporated into the final models. Sharing of senones is not necessary in the digit task because with the small number of phones (22) and possible contexts (again limited by the size of the dictionary) it should be possible to individually train all senones. Hence untying some of these dependencies should improve the performance of SPHINX II on the digits task.

7.1 System Description

A new phone set was defined to train a system that is close to a word-based continuous system. For all the words in the dictionary the phones were made word-dependent. For example in the original dictionary based on the original phone set the representation for “SEVEN” is

S EH V AX N

In the new dictionary the representation is

S_7 EH_7 V_7 AX_7 N_7

Here we see that the word “SEVEN” is now made up of unique phones that are not shared by any other word in the vocabulary. The phone “S” that used to be common to SIX also now appears as “S_6” in the word SIX and “S_7” in the word SEVEN. Hence by untying the individual phones at the word level the number of phones has increased from 22 to 33. Since the vocabulary is small in this task, making the phones word-dependent does not increase their number drastically. We still have fewer phones than the number required by SPHINX II to model all the sound in English.

The use of this phone structure ensures that triphones are not shared across words. Untying of this nature is only possible in a small-vocabulary system as it would greatly increase the number of parameters to be learned in a larger system. Learning the increased number of parameters from the same amount of training data is more likely to be detrimental than beneficial in a larger system.

Another change that was made was that the number of senones was not minimized. Each senone was allowed to form a cluster containing just itself. However for optimum usage of the training data some clustering of the senones did take place.

Now that identical phones that had been shared by two words have been differentiated in the new phone set, it was easy to set up Gaussian sets for phones from single words and to ensure that there is no sharing of distributions between different words. All the phones that make up a single word have tied output distributions.

7.2 Training

As the word-based system uses a new phone set, for which no initial models exist, a training method to generate initial models must be selected.

One method would be to use existing models as initial models. Thus the model for N would be an initial model for N_1, N_7 and N_9. The advantage of this method is that training could proceed at a rapid rate as the existing models would just be fine tuned to the different contexts for every model. Instead of training each new model from the ground up just a few iterations of Baum-Welch retraining should provide us with fully-trained models. The disadvantage, however, is that because each of the new models are seeded from the same initial models, the new models may inherit enough of the properties of the initial models that they would remain mutually confusable.

A second method would be to use the data-driven approach and train the new phone set from scratch. This method has already been shown to be successful in training models for the MALL88 and the MALL91 databases, albeit using the traditional phone set. This process has the additional advantage of facilitating comparisons between systems based on the original and word-based phone sets. The disadvantage is that this approach is computationally expensive, especially for the MALL91 database.

7.2.1 The MALL88 Database

The training procedure followed a similar path as in training from scratch using the original phone set. Vector-quantized data were used to bootstrap the CI-DHMMs. These models eventually led to the CD-SCHMMs by the way of segmentation of the training data by CI-SCHMMs.

7.2.1.1 Context-Independent Semi Continuous HMMs

Acoustic and lexical features were extracted as described above. As before five iterations of maximum likelihood training was used to generate CI-DHMMs models starting from flat distributions. With these models as the initial point maximum likelihood training was performed to generate acceptable CI-SCHMMs.

The performance of these models on the tests sets MALL88 and MALL91 is shown in Table 12. For comparison, results obtained using the original phone-set have also been included. It should be noted that even though these models are context independent, a degree of context dependency has been built into the models because each phone depends on the word of which it is a part. Hence these models already have some of the advantages of intra word context modelling.

MALL88 models	MALL88 test set	MALL91 test set
Original phone-set	3.7	4.1
Word-based phone-set	3.2	4.2

Table 12: Word Error Rates for MALL88 CI-SCHMMs

This context dependency is one of the reasons for the 13.5% improvement in performance over the original phone models using the MALL88 test set. The 2.4% increase in errors observed for the MALL91 data suggests that as these models become more fine tuned to the subtleties of the of the MALL88 database they lose some of their generality.

7.2.1.2 Context-Dependent Semi-Continuous HMMs

The next step towards context dependent modelling is the segmentation of the training data. In cases such as the current one where a new system is being built (new phone set) no previous models exist. This is where the advantage of training from scratch is evident; it provides us with initial models to perform the segmentation.

The segmentation process resulted in a list of 375 trainable triphones for this task. The 375 triphones are made up of a total of 1875 unclustered senones. In this system we attempt to minimize the tying between these senones so as to approach the spirit of a fully-continuous system. The tying is now performed not at the senonic level but only at the output distribution level. Later attempts will be made to further untie the output distributions of the HMMs. It should be noted that each untying increases the amount of training data required, as the number of free parameters in the system increases.

Once the senone mapping table was created, three iterations of Baum-Welch training were required to obtain context-dependent SCHMMs. Further iterations did not improve the performance of the models. Once again these models were tested on both the MALL88 and the MALL91 test sets.

MALL88 models	MALL88 test set	MALL91 test set
Original phone set	2.2	2.5
Word based phone set	2.1	2.6

Table 13: Word Error Rates for MALL88 CD-SCHMMs

The results obtained with the models trained with the original phone set have been presented along with the results for the new models. The new models perform slightly (4.5%) better than the existing models on the MALL88 test set and slightly worse on the MALL91 test set. By this stage the models based on the original phone set are context dependent and hence have a performance close to the word-based phone models. The advantage of the word-based models is that each triphone (except the inter-word triphones) has been trained individually from scratch, and information from triphones sharing the same central phone has not been shared. This has allowed for cleaner triphone modelling, as in this case there were enough training data to train the triphones from scratch. In cases where there is limited training data, some amount of tying is necessary to obtain robust estimates for the triphone. In our case each triphone model is bootstrapped with the model of its central phone.

In the cross-condition test we can see the limits of the robustness of this method. As a set of models are fine-tuned to a certain training set, their adaptability to cross conditions deteriorates. These models have been fine tuned to the MALL88 database, so their performance on the MALL91 database is worse than models using the original phone set. This trend was also evident with the CI-SCHMMs.

7.2.1.3 Further Processing

At this point we have generic context-dependent, SCHMMs for the word-based phones. Further processing for these models includes sex dependent and Multiple Gaussian Set training. Both these procedure will be discussed once the creation of context-dependent, SCHMMs for the Word-based phones for the MALL91 database has been described.

7.2.2 The MALL91 Database

The training for the MALL91 word-based phone models closely followed the training of the original phone models.

7.2.2.1 Context-Independent Semi-Continuous HMMs

Once the acoustic and lexical features were extracted as described above, Context Independent HMMs were trained. The performance of the CI-SCHMMs are shown in Table 14.

MALL91 models	MALL88 test set	MALL91 test set
Original phone set	8.4	7.8
Word-based phone set	8.3	7.1

Table 14: Word Error Rates for MALL91 CI-SCHMMs

As stated earlier word-based phones already have a degree of context dependence built into them. This context dependency provides the 8.9% improvement on the original phone context-independent system, on the MALL91 test set (matched conditions). However the improvement on the MALL88 test set (cross set conditions) is minimal. The context dependencies of the MALL91 data set have not improved performance on to the MALL88 task. This is another indicator to the differences in these two data sets.

7.2.2.2 Context Dependent Semi-Continuous HMMs

The CI-SCHMMs that had just been trained were used for segmenting the training data. The segmentation process resulted in a list of 385 trainable triphones for the MALL91 database. The 385 triphones are made up of a total of 1925 unclustered senones. The clustering between these senones was minimized so that each senone would be individually trained hence approximating a continuous system.

Once the mapping table had been created, context-dependent models were trained. As with the original phone system the first pass generic CD-SCHMMs were used to resegment the training data and obtain a more accurate senonic mapping table. This mapping table was used with generic MALL88 word based, context-dependent, phone models as initial models to train generic MALL91 word based, context-dependent, phone models. The performance of these models on the MALL88 and the MALL91 test sets is shown below.

MALL91 models	MALL88 test set	MALL91 test set
Original phone set	2.4	1.9
Word based phone set	2.5	2.2

Table 15: Word Error Rates for MALL91 CD-SCHMMs

We observe that the Word-based phone model have not achieved the level of performance of the original phone models. We assume that the sharing of parameters allowed the distributions in the original phone models to be better trained. The training data in the MALL91 database does not seem sufficient to train word-based models to the same level of performance. We however assume that the performance of the word-based models will improve further with the processing described in the following chapter.

7.2.2.3 Further Processing

We described the training of generic context-dependent SCHMMs for the MALL91 word-based phone system. The next step in the training procedure would be to train gender-dependent models (fine tune the CD-SCHMMs to the male and female speakers separately) and multiple Gaussian set system. The work on these extensions will be described in the next chapter.

7.3 Summary

We have described the training of word-based phone systems for the MALL88 and the MALL91 task. The system development in both cases proceeded as in the case of the corresponding systems using the original phone set. The performance using word-based phones is better than that obtained using the original phone set. This gain is primarily due to the fact that we have minimized the amount of acoustic training data shared by the within-word triphones, thereby reducing the acoustic confusability of the resulting triphone models. It should be noted that this method presupposes the availability of enough training data to accurately model the new phones and the increased number of triphones and may reduce the generality of the models obtained.

Chapter 8

Towards a Word-based System using an Approximation to Continuous HMMs

The training of the word-based systems up to this point has followed closely the development of the original phone systems that had been trained from scratch. Further processing in the case of the normal phone systems implied gender dependent models and some degree of parameter optimization. The advantages of the word-based system are that we can carry the processing a step further towards a continuous system that uses word models to model the vocabulary. Both procedures should further improve performance.

In a system that uses word models, each word in the vocabulary is modelled by a unique HMM. The advantage is that within-word transitions between phonemes are perfectly modelled as training data are not shared between words. Inter-phoneme transitions are fine tuned to the single context that is found in the word being modelled. Hence each final model perfectly models its individual word to the extent that it can be trained.

SPHINX II is a semi-continuous HMM system. Its state output probabilities are mixture Gaussians consisting the top four distributions from a set of 256 Gaussians. Hence the modelling of the output probabilities are inherently limited to the combinations possible from the given set. In a fully continuous system the output distributions are made up of untied (unshared) Gaussians. This limitation, in SPHINX II, can be removed by increasing the number of Gaussian distributions to choose from. SPHINX II achieves this in a systematic manner. Phones are grouped into different acoustic classes. Each phone class then shares a completely independent set of 256 Gaussian distributions to generate output probabilities for all the senones in that class. To move SPHINX II more towards a continuous system it was decided to group the phones according to the words from which they were derived rather than according to their acoustic classes. We are able to do this because each word in our dictionary is made up of unique phones. This ensures that training data for one digit are not used to train the output distributions for a senone from another digit. In this manner we achieve fully-continuous word modelling using the SPHINX II system. If enough training data exists, the word-based classes can be further divided according to acoustical characteristics

In this chapter we will first describe the final stages of the development of the word-based system. Once this system has been trained we will attempt to further fine tune the system development by training gender dependent models for both the MALL88 and the MALL91 systems. It should be noted that both these methods increase the demand on the amount of training data required as they increase the number of parameters to be learnt. This increased data requirement remains a limiting factor in the eventual performance of these systems.

8.1 Training using Multiple Gaussian Set (MGS)

The set of Gaussian distributions that are shared and make up the output distributions of the senones is referred to as a Gaussian set (GS). Traditionally a Gaussian set in SPHINX II is made up of 256 unique Gaussian distributions. Up to this point the systems consist of only one Gaussian set. The following systems have individual Gaussian sets for each of the digits. All the senones, of these systems, that make up the phones of a particular digit derive their output distributions from the single Gaussian set associated with the digit. Hence there is no sharing of Gaussian distributions among the digits.

The initial models used were generic models using a single Gaussian set. The distributions in these models were used as a starting point for the 11 distribution sets that make up the new models. The training was stopped after two iterations since further iterations did not improve performance on the test set. The models were tested on the MALL88 and MALL91 test sets. The results are shown below in Table 16.

	MALL88 SGS models	MALL88 MGS models	MALL91 SGS models	MALL91 MGS models
MALL88 test set	2.1	1.8	2.5	2.5
MALL91 test set	2.6	2.2	2.2	1.9

Table 16: Word Error Rates for MGS CD-SCHMMs

These results show that the MGS models perform better than the previous SGS models. They perform better on the matched conditions. They improve performance in the cross conditions for the MALL88 models while no change is observed for the MALL91 models. The MALL88 models improve performance by 14% in matched conditions and 15% in cross conditions. Similarly the MALL91 models show a 13.6% improvement in the matched conditions.

8.2 Gender-dependent training

Gender dependent training is a common way to further fine tune the models to a particular set of speakers. Here the models are fine tuned to the gender of the test speaker. The training corpus is divided according to gender with the male training speakers used to fine tune the male models and the female training speakers used to fine tune the female models. It should be remembered that the training available for the male/female models is about half of what was available for the generic models. Hence the performance of the gender-dependent models is closely linked to sufficiency of the training data to suitably estimate the gender dependent parameters.

Using the MGS generic models as initial models gender dependent training was carried out. These models were then tested in matched and cross conditions. The results are shown below in Table 17.

	MALL88	MALL91
MALL88 models	1.6	2.3
MALL91 models	3.1	2.2

Table 17: Word Error Rates for gender dependent MALL MGS CD-SCHMMs

The MALL88 models show an improved performance in the matched conditions with a marginal drop in performance in the cross conditions. This is a trend that has been observed throughout this work: as the models get fine tuned to one database their performance on the other database degrades. The MGS gender-dependent models provide the best performance so far on the MALL88 test data. The performance on the MALL91 models has however degraded in both conditions. This show us that the training data was not able to successfully trained the increased parameters caused by moving to a MGS gender-dependent system.

8.3 Summary

We have explored the possibility of improving the performance of the generic CD-SCHMMs word based phone models that were trained. The first technique used was to reduce parameter sharing between the models. Gaussians that make up the output distributions were shared between states that made up a single word, instead of across all trained states. This ensured that the training data for a specific word only trained parameters for that word, and was not used to jointly estimate parameters for other words in the dictionary. We observed improvements in performance for both the MALL88 and the MALL91 database when tested in matched conditions. The MALL88 models also improved performance in the cross conditions. This suggests that by reducing the amount of parameter sharing we have increased the discerning power of the models. Parameter sharing causes the models to be similar to each other and hence more confusable.

The second technique, gender dependent training, has been explored with previous models. In this case we observed improved performance only for the MALL88 models. Gender-dependent training doubles the number of parameters being trained and the MALL91 training data was not able to successfully estimate these increased parameters. We see that the improvements obtained from gender dependent training add on those obtained from reduced parameter sharing. However with the caveat that there should be enough training data to train the larger set of independent parameters

Chapter 9

Environmental Adaptation

Recognition systems perform best when the testing environment closely matches the training environment. This is especially true for recognition systems based on HMMs. To increase robustness to environmental differences between testing and training conditions algorithms such as CDCN (Codeword-Dependent Cepstral Normalization) [1] are used. These algorithms are designed to work on the cepstral vectors before they are used for recognition and normalize them (bring them closer) to the training vectors.

Both the MALL88 and the MALL91 databases consists of two data recorded at 2 different sites, namely Long Island and Boston. These parts during training have been considered similar enough to be used together for training purposes. We now look at the acoustic differences in each of these parts closely and study our ability to adapt to the acoustic differences that might exist between the data collected from the two sites.

Another source of acoustic mismatch that usually occurs, to a lesser degree, is the found between the training and testing utterances. We would like our system to be robust to these differences as they degrade the system performance in addition to the degradation extraneous sources such as additive noise.

9.1 CDCN

CDCN is a technique for dealing jointly with additive noise and channel equalization. It does not require stereo training data. Hence it is suitable for a task such as recognition on the telephone where no stereo data exist.

9.2 Cross Environment Normalization

In this section we look at acoustic differences across the recording site dimension. We look at CDCN's ability to normalize these acoustic mis-matches and reduce the difference in performance between matched and cross site recognition results. The experiments are performed separately for the MALL88 and the MALL91 databases.

9.2.1 MALL88

The entire MALL88 training set was divided into two training sets - the Long Island set and the Boston set. A similar division was performed on the test set. Using the new training set models were trained for the Long Island and for the Boston sets. The models trained were SGS CD-SCHMM word based phone models. Only generic models were trained as the training data had already been halved and further division for gender-dependent or MGS models were likely to have been poorly estimated.

The initial models used in the training procedure were the CI-SCHMM word-based phone models. Three iterations of Baum Welch were performed for both the Long Island and Boston data. Table 18 shows the results obtained, testing separately on speech recorded in Boston and Long Island.

	LI test set	BOS test set
LI models	2.3	1.9
BOS models	3.8	2.0

Table 18: Baseline Word Error Rates

The Long Island models perform better because there is 30% more training data from LI than from Boston. The differences in matched and cross conditions do suggest a certain amount of acoustic differences between the two sets. We see that both models perform better on the Boston test set. This was expected for the Boston models, but it is surprising for the LI models. It is possible that the Boston part of the MALL88 database has more clearly-enunciated utterances though these differences are not apparent by casual listening. This would explain the better performance of both models on the Boston test set.

CDCN distributions were now trained separately on the Boston and LI training sets. The test sets were processed with CDCN and recognition with the above models was rerun. The following results were obtained after CDCN was used to process the test data.

	LI test set	BOS test set
LI models + LI trained CDCN	2.0	2.0
BOS models + BOS trained CDCN	3.6	2.0

Table 19: Word Error Rates after CDCN

We see that CDCN has brought the LI test set acoustically closer to the LI models. We see an 13% improvement on the LI test set. On the other hand CDCN did not improve the performance of the Boston models on the Boston test set. This was expected as we assumed that the Boston test set was already quite similar to the Boston training set and normalization was not required. The cross condition results were not significantly changed.

9.2.2 MALL91

As with the MALL88 database, SGS CD SCHMM word-based phone models were trained for the Long Island and for the Boston sets, after splitting the database.

The models were then tested on the Boston and Long Island test sets. The results obtained are shown in Table 20.

	LI test set	BOS test set
LI models	4.1	6.3
BOS models	4.9	5.6

Table 20: Baseline Word Error Rates

The differences in matched and cross conditions for this MALL91 data are greater than those observed in the MALL88 database. This suggests that there is a greater degree of acoustic differences between the two, Boston and LI, sets. In this case both models perform better on the LI test set. This again suggests that in the MALL91 database the LI set is acoustically cleaner and less confusable.

CDCN distributions were trained separately from the MALL91 Boston and LI training sets. The test sets were processed with CDCN and recognition with the above models was rerun. The following are the results obtained after test data has been processed with CDCN.

	LI test set	BOS test set
LI models + LI trained CDCN	3.8	5.7
BOS models + BOS trained CDCN	4.2	5.1

Table 21: Word Error Rates after CDCN

CDCN has improved the performance of the models both in matched conditions and cross conditions. The performance of the LI models improved by 7% on the LI test set and 9.5% on the Boston test set. The performance of the Boston models improved by 9% on the Boston test set and 14% on the LI test set. We see that the gains have been more in the cross conditions. These results show us that there is a greater amount of mismatch in the MALL91 database than there was in the MALL88 database. This was expected as the training for the MALL91 database was always more involved in order to reach a comparable (to MALL88) level of performance. Furthermore we see that the gains obtained by CDCN are more across the different environment rather than across the training and testing sets.

9.3 Test set Normalization

In this section we look the ability of CDCN to bring the test set closer to the training set by applying CDCN to the test set. The CDCN distribution are trained on the entire training set. These experiments for the MALL88 and the MALL91 databases were performed using the models that provided the best recognition accuracies. We did this in order to learn how much CDCN processing would help when the most optimized models were being used. The performance of unoptimized models can be improved by CDCN and by further training thus obscuring the real cause for the gain in performance.

Following are the baseline (best system) and post CDCN (best system + CDCN) results for the MALL88 and MALL91 databases. MGS sex dependent models were used for the MALL88 tests and MGS generic models were used for the MALL91 tests. Cross-database tests were not performed as these databases differ along more than just the environment dimension.

Train & Test Sets	No CDCN	Test Set CDCN Processed
MALL88	1.6	1.6
MALL91	1.9	1.8

Table 22: Comparison of baseline WERs to those obtained after CDCN was applied to the test sets.

We observe no significant improvement in the performance with the use of CDCN. This suggests that the models have completely modelled all the environmental variability in the training set and as the test set is close to the training set further normalization does not improve performance further.

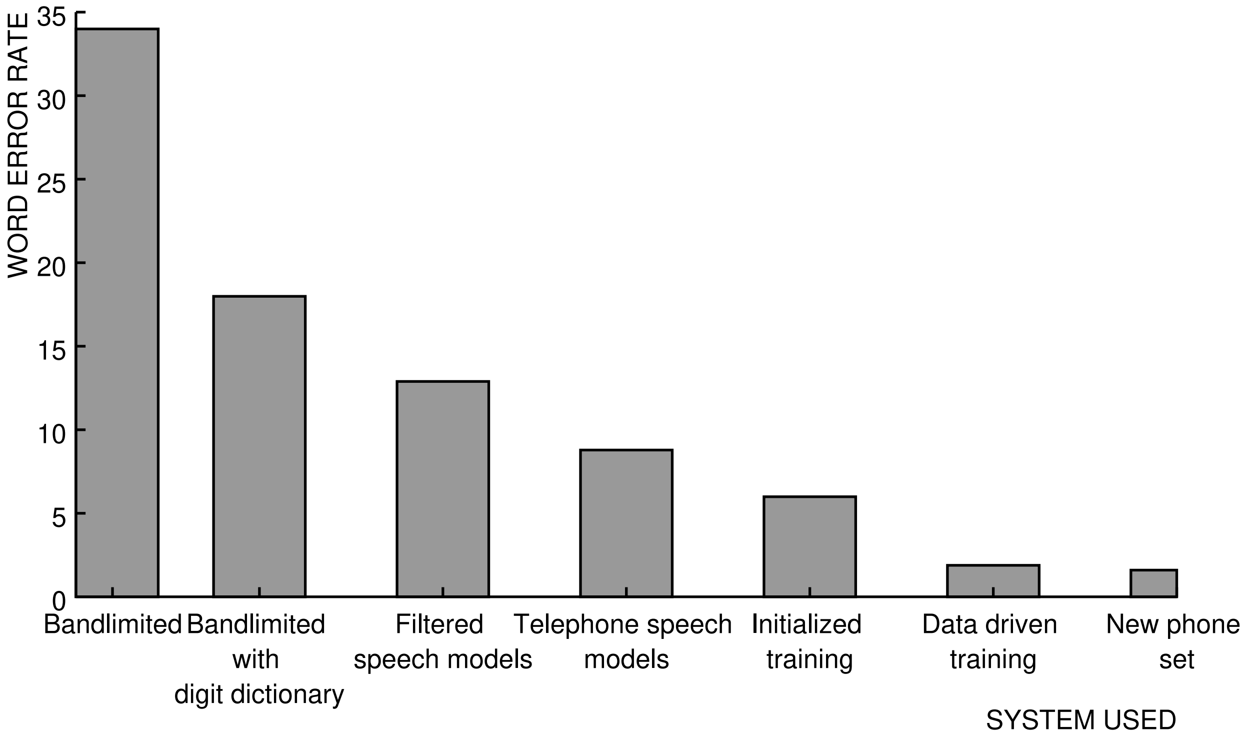
9.4 Summary

The following are the facts that were gleaned about CDCN's ability to achieve channel and environment normalization. CDCN improved performance to a limited degree due to channel normalization. CDCN appears to help when the models have not reached the optimum performance level. However, CDCN does not improve performance once the models have been optimized to closely model the data. There are two reasons for the lack of performance gain provided by CDCN. CDCN works best in conditions where there is a mismatch in the acoustic condition os the training and testing data. These mismatches are modelled as the test data being corrupted by a linear channel and additive noise. In this task we reduce mismatch between training and testing data by ensuring that the training data resembles the test data as closely as possible. Furthermore the linear channel assumption is a good model for the telephone channel. We believe that better results can be obtained by using a normalization technique that models the nonlinearities found in telephone channels.

Chapter 10

Conclusion

In this chapter we discuss the results we obtained in this work. Finally we close by suggesting some directions for future work.



Overall word error rate reduction

In this work we have shown that a large vocabulary system, SPHINX II, can be successfully adapted and optimized for the task of digit recognition. As we see in the diagram above, which shows the reduction in WER obtained for the MALL88 test set, we started with a system that had been optimized for large vocabulary recognition and by exploring new training paradigms and adapting the acoustic models we were able eventually to achieve respectable results for the digit recognition task. We summarize the advantages obtained from the different solutions explored

- Training models as close to the testing data as possible
- Making the system completely digit oriented
- Increasing the model size, from phone-level models to word level models
- Reducing the amount of parameter sharing

10.1 Training models closer to the training data

The digit database that we were trying to recognize had been collected over long distance telephone channels. Starting from models trained on bandlimited speech we improved the performance by testing on models trained on speech filtered by the average telephone channel. This improvement was due to the additional modelling of the spectral coloration of the channel. Further improvements were then obtained by testing using models that were trained on real telephone speech. This improvement reflects the fact that speech recognition systems are sensitive to differences between the training and testing environments. Furthermore we observe that the effect of the telephone channel is more than that of bandlimiting and attenuating some frequencies of the speech signal.

10.2 Making the system completely digit oriented

Large vocabulary systems are optimized to work with dictionaries with thousands of words. These systems then depend on language model scores, along with acoustic scores to prune the search space caused by the large vocabulary. In the digit task we did not have the added benefit that is derived from language modelling. The search space was reduced by limiting the dictionary to the digits to be recognized. This improved the performance by reducing the amounts of substitutions in the hypotheses.

Further improvements were obtained by training models using the digit training database. To this end we explored two training paradigms. The first, using existing models to initialize the training data, is the method routinely used when training models for a large-vocabulary task. This method did not improve performance much further than the initializing models. We believe this was because the initializing models caused the training to settle in a local maximum which was close to the maxima that the models themselves had been trained too. We require a completely fresh way of initializing the training data if we are to move away from the local maxima of the initializing models.

The second method attempted to move the training procedure away from local maxima around previously trained models by not using them to initialize the data. In this case the models were trained using only the training data and not using any extraneous information or assumptions. This method provide models that showed a drastic improvement in performance. This method is applicable to a small vocabulary task with automatic clustering as the models to be learned are small in number and the training data exhibits limited variability. This method would not be successful in a large-vocabulary tasks as the amount of variability is higher and the training data cannot be clustered automatically without substantial errors. Hence for these tasks the training data is initialized by using existing models.

10.3 Increased models size and reduced parameter sharing

Small-vocabulary systems by their very nature have fewer parameters to train. This alleviates the problem of limited training data that one faces when training large vocabulary systems. The fewer parameters allow us to reduce the parameter sharing and eventually train sharper models that closer model the acoustic data. The existence of enough training data also allows us to increase the model size (from phone to word models) which is a advantage in continuous speech recognition systems. Larger models allow closer modelling of coarticulation effects. We achieved both of these ends by defining phones that were dependent on words, word based phones. In this work we have shown that these models improve performance over the usual phone models. This leads us to believe that greater improvements can be derived by training pure word models.

10.4 Normalization

Finally we explored the possibility of bringing the training set and the test set closer by channel and test set normalization. To this end we used the CDCN algorithm which works in situations where there is no stereo (simultaneous recording of clean and noisy) data. This algorithm assumes that the speech has been corrupted by a linear channel and additive noise. Because linear filtering and noise effects do not appear to be the limiting factors for digit recognition in the present system, we did not observe improvements in performance using CDCN on our databases. We believe that better results could have been obtained by modelling the nonlinearities found in telephone channels.

10.5 Future Work

Further improvements in performance in the digit recognition task can be obtained by

10.5.1 Power variance training

Here we increase the parameter set to also model the variance in the power of our feature vectors. This method has shown some improvements in recognition of telephone and low SNR speech.

10.5.2 Silence removal

We should make sure that the training and testing corpus do not have large amounts of non-speech events at the beginning and the end of the utterances. This is particularly important when we try to automatically initialize the training data, as noise events would cause incorrect clustering of the training data. Pilot experiments showed improved performance.

References

[1] A. Acero, “Acoustical and Environmental Robustness in Automatic Speech Recognition”, Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Sept., 1990.

[2] F. Alleva, X. Huang, and M. Hwang, “An Improved Search Algorithm for Continuous Speech Recognition”, IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. II 307-310, May, 1993.

[3] R. Bakis, “Continuous Speech Recognition via Centisecond Acoustic States”, 91st Meeting of the Acoustical Society of America, April, 1976.

[4] L. Baum, “An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes”, Inequalities 3:1-8, 1972.

[5] J. Bernstein and K. Taussig., “Macrophone: An American English Telephone Speech Corpus for the Polyphone Project”. International Conference on Acoustics, Speech, and Signal Processing, pp. I 81-84, May, 1994.

[6] R. Gray, “Vector Quantization”, IEEE Transactions onAcoustics, Speech, and Signal Processing 1(2):4-29, April, 1984.

[7] X. Huang, Y. Ariki, and M. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, U.K., 1990.

[8] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, R. Rosenfeld, “The SPHINX-II Speech Recognition System: An Overview”, Computer Speech and Language, vol. 2, pp. 137-148, 1993.

[9] M. Hwang, “Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition”, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Dec., 1993.

[10] K. Lee, “Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System”, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, April, 1988.

[11] Y. Linde, A. Buzo, R. Gray, “An Algorithm for Vector Quantization Design”, IEEE Transactions on Communication COM-28(1): 84-95, Jan., 1980.

[12] P. J. Moreno, M. A. Siegler, U. Jain, R. M. Stern, " Continuous Recognition of Large-Vocabulary Telephone-Quality Speech", Proceedings of the Spoken Language Systems Technology Workshop, pp 70-73, Jan., 1995.

[13] D. Paul, and J. Baker, "The Design of the Wall Street Journal-based CSR Corpus", Proceedings of ARPA Speech and Natural Language Workshop, pp. 357-362, Feb., 1992.

[14] L. Rabiner, and B. Juang, Fundamentals of Speech Recognition, Prentice-Hall International, New Jersey, U.S.A., 1993.

[15] R. Stern, F. Liu, Y. Ohshima, T. Sullivan, and A. Acero, "Multiple Approaches to Robust Speech Recognition", Proceedings of DARPA Speech and Natural Language Workshop, pp. 274-279, Feb., 1992.