

LINEAR TRANSFORMS IN AUTOMATIC
SPEECH RECOGNITION: ESTIMATION
PROCEDURES AND INTEGRATION OF
DIVERSE ACOUSTIC DATA

Stavros Tsakalidis

A dissertation submitted to the Johns Hopkins University in conformity with the
requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

2005

Copyright © 2005 by Stavros Tsakalidis,
All rights reserved.

Abstract

Linear transforms have been used extensively for both training and adaptation of Hidden Markov Model (HMM) based automatic speech recognition (ASR) systems. Two important applications of linear transforms in acoustic modeling are the decorrelation of the feature vector and the constrained adaptation of the acoustic models to the speaker, the channel, and the task.

Our focus in the first part of this talk is the development of training methods based on the Maximum Mutual Information (MMI) and the Maximum A Posteriori (MAP) criterion that estimate the parameters of the linear transforms. We integrate the discriminative linear transforms into the MMI estimation of the HMM parameters in an attempt to capture the correlation between the feature vector components. The transforms obtained under the MMI criterion are termed Discriminative Likelihood Linear Transforms (DLLT). Experimental results show that DLLT provides a discriminative estimation framework for feature normalization in HMM training for large vocabulary continuous speech recognition tasks that outperforms its Maximum Likelihood counterpart. Then, we propose a structural MAP estimation framework for feature-space transforms. Specifically, we formulate, based on MAP estimation, a Bayesian counterpart of the Maximum Likelihood Linear Transforms (MLLT). Prior density estimation issues are addressed by the use of a hierarchical tree structure in the transform parameter space.

In the second part we investigate the use of heterogeneous data sources for acoustic training. We propose an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data. The approach is an application of model-based acoustic normalization techniques to map the out-of-domain feature space onto the in-domain data. A larger in-domain training set is created by effectively transforming the out-of-domain data before incorporation in training. We put the cross-corpus normalization procedure into practice by investigating the use of diverse Mandarin speech corpora for building a Mandarin Conversational Telephone Speech ASR system. Performance is measured by improvements on the in-domain test set.

Advisor: Prof. William J. Byrne

Readers: Prof. William J. Byrne and Prof. Paul Sotiriadis

Thesis Committee: Prof. William J. Byrne, Prof. Sanjeev Khudanpur,
Prof. Trac Duy Tran and Prof. Paul Sotiriadis

Acknowledgements

I would like to thank all those people who made this thesis possible and an enjoyable experience for me.

First of all I wish to express my sincere gratitude to William Byrne, who guided this work and helped whenever I was in need.

I am also indebted to Sanjeev Khudanpur and Frederick Jelinek for the opportunity to work at the CLSP.

I am grateful to all the members of CLSP for their support and their comradeship.

Finally, I would like to express my deepest gratitude for the constant support, understanding and love that I received from my my family and friends during the past years.

To my family

Contents

List of Tables	viii
List of Figures	x
1 An Overview of Automatic Speech Recognition (ASR)	1
1.1 The Speech Recognition Problem	1
1.2 Evaluation Measures	2
1.3 A Mathematical formulation for the Decoder	3
1.4 Language Modeling	4
1.5 Acoustic Modeling	5
1.5.1 The Hidden Markov Model	5
1.5.2 Output Distributions	7
1.6 Estimation of the HMM parameters	8
1.6.1 Maximum Likelihood Estimation	9
1.6.2 Maximum Mutual Information Estimation	11
1.6.3 Maximum A Posteriori Estimation	13
1.7 Linear Transformations of the Feature Vector	16
2 Linear Transforms in Hidden Markov Model-Based Automatic Speech Recognition	20
2.1 Acoustic Adaptation	20
2.2 Acoustic Normalization	23
2.3 Correlation Modeling	25
2.3.1 Model-space Schemes	26
2.3.2 Feature-based Schemes	27
2.4 Discussion	28
3 Overview and Objectives	29
4 Discriminative Likelihood Linear Transforms (DLLT)	32
4.1 Previous developments leading to Discriminative Linear Transforms	32
4.2 Discriminative Likelihood Linear Transforms for Acoustic Normalization	33
4.2.1 DLLT Estimation	35
4.2.2 Gaussian Parameter Estimation	38
4.2.3 The DLLT Algorithm	40

4.3	Discussion	41
5	DLLT Performance in Large Vocabulary Conversational Speech Recognition	43
5.1	Validating DLLT	44
5.1.1	System Description	44
5.1.2	Effective DLLT Estimation	45
5.1.3	DLLT Results	46
5.2	DLLT Performance on SWITCHBOARD	49
5.2.1	2002 JHU LVCSR System Description	50
5.2.2	DLLT results on SWITCHBOARD	50
5.3	Summary	52
6	Structural Maximum-A-Posteriori (MAP) Linear Transforms	54
6.1	MLLR Adaptation	55
6.2	MAP Linear Regression	56
6.3	Structural MAP Linear Regression	58
6.4	MAP Feature-Space Transforms	60
6.4.1	MAP Estimation of Feature-Space Transforms	61
6.4.2	Relationship Between MAP and ML Feature-Space Transforms	64
6.5	Structural MAP Feature-Space Transforms	65
6.6	Discussion	67
7	Cross-Corpus Normalization Of Diverse Acoustic Data	69
7.1	Acoustic Training from Heterogeneous Data Sources	70
7.2	Cross-Corpus Normalization	71
7.2.1	Corpus-Normalizing Transform Estimation	73
7.2.2	Gaussian Parameters Estimation	74
7.3	Modelling Speaker Variation within Cross-Corpus Normalization . . .	76
7.3.1	Maximum Likelihood Speaker-to-Corpus Normalization	77
7.3.2	Structural MAP Speaker-to-Corpus Normalization	78
7.4	Summary	79
8	Cross-Corpus Normalization of Mandarin Speech Corpora	82
8.1	Mandarin Speech Corpora Description	82
8.2	ASR System Description	85
8.3	Unnormalized Out-of-Domain Acoustic Data	86
8.4	Cross-Corpus Normalized Out-of-Domain Acoustic Data	87
8.5	Speaker-to-Corpus Normalized Out-of-Domain Acoustic Data	88
8.5.1	Distance Measures Between Model Sets	90
8.6	Speaker Adaptive Training on Normalized Out-of-Domain Acoustic Data	93
8.7	Summary	94

9	Minimum Risk Acoustic Clustering for Acoustic Model Combination	97
9.1	Multilingual Acoustic Modeling	98
9.2	Log-linear Combination of Multiple Information Sources	99
9.2.1	Static combination	99
9.2.2	Dynamic combination	100
9.2.3	Optimization issues	101
9.3	Multilingual Acoustic Model Combination	101
9.3.1	Database description	101
9.3.2	Knowledge based partition	102
9.3.3	Searching for optimal partition of the parameter space	103
9.4	Discussion	105
10	Conclusions and Future Work	108
10.1	Thesis Summary	108
10.2	Suggestions For Future Work	111
A	Measuring the Distance Between Gaussian Densities Based on Kullback-Leibler Divergence	113
	Bibliography	115

List of Tables

5.1	Word Error Rate (%) of systems trained with MLLT and DLLT and tested on the SWBD1 and SWBD2 test sets. The HMM Gaussian parameters are kept fixed at their ML values throughout transform updates.	46
5.2	Word Error Rate (%) of systems trained with MLLT and DLLT and tested on the SWBD1 and SWBD2 test sets for different number of classes. DLLT systems are seeded from well trained MLLT systems, indicated by asterisks.	47
5.3	Word Error Rate (%) of systems trained with DLLT and tested on the SWBD1 and SWBD2 test sets for two different initialization points.	48
5.4	Word Error Rate (%) of systems trained with MLLT+MMIE and DLLT+MMIE is seeded from models found after 6 MLLT iterations.	49
5.5	The value of the CML objective function as a function of the iteration number for the DLLT-467 system. Iteration 0 indicates the MLLT baseline.	49
5.6	Word Error Rate (%) of DLLT system trained from the full SWITCHBOARD data and tested on the SWBD1F, SWBD2F and CELL test sets. Results are reported with unsupervised MLLR speaker adaptation.	51
8.1	Mandarin data sources used in acoustic and language model training.	85
8.2	Weights used for each linearly interpolated language model. The interpolation weights were chosen so as to minimize the perplexity on held-out CallFriend transcriptions.	85
8.3	Character Error Rate (%) of baseline systems trained from various corpus combinations as evaluated on the CF test set. Results are reported with and without unsupervised MLLR speaker adaptation.	86
8.4	Character Error Rate (%) of systems by normalizing out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that a source was included in training with / without normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation.	87

8.5	Character Error Rate (%) of systems by normalizing on the speaker level out-of-domain acoustic training data relative to in-domain data. In the first system the transforms were estimated under the ML criterion; in the second under the MAP criterion. An ‘T’ / ‘I’ indicates that each speaker in the source was included in training with / without normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation.	88
8.6	Character Error Rate (%) of SAT derived systems from unnormalized and normalized out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that a source was included in speaker adaptive training with / without cross-corpus normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation.	93
8.7	Summary of Character Error Rate (%) of systems by normalizing out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that each speaker in the source was included in training with / without normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation. Systems (a)-(c) are described in Section 8.3, system (d) is described in Section 8.4, and systems (e)-(f) in Section 8.5.	95
9.1	Combination of English and Czech acoustic models using different acoustic classification schemes.	102

List of Figures

1.1	Source-channel representation of the speech recognition problem. . . .	2
1.2	A 3 state, left-to-right HMM typical used as a phonetic sub-word model.	6
6.1	Tree-structured SMAPLR algorithm. The adaptation data associated to a node i is denoted (\hat{o}_i, \hat{w}_i) . The corresponding prior density is denoted $p(T_i)$. For each children i of parent j , the prior density $p(T_i)$ is specified by parameters estimated under the posterior distribution $p(T_j \hat{o}_j, \hat{w}_j)$ of the parent. From Siohan et al. [119].	59
7.1	Schematic diagram of the cross-corpus acoustic normalization. Each out-of-domain feature space is transformed before being used in training.	71
7.2	Tree structure for training speakers in out-of-domain corpus. The root node contains all the speakers in the out-of-domain corpus, and the leaf nodes contain each distinct speaker.	79
8.1	Histogram of the amount of data for each speaker in each Mandarin data source used in acoustic model training.	84
8.2	Three-level tree structure for training speakers in out-of-domain corpus c . The root node contains all the speakers in the out-of-domain corpus, the second level divides the speakers by their gender and the leaf nodes contain each distinct speaker.	89
8.3	Average Kullback-Leibler (KL) divergence $D(T_{(MAP)}^k, T_{(ML)}^k)$, as defined in equation (8.1), of every pair of transforms $(T_{(MAP)}^k, T_{(ML)}^k)$ that corresponds to each speaker k in the out-of-domain corpora, plotted against the amount of data available for the estimation of the speaker-dependent transforms $T_k^{(MAP)}$ and $T_k^{(ML)}$. For presentation purposes a logarithmic scale was used for both axes.	92
9.1	The binary tree partition constructed by the automatic partition algorithm. The class weights are shown in each leaf.	106
9.2	Word Error Rate (%) of systems derived with the knowledge based partition and the automatic partition algorithm of Section 9.3.3, and tested on both the training and test data as a function of the number of classes. For the knowledge based partition system each phone model has its own weight (71 classes in total).	107

Chapter 1

An Overview of Automatic Speech Recognition (ASR)

This chapter gives an overview of Automatic Speech Recognition (ASR) systems and introduces the concept of linear transforms along with their elementary properties. We will also introduce terminology that will be used throughout the rest of this thesis.

1.1 The Speech Recognition Problem

The goal of an Automatic Speech Recognition (ASR) system is to transcribe speech to text. The speech recognition problem can be described using the source-channel framework as shown in the schematic diagram of Figure 1.1. Under this framework, a spoken string W of words is generated by the vocal apparatus of the speaker and, subsequently, passes through an acoustic channel. The acoustic channel consists of two parts: the acoustic processor and the linguistic decoder. The acoustic processor (*front end*) uses signal processing techniques to extract important features from the speech signal and transform it into a sequence of real-valued vectors O (referred as feature or observation vectors). The task of the linguistic decoder is to recover the original word sequence W from the noisy feature sequence O . In order to do this, the decoder uses the feature vectors and a stochastic model of the acoustic channel and outputs a word string W^* .

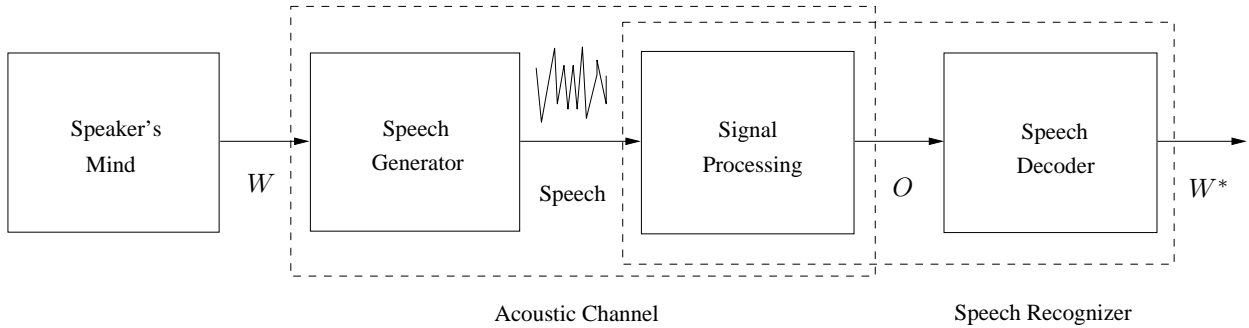


Figure 1.1: Source-channel representation of the speech recognition problem.

1.2 Evaluation Measures

To measure the accuracy of the speech recognizer we compare the true transcription W , which is obtained manually from humans, to the output of the recognizer \hat{W} (referred as *hypothesis*). This comparison is usually performed on the sentence level using the Sentence Error Rate (SER) metric or on the word level using the Word Error Rate (WER) metric.

The SER is defined as the percentage of correctly transcribed sentences. The accuracy of a speech recognizer at the sentence level may not be a good indicator of performance. This is because the SER metric yields the same value without taking into account whether there is one or more incorrectly hypothesized words in a sentence between the true transcription and the hypothesis. The WER is defined as the minimal cost involved to transform the true word string into the hypothesis by using three elementary operations: deletion, insertion and substitution of a word. This measure which is also known as the Levenstein edit distance [77] can be efficiently calculated using a dynamic programming algorithm [109]. The WER is more informative than the SER since it allows to identify specific ASR errors, i.e. the most frequently confusable words.

Part of this thesis involves the development of Chinese (Mandarin) ASR systems. It is well known that Chinese text is written as a string of ideograms with no specific word delimiter in contrast to languages like English. There is no standard definition of words in Chinese; words are identified based on the context in which they appear. Thus, any given definite Chinese syllable string could correspond to many Chinese word strings due to the uncertainty of the Chinese word boundaries [139].

This distinguishing feature of the Chinese language and a few other Asian languages demands a departure from the conventional WER metric. Therefore, Chinese ASR transcriptions are usually evaluated under the Character Error Rate (CER) rather than the WER. Its is defined similarly to the WER but the transcriptions are compared at the Chinese character level.

1.3 A Mathematical formulation for the Decoder

The goal of the decoder is to output a word sequence \hat{W} that matches best with the speech produced by the speaker. Let $P(W|O)$ denote the posterior probability of the word sequence W , given the acoustic evidence O . The Maximum a Posteriori (MAP) decoder chooses the word sequence W^* with the highest posterior probability [8]:

$$W^* = \operatorname{argmax}_{W \in \mathcal{W}} P(W|O) \quad (1.1)$$

where \mathcal{W} represents all possible word strings. The MAP criterion is the optimum decoding criterion when performance is measured under the Sentence Error Rate criterion. This is the rule we will use for recognition for the purposes of this thesis.

Using Bayes' rule, the posterior probability $P(W|O)$ can be written as

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \quad (1.2)$$

where $P(O|W)$ is the probability that the acoustic evidence O was produced from the word sequence W ; $P(W)$ is the prior probability of the word sequence; and $P(O) = \sum_{W'} P(O|W')P(W')$ is the average probability that O will be observed.

For a known sequence of observations, the marginal distribution $P(O)$ is constant and therefore does not affect the MAP criterion of equation (1.1). It follows that the decoder decides in favor of the word string \hat{W} that satisfies

$$W^* = \operatorname{argmax}_{W \in \mathcal{W}} P(O|W)P(W) \quad (1.3)$$

To compute the product $P(O|W)P(W)$ we employ stochastic models of the acous-

tic and linguistic properties of speech. Hence, the values $P(O|W)$ and $P(W)$ are provided from the *acoustic model* and the *language model* respectively. In the next sections we will briefly introduce the commonly used *n-gram* language model and describe the usual acoustic model employed in state-of-the-art ASR systems.

1.4 Language Modeling

The function of the language model is to assign a probability $P(W)$ to a given sequence of words $W = w_1^n = w_1, \dots, w_n$. The language model guides and reduces the search space of the best word sequence and improves ASR performance by providing contextual information.

Using the chain rule, $P(w_1^n)$ can be decomposed to

$$P(w_1^n) = \prod_{i=1}^n P(w_i | w_1, \dots, w_{i-1}) \quad (1.4)$$

where $P(w_i | w_1, \dots, w_{i-1})$ is the probability of w_i to follow a given word string w_1, \dots, w_{i-1} . The past w_1, \dots, w_{i-1} is usually referred to as history.

ASR systems usually employ an *n-gram* language model. This model makes the assumption that only the previous $n - 1$ words have any effect on the probabilities for the next word. A typical value for n is three (hence the term *trigram model*). Under the trigram model equation (1.4) becomes

$$P(w_1^n) = P(w_1)P(w_2|w_1) \prod_{i=1}^n P(w_i | w_{i-2}, \dots, w_{i-1}) \quad (1.5)$$

where the first two terms on the right-hand side of equation (1.5) are called *unigram* and *bigram*, respectively.

Regardless of the type of the language model employed, it is important that any language model should be trained on data that are closely related to the task to which the ASR system will be applied.

1.5 Acoustic Modeling

Speech is a non-stationary process but, it is assumed to be 'quasi-stationary', meaning that over a short period of time the statistical properties of speech do not change from sample to sample. Hence, the task of acoustic modeling is to provide a stochastic model that captures both the acoustic and temporal characteristics of speech. Hidden Markov models (HMMs) [13, 12] have proven to be well suited to this task [8, 10, 11, 78, 104]. We will introduce the concept of HMM and we will show how HMMs are used for acoustic modeling.

1.5.1 The Hidden Markov Model

The HMM is a stochastic finite state machine, in which each state is associated with an output (emission) probability distribution. The HMM changes state once every time unit according to a probabilistic transition function and each time τ a state i is entered, an output symbol is generated from the output probability distribution. Formally, an HMM is defined by

- A set of states $\mathcal{S} = \{s_1, \dots, s_N\}$
- A probability distribution of transitions between states $p(s_i|s_j)$, $1 \leq i, j \leq N$
- The number of observation symbols in the alphabet, M . If the observations are continuous then M is infinite
- An output probability distribution $q(\cdot|s_i)$ associated with each state s_i , $1 \leq i \leq N$

Hidden Markov modeling assumes that the sequence of feature vectors is a piecewise stationary process for which each stationary segment is associated with a specific HMM state. The set of state output processes models the (locally) stationary feature vectors. Furthermore, since transitions among states are governed by the set of transition probabilities the HMM can model varying temporal sequences stochastically.

When using HMMs to model speech we assume that the feature vectors follow a first order Markov process, i.e. the distribution of the vectors depends only on the state and not on other vectors. Additionally, the transition probability of the

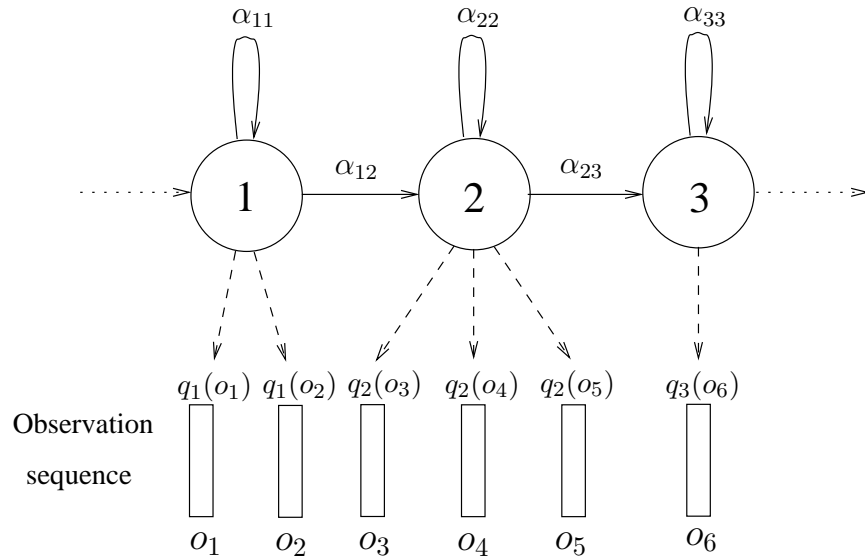


Figure 1.2: A 3 state, left-to-right HMM typical used as a phonetic sub-word model.

system to the next state depends only on the current state regardless of the earlier transition history. The sequence of the feature vectors is the observed output of the HMM. The state sequence is not observed, it is hidden.

Depending on the task, an HMM can be used to model speech at several linguistic levels e.g. words, syllables, phonemes, or phonemes with context information. The standard choice for large vocabulary ASR systems is the *triphones*, which are phonemes in right and left context. Context-dependent phonetic models have the advantage of capturing coarticulation effects between adjacent speech units. Each triphone is associated with a Markov model made up from states from \mathcal{S} according to a predefined topology. Figure 1.2 is an typical example of a triphone (phonetic sub-word) HMM, as it is used in this thesis. The topology is strictly left-to-right. Each state is allowed to take transitions to itself (loop) or to the following one (forward). The HMM for a word is then obtained by concatenation of the phonetic sub-word HMMs which form the word.

Let $\mathcal{M} = \{m_1, \dots, m_U\}$ represent the set of possible elementary phonetic sub-word HMMs and $\theta = \{\theta_1, \dots, \theta_U\}$ the set of associated parameters. In the following, M_w will represent the HMM associated with a specific word w and obtained by concatenating elementary HMMs of \mathcal{M} associated with the triphone speech units constituting w .

We can now give an example of how the sequence of feature vectors is generated by the HMM. According to Figure 1.2, the model traverses through the state sequence $S = 1, 1, 2, 2, 2, 3$ in order to generate the observation sequence o_1 to o_6 . Observations $\{o_1, o_2\}$ are generated by state $s = 1$, observations $\{o_3, o_4, o_5\}$ by state $s = 2$ and $\{o_6\}$ by state $s = 3$.

Given that the state sequence s_1^l is hidden (unknown), the probability of observing an HMM output sequence of feature vectors o_1^l , given some model M , is computed by summing over all possible state sequences. That is

$$P(o_1^l|M) = \sum_{s_1^l} P(o_1^l, s_1^l|M) = \sum_{s_1^l} \prod_{\tau=1}^l p(s_\tau|s_{\tau-1})p(o_\tau|s_\tau) \quad (1.6)$$

1.5.2 Output Distributions

In general, the state output distributions can be either discrete or continuous. Discrete density HMMs allow only discrete observations of a fixed codebook. The output distribution in each emitting state consists of a separate discrete probability $b_i(k)$ for each observation symbol c_k . To make this possible, the continuously valued feature vectors must first be (vector) quantized (VQ) to the nearest codeword in the codebook. One advantage of such representation is that discrete distributions do not assume any specific form of the density functions and therefore any arbitrary distribution in the feature space can be approximated. At the same time the calculation of the observation probabilities in the HMM states reduces to simple table look-ups, which makes the calculation of $P(O|W)$ computationally efficient. Unfortunately, the discretization of continuously valued vectors introduces quantization errors. These modelling inaccuracies, inherited in the VQ, degrade the recognition accuracy of the system.

More sophisticated methods of VQ can alleviate such loss [9, 31]. These quantization methods allow the development of observation distributions for HMMs that combine subvector quantization and mixtures of discrete distributions [32]. The discrete-mixture HMMs were found to perform at least as well as continuous-mixture-density HMMs at much faster decoding speeds.

Continuous density HMMs use parametric distributions of a predetermined form. Most ASR systems employ weighted sums (mixtures) of multivariate Gaussian densities

$$q(o|s; \{c_{s,m}, \mu_{s,m}, \Sigma_{s,m}\}) = \sum_{m=1}^M c_{s,m} \mathcal{N}(o; \mu_{s,m}, \Sigma_{s,m}) \quad (1.7)$$

$$= \sum_{m=1}^M c_{s,m} \frac{1}{\sqrt{(2\pi)^d |\Sigma_{s,m}|}} e^{-\frac{1}{2}(o-\mu_{s,m})^T \Sigma_{s,m}^{-1} (o-\mu_{s,m})}. \quad (1.8)$$

where M is the number of mixture components and $c_{s,m}$ is the mixture weight for the m^{th} mixture component in state s ; $\mu_{s,m}$ and $\Sigma_{s,m}$ are the mean vector and covariance matrix of the m^{th} Gaussian density at state s . To reduce the number of free parameters, the components of the feature vector are assumed to be uncorrelated, that is the covariance matrix is assumed to be diagonal. The mixture weights must satisfy

$$\sum_{m=1}^M c_{s,m} = 1, \quad 0 \leq c_{s,m} \leq 1 \quad (1.9)$$

In theory, mixtures of Gaussian densities can assume arbitrary shapes. Therefore, the use of Gaussian mixture densities allows the modeling of data with non-Gaussian nature and implicitly of the correlations between the components of the feature vector.

Usually, an incremental approach is followed to acoustic model building based on the idea of gradually increasing model resolution. The HMMs are initialized by a single component density per state and, subsequently, the number of mixture components in each state is increased by a succession of reestimation and mixture splitting.

1.6 Estimation of the HMM parameters

The parameters of the models must be estimated before the HMMs can be used for recognition. The parameters to be estimated in an HMM-based ASR system are the state transition probabilities and the parameters related to the emission probabilities. The most popular estimation techniques for ASR systems are the Maximum Likelihood (ML), Maximum Mutual Information (MMI) and Maximum A Posteriori (MAP) criteria. The following sections give a brief overview of these estimation criteria and discuss their strengths and weaknesses.

1.6.1 Maximum Likelihood Estimation

Maximum Likelihood estimation attempts to find the HMM parameter set θ^* which maximizes the likelihood of the observed acoustic data $\hat{o}_1^{\hat{l}}$ given the model $M_{\hat{w}_1^{\hat{n}}}$ corresponding to the correct transcription $\hat{w}_1^{\hat{n}}$

$$\theta^* = \operatorname{argmax}_{\theta} P(\hat{o}_1^{\hat{l}} | M_{\hat{w}_1^{\hat{n}}}; \theta) \quad (1.10)$$

Let \mathcal{S} be the set of all possible \hat{l} -length state sequences (paths) that can represent $\hat{w}_1^{\hat{n}}$, i.e. paths that begin at the initial state, end at the final state, and traverse exactly \hat{l} arcs, resulting in exactly \hat{l} outputs. Then equation (1.10) can be written as

$$\begin{aligned} \theta^* &= \operatorname{argmax}_{\theta} P(\hat{o}_1^{\hat{l}} | M_{\hat{w}_1^{\hat{n}}}; \theta) \\ &= \operatorname{argmax}_{\theta} \sum_{s_1^{\hat{l}} \in \mathcal{S}} P(\hat{o}_1^{\hat{l}}, s_1^{\hat{l}} | M_{\hat{w}_1^{\hat{n}}}; \theta) \end{aligned} \quad (1.11)$$

In general, it is not possible to estimate θ^* directly from equation (1.11). However, an iterative procedure known as the Expectation-Maximization (EM) algorithm may be used to locally maximize the likelihoods in equation 1.11. The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when data are incomplete or have missing values. It is assumed that the estimation problem is to find a parameterized joint probability density on a *complete* random variable X , based on observations of an *incomplete* random variable $Y = y(X)$, where $y(\cdot)$ is a many-to-one mapping. This algorithm exploits the fact that the complete-data likelihood is analytically easier to maximize than the likelihood of the incomplete data.

The estimation of the parameters of an HMM, as used in acoustic modeling, is an example of estimation from incomplete data. In this case, the complete random variable X consists of the arbitrary length word sequence W_1^N , and the observation sequence O_1^L and state sequence S_1^L of equal but arbitrary length. The incomplete random variable is the word sequence and the observation sequence. The pair $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$ denotes observed values of these random variables, i.e. the training data.

Thus, we can observe W_1^N and O_1^L but, not the corresponding state sequence S_1^L ; the state sequence is *hidden*.

The EM algorithm is a two-step iterative procedure. The first step, called the expectation (E step), computes the *auxiliary function* $Q(\bar{\theta}, \theta)$ defined as the expected value of the complete-data log-likelihood $\log p(O, S, W; \theta)$ with respect to the hidden state sequence given the observed data and the current parameter estimates. We define

$$Q(\bar{\theta}, \theta) = E[\log p(\hat{o}_1^i, s_1^i, M_{\hat{w}_1^{\hat{n}}}; \bar{\theta}) | \hat{o}_1^i, M_{\hat{w}_1^{\hat{n}}}; \theta] \quad (1.12)$$

$$= \sum_{s_1^i \in \mathcal{S}} p(s_1^i | \hat{o}_1^i, M_{\hat{w}_1^{\hat{n}}}; \theta) \log p(\hat{o}_1^i, s_1^i, M_{\hat{w}_1^{\hat{n}}}; \bar{\theta}). \quad (1.13)$$

Here, θ are the current parameter estimates that we used to evaluate the expectation and $\bar{\theta}$ are the new parameters that we optimize to increase $Q(\bar{\theta}, \theta)$.

The second step, called the maximization step (M step), maximizes the expectation computed in the first step, i.e.

$$\theta^* = \operatorname{argmax}_{\bar{\theta}} Q(\bar{\theta}, \theta) \quad (1.14)$$

These two steps are repeated as necessary. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function. That is, if a value $\bar{\theta}$ satisfies $Q(\bar{\theta}, \theta) > Q(\theta, \theta)$, then $P(O|M_W; \bar{\theta}) > P(O|M_W; \theta)$.

The auxiliary function $Q(\bar{\theta}, \theta)$ can be simplified by introducing the HMM conditional independence assumptions, mentioned in Section 1.5.1, as

$$Q(\bar{\theta}, \theta) = \sum_{s \in \mathcal{S}} \sum_{\tau=1}^{\hat{l}} \gamma_s(\tau; \theta) (\log \bar{\alpha}_{s_\tau s_{\tau+1}} + \log q(o_\tau | s_\tau; \bar{\mu}_s, \bar{\Sigma}_s)) \quad (1.15)$$

Here, $a_{s_\tau s_{\tau+1}} \triangleq p(s_{\tau+1} | s_\tau)$ denotes the Markov transition probability from state s_τ to state $s_{\tau+1}$ and $\gamma_s(\tau; \theta) \triangleq q_{s_\tau}(s | M_{\hat{w}_1^{\hat{n}}}, \hat{o}_1^i; \theta)$ denotes the conditional occupancy probability of being in state s at time τ given the training acoustics \hat{o}_1^i and the model $M_{\hat{w}_1^{\hat{n}}}$ corresponding to the transcription $\hat{w}_1^{\hat{n}}$. The probability $\gamma_s(\tau; \theta)$ is computed using the *forward-backward* algorithm [104].

The ML criterion is the commonest objective criterion used for HMM training. Its wide-spread adoption is a result of empirically good performance and the availability of the fast, globally convergent EM training algorithm. Moreover, ML estimators exhibit desirable statistical properties such as sufficiency, consistency and efficiency [42, 76, 130].

However, there are two fundamental limitations with the MLE approach. First, one of the most commonly cited problems is the violation of the model correctness assumption. It is well known that an HMM is not the true generator for speech; in fact the true generator is unknown. Thus, parameterized models obtained via MLE can be employed optimally for detection and classification if the data encountered are generated by some distribution from the model family. The problem arises due to the various conditional independence assumptions that underlie HMM models, as we have already mentioned. Given these assumptions, it is unlikely that the processes that actually generate speech can be closely modelled by HMMs. Second, by inspection of the ML criterion (equation (1.10)) we can see that MLE is only concerned with maximizing the a posteriori probability of the training data given the model corresponding to the data. The models from other classes do not participate in the parameter re-estimation. Therefore, under MLE each HMM is trained only to generate high probabilities for its own class, without discriminating against rival models. Consequently, the MLE objective function is not directly related to the objective of reducing the error rate.

As an alternative to relying on the asymptotic behavior of ML estimation under the model correctness assumption, there are discriminative estimation procedures that directly attempt to optimize recognition performance. In general, discriminative criteria not only try to maximize the class conditional probability of the acoustic evidence given the correct classes but, also try to minimize the class conditional probabilities of the corresponding competing classes. In the next section we briefly describe one of the most popular discriminative criterion, i.e. the Maximum Mutual Information criterion

1.6.2 Maximum Mutual Information Estimation

Maximum Mutual Information estimation [7, 99, 134] attempts to find acoustic model parameters θ that maximize the mutual information $I(W, O; \theta)$ between the

training word sequence and the corresponding observation sequence

$$I(W, O; \theta) = \frac{P(O, W; \theta)}{P(O; \theta)P(W)} \quad (1.16)$$

If the language model is assumed independent of the acoustic model parameters θ , maximizing the MMI criterion is equivalent to maximizing the Conditional Maximum Likelihood (CML) criterion of Nadas [93, 94]

$$P(W|O; \theta) = \frac{P(O, W; \theta)}{P(O; \theta)} = \frac{P(O|W; \theta)P(W)}{\sum_{W'} P(O|W'; \theta)P(W')} \quad (1.17)$$

Therefore, for the remaining part of this thesis the terms MMI and CML will be used interchangeably.

The CML criterion tries to improve the a posteriori probability of the correct sentence hypothesis given the acoustic evidence. Since this is also the probability used in MAP decoding (equation (1.1)), the CML objective is directly related to reducing the Sentence Error Rate on the acoustic training set.

Equation (1.16) shows the difference between MMIE and MLE. Maximization of the MMIE criterion over \mathcal{W} , which is usually taken to be the set of all word strings allowed in the language, leads to discriminant models since it implies that the contribution of the numerator (correct word sequence) should be enhanced while the contribution of the denominator (all possible alternative models) should be reduced. Therefore, under MMIE the parameters of all HMMs are adjusted simultaneously. In contrast, under MLE the HMMs are trained independently of each other. The computation and processing of the alternate word hypothesis defining the competing models in the denominator is the most computationally expensive step in MMI training. In practice, the space \mathcal{W} is approximated either by N-best lists [22] or lattices [134]. The N-best lists or the lattices are generated via a recognition pass on each of the training utterances.

The MMI criterion has a solid theoretic foundation derived from an information theoretic approach to classifier design. However, the successful application of MMIE - and in general any other estimation procedure - to HMMs depends on the availability of a fast and effective training algorithm. Initially, MMI training made use of gradient descent optimization that converged slowly and gave little improvement [7].

Later, the *extended* Baum-Welch algorithm [52], an extension to the standard Baum-Welch algorithm [12], for optimizing rational functions was introduced. The *extended* Baum-Welch algorithm was first developed for discriminative training of HMMs with discrete output distributions. Then, it was extended to the optimization of HMMs with continuous Gaussian densities [98]. The idea was to create a discrete approximation of the Gaussian density so that the *extended* Baum-Welch algorithm can be applied. Here, we use a simplified derivation of the CML estimation algorithm that does not require discretization of the continuous observation process.

The Conditional Maximum Likelihood Algorithm

Gunawardana’s derivation of the extended Baum-Welch procedure for CML estimation of HMMs [56] provides an auxiliary function for discriminative model estimation that is analogous to the EM auxiliary function used to derive Maximum Likelihood modeling procedures. It does not require discrete density approximations and extends the extended Baum-Welch algorithm to arbitrary continuous density HMMs with arbitrary parameterizations. This is particularly useful for the purposes of this thesis since we are interested in discriminative estimation procedures of Gaussian observation distributions with constrained parameterizations.

Gunawardana’s iterative CML algorithm chooses $\theta^{(p+1)}$ given a parameter iterate $\theta^{(p)}$ according to

$$\theta^{(p+1)} \in \operatorname{argmax}_{\theta \in \Theta} \sum_{s_1^{\hat{i}}} \left[q(s_1^{\hat{i}} | M_{\hat{w}_1^{\hat{n}}}, \hat{o}_1^{\hat{i}}; \theta^{(p)}) - q(s_1^{\hat{i}} | \hat{o}_1^{\hat{i}}; \theta^{(p)}) \right] \log q(\hat{o}_1^{\hat{i}} | s_1^{\hat{i}}; \theta) + \sum_{s_1^{\hat{i}}} d'(s_1^{\hat{i}}) \int q(o_1^{\hat{i}} | s_1^{\hat{i}}; \theta^{(p)}) \log q(o_1^{\hat{i}} | s_1^{\hat{i}}; \theta) do_1^{\hat{i}} \quad (1.18)$$

Here, the pair $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{i}})$ denotes observed values of the observation vector sequence O and corresponding word sequence W random variables, i.e. the training data. $d'(s_1^{\hat{i}})$ leads to the MMI constant as $D_s = \sum_{s_1^{\hat{i}}:s_{\tau}=s} d'(s_1^{\hat{i}})$.

1.6.3 Maximum A Posteriori Estimation

One of the most commonly used Bayesian learning criterion is the maximum a posteriori (MAP) criterion. The MAP estimate can be seen as a Bayes estimate when

the loss function is not specified [28]. The MAP estimation framework provides a way of incorporating prior knowledge about the model parameters into the training process. Specifically, in MAP the acoustic model parameters to be estimated θ are assumed to be random variables whose range is contained in Θ , and therefore are described by their probability density function $p(\theta)$, called prior density. The MAP estimate is defined as the mode of the posterior density of θ , that is

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} p(\theta | \hat{\delta}_1^l, \hat{w}_1^{\hat{n}}) \quad (1.19)$$

Using Bayes' rule, we rewrite $p(\theta | \hat{\delta}_1^l, \hat{w}_1^{\hat{n}})$ as

$$\begin{aligned} p(\theta | \hat{\delta}_1^l, \hat{w}_1^{\hat{n}}) &= \frac{p(\theta)p(\hat{\delta}_1^l, \hat{w}_1^{\hat{n}} | \theta)}{p(\hat{\delta}_1^l, \hat{w}_1^{\hat{n}})} \\ &= \frac{p(\theta)p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}}; \theta)p(\hat{w}_1^{\hat{n}} | \theta)}{p(\hat{\delta}_1^l, \hat{w}_1^{\hat{n}})} \end{aligned} \quad (1.20)$$

Since the language model is independent of the acoustic model parameters θ , equation (1.20) becomes

$$\begin{aligned} p(\theta | \hat{\delta}_1^l, \hat{w}_1^{\hat{n}}) &= \frac{p(\theta)p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}}; \theta)p(\hat{w}_1^{\hat{n}} | \theta)}{p(\hat{\delta}_1^l, \hat{w}_1^{\hat{n}})} \\ &= \frac{p(\theta)p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}}; \theta)p(\hat{w}_1^{\hat{n}})}{p(\hat{\delta}_1^l, \hat{w}_1^{\hat{n}})} \\ &= \frac{p(\theta)p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}}; \theta)}{p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}})} \end{aligned}$$

Given the above expression for $p(\theta | \hat{\delta}_1^l, \hat{w}_1^{\hat{n}})$, we can see that maximizing equation (1.19) with respect to θ is the same as solving (1.21) since $p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}})$ is constant with respect to θ .

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} p(\theta)p(\hat{\delta}_1^l | \hat{w}_1^{\hat{n}}; \theta) \quad (1.21)$$

Hence, the MAP objective function is given by

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} p(\theta)p(\hat{o}_1^l | M_{\hat{w}_1^{\hat{n}}}; \theta) \quad (1.22)$$

where $M_{\hat{w}_1^{\hat{n}}}$ is the composite HMM model corresponding to the word sequence $\hat{w}_1^{\hat{n}}$ (see Section 1.5.1) and $p(\hat{o}_1^l | M_{\hat{w}_1^{\hat{n}}}; \theta)$ is the likelihood function of the observed acoustic data given the model corresponding to the transcription.

As we previously discussed, in Section 1.6.1, the likelihood function $p(\hat{o}_1^l | M_{\hat{w}_1^{\hat{n}}}; \theta)$ in equation (1.22) is the function to be maximized under the ML criterion (see equation (1.10)). By inspection of the MAP criterion (equation (1.22)) and ML criterion (equation (1.10)) we can see that the MAP objective function has an extra term, i.e. the prior distribution $p(\theta)$ of the model parameters θ . Hence, the difference between the ML and MAP estimation procedure lies in the assumption of an appropriate prior distribution of the parameters θ . If θ is assumed to be fixed but unknown, which is equivalent to assuming a *noninformative* or *improper* prior [79], then equation (1.22) reduces to the corresponding ML equation (1.10).

The MAP estimate can also be formulated in an EM algorithm as described by Dempster et al. [29]. Thus, an EM approach can be used to estimate the mode of the posterior density by maximizing the following auxiliary function

$$R(\bar{\theta}, \theta) = Q(\bar{\theta}, \theta) + \log p(\bar{\theta}) \quad (1.23)$$

where $Q(\bar{\theta}, \theta)$ is the conventional ML auxiliary function (1.13). It follows that maximizing $R(\bar{\theta}, \theta)$ leads to improvements in the a posteriori probability density function of θ . The prior density adds a regularization term to the reestimation equation, and penalizes estimates that deviate from the prior.

An interesting property of MAP estimation is its asymptotic behavior: as the amount of adaptation data increases, the likelihood function becomes dominant whereas the contribution of the prior becomes negligible. Thus, as long as the initial MAP and ML estimates are identical, the MAP estimate converges to the same point as ML when the amount of training data approaches infinity.

The major shortcoming of MAP estimation is that it is strongly related to the choice of the prior distribution family. The common practice is to use *conjugate* priors [49] so that the posterior distribution has the same functional form as the

prior. In this way, the problem is reduced to estimate the parameters in a known form of density. However, it is not always possible to find a conjugate prior density.

Furthermore, additional parameters (referred as hyperparameters) are needed to describe the prior distribution. In a strict Bayesian approach, these additional parameters are assumed known, based on common or subjective knowledge about the stochastic process. Because in most cases the hyperparameters cannot be derived from such knowledge, an alternative solution is to adopt an empirical Bayes approach [106] in which the hyperparameters are estimated directly from the data.

Therefore, the use of priors in MAP makes the parameter estimation more complex than ML. On the other hand, MAP estimation is particularly useful when training data are sparse because, in this case, the ML approach can give inaccurate estimates. Hence, if we know what the parameters of the model are likely to be - before observing any data - we might get a decent estimate given only a limited amount of training data.

1.7 Linear Transformations of the Feature Vector

As we mentioned in Section 1.1, the observation (feature) vectors, which are the output of the acoustic processor, contain important features from the speech signal. Linear transformations of the feature vectors is the main topic of this thesis. In this last section we will review the properties of linear transformations and present their elementary properties, in order to serve as a reference for later chapters in this thesis.

Let X and Y be d -dimensional real-valued random vectors. A function $g(\cdot)$ is called an *affine transformation* of \mathbb{R}^d if there exists a $d \times d$ matrix A and a d -dimensional vector b such that

$$g(X) = Y = AX + b. \quad (1.24)$$

If $b = 0$, $g(\cdot)$ is called a *linear transformation*. The function $g(\cdot)$ is one-to-one if, and only if, A is nonsingular and then,

$$g^{-1}(Y) = A^{-1}(y - b), \quad (1.25)$$

where A_{-1} is the inverse of A .

Now, let $\mu_X = E[X]$ denote the expected vector of X and $\Sigma_X = E[(X - \mu_X)(X - \mu_X)^T]$ the covariance matrix of X . Then, the expected vector and covariance matrix of Y are

$$\begin{aligned}\mu_Y &= E[Y] = AE[X] + b = A\mu_X + b, \\ \Sigma_Y &= E[(Y - \mu_Y)(Y - \mu_Y)^T] \\ &= E[(AX + b - A\mu_X - b)(AX + b - A\mu_X - b)^T] \\ &= AE[(X - \mu_X)(X - \mu_X)^T]A^T \\ &= A\Sigma_X A^T\end{aligned}$$

The feature vectors in the HMM are modelled by the state dependent Gaussian densities, that is we assume that $X \sim \mathcal{N}(\cdot; \mu_X, \Sigma_X)$. Assume that we apply a linear transform to the feature vector X , given by equation (1.24). Then, the density of the random vector $Y = AX + b$ is given by the following theorem [3].

Theorem 1 *Let X be a d -dimensional normal random vector with mean vector μ_X and positive definite covariance matrix Σ_X . Let A be a nonsingular $d \times d$ matrix and b a d -dimensional vector. Then $Y = AX + b$ is a d -dimensional normal random vector with mean vector $\mu_Y = A\mu_X$ and covariance matrix $\Sigma_Y = A\Sigma_X A^T$.*

Proof The density of Y is given by [15]

$$p_Y(y) = \frac{p_X(g^{-1}(y))}{|J_g(g^{-1}(y))|} \quad (1.26)$$

where $J_g(g^{-1}(y))$ is the *Jacobian* of g evaluated at $g^{-1}(y)$, that is $J_g(g^{-1}(y)) = |A|$.

Hence

$$p_Y(y) = |A|^{-1} \frac{1}{\sqrt{(2\pi)^d |\Sigma_X|}} \exp \left(-\frac{1}{2} (A^{-1}(y-b) - \mu_X)^T \Sigma_X^{-1} (A^{-1}(y-b) - \mu_X) \right) \quad (1.27)$$

$$= \frac{1}{\sqrt{(2\pi)^d |A| |\Sigma_X|^{1/2}}} \exp \left(-\frac{1}{2} (y - (A\mu_X + b))^T (A^{-T} \Sigma_X^{-1} A^{-1}) (y - (A\mu_X + b)) \right) \quad (1.28)$$

$$= \frac{1}{\sqrt{(2\pi)^d |A \Sigma_X A^T|^{1/2}}} \exp \left(-\frac{1}{2} (y - (A\mu_X + b))^T (A \Sigma_X A^T)^{-1} (y - (A\mu_X + b)) \right) \quad (1.29)$$

But $A\mu_X + b = \mu_Y$ and $A\Sigma_X A^T = \Sigma_Y$. Hence, equation can be rewritten as

$$p_Y(y) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_Y|}} \exp \left(-\frac{1}{2} (y - \mu_Y)^T \Sigma_Y^{-1} (y - \mu_Y) \right). \quad (1.30)$$

■

The modelling techniques developed in this thesis apply affine transforms to the observation vector X so that a normalized feature vector is found by equation (1.24). A common approach is to tie the transforms across sets of Gaussians in order to reduce the number of parameters to be estimated from the training data. Therefore, we assume that we group the HMM states into R classes. Then each affine transform $(A_{\mathcal{R}(s)}, b_{\mathcal{R}(s)})$ is associated with a group of states $S_r = \{s | \mathcal{R}(s) = r\}$ for classes $r = 1, \dots, R$. Let $T_{\mathcal{R}(s)}$ to denote the extended transformation matrix

$$T_{\mathcal{R}(s)} \triangleq [b_{\mathcal{R}(s)} \ A_{\mathcal{R}(s)}] \quad (1.31)$$

and ζ the extended observation vector

$$\zeta \triangleq [1 \ X^T]^T \quad (1.32)$$

The emission density of the HMM state s is assumed to be Gaussian. Using equation (1.26), we can see that the Gaussian densities are reparameterized as

$$q(\zeta|s; \theta) = \frac{|A_{\mathcal{R}(s)}|}{\sqrt{(2\pi)^m |\Sigma_s|}} e^{-\frac{1}{2}(T_{\mathcal{R}(s)}\zeta - \mu_s)^T \Sigma_s^{-1} (T_{\mathcal{R}(s)}\zeta - \mu_s)}. \quad (1.33)$$

Here, μ_s and Σ_s are the mean and variance for the observation distribution of state s . The reparametrization of the emission density augments the usual set of HMM parameters with the parameters of the transform. The entire parameter set is defined as $\theta = (T_{\mathcal{R}(s)}, \mu_s, \Sigma_s)$.

Given Theorem 1, we can establish the following property of normal vectors. By inspection of equations (1.27) and (1.29), it is easy to confirm that

$$|A| \mathcal{N}(AX + b, \mu_X, \Sigma_X) = \mathcal{N}(X, A^{-1}(\mu_X - b), A^{-1}\Sigma_X A^{-T}) \quad (1.34)$$

That is an affine transformation of a normal vector X is equivalent to an affine transformation of the corresponding Gaussian means and the covariances since the respective Gaussian likelihoods are equal. This property of normal vectors allows Gaussian observation distributions with constrained parameterizations to be implemented as a transformation of the vector and a scaling of the likelihood by the Jacobian term $|A|$.

Chapter 2

Linear Transforms in Hidden Markov Model-Based Automatic Speech Recognition

In the previous chapter we reviewed the basic statistical HMM-based acoustic model for Automatic Speech Recognition (ASR) and the estimation of the acoustic model parameters under the Maximum Likelihood, Maximum Mutual Information, and Maximum A Posteriori criteria. We have also reviewed the use of linear transforms and presented their elementary properties. This chapter provides an overview of the application of linear transforms in HMM-based ASR systems.

Linear transforms have been used extensively for HMM-based acoustic modeling of ASR systems as a powerful tool for increased classification accuracy and improved descriptive power. The application of linear transforms in acoustic modeling can be grouped broadly in three categories: *Acoustic Adaptation*, *Acoustic Normalization* and *Correlation Modeling*.

2.1 Acoustic Adaptation

The parameters of the acoustic models are usually estimated from large amounts of training data collected from many speakers. Then, the estimated models are used to transcribe speech from previously unseen (test) data. ASR systems suffer a degradation in performance when there is a mismatch between training and testing conditions. The mismatch is characterized by a number of contributing factors, such

as

- Language
- Dialect
- Domain or topic
- Acoustic channel and environment
- Sampling rate
- Speaking style
- Speaker age, education and fluency

The goal of *Acoustic Adaptation* is to compensate for the differences between the speech on which the system was trained and the speech which it has to recognize. Transformation-based model adaptation can be summarized as follows: assuming that the estimated models have been trained in a given condition leading to a set of parameters θ , a new set of model parameters θ' is obtained for a new condition with a suitable parameter transformation procedure $\theta' = F(\theta)$ using some training data (called adaptation data) in the new condition. The amount of data used for the estimation of the transformation parameters is usually much lower than the one used for training the initial models. To avoid introducing more parameters than can be reliably estimated, transforms are tied across groups of parameters.

Parameter transformation procedures $F(\cdot)$ can be classified into linear and non-linear transformations. Linear transformation procedures adapt the parameters by means of an affine transformation, as given by formula (1.24), that maps the original space to the space of interest. Nonlinear procedures that perform the same operation with nonlinear mapping [102] are out of the scope of this thesis.

One popular linear transformation-based approach is the maximum likelihood linear regression (MLLR) [74, 75], which estimates a set of linear transformations for the mean parameters of a mixture Gaussian HMM system to maximize the likelihood of the adaptation data. The basic MLLR approach was later extended to be able to compensate the variance parameters of the Gaussians in addition to the means [45]. Under this approach, which is termed unconstrained model-space adaptation, the transforms that act on the means and the covariances are unrelated

to each other. That is, the general linear transform of the Gaussian mean μ is given by

$$\mu' = A\mu + b \quad (2.1)$$

where A and b is the transformation matrix and bias vector. The Gaussian covariance matrices are transformed as

$$\Sigma' = LHL^T \quad (2.2)$$

where L is the Choleski factor of the original covariance matrix Σ , or

$$\Sigma' = H\Sigma H^T \quad (2.3)$$

In both cases H is the transformation matrix applied to the covariance.

Constrained adaptation of the Gaussian parameters has also been considered [33], in which the transformation applied to the mean must correspond to the transformation applied to the covariance. Using a similar notation to the above, the transformed mean and covariance is given by

$$\mu' = A\mu + b \quad (2.4)$$

$$\Sigma' = A\Sigma A^T. \quad (2.5)$$

Here, we can see that the same transform A applies to both mean and covariance.

Transformation-based adaptation techniques can be applied in a flexible manner, depending on the amount of adaptation data that is available. Typically, the amount of adaptation data is significantly lower than the one used for the training of the initial models. Tying of transformation matrices across sets of Gaussians reduces the number of parameters to be estimated from the adaptation data. This results in more robust estimates and can provide adaptation for the Gaussian distributions which there were no observations at all in the adaptation data. The ability of transformation-based adaptation techniques to adapt every parameter in

the model is a major advantage over the conventional maximum a posteriori (MAP) based adaptation [49]. In MAP adaptation the reestimation formulae only applies to individual model parameters and therefore if a Gaussian component is not observed in the adaptation data it cannot be adapted. However, transformation-based adaptation techniques generally suffer from poor asymptotic properties leading to a quick saturation in performance as the amount of adaptation data increases.

Transformation-based adaptation techniques have also been combined with Bayesian learning methods. One of the most commonly used Bayesian learning criterion is the maximum a posteriori (MAP) criterion. The MAP estimation framework provides a way of incorporating prior knowledge about the model parameters into the training process. Siohan et al. [118] reformulated the MLLR algorithm under the Bayesian framework by introducing a prior distribution for the affine transformation matrices leading to the maximum a posteriori linear regression (MAPLR) algorithm. The prior distribution was modelled by the Normal-Wishart distribution, and unlike MLLR, MAPLR did not have a closed-form solution. Chou [20] proposed the elliptically symmetric matrix variate distribution [58] as the prior distribution for MAPLR. A closed-form solution for MAPLR was obtained under these elliptically symmetric priors. Later, Chou and He [21] developed a MAPLR adaptation framework for the variance parameters as well. The prior distribution of the transformation matrix, which is applied to the the Gaussian covariances, was selected from the family of vector normal distributions. A closed-form solution of MAPLR for variance adaptation under this prior was derived from its Expectation-Maximization formulation.

Linear transforms have also been used for Speaker Adaptive Training (SAT) [2]. The goal of SAT is to reduce inter-speaker variability within the training set. SAT is an iterative procedure that generates a set of speaker-independent state observation distributions along with matched speaker-dependent transforms for all speakers in the training set. Standard SAT uses unconstrained model-space transformations of the means and the variances in both training and testing.

2.2 Acoustic Normalization

Rather than adapting the model parameters to a new speaker or in general to a new condition, it is also possible to transform (normalize) the acoustic features. *Acoustic normalization* techniques transform the acoustic features in an attempt to

reduce variations between speech frames. A broad family of acoustic normalization techniques are based on signal processing concepts. The simplest of these is cepstral mean normalization (CMN) [5, 40] in which the long-term average of the cepstrum is subtracted from the cepstrum of each speech frame. As a result CMN removes mismatches due to channels and additive noise effects. It is known that additive noise shifts the mean and reduces the variance of the cepstral features [65]. Therefore, in addition to mean normalization, cepstral variance normalization (CVN) has also been applied to compensate this effect by normalizing all feature components to have variance of 1.0.

Historically, the usefulness of linear feature transformations was first demonstrated in *template-based* and discrete-HMM speech recognition systems [54, 55, 66, 60, 25, 114]. For a detailed discussion of these techniques the reader is referred to [73]. Zhao [138] extended the idea of feature transformations for acoustic normalization to continuous HMM-based ASR systems. He proposed a decomposition of the spectral variations into two sources: one acoustic and the other phone-specific. The acoustic source is attributed to speaker's physical individualities that cause spectral variations independent of phone units. The phone-specific source is attributed to personal speaking idiosyncrasies (e.g. accent and pronunciation). The acoustic and phone-specific variation sources are modelled as two cascaded linear transforms.

Sankar and Lee [108] presented a stochastic matching algorithm based on the ML criterion for robust speech recognition but, only in a very restricted form. The goal was to reduce the mismatch between the training and test speech data during recognition. The mismatch was modelled by a bias vector both in the feature-space and in the model-space. Gopinath [53] applied the idea of maximum likelihood linear transforms for the training of Gaussian mixture distributions. The maximum likelihood estimates of the transform parameters were obtained using a conjugate gradient method with analytic gradient supplied.

Gales [46] extended the idea of feature-based maximum likelihood linear transforms from the simple diagonal case [33] to the full matrix case. The maximum likelihood estimates of the transform parameters were obtained by an iterative closed-form procedure in which each row of the transform matrix is optimized given the current value of all the other rows. In the same work [46] Gales proposed the use of feature-based transforms for speaker adaptive training as an alternative to the standard SAT [2] scheme which uses model-space transformations of the means and/or

the variances. Using feature-based transforms in SAT, as proposed by Gales [46], the re-estimation formulae become almost identical to the standard ML-based mean and variance re-estimation formulae. Thus, SAT with feature-based transforms is simpler, computationally efficient and requires minimum alternation to the standard code.

Another commonly used normalization technique that stems from the physical model of the vocal tract is the vocal tract length normalization (VTLN). VTLN addresses the shift of the formant frequencies of the speakers caused by their different vocal tract lengths. VTLN is typically performed by warping the frequency axis of the spectrum using a suitable parameterized function. The parameter values are estimated individually for each speaker in an attempt to compensate for differences in vocal tract length between speakers. Eide and Gish [38] proposed a nonlinear warping for VTLN. McDonough et al. [89] used the *bilinear transform* that results as a matrix transformation directly in the cepstral domain. Uebel and Woodland [124] investigated a matrix approximation technique to learn the mapping between the unwarped and warped cepstra by applying a least squares estimation of the transform matrix.

Although the terms *Acoustic Normalization* and *Acoustic Adaptation* have been categorized separately here, they have been used loosely and in different ways in the literature. This is because for certain types of linear transformations, it is possible to show that acoustic normalization of feature vectors and acoustic adaptation of model parameters are equivalent. An example is presented in Section 2.4.

2.3 Correlation Modeling

It is well known that explicit modelling of correlations between spectral parameters in speech recognition results in increased classification accuracy and improved descriptive power (higher likelihoods) [81]. For example, let us compare a Gaussian density with full covariance matrix $\mathcal{N}(X, \mu_X, \Sigma_X)$ to a Gaussian density with diagonal covariance $\mathcal{N}(X, \mu_X, \text{diag}(\Sigma_X))$. When a full covariance matrix is used then all of the correlations are explicitly modelled. On the other hand, when a diagonal covariance is used, there is an underlying assumption that the components of the feature vector are not correlated. It is easy to show that a full covariance Gaussian model results in higher likelihoods [53]

$$\mathcal{N}(X, \mu_X, \text{diag}(\Sigma_X)) \leq \mathcal{N}(X, \mu_X, \Sigma_X) \quad (2.6)$$

However, computational, storage and robust estimation considerations make the use of unconstrained, full covariance matrices impractical. Methods to address this problem by the use of linear transforms can be divided in two classes, model-space schemes and feature-based schemes.

2.3.1 Model-space Schemes

Model-space schemes use Gaussian mixture components with richer covariance structure than the commonly used diagonal covariances. However, using a complex covariance structure is computationally expensive, since covariance matrixes need to be inverted for likelihood calculation. Thus, most of the model-space schemes model the inverse covariance Σ^{-1} (precision matrix) instead. Semi-tied covariances (STC) [47] express the $d \times d$ precision matrix P_g of a Gaussian m as

$$P_g = \Sigma^{-1} = A^T \Lambda_g A$$

where Λ_g is a diagonal matrix whose elements consist of the inverse variances $\lambda_{i,i} = 1/\sigma_i^2$ and A is a $d \times d$ transformation matrix (called as the semi-tied transform). This semi-tied transform can be either the same for all Gaussian components (global) or can be class-dependent, and therefore, tied over a set of Gaussian components. The parameters of the STC approach are estimated using an iterative closed-form update formulae.

Extensions to the STC approach model the precision matrix by superimposing multiple bases. For example, the extended maximum likelihood linear transforms (EMLLT) [100] allow the semi-tied transform A to be non-square. This approach constrains the precision matrices to be a linear combination of D rank-1 matrices, where $d \leq D \leq d(d+1)/2$. Therefore the precision matrix can be expressed as

$$P_g = A^T \Lambda_g A = \sum_{i=1}^D \lambda_i^{(g)} \alpha_i \alpha_i^T$$

where the d -dimensional basis vector α_i^T is the i th row of the $D \times d$ matrix A . When

$D = d$ the EMLLT scheme reduces to the STC scheme, and when $D = d(d + 1)/2$ the total number of parameters in the EMLLT approach is the same as that of a full covariance matrix model.

A generalization of the EMLLT approach is the Subspace for Precision and Mean (SPAM) [6] approach. SPAM constrains the mean vector and the precision matrix to lie in a D -dimensional subspace of the space of symmetric matrices, which is spanned by a collection of matrices S_i where $1 \leq i \leq D$. The precision matrices are given by

$$P_g = \sum_{i=1}^n \lambda_i^{(g)} S_i$$

where the matrices S_i are symmetric $d \times d$ matrices. In the SPAM model D can range from 1 to $d(d + 1)/2$. In contrast to the STC scheme, there is no closed-form solution for the update of the EMLLT and SPAM parameters. Thus, gradient-based optimization algorithms are employed for the update of the EMLLT and SPAM parameters.

2.3.2 Feature-based Schemes

Feature-based schemes transform the feature vector to better satisfy the constraints of diagonal-covariance matrices. A common technique is the discrete cosine transform (DCT) used in the final step of cepstrum coefficient extraction [27]. Other techniques include the Karhunen-Loève transform or principal component analysis (PCA) [37, 43], linear discriminant analysis (LDA) [34, 37, 43] and heteroscedastic linear discriminant analysis (HLDA) [71].

LDA finds a linear subspace that maximizes class separability among the feature vector projections in this space. Thus, LDA reduces the feature dimension by choosing a linear p -dimensional subspace of the d -dimensional feature space and by rejecting a $(d - p)$ -dimensional subspace. The implicit assumption is that the rejected subspace does not carry any classification information. For Gaussian models, this assumption is equivalent to the assumption that the means and variances of the class distributions are the same for all classes, in the rejected subspace. Furthermore, LDA assumes that the within-class variances are equal for all the classes. HLDA is a generalization of the LDA scheme, which relaxes the assumption of equal

variance in the parametric model.

Finally the maximum likelihood linear transform (MLLT) [53, 46] can also be employed for the purpose of feature decorrelation. Under this approach, MLLT applies a linear transform to the acoustic features in an attempt to capture the correlation between the feature vector components.

2.4 Discussion

The application of linear transforms in acoustic modeling for *Acoustic Adaptation*, *Acoustic Normalization* and *Correlation Modeling* has been summarized in this chapter. Numerous applications of linear transforms for each of these categories have been presented. The transforms vary from a simple bias to affine transforms and from a scalar factor to full matrixes. Moreover, the transforms can be applied to the feature space, as an affine transformation of the feature vector, and to the model space, for the transformation of the mean and/or the variance parameters of Gaussian HMM systems. The most effective form of transforms depends on the application.

Chapter 3

Overview and Objectives

In Chapter 1, we presented an overview of Automatic Speech Recognition (ASR) systems followed by prior research into the application of linear transforms in ASR systems in Chapter 2. As we saw in the previous chapter, the transforms are described as applied in either the model-space or feature-space [108]. This dissertation focuses on the use of feature-space linear transforms (also known as *constrained* model-space transforms [46]) with the aim of improving the overall recognition performance of ASR systems. The choice of feature-space transforms is motivated by the following observations:

- It is possible to create a new set of acoustic data using the apparatus of feature-based transforms. The new set can be used by other modelling schemes and estimation procedures requiring no modification to the model structure or estimation formulae.
- Feature-space transforms are quite versatile since they can be applied in all three major applications of linear transforms in acoustic modelling, which are, Acoustic Adaptation, Acoustic Normalization and Correlation Modeling.
- There is a duality between feature-space and constrained model space transforms, as shown in equation (1.34). That is, an affine transformation of the feature vector is equivalent to an affine transformation of the Gaussian means and the covariances since the respective Gaussian log-likelihoods are equal.
- By inspection of the left-hand side of equation (1.34) we can see that feature-space transforms can be implemented as a transformation of the feature space

and a simple multiplication of the Jacobian term. Thus, there is no need to alter the model parameters, as required by the right-hand side of equation (1.34), which in some circumstances may be computationally expensive or even undesired. For example in *online* adaptation mode [19, 30, 137], adaptation is performed incrementally as enrollment data become available. As a consequence, a model-space adaptation would require the models to be repeatedly updated. Another example can be given in the context of adaptive training. The standard SAT [2] uses a model-space transformation of the means and/or the variances. In order to estimate the means it is necessary to store a full matrix for each Gaussian component in the system. Furthermore it requires two passes over the training data, one for the means and one for the variances to update the Gaussian parameters. On the other hand, by the use of feature-based transforms in SAT, as proposed by Gales [46], the re-estimation formulae for the Gaussian means and variances become almost identical to the standard ML-based mean and variance re-estimation formulae. Thus, SAT with feature-based transforms is computationally efficient and requires minimum alternation to the standard code.

Given the above motivation for the use of feature-space linear transforms, this thesis focuses on the following three research areas:

- The development of estimation procedures for the feature-space linear transforms based on different estimation criteria.
- The development of an acoustic normalization procedure for the integration of diverse acoustic data.
- The development of procedures for combining multiple acoustic models, obtained using training corpora from different languages, in order to improve ASR performance in languages and domains for which large amounts of training data are not available.

Specifically, the following novel approaches will be presented:

Discriminative Likelihood Linear Transforms (DLLT) We will develop, in Chapter 4, novel estimation procedures that find Discriminative Linear Transforms jointly with MMI for feature normalization and we will present reestimation formulae

for this training scenario. These fully discriminative procedures will be derived by maximizing the Conditional Maximum Likelihood (CML) auxiliary function. Then, in Chapter 5, we will show how this estimation criterion can be used for feature normalization in HMM training for the transcription of large vocabulary conversational speech tasks.

Structural Maximum-A-Posteriori (MAP) Linear Transforms We will develop, in Chapter 6, a novel structural maximum a posteriori estimation framework for feature-space transforms. Specifically, a Bayesian counterpart of the maximum likelihood linear transforms (MLLT) will be formulated based on MAP estimation. Prior density estimation issues will be addressed by the use of a hierarchical tree structure in the transform parameter space.

Cross-Corpus Normalization Of Diverse Acoustic Data We will investigate, in Chapter 7, the use of heterogeneous data sources for acoustic training. We will describe an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data. Under this procedure, a larger in-domain training set can be created by transforming the out-of-domain data before incorporation in training. Then, in Chapter 8, we will put the cross-corpus normalization procedure into practice by investigating the use of diverse Mandarin speech corpora for building a Mandarin Conversational Telephone Speech ASR system.

Minimum Risk Acoustic Clustering for Multilingual Acoustic Model Combination We will present, in Chapter 9, a new approach for sub-word multilingual acoustic model combination. This approach aims at an optimal integration of all possible information sources into one log-linear posterior probability distribution. To this end, we will develop an automatic approach to find the optimal partitioning of model scores into classes.

Finally, in Chapter 10, we will provide a summary of this thesis identifying specific research contributions and present some suggestions for future work in this area.

Chapter 4

Discriminative Likelihood Linear Transforms (DLLT)

In Chapter 2 we reviewed the application of linear transforms in ASR systems. We saw that the transforms can be used for various modelling issues, such as acoustic adaptation, acoustic normalization and correlation modelling, can be applied in the feature or the model space, and can vary in form. Until recently these transformation techniques have been based on the maximum likelihood (ML) parameter estimation framework. As we discussed in Section 1.6.1, under realistic conditions ML estimation of HMMs cannot be relied upon to yield models that are optimum for ASR. As an alternative to ML estimation, the Maximum Mutual Information (MMI) criterion has been proposed which directly attempts to optimize performance on the acoustic training. In this chapter we will show how the MMI criterion can be used for feature normalization in HMM training. A fully discriminative training procedure is obtained that estimates both the linear transforms and Gaussian parameters under the MMI criterion.

4.1 Previous developments leading to Discriminative Linear Transforms

Until recently, linear transformation techniques have been based on the maximum likelihood parameter estimation framework. For example, in MLLT [46, 53, 71, 108, 138] the feature-space transforms are estimated under the ML criterion.

Discriminative training under the Maximum Mutual Information (MMI) [99] criterion has recently been shown to be useful in large vocabulary conversational speech recognition (LVCSR) tasks [134, 36]. Its success has triggered an interest in the use of linear transforms estimated under the MMI criterion rather than via ML estimation. These are called Discriminative Linear Transforms (DLT) [125].

One approach to the discriminative training of DLTs is Maximum Mutual Information Linear Regression (MMILR) which was introduced by Uebel and Woodland [125, 126], who showed that it can be used for supervised speaker adaptation. Gunawardana and Byrne [57] introduced the Conditional Maximum Likelihood Linear Regression (CMLLR) algorithm and showed that CMLLR can be used for unsupervised speaker adaptation. Both adaptation algorithms, MMILR and CMLLR, consider a model-space transformation of the mean and/or the variance parameters.

Maximum likelihood linear transforms have also been incorporated with MMI training, in a hybrid ML/MMI modelling approach. McDonough et al. [87] combined Speaker Adaptive Training (SAT) with MMI by estimating speaker dependent linear transforms under ML and subsequently using MMI for the estimation of the speaker independent HMM Gaussian parameters. Similarly, Ljolje [82] combined MLLT with the MMI estimation of HMM Gaussian parameters. These transforms were found using ML estimation techniques and were then fixed throughout the subsequent iterations of MMI model estimation.

In this chapter, we propose training methods based on the MMI criterion that estimate both HMM acoustic parameters and linear transforms. We obtain fully discriminative procedures for feature normalization in HMM training under the MMI criterion. These procedures are derived by maximizing Gunawardana’s Conditional Maximum Likelihood (CML) auxiliary function (equation 4, [56]). We will show in the subsequent sections how this estimation criterion can be used for feature normalization in HMM training.

4.2 Discriminative Likelihood Linear Transforms for Acoustic Normalization

Feature-space linear transforms in acoustic modelling [46, 53, 71, 108, 138] has been a common ingredient for building ASR systems [39]. This modelling tech-

nique applies affine transforms to the m dimensional observation vector o so that a normalized feature vector is found as $Ao + b$, where A is a nonsingular $m \times m$ matrix and b is a m dimensional vector. The emission density of state s , which is assumed to be Gaussian, is therefore reparametrized as shown in equation 1.33. The reparametrization of the emission density augments the usual set of HMM parameters with the parameters of the transform. The entire parameter set is defined as $\theta = (T_{\mathcal{R}(s)}, \mu_s, \Sigma_s)$.

Our goal is to estimate discriminative likelihood linear transforms and HMM parameters under the CML criterion, which was introduced in Section 1.6.2. The transforms obtained under this criterion are termed Discriminative Likelihood Linear Transforms (DLLT). This estimation is performed as a two-stage iterative procedure. We first maximize the CML criterion with respect to the affine transforms while keeping the Gaussian parameters fixed. Subsequently, we compute the Gaussian parameters using the updated values of the affine transforms. All these estimation steps are done under the CML criterion.

These procedures are derived by maximizing Gunawardana's Conditional Maximum Likelihood (CML) auxiliary function (equation 1.18). Using calculus to do the maximization yields the following update rule to be satisfied by the parameter estimation procedures: given a parameter estimate θ , a new estimate $\bar{\theta}$ is found so as to satisfy

$$\begin{aligned} \bar{\theta} : \sum_{s_1^i} \left[q(s_1^i | M_{\hat{w}_1^i}, \hat{o}_1^i; \theta) - q(s_1^i | \hat{o}_1^i; \theta) \right] \nabla_{\theta} \log q(\hat{o}_1^i | s_1^i; \bar{\theta}) \\ + \sum_{s_1^i} d'(s_1^i) \int q(o_1^i | s_1^i; \theta) \nabla_{\theta} \log q(o_1^i | s_1^i; \bar{\theta}) do_1^i = 0. \quad (4.1) \end{aligned}$$

Here, O is the acoustic observation vector sequence and W is the corresponding word sequence. The pair $(\hat{w}_1^i, \hat{o}_1^i)$ denotes observed values of these random variables, i.e. the training data. $d'(s_1^i)$ leads to the well-known MMI constant as $D_s = \sum_{s_1^i: s_{\tau}=s} d'(s_1^i)$.

4.2.1 DLLT Estimation

The DLLT modeling approach incorporates Gales' [46] treatment of MLLT and Gunawardana's CMLLR derivation [57]. In the first part of the two-stage estimation procedure we fix the HMM means and variances and maximize the CML criterion with respect to the affine transforms.

The parameter update relationship of equation (4.1) can be simplified by using the Markov assumptions and noticing that each of the states is uniquely assigned to one of R disjoint transform classes S_r . Therefore we can write

$$\log q(\hat{\zeta}_1^{\hat{l}} | s_1^{\hat{l}}; \bar{\theta}) = \sum_s \sum_{\tau=1}^{\hat{l}} \log q(\hat{\zeta}_\tau | s; \bar{T}_r) 1_s(s_\tau) 1_r(\mathcal{R}(s))$$

where $1_s(s_\tau) = 1$ if $s_\tau = s$, 0 otherwise and similarly, $1_r(\mathcal{R}(s)) = 1$ if $r = \mathcal{R}(s)$, 0 otherwise. We can then express equation (4.1) as:

$$\begin{aligned} [\bar{T}_r]_i : \sum_{s \in S_r} \sum_{\tau=1}^{\hat{l}} \gamma'_s(\tau; \theta) \cdot \nabla_{[T_r]_i} \log q(\hat{\zeta}_\tau | s; \bar{T}_r) \\ + \sum_{s \in S_r} D_s \int q(\zeta; T_r) \nabla_{[T_r]_i} \log q(\zeta | s; \bar{T}_r) d\zeta = 0 \quad i = 1, \dots, m \end{aligned} \quad (4.2)$$

where $[T_r]_i$ denotes the i^{th} row of T_r and $\gamma'_s(\tau; \theta) = \gamma_s(\tau; \theta) - \gamma_s^g(\tau; \theta)$. Here, $\gamma_s(\tau; \theta) = q_{s_\tau}(s | M_{\hat{w}_1^{\hat{l}}}, \hat{\delta}_1^{\hat{l}}; \theta)$ is the conditional occupancy probability of state s at time τ given the training acoustics and the model that corresponds to the transcription, $\gamma_s^g(\tau; \theta) = q_{s_\tau}(s | \hat{\delta}_1^{\hat{l}}; \theta)$ is the conditional occupancy probability of state s at time τ given only the training acoustic data, and $D_s = \sum_{s_1^{\hat{l}}: s_\tau = s} d'(s_1^{\hat{l}})$.

With the HMM means and variances fixed, the logarithm of the reparametrized conditional density $\log q(\zeta | s; \theta)$ is given by (ignoring all terms independent of T_r):

$$\log q(\zeta | s; \theta) = \log(|A_r|) - \frac{1}{2} \sum_{i=1}^m ([T_r]_i Z_{s,i} [T_r^T]_i - 2[T_r]_i w_{s,i}^T)$$

where $\mathcal{R}(s) = r$ and

$$Z_{s,i} = \frac{1}{\sigma_{s,i}^2} \zeta \zeta^T$$

$$w_{s,i} = \frac{\mu_{s,i}}{\sigma_{s,i}^2} \zeta^T$$

$\mu_{s,i}$ and $\sigma_{s,i}$ are the i th elements of the mean and variance vector, for state s .

The gradient of $\log q(\zeta|s; \theta)$ with respect to the parameter component $[T_r]_i$ is given by

$$\nabla_{[T_r]_i} \log q(\zeta|s; \theta) = \frac{p_i}{p_i [\bar{T}_r^T]_i} - [T_r]_i Z_{s,i} + w_{s,i}$$

where p_i is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, ($c_{ij} = \text{cof}(A_{ij})$).

Substituting the above expression for the gradient into equation (4.2) yields

$$\begin{aligned} \sum_{s \in S_r} \sum_{\tau=1}^{\hat{l}} \gamma'_s(\tau; \theta) \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - [\bar{T}_r]_i \hat{Z}_{s,i} + \hat{w}_{s,i} \right) \\ + \sum_{s \in S_r} D_s \int q(\zeta; T_r) \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - [\bar{T}_r]_i Z_{s,i} + w_{s,i} \right) d\zeta = 0. \end{aligned} \quad (4.3)$$

The calculation of the integral in equation (4.3) proceeds as:

$$\begin{aligned} \int q(\zeta; T_r) \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - [\bar{T}_r]_i Z_{s,i} + w_{s,i} \right) d\zeta \\ = \frac{p_i}{p_i [\bar{T}_r^T]_i} - \frac{1}{\sigma_{s,i}^2} [\bar{T}_r]_i \int q(\zeta; T_r) \zeta \zeta^T d\zeta + \frac{\mu_{s,i}}{\sigma_{s,i}^2} \int q(\zeta; T_r) \zeta^T d\zeta \\ = \frac{p_i}{p_i [\bar{T}_r^T]_i} - \frac{1}{\sigma_{s,i}^2} [\bar{T}_r]_i J_s + \frac{\mu_{s,i}}{\sigma_{s,i}^2} [J_s]_1 \end{aligned}$$

where J_s is defined as the matrix

$$\begin{bmatrix} 1 & [A_r^{-1}(\mu_s - b_r)]^T \\ A_r^{-1}(\mu_s - b_r) & A_r^{-1}[\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T]A_r^{-T} \end{bmatrix}. \quad (4.4)$$

Equation (4.3) can then be written as

$$\sum_{s \in S_r} \sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - [\bar{T}_r]_i \hat{Z}_{s,i} + \hat{w}_{s,i} \right) + \sum_{s \in S_r} D_s \left(\frac{p_i}{p_i [\bar{T}_r^T]_i} - \frac{1}{\sigma_{s,i}} [\bar{T}_r]_i J_s + \frac{\mu_{s,i}}{\sigma_{s,i}^2} [J_s]_1 \right) = 0$$

Rearranging yields

$$\beta \frac{p_i}{p_i [\bar{T}_r^T]_i} = [\bar{T}_r]_i G_i - k_i \quad (4.5)$$

where

$$G_i = \sum_{s \in S_r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T + D_s J_s \right) \quad (4.6)$$

$$k_i = \sum_{s \in S_r} \frac{\mu_{s,i}}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) \hat{\zeta}_\tau^T + D_s [J_s]_1 \right) \quad (4.7)$$

$$\beta = \sum_{s \in S_r} \left(\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \theta) + D_s \right) \quad (4.8)$$

An iterative solution to the optimization of equation (4.5) is described by Gales [46], where each row of T_r is optimized given the current value of all the other rows. It can be shown that the i^{th} row of the transformation matrix is found by

$$[\bar{T}_r]_i = (\alpha p_i + k_i) G_i^{-1} \quad (4.9)$$

where α satisfies a quadratic expression (Equation B1.8, [46]).

4.2.2 Gaussian Parameter Estimation

This section describes the estimation procedure for the state dependent Gaussian means and variances under the CML criterion. With the transforms estimated as described in Section 4.2.1, we denote the parameter set as $\tilde{\theta} = (\bar{T}_r, \mu_s, \Sigma_s)$. Using the Markov assumptions, we can write $\log q(\hat{\zeta}_1^i | s_1^i; \tilde{\theta})$ as $\sum_s \sum_{\tau=1}^i \log q(\hat{\zeta}_\tau | s; \tilde{\theta}) 1_s(s_\tau)$ and simplify equation (4.1) as:

$$\bar{\theta} : \sum_{\tau=1}^i \gamma'_s(\tau; \tilde{\theta}) \cdot \nabla_{\tilde{\theta}} \log q(\hat{\zeta}_\tau | s; \tilde{\theta}) + D_s \int q(\zeta; \tilde{\theta}) \nabla_{\tilde{\theta}} \log q(\zeta; \tilde{\theta}) d\zeta = 0 \quad (4.10)$$

where we define $\gamma'_s(\tau; \tilde{\theta}) = \gamma_s(\tau; \tilde{\theta}) - \gamma_s^g(\tau; \tilde{\theta})$. Here, the posteriors $\gamma_s(\tau; \tilde{\theta})$ and $\gamma_s^g(\tau; \tilde{\theta})$ are estimated for each state using the new transform estimates and old Gaussian model parameters. To simultaneously update the Gaussian means and variances we take the derivative of the state dependent emission density with respect to μ_s and Σ_s^{-1} , substitute the result in equation (4.10) and solve for μ_s and Σ_s .

Mean estimation

The gradient of $\log q(\zeta | s; \tilde{\theta})$ with respect to the parameter component μ_s is given by

$$\begin{aligned} \nabla_{\mu_s} \log q(\zeta | s; \tilde{\theta}) &= \nabla_{\mu_s} \left(-\frac{1}{2} (\bar{T}_r \zeta - \mu_s)^T \Sigma_s^{-1} (\bar{T}_r \zeta - \mu_s) \right) \\ &= \Sigma_s^{-1} (\bar{T}_r \zeta - \mu_s) \end{aligned}$$

Substituting into equation (4.10) and rearranging gives

$$\sum_{\tau=1}^i \gamma'_s(\tau; \tilde{\theta}) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) + D_s \left(\int q(\zeta; \tilde{\theta}) \bar{T}_r \zeta d\zeta - \int q(\zeta; \tilde{\theta}) \bar{\mu}_s d\zeta \right) = 0$$

Calculating the integral yields

$$\sum_{\tau=1}^i \gamma'_s(\tau; \tilde{\theta}) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) + D_s (\mu_s - \bar{\mu}_s) = 0$$

Finally the update equation for μ_s is given by

$$\bar{\mu}_s = \frac{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau + D_s \mu_s}{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) + D_s} \quad (4.11)$$

Variance estimation

The gradient of $\log q(\zeta|s; \tilde{\theta})$ with respect to Σ_s^{-1} is given by

$$\begin{aligned} \nabla_{\Sigma_s^{-1}} \log q(\zeta|s; \tilde{\theta}) &= \nabla_{\Sigma_s^{-1}} \left(\log |\Sigma_s| - (\bar{T}_r \zeta - \mu_s) \Sigma_s^{-1} (\bar{T}_r \zeta - \mu_s)^T \right) \\ &= \Sigma_s - (\bar{T}_r \zeta - \mu_s) (\bar{T}_r \zeta - \mu_s)^T \end{aligned}$$

Substituting into equation (4.10) gives

$$\begin{aligned} \sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s)^T \right) \\ + D_s \int q(\zeta|s; \tilde{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \zeta - \bar{\mu}_s) (\bar{T}_r \zeta - \bar{\mu}_s)^T \right) d\zeta = 0. \quad (4.12) \end{aligned}$$

Calculating the integral in the previous equation gives

$$\begin{aligned} \bar{\Sigma}_s - \int q(\zeta|s; \tilde{\theta}) (\bar{T}_r \zeta - \bar{\mu}_s) (\bar{T}_r \zeta - \bar{\mu}_s)^T d\zeta \\ = \bar{\Sigma}_s - \bar{\mu}_s \bar{\mu}_s^T - \int q(\zeta|s; \tilde{\theta}) (\bar{T}_r \zeta \zeta^T \bar{T}_r^T - \bar{T}_r \zeta \bar{\mu}_s^T - \bar{\mu}_s \zeta^T \bar{T}_r^T) d\zeta \\ = \bar{\Sigma}_s - \Sigma_s - \mu_s \mu_s^T + \mu_s \bar{\mu}_s^T + \bar{\mu}_s \mu_s^T - \bar{\mu}_s \bar{\mu}_s^T \end{aligned}$$

Substituting the integral into equation (4.12) yields

$$\begin{aligned} \sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s)^T \right) \\ + D_s (\bar{\Sigma}_s - \Sigma_s - \mu_s \mu_s^T + \mu_s \bar{\mu}_s^T + \bar{\mu}_s \mu_s^T - \bar{\mu}_s \bar{\mu}_s^T) = 0 \end{aligned}$$

Using the fact that $\bar{\mu}_s$ is given by equation (4.11) we can obtain the reestimation formula for the new estimate of Σ_s as

$$\bar{\Sigma}_s = \frac{\sum_{\tau=1}^{\hat{I}} \gamma'_s(\tau; \tilde{\theta}) \bar{T}_\tau \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}_\tau^T + D_s (\Sigma_s + \mu_s \mu_s^T)}{\sum_{\tau=1}^{\hat{I}} \gamma'_s(\tau; \tilde{\theta}) + D_s} - \bar{\mu}_s \bar{\mu}_s^T \quad (4.13)$$

In the implementation of the procedures reported here, the D_s values for equations (4.9), (4.11) and (4.13) are calculated as described by Woodland and Povey [134] (Sec. 3, schemes *i* & *ii* with $E = 2$).

4.2.3 The DLLT Algorithm

This concludes the estimation procedure for the parameters of the DLLT model. The derivation described above is an iterative procedure that can be summarized in the following steps:

1. For each training utterance, create the numerator and denominator lattices [127].
2. Calculate the conditional occupancy probabilities $\gamma_s(\tau; \theta)$ and $\gamma_s^g(\tau; \theta)$ and collect numerator/denominator statistics for the transform update equations.

The numerator statistics for state s are

$$\begin{aligned} & - \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; \theta) \\ & - \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; \theta) \hat{\zeta}_\tau^T \\ & - \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T \end{aligned}$$

The denominator statistics are

$$\begin{aligned} & - \sum_{\tau=1}^{\hat{I}} \gamma_s^g(\tau; \theta) \\ & - \sum_{\tau=1}^{\hat{I}} \gamma_s^g(\tau; \theta) \hat{\zeta}_\tau^T \\ & - \sum_{\tau=1}^{\hat{I}} \gamma_s^g(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T. \end{aligned}$$

3. Estimate the transforms via equation (4.9), which is iterated until the $[T_r]_i$ parameters converge, while keeping the Gaussian parameters fixed at their current values.

4. Calculate the conditional occupancy probabilities $\gamma_s(\tau; \tilde{\theta})$ and $\gamma_s^g(\tau; \tilde{\theta})$, using the updated values of the transforms, and collect numerator/denominator statistics for the mean and variance update equations. The numerator statistics for state s are

$$\begin{aligned}
& - \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; \tilde{\theta}) \\
& - \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau \\
& - \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}_r^T
\end{aligned}$$

The denominator statistics are

$$\begin{aligned}
& - \sum_{\tau=1}^{\hat{I}} \gamma_s^g(\tau; \tilde{\theta}) \\
& - \sum_{\tau=1}^{\hat{I}} \gamma_s^g(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau \\
& - \sum_{\tau=1}^{\hat{I}} \gamma_s^g(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}_r^T
\end{aligned}$$

5. Estimate the Gaussian parameters via equations (4.11) and (4.13) using the updated values of the transforms.

These steps can be repeated as necessary to achieve the required convergence.

4.3 Discussion

This chapter described the integration of Discriminative Linear Transforms into MMI estimation for Large Vocabulary Speech Recognition. We developed estimation procedures that find Discriminative Linear Transforms jointly with MMI for feature normalization and we presented reestimation formulae for this training scenario. We obtained fully discriminative procedures which were derived by maximizing the Conditional Maximum Likelihood (CML) auxiliary function.

There are numerous maximum likelihood modeling techniques that incorporate linear transforms into acoustic normalization and adaptation, as discussed in Chapter 2 and in [48]. Many of those techniques can be transformed into discriminative modeling techniques in a manner analogous to the approach we have taken here with the maximum likelihood linear transformation for feature normalization.

For example, the unconstrained model-based transforms [2] and constrained model-based transforms [46] used in Speaker Adaptive Training are based on Max-

imum Likelihood estimation. As an alternative to these ML-SAT estimation procedures, the input transform and the Gaussian model parameters can be estimated using the CML auxiliary function. Doumpiotis et al. [35, 122, 123] proposed DSAT, a fully discriminative procedure for unconstrained model-based SAT. This training procedure estimates both the HMM Gaussian parameters and the unconstrained linear transforms used in SAT via the CML criterion. McDonough and Waibel [88] presented a similar technique for performing SAT using the MMI criterion. Analogously, we note that the DLLT framework presented here can easily be modified and applied for MMI estimation of the constrained model-space transforms and HMM parameters for SAT [46].

Discriminative criteria have also been combined with Linear Discriminative Analysis [62]. LDA is a common feature extraction method for speech recognition where the transforms are estimated under a class separability criterion. Schlüter [111] developed the Linear MMI Analysis (LMA) in which he replaced the class separability criterion of LDA with a MMI criterion. Experimental results showed that although for single densities a relative improvement in word error rate could be observed for LMA in comparison to LDA, the prominence of LMA diminished with increasing parameter numbers.

Finally, Goel et al. [51] combined the Subspace Precision and Mean (SPAM) model [6] with the MMI criterion. The SPAM model which places a general subspace constraint on the precision matrices (inverse covariance) and means of all the Gaussians in a HMM, was initially based on the Maximum Likelihood criterion. The discriminative counterpart of SPAM outperformed the ML-based SPAM model on a small digit recognition task.

In the next chapter we will show how the discriminative likelihood linear transforms (DLLT) can be used for feature normalization in HMM training for Large Vocabulary Conversational Speech Recognition (LVCSR) tasks. Specifically, we will validate DLLT as an estimation procedure and compare the proposed discriminative training technique to its Maximum Likelihood counterpart.

Chapter 5

DLLT Performance in Large Vocabulary Conversational Speech Recognition

In Chapter 4 we described the integration of Discriminative Linear Transforms into MMI estimation and we developed estimation procedures that find DLTs jointly with MMI for feature normalization. In this chapter we will show how this estimation criterion can be used for feature normalization in HMM training for the transcription of large vocabulary conversational speech tasks.

The transcription of spontaneous conversational telephone speech is one of the most challenging tasks for speech recognition technology. Conversational speech is quite different in tone, speed, inflection, and overall style than scripted speech [116, 117]. Elements such as poor pronunciation, pauses and other disfluencies have major impact in recognition performance. As a result state-of-the-art systems yield higher error rates on large vocabulary conversational speech recognition (LVCSR) tasks than scripted speech and small vocabulary tasks.

The experimental results of this chapter are carried out on such LVCSR task, i.e. the SWITCHBOARD corpus [50]. It consists of spontaneous conversations from every major dialect of American English collected over standard telephone lines. We will first validate the proposed discriminative procedure of Chapter 4 on a subset of the SWITCHBOARD corpus for fast turnaround. Subsequently, we will present experimental results using the full SWITCHBOARD corpus.

5.1 Validating DLLT

In this section a series of experiments will validate DLLT as an estimation procedure and explore the sensitivity of DLLT to different initialization points. Furthermore, we will compare the proposed discriminative training technique to its Maximum Likelihood counterpart. We will begin with a description of the system we worked on and a discussion of the practical DLLT estimation.

5.1.1 System Description

The system is a speaker independent continuous mixture density, tied state, cross-word, gender-independent, triphone HMM system. The baseline acoustic models used as seed models for our experiments were built using HTK [136] from 16.4 hours of Switchboard-1 and 0.5 hour of Callhome English data. This collection defined the development training set for the 2001 JHU LVCSR system [16]. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration components [59]. Cepstral mean and variance normalization was performed over each conversation side. The acoustic models used cross-word triphones with decision tree clustered states [136], where questions about phonetic context as well as word boundaries were used for clustering. There were 4000 unique triphone states with 6 Gaussian components per state. Lattice rescoring experiments were performed using the AT&T Lange Vocabulary Decoder [92], using a 33k-word trigram language model provided by SRI [121].

The recognition tests were carried out on a subset of the 2000 Hub-5 Switchboard-1 evaluation set (SWBD1) [85] and the 1998 Hub-5 Switchboard-2 evaluation set (SWBD2) [84]. The SWBD1 test set was composed of 866 utterances consisting of 10260 words from 22 conversation sides, and the SWBD2 test set was composed of 913 utterances consisting of 10643 words from 20 conversation sides. The total test set was 2 hours of speech.

To define the number of transforms and assign the Gaussians in the model set to clusters we employed a variation of the HTK regression class tree implementation [136]. The basic implementation uses a centroid-splitting algorithm in which similar states are determined via an Euclidian distance measure between distributions. Here, all states of every context-dependent phone associated with the same monophone were assigned to the same initial class. The HTK splitting algorithm was then

applied to each of the initial classes with the additional constraint that all the mixture components associated with the same state belong to the same regression class.

Discriminative training requires alternate word sequences that are representative of the recognition errors made by the decoder. These are obtained via triphone lattices generated on the training data. Our approach is based on the MMI training procedure developed by Woodland and Povey [134]. However, rather than accumulating statistics via the Forward-Backward procedure at the word level, we use the Viterbi procedure over triphone segments. These triphone segments are fixed throughout MMI training.

5.1.2 Effective DLLT Estimation

By inspecting the definition of G_i (equation (4.6)), one can see that large values of D_s make the resulting transform to have dominant diagonal terms when the covariance Σ_s in J_s is diagonal. As we showed in Section 4.2.1, the matrix J_s is defined as the expectation of $\zeta\zeta^T$ under the reparameterized Gaussian emission density of state s , given by equation (1.33). That is

$$J_s \triangleq E_q[\zeta\zeta^T] = \int q(\zeta; T_r)\zeta\zeta^T d\zeta$$

For presentation purposes, we also repeat the resulting expression for the J_s matrix, as given by equation (4.4):

$$J_s = \begin{bmatrix} 1 & [A_r^{-1}(\mu_s - b_r)]^T \\ A_r^{-1}(\mu_s - b_r) & A_r^{-1}[\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T]A_r^{-T} \end{bmatrix}.$$

Now, let us focus on the terms of the submatrix $\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T$ of the J_s matrix. For brevity we drop the subscripts in the previous expression. The diagonal terms are $\sigma_{i,i} + (\mu_i - b_i)^2$ whereas the off-diagonal terms are $\sigma_{i,j} + (\mu_i - b_i)(\mu_j - b_j)$ for $j \neq i$. If Σ_s is assumed to be diagonal then $\sigma_{i,j} = 0$. Hence, the diagonal terms dominate slightly when Σ_s is diagonal. Recall that the J_s matrix is multiplied by the MMI factor D_s (see equation (4.6)). The large values of D_s as used in MMI further exaggerate this effect. In this case, the resulting DLLT transform is effectively

Transform Reestimation Only (Mean & Variance Fixed)		
	SWBD1	SWBD2
ML	41.1	51.1
ML+MLLT-1it	39.1	50.3
ML+MLLT-2it	39.4	50.4
ML+DLLT-1it	38.5	49.7
ML+DLLT-2it	38.3	49.9

Table 5.1: Word Error Rate (%) of systems trained with MLLT and DLLT and tested on the SWBD1 and SWBD2 test sets. The HMM Gaussian parameters are kept fixed at their ML values throughout transform updates.

identity, i.e. $T_r \approx I$. We note that MLLT does not have this problem since it has no D_s or J_s terms.

We have found it effective to replace Σ_s in J_s by the estimate of its *full covariance* matrix as found from the most recently computed statistics. This is because, the use of the full covariance form in J_s prevents the diagonal terms from dominating the new transform. We stress however that the full covariance is not used elsewhere; it is not used in the estimation of the Gaussian emission densities (see update equation (4.13)).

5.1.3 DLLT Results

We conducted a series of experiments to compare DLLT to MLLT. Throughout the MLLT experiments [46], we perform one update of the transforms followed by one update of the Gaussians using statistics, obtained from the Viterbi alignment. These alignments were obtained from the ML baseline and were kept fixed throughout the MLLT experiments. Similarly, during DLLT iterations, we perform one update of the transforms followed by one update of the Gaussians, as described in Section 4.2.3. The triphone posterior probabilities used during MMI based training were calculated from the ML baseline and were kept fixed throughout the DLLT updates.

Our first experiment kept the parameters of the HMM observation distributions fixed at their ML values. Results are reported in Table 5.1. Throughout these experiments we used a fixed set of 467 transform classes generated by the above described clustering algorithm. The SWBD1 ML baseline Word Error Rate is 41.1%. The first and second iteration of MLLT yield Word Error Rate of 39.1% and 39.4%, showing overtraining at the second iteration. DLLT yields Word Error Rate of

			Iteration								
#classes			0	1	2	3	4	5	6	7	8
48	MLLT	SWBD1	41.1	39.4	39.2	39.2	39.0	38.9	38.8	38.7*	38.8
		SWBD2	51.1	50.0	50.3	50.4	50.3	50.2	50.2	50.2*	50.1
	DLLT	SWBD1	*	37.6	37.4						
		SWBD2	*	49.5	48.8						
223	MLLT	SWBD1	41.1	38.9	38.7	38.7	38.2	37.9	37.8*	38.0	37.9
		SWBD2	51.1	49.8	49.8	49.6	49.5	49.5	49.3*	49.1	49.1
	DLLT	SWBD1	*	37.0	37.0						
		SWBD2	*	48.6	48.6						
467	MLLT	SWBD1	41.1	38.4	38.2	38.2	37.7	37.9	37.8*	37.9	37.8
		SWBD2	51.1	49.6	49.5	49.3	49.2	49.0	49.0*	49.0	49.1
	DLLT	SWBD1	*	37.1	36.9						
		SWBD2	*	48.4	48.8						

Table 5.2: Word Error Rate (%) of systems trained with MLLT and DLLT and tested on the SWBD1 and SWBD2 test sets for different number of classes. DLLT systems are seeded from well trained MLLT systems, indicated by asterisks.

38.5% and 38.3% at the first and second iteration. Similar performance was found on SWBD2. These experiments show that discriminative estimation of linear transforms improves over ML estimation for feature normalization.

Our next objective was to see the improvements obtained by varying the number of transformation classes. For this purpose, we trained three MLLT systems with 48, 223 and 467 transformation classes, all initialized from the ML baseline. The 48 class MLLT system has transforms tied to each state of every context-dependent phone associated with the same monophone. For the 223 and 467 class MLLT systems we allow a maximum of 6 and 10 divisions of the original classes, respectively. Results are reported in Table 5.2. In these MLLT experiments we observe significant performance gains in using the larger number of transformation classes.

We then performed DLLT, for these same three collections of transformation classes. In each case, the DLLT system is seeded from a well-trained MLLT system (indicated by *). We note that in all cases the DLLT systems outperform the MLLT systems. Best performance is obtained with the larger number of transformation classes, although the advantages are not as great as with MLLT. This suggests that DLLT can be effective with fewer transforms than MLLT. For subsequent experiments we use the 467 transformation class set.

In the next set of experiments we explore the sensitivity of DLLT to different

Iteration	MLLT		DLLT		DLLT	
	SWBD1	SWBD2	SWBD1	SWBD2	SWBD1	SWBD2
0	41.1	51.1	†	†	*	*
1†	38.4	49.6	38.2	49.2	37.1	48.4
2	38.2	49.5	37.3	48.9	36.9	48.8
3	38.2	49.3	37.8	48.8	-	-
4	37.7	49.2				
5	37.9	49.0				
6*	37.8	49.0				
7	37.9	49.0				
8	37.8	49.1				

Table 5.3: Word Error Rate (%) of systems trained with DLLT and tested on the SWBD1 and SWBD2 test sets for two different initialization points.

initialization points. These experiments are shown in Table 5.3. The DLLT system is seeded after 1 iteration of MLLT experiments (indicated by †). After two iterations, DLLT performance is (37.3%/48.9%). For convenience we include in Table 5.3 the 467 class MLLT experiments reported in Table 5.2. We also include the 467 class DLLT experiments of table Table 5.2 in which estimation was initialized after 6 iterations of MLLT experiments. After two iterations, DLLT performance is (36.9%/48.8%). We find that the latter is superior to performing DLLT after only 1 iteration of MLLT. This shows the importance of a proper initialization of the DLLT procedure.

We also observe, that DLLT converges in fewer iterations than MLLT. After two iterations, DLLT yields better performance (37.3%/48.9%) than six iterations of MLLT (37.8%/49.0%). Moreover, DLLT consistently outperforms MLLT.

We next study the application of MMIE estimation of the Gaussian means and variances in a system with fixed transforms estimated via MLLT. We call this technique MLLT+MMIE following Ljolje [82], in Table 5.4. We compare this to a similar approach in which the transforms are estimated via 1 iteration of DLLT and then fixed prior to further MMIE Gaussian estimation. We call this technique DLLT+MMIE in Table 5.4. Both procedures were initialized with the well trained MLLT system of Table 5.3 found at iteration 6. The similar performance found in these scenarios is not surprising, in that enough in domain data are available to allow discriminative estimation of the unconstrained observation distributions. This clearly dominates the intermediate calculations of the underlying transforms.

Finally we calculate the value of the CML objective function (equation 1.17) as

Iteration	MLLT+MMIE		DLLT+MMIE	
	SWBD1	SWBD2	SWBD1	SWBD2
0	*	*	*	*
1	37.6	48.7	36.7	48.6
2	37.0	48.4	36.8	48.9
3	36.9	48.7	-	-

Table 5.4: Word Error Rate (%) of systems trained with MLLT+MMIE and DLLT+MMIE is seeded from models found after 6 MLLT iterations.

Iteration	CML Objective Function $\log P(W O; \theta)$
0	-2.15E05
1	-1.80E05
2	-1.53E05

Table 5.5: The value of the CML objective function as a function of the iteration number for the DLLT-467 system. Iteration 0 indicates the MLLT baseline.

a function of the iteration number for the DLLT-467 system. Results are reported in Table 5.5. In iteration 0 (the MLLT baseline) the value of the CML objective function is $-2.15E05$, in DLLT-467 iteration 1 is $-1.80E05$ and in DLLT-467 iteration 2 is $-1.53E05$. These results confirm that the CML objective function is increasing under the estimation procedure for DLLT.

The experimental results of this section demonstrated the effectiveness of the proposed discriminative training procedure relative to maximum likelihood for feature normalization.

5.2 DLLT Performance on SWITCHBOARD

The experiments described in this section were conducted during the development of the 2002 JHU LVCSR system [103] which was build for the 2002 NIST Rich Transcription evaluation of English conversational telephone speech transcription. The 2002 JHU LVCSR system and the system of Section 5.1 have mostly the same architecture. However, the former was built from the full SWITCHBOARD corpus which is almost ten times bigger than the development training set used to validate DLLT. Furthermore, the 2002 JHU LVCSR system incorporates Vocal Track Normalization (VTN). Thus, in this section we explore the effectiveness of DLLT trained from a much bigger acoustic set and in conjunction with VTN.

5.2.1 2002 JHU LVCSR System Description

The acoustic model of the 2002 JHU LVCSR system was trained from 150 hours of Switchboard-1 and 15 hours of CallHome transcribed acoustic data. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration coefficients. Cepstral mean and variance normalization was performed over each conversation side. Initial estimates of Vocal Tract Normalization (VTN) parameters were obtained using Gaussian Mixture Models. VTN was based on Sine-Log All-Pass (SLAPT-2) [86] parameterized transformations applied directly to the PLP cepstra. The acoustic models used cross-word triphones with decision tree clustered states [136]. There were 8340 unique triphone states with 16 Gaussian components per speech state. The language model used was the SRI 33K word back-off trigram language model that contained multiwords [121]. A detailed description of the system is available [103].

The recognition tests were carried out on the development set used during the training of the 2002 JHU LVCSR system. This set consisted of the 2000 Hub-5 Switchboard-1 evaluation set (SWBD1F) [85] the 1998 Hub-5 Switchboard-2 evaluation set (SWBD2F) [84] and the Switchboard-2 Cellular set (CELL). The total development set was approximately 6 hours of speech. All the experimental results reported here include unsupervised MLLR speaker adaptation.

5.2.2 DLLT results on SWITCHBOARD

Acoustic model training was performed in stages with successively more complex acoustic model being applied in later stages. Initial acoustic models (HMM Set A) were trained under the maximum likelihood criterion using cepstral mean and variance normalized data. Then three iterations of single-pass retraining [136] were performed to derive a set of HMMs (HMM Set B) for the SLAPT-2 VTN normalized acoustic training data. One iteration of Maximum Likelihood Linear Transform (MLLT) estimation were performed to refine HMM Set B. The estimation of the linear transforms and Gaussian parameters was performed as a two-stage procedure. The affine transforms were first estimated while keeping the Gaussian parameters fixed. Subsequently, the Gaussian parameters were computed using the updated values of the affine transforms. All these estimation steps were done under the ML criterion. A set of 1500 regression classes were used for the affine transforms.

System	SWBD1F	SWBD2F	CELL
ML	25.0	40.2	39.9
MLLT	24.7	39.2	39.8
MLLT+MMIE	24.2	38.7	39.2
DLLT	24.0	38.7	38.8

Table 5.6: Word Error Rate (%) of DLLT system trained from the full SWITCHBOARD data and tested on the SWBD1F, SWBD2F and CELL test sets. Results are reported with unsupervised MLLR speaker adaptation.

One iteration of Discriminative Likelihood Linear Transform estimation (DLLT) was then performed. The estimation of the linear transforms and Gaussian parameters was performed again as a two-stage iterative procedure. The Conditional Maximum Likelihood (CML) criterion was first maximized with respect to the affine transforms while keeping the Gaussian parameters fixed. Subsequently, the Gaussian parameters were computed using the updated values of the affine transforms. All these estimation steps were done under the CML criterion. The same set of 1500 regression classes, used in MLLT, were used for the DLLT transforms as well. Moreover, since in Section 5.1 we found that DLLT works best when initialized by MLLT, the DLLT transforms were initialized by the MLLT transforms. Finally, two iterations of MMI was applied incorporating the results of DLLT estimation.

For comparison we have also trained a hybrid ML/MMI acoustic model. This hybrid model, termed MLLT+MMIE, combined the MLLT transforms with the MMI estimation of HMM Gaussian parameters. These transforms were found using ML estimation techniques and were then fixed throughout the subsequent iterations of MMI model estimation.

The recognition results on the development set with unsupervised MLLR speaker adaptation for the aforementioned acoustic models are shown on Table 5.6. We first observe that feature normalization obtained via the ML criterion in HMM training improves the performance of the baseline ASR system (ML vs. MLLT). Then we observe that discriminative training of the Gaussian means and variances further improves performance (MLLT vs. MLLT+MMIE). Finally, the fully discriminative procedure which estimates both HMM acoustic parameters and linear transforms under the CML criterion outperforms both ML training and the hybrid ML/MMI training (MLLT vs. DLLT and MLLT+MMIE vs. DLLT).

The results reported here are in accord with the results of Section 5.1 in which we

validated the DLLT procedure using a ten times smaller training set. Furthermore, it is worth noticing that the DLLT normalization procedure is still effective for an ASR system that incorporates Vocal Track Normalizing (VTN) transforms and unsupervised MLLR speaker adaptation on the test side.

5.3 Summary

In this chapter we investigated the use of discriminative linear transforms for feature normalization in HMM training for Large Vocabulary Conversational Speech Recognition (LVCSR) tasks. The experimental results were carried out on the SWITCHBOARD corpus [50] which consists of spontaneous conversations from every major dialect of American English collected over standard telephone lines.

We first discussed modeling approximations needed for their effective implementation. Specifically, we found that when the covariance matrix Σ_s in J_s (equation 4.4) is diagonal, then the resulting DLLT transform had dominant diagonal terms and was effectively the identity matrix. Replacing Σ_s in J_s by the estimate its full covariance matrix prevented the diagonal terms from dominating the transform. We emphasize however that the full covariance matrix was not used in the estimation of the Gaussian emission densities.

This new fully discriminative training procedure was first validated on a small subset of the Switchboard corpus and gave approximately 0.8% absolute Word Error Rate improvement over the ML estimation procedure. Another goal of the experiments was to identify a proper initialization point for the proposed discriminative technique. It turns out that DLLT works better when initialized by MLLT instead of the identity transform. We have also confirmed that the MMI objective function is increasing under the DLLT estimation procedure.

We then applied the DLLT training procedure on the 2002 JHU LVCSR system, which was build for the 2002 NIST Rich Transcription evaluation of English conversational telephone speech transcription. This ASR system apart from the DLLT transforms included another normalization procedure, i.e. Vocal Track Normalization. Furthermore, decoding was performed with unsupervised MLLR speaker adaptation on the test side. The experimental results conducted using the 2002 JHU LVCSR system reconfirmed that fully discriminative procedures outperform both ML training and hybrid ML/MMI training. We also note that iterative estimation

of the DLLT and HMM parameters yielded optimum results.

We conclude that the DLLT provides a discriminative estimation framework for feature normalization in HMM training for LVCSR tasks that outperforms in recognition performance its Maximum Likelihood counterpart.

Chapter 6

Structural Maximum-A-Posteriori (MAP) Linear Transforms

In Chapter 4 we described how the maximum mutual information (MMI) criterion can be used for feature normalization in HMM training. A fully discriminative training procedure that estimates both the linear transform and Gaussian parameters under the MMI criterion was obtained. In this chapter we investigate the estimation of linear transforms for feature normalization under another popular criterion: the maximum a posteriori (MAP) criterion.

As we discussed in Section 1.6.3, the MAP estimation framework provides a way of incorporating prior information about the model parameters in the training process. This is particularly useful when training data are sparse and, hence, the ML approach can give inaccurate estimates. Prior density estimation issues are also addressed by the use of a hierarchical tree structure in the transform parameter space. A hierarchical tree structure can provide a better use of the adaptation data since transformations are hierarchically derived; the global transformations are being used to constrain the estimation of the more local transformations.

First, in Section 6.2, we will review the maximum a posteriori linear regression (MAPLR) [118] adaptation algorithm. MAPLR is a Bayesian counterpart of the maximum likelihood linear regression (MLLR) [74, 75] adaptation. While MLLR is based on ML estimation, MAPLR is based on MAP estimation. We will also review in Section 6.3, the structural maximum a posteriori linear regression (SMAPLR) [119] adaptation algorithm. SMAPLR is a structural MAP approach to improving the MAPLR estimates of the model-space linear regressions. Finally in

Section 6.4, we will propose a novel estimation framework for feature-space transforms based on the MAP criterion inspired by the structural MAP estimation treatment of model-space transforms presented by Siohan et al. [119].

6.1 MLLR Adaptation

The goal of *Acoustic Adaptation* techniques (see Section 2.1) is to compensate for the differences between the speech on which the system was trained and the speech which it has to recognize. The maximum likelihood linear regression (MLLR) [74, 75] adaptation scheme parameterizes the state-dependent Gaussian densities of the acoustic HMMs by affine transformations of the means μ :

$$\mathcal{N}(\cdot, A\mu + b, \Sigma) = \mathcal{N}(\cdot, T\tilde{\mu}, \Sigma)$$

where $T=[A \ b]$ and $\tilde{\mu} = [1 \ \mu]^T$ is the extended mean vector.

The transformation parameters T are estimated via the ML criterion (see Section 1.6.1) from the test data $(\hat{o}_1^l, \hat{w}_1^{\hat{n}})$, called adaptation data. \hat{o}_1^l are the observed acoustic data \hat{o}_1^l and $\hat{w}_1^{\hat{n}}$ the corresponding transcription. When the correct word-string transcription is known, the method is a *supervised* adaptation scheme. In *unsupervised* adaptation the correct word-string transcription is not known. In this case, the usual approach is to first recognize the adaptation data to get the best word-string transcription. According to the ML criterion, given by formula (1.10), the transformation parameters are found as

$$\bar{T}^{(ML)} = \arg \max_T P(\hat{o}_1^l | M_{\hat{w}_1^{\hat{n}}}; T, \theta) \quad (6.1)$$

where θ are the usual HMM parameters and $P(\hat{o}_1^l | M_{\hat{w}_1^{\hat{n}}}; T, \theta)$ is the likelihood function of the adaptation data given the transformed model.

Usually, the amount of adaptation data is significantly less than what is used for the training of the initial models. To avoid introducing more parameters than can be reliably estimated, transforms are tied across sets of states. The sets of states and, therefore, the clusters of mean parameters that share a common transformation, are typically obtained by dynamically controlling the number of clusters based on the

available amount of adaptation data.

MLLR makes use of a regression class tree to group the state-dependent emission densities in the model set. The tree is constructed so that states that are close in acoustic space are clustered together and, thus, similar states are transformed in a similar way. The tying of each transform across a number of states makes possible the adaptation of all state-dependent distributions even for those that there are no observations at all. The number of transformations is controlled by the use of a minimum occupancy threshold. If only a small amount of data is available then a *global* transform can be generated. However, as more adaptation data become available, the occupancy in lower layers of the tree can exceed the predetermined threshold and, therefore, more transforms may be generated. Each transformation is now more specific and applied to finer groups of states.

The effect of this transformation is to adjust the means in the initial model so that each HMM state is more likely to have generated the adaptation data. Under this description, MLLR finds transformations that put as much probability mass as possible on the adaptation data. However, when only a small amount of adaptation data is available it may not be representative of the statistical properties of the target acoustics and may lead to the well known overtraining property of the maximum likelihood estimator. To overcome overtraining MLLR can be combined with Bayesian methods. In general, Bayesian methods introduce some constraints on the values of the parameters to be estimated. Such a technique is discussed in the next section.

6.2 MAP Linear Regression

Here, a Bayesian counterpart of the MLLR adaptation is discussed based on maximum a posteriori estimation. As we discussed in Section 1.6.3, in MAP the parameters to be estimated are assumed to be random variables and, therefore, are described by their prior density function. Under the linear regression adaptation scheme, the parameters to be estimated are the transforms. Thus, we assume that T are random variables having $p(T)$ as a prior density.

According to the MAP criterion, expressed by formula (1.22), given some adaptation data $(\hat{o}_1^j, \hat{w}_1^{\hat{n}})$ the MAP estimate of T is defined as the mode of the posterior pdf of T . That is

$$\bar{T}^{(MAP)} = \arg \max_T p(\hat{\delta}_1^l | M_{\hat{\omega}_1^{\hat{p}}}, T; \theta) p(T) \quad (6.2)$$

where θ are the usual HMM parameters and $p(\hat{\delta}_1^l | M_{\hat{\omega}_1^{\hat{p}}}, T; \theta)$ is the likelihood function of the adaptation data given the transformed model.

Note that the likelihood function $p(\hat{\delta}_1^l | M_{\hat{\omega}_1^{\hat{p}}}, T; \theta)$ in equation (6.2) is the objective function for the MLLR estimate (see equation (6.1)). Thus, the difference between the MAP and ML objective functions is the prior term in the MAP objective function. If T is assumed to be fixed but, unknown, which is equivalent to assuming a *noninformative* or *improper* prior [79], then equation (6.2) reduces to the corresponding ML equation (6.1).

Under this approach, the traditional MLLR [74] algorithm is reformulated under the Bayesian framework by introducing a prior distribution for the affine transformation matrices leading to the maximum a posteriori linear regression (MAPLR) algorithm [118]. The MAP solution is strongly related to the choice of the prior distribution family and the specification of the parameters for the prior densities. As we discussed in Section 1.6.3, the common practice in Bayesian learning is to use *conjugate* priors. However, unlike MAP estimation of HMM parameters [49], there is no conjugate prior distribution for the transformation parameters [20]. The MAPLR algorithm uses a prior density from the elliptically symmetric distribution family as suggested by Chou [20], i.e. the matrix variate normal distribution [58]:

$$g(T; \eta) \propto |\Phi|^{-(m+1)/2} |\Psi|^{-m/2} \exp\left\{-\frac{1}{2} \text{tr}(\Phi^{-1}(T - \Lambda)\Psi^{-1}(T - \Lambda)^T)\right\} \quad (6.3)$$

where $T, \Lambda \in \mathbb{R}^{m \times (m+1)}$, $\Phi \in \mathbb{R}^{m \times m}$, $\Phi > 0$, $\Psi \in \mathbb{R}^{(m+1) \times (m+1)}$, $\Psi > 0$. The parameters of the matrix variate normal distribution $g(\cdot; \eta)$ are $\eta = \{\Lambda, \Psi, \Phi\}$. Using $g(\cdot; \eta)$, as a prior for the transformation parameters, the MAP estimate of the transformation can be derived under closed form via the EM algorithm.

The matrix variate normal distribution, which is defined above in equation (6.3), is related to the location-scale family of distributions [90]. A random variable X belongs to the location-scale family of distributions if its cumulative distribution function $F(\cdot)$ can be expressed as

$$Pr(X \leq x) = F(x; \mu, \sigma) = \Omega\left(\frac{x - \mu}{\sigma}\right), \quad (6.4)$$

where Ω does not depend on any unknown parameters. In this case we say that $-\infty < \mu < \infty$ is a location parameter and that $\sigma > 0$ is a scale parameter. Substitution shows that Ω is the cumulative distribution function of X when $\mu = 0$ and $\sigma = 1$. Also, Ω is the cumulative distribution function of $(X - \mu)/\sigma$. For the matrix variate normal distribution (see equation (6.3)), Λ is the location parameter and Ψ and Φ are the scale parameters.

Although the MAP estimation approach provides a framework to overcome estimation problems posed by sparse training data, the need for a prior distribution makes the parameter estimation more complex. This is because additional parameters, the hyperparameters, are needed to describe the prior distribution itself. The solution proposed in MAPLR is to derive the parameters of the prior distribution from the speaker independent models. Despite providing more robust estimates than MLLR, this ad-hoc approach might not give an accurate prior distribution. As an alternative, prior densities for the transformation parameters can be hierarchically structured in a tree, as proposed by Siohan et al. [119]. This structured prior information approach is reviewed in the next section.

6.3 Structural MAP Linear Regression

Siohan et al. [119] proposed the structural maximum a posteriori linear regression (SMAPLR) adaptation technique. SMAPLR extends the MAPLR technique by adding structure to the transformation priors. In MAPLR, the regression tree is only used to define the transformation clusters. The same tree can also be used to establish a hierarchical prior framework over the transforms used in MAP estimation. In SMAPLR, it is assumed that the prior density $p(T_k)$ for estimating a transform at tree layer k , is based on some knowledge about the corresponding parent node at layer $k - 1$. Specifically, the prior density $p(T_k)$ is selected from the matrix variate normal family $g(\cdot)$ (see equation 6.3). Then, the prior $p(T_k)$ is specified by parameters estimated under the posterior distribution $p(T_{k-1}|\hat{o}_{k-1}, \hat{w}_{k-1})$ of the corresponding parent node at layer $k - 1$. The pair $(\hat{o}_{k-1}, \hat{w}_{k-1})$ denotes the training data associated to the node at tree layer $k - 1$. Following a top-down process, as

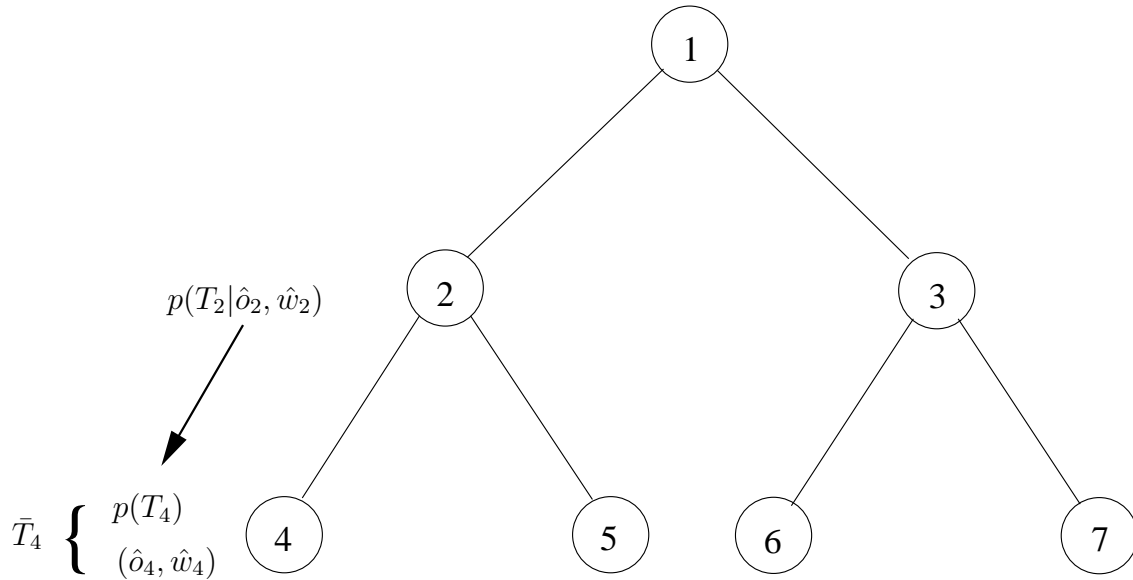


Figure 6.1: Tree-structured SMAPLR algorithm. The adaptation data associated to a node i is denoted $(\hat{\omega}_i, \hat{w}_i)$. The corresponding prior density is denoted $p(T_i)$. For each children i of parent j , the prior density $p(T_i)$ is specified by parameters estimated under the posterior distribution $p(T_j|\hat{\omega}_j, \hat{w}_j)$ of the parent. From Siohan et al. [119].

illustrated in Figure 6.1, the transforms associated with each node in the tree can be estimated using the same prior/posterior principle.

Under this hierarchical tree structure, the estimation in a given node is constrained via the prior density, which is specified by parameters estimated under the posterior distribution of the corresponding parent node. As we move toward the leaf nodes, the data associated with the corresponding nodes become sparser. The prior densities become dominant and therefore the posterior densities are less modified. This prevents overtraining even for transforms derived close to leaf nodes, since the prior information has been obtained from the top nodes.

Given this hierarchical prior framework of the SMAPLR scheme, the selection or estimation of the parameters of the matrix variate normal priors $\eta = \{\Lambda, \Psi, \Phi\}$ is briefly reviewed next. As we mentioned above, the prior distribution $p(T_k)$, which is modelled by $g(\cdot)$, at tree layer k is specified by parameters estimated under the posterior distribution $p(T_{k-1}|\hat{\omega}_{k-1}, \hat{w}_{k-1})$ of the corresponding parent node at layer $k-1$. Specifically, the mode Λ_k of the prior $p(T_k)$, which is also the location-parameter as we mentioned in the previous section, is set as the mode of the posterior distribution

$p(T_{k-1}|\hat{\theta}_{k-1}, \hat{w}_{k-1})$. The mode of $p(T_{k-1}|\hat{\theta}_{k-1}, \hat{w}_{k-1})$ is the MAPLR estimate $\bar{T}_{k-1}^{(MAP)}$, as given by formula (6.2). Hence, the Λ_k hyperparameter is set as

$$\Lambda_k = \bar{T}_{k-1}^{(MAP)} \quad (6.5)$$

We next estimate the Ψ_k and Φ_k hyperparameters. One of the advantages of using matrix variate normal priors is that they are related to the location-scale family. As we discussed in the previous section, Λ_k is the location parameter and Ψ and Φ are the scale parameters. Thus, once the location parameter Λ_k is determined the scale of the prior distribution is controlled by the two hyperparameters Ψ_k and Φ_k . In SMAPLR the estimation of T_k is simplified by fixing Ψ_k to the identity matrix, $\Psi_k = I$, and setting Φ_k to a scaled identity matrix, $\Phi_k = c \cdot I$. Given these approximations and equation (6.5), the prior distribution (equation (6.3)) at tree layer k reduces to

$$p(T_k) \propto c^{-m(m+1)/2} \exp\left\{-\frac{1}{2} \text{tr}\left(c^{-1}(T_k - \bar{T}_{k-1}^{(MAP)})(T_k - \bar{T}_{k-1}^{(MAP)})^T\right)\right\} \quad (6.6)$$

since

$$\begin{aligned} \Psi_k^{-1} &= I^{-1} = I \\ |\Psi_k| &= |I| = 1 \\ \Phi_k^{-1} &= (c \cdot I)^{-1} = c^{-1} \cdot I \\ |\Phi_k| &= |c \cdot I| = c^m \cdot |I| = c^m \end{aligned}$$

Therefore, the scale of the prior distribution is controlled only by the scalar coefficient c .

6.4 MAP Feature-Space Transforms

We have reviewed so far an HMM adaptation technique using linear regression based on maximum a posteriori estimation. The transformation parameters were

constrained using a prior distribution. A hierarchical tree structure in the transformation parameter space was assumed. Each node in the tree represents a cluster of Gaussian emission densities sharing a common transformation for their mean parameters. The posterior distributions for the transformation parameters at one level were used to specify the prior densities for the transformation parameters at adjacent levels.

In MAP estimation the use of prior information for the transformation parameters reduces significantly the risk of overfitting the adaptation data. Moreover, the hierarchical tree structure in the transformation parameter space provides a better use of the adaptation data since transformations are hierarchically derived; the global transformations are being used to constrain the estimation of the more local transformations, i.e transformations at lower levels of the tree structure.

We now turn our attention to the estimation of affine transformations of the feature vector for feature normalization. In MLLT the feature-space transforms are derived under the ML criterion. Thus, the well known overfitting property of the ML estimators affects the MLLT estimates as well. Our goal is to derive a structural MAP estimation framework for the transformations applied to the feature vectors. This MAP framework will serve as an alternative to the ML estimation of linear transforms for feature normalization.

6.4.1 MAP Estimation of Feature-Space Transforms

As we discussed in Section 1.7, feature normalization applies affine transforms to the m dimensional observation vector o so that a normalized feature vector is found by equation (1.26). In this section, we will provide a detailed derivation for the estimation of feature-space transforms under the MAP criterion, as given by formula (6.2).

In the remainder of this section, we derive the EM update equation for the parameters of the transforms. In the acoustic hidden Markov model (HMM) framework, the HMM state sequence is unknown and usually the EM algorithm [29] is employed to derive Maximum Likelihood estimates. For example, in MLLT the parameters of the transforms are found using an EM approach [46]. Here, we are interested in the MAP estimates of the transforms instead of the ML estimates. As we mentioned in Section 1.6.3, the MAP estimate can also be formulated in an EM algorithm

as described by Dempster et al. [29]. Therefore, the MAP estimate is found by maximizing the auxiliary function of equation (1.23).

In the case of acoustic HMMs we can substitute equation (1.15) into equation (1.23). This yields

$$R(\bar{T}, T) = C + \sum_s \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; T, \theta) \log q(\hat{\zeta}_\tau | s; \theta_s, \bar{T}) + \log p(\bar{T}) \quad (6.7)$$

where the constant C is independent of the transformation parameters T ; $\gamma_s(\tau; T, \theta) = q_{s\tau}(s | M_{\hat{w}_1^{\hat{n}}}, \hat{\theta}_1^{\hat{I}}; T, \theta)$ is the conditional occupancy probability of state s at time τ given the training acoustics $\hat{\theta}_1^{\hat{I}}$ and the composite model $M_{\hat{w}_1^{\hat{n}}}$ that corresponds to the transcription $\hat{w}_1^{\hat{n}}$ and $q(\hat{\zeta}_\tau | s; \theta_s, T)$ is the reparameterized state-dependent Gaussian density of equation (1.33).

Hence, substituting equation (1.33) in the above expression gives

$$R(\bar{T}, T) = \sum_s \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; T, \theta) \left(\log(|\bar{A}|) - \frac{1}{2} (\bar{T}\zeta_\tau - \mu_s)^T \Sigma_s^{-1} (\bar{T}\zeta_\tau - \mu_s) \right) + \log p(\bar{T}) \quad (6.8)$$

In accordance with the MAPLR technique [118], we select a prior density $p(T)$ for the matrix T from the elliptically symmetric distribution family. Specifically, we use the matrix variate generalization of the multivariate normal distribution, *i.e.*, matrix variate normal distribution which is shown in equation (6.3). Hence, substituting the matrix variate normal prior (6.3) into the auxiliary function (6.8) yields (ignoring all terms independent of T):

$$R(\bar{T}, T) = \sum_s \sum_{\tau=1}^{\hat{I}} \gamma_s(\tau; T, \theta) \left(\log(|\bar{A}|) - \frac{1}{2} (\bar{T}\zeta_\tau - \mu_s)^T \Sigma_s^{-1} (\bar{T}\zeta_\tau - \mu_s) \right) - \frac{1}{2} \text{tr} (\Phi^{-1} \bar{T} \Psi^{-1} \bar{T}^T - \Phi^{-1} \Lambda \Psi^{-1} \bar{T}^T - \Phi^{-1} \bar{T} \Psi^{-1} \Lambda^T) \quad (6.9)$$

Given that Σ_s are constrained to be diagonal covariance matrices, we rewrite the auxiliary function (6.9) as

$$\begin{aligned}
R(\bar{T}, T) &= \beta \log(p_i [\bar{T}^T]_i) - \frac{1}{2} \sum_{i=1}^m ([\bar{T}]_i G_i [\bar{T}^T]_i - 2[\bar{T}]_i k_i^T) \\
&\quad - \frac{1}{2} \text{tr} (\Phi^{-1} \bar{T} \Psi^{-1} \bar{T}^T - \Phi^{-1} \Lambda \Psi^{-1} \bar{T}^T - \Phi^{-1} \bar{T} \Psi^{-1} \Lambda^T)
\end{aligned} \tag{6.10}$$

where $[T]_i$ denotes the i^{th} row of T ; p_i is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, ($c_{ij} = \text{cof}(A_{ij})$) and

$$G_i = \sum_s \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^{\hat{i}} \gamma_s(\tau; T, \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T \right) \tag{6.11}$$

$$k_i = \sum_s \frac{\mu_{s,i}}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^{\hat{i}} \gamma_s(\tau; T, \theta) \hat{\zeta}_\tau^T \right) \tag{6.12}$$

$$\beta = \sum_s \left(\sum_{\tau=1}^{\hat{i}} \gamma_s(\tau; T, \theta) \right) \tag{6.13}$$

Differentiating the auxiliary function (6.10) with respect to each row vector of $[T]_i$ and equating the result to zero, we obtain:

$$\begin{aligned}
\beta \frac{p_i}{p_i [\bar{T}^T]_i} &= [\bar{T}]_i \left(G_i + \frac{1}{2} (\phi^{-1})_{i,i} (\Psi^{-1} + \Psi^{-T}) \right) \\
&\quad - \left(k_i + \frac{1}{2} [\Phi^{-1} \Lambda \Psi^{-1}]_i + \frac{1}{2} [\Phi^{-T} \Lambda \Psi^{-T}]_i \right) \\
&\quad + \frac{1}{2} \sum_{j \neq i} [T]_j \left((\phi^{-1})_{i,j} \Psi^{-1} + (\phi^{-1})_{j,i} \Psi^{-T} \right)
\end{aligned} \tag{6.14}$$

where $(\phi^{-1})_{i,j}$ is the (i, j) element of the matrix Φ^{-1} . Rewriting equation (6.14) in a compact form

$$\beta \frac{p_i}{p_i [\bar{T}^T]_i} = [\bar{T}]_i \tilde{G}_i - \tilde{k}_i \tag{6.15}$$

where

$$\tilde{G}_i = G_i + \frac{1}{2}(\phi^{-1})_{i,i}(\Psi^{-1} + \Psi^{-T}) \quad (6.16)$$

and

$$\begin{aligned} \tilde{k}_i = k_i &+ \frac{1}{2}[\Phi^{-1}\Lambda\Psi^{-1}]_i + \frac{1}{2}[\Phi^{-T}\Lambda\Psi^{-T}]_i \\ &- \frac{1}{2}\sum_{j \neq i} [T]_j \left((\phi^{-1})_{i,j}\Psi^{-1} + (\phi^{-1})_{j,i}\Psi^{-T} \right) \end{aligned} \quad (6.17)$$

reduces to the familiar form of the MLLT equation (Equation B1.1, [46]). An iterative solution to the optimization of equation (6.15) is described by Gales [46], in which each row of T is optimized given the current value of all the other rows. It can be shown that the i^{th} row of the transformation matrix is found by

$$[\bar{T}]_i = \left(\alpha p_i + \tilde{k}_i \right) \tilde{G}_i^{-1} \quad (6.18)$$

where α satisfies a quadratic expression (Equation B1.8, [46]).

By inspecting the update equation (6.18) and the definition of \tilde{k}_i (equation (6.17)), we can see that each row of T , $[T]_i$, is optimized given the current value of all other rows and the cofactor vector p_i . Specifically, to reestimate the i^{th} row of T we need to compute the \tilde{G}_i and \tilde{k}_i accumulators and the cofactor vector p_i . Note that the computation of \tilde{k}_i (equation (6.17)) involves a weighted sum over the current values of all other rows, i.e. $[T]_j$ for all $j \neq i$, of the transform T . Also, the cofactor vector p_i in the update formula (6.18) should correspond to the current estimate of the transform. Therefore, in order for the new row $[T]_i$ to be optimized it is necessary to update \tilde{k}_i and the cofactor vector p_i , given the current value of all other rows.

6.4.2 Relationship Between MAP and ML Feature-Space Transforms

In MLLT [46] the i^{th} row of the transformation matrix is given by

$$[\bar{T}]_i = (\alpha p_i + k_i) G_i^{-1} \quad (6.19)$$

where G_i and k_i are given by equations (6.11) and (6.12) respectively.

By inspecting the MAP update equation (6.18) and the ML update equation (6.19) we can see that they are similar, except for the additional terms in the definition of \tilde{G}_i and \tilde{k}_i . These additional terms are related to the prior density. When the amount of training data is large, these terms become negligible and the MAP estimate becomes equivalent to the ML estimate. On the other hand, if only a small amount of data is available, the prior-related terms act as regularization terms to the MAP reestimation formula. As a result, it can be ensured that estimates that deviate from the prior are penalized, and are therefore robust.

6.5 Structural MAP Feature-Space Transforms

In order to carry out the MAP estimation of feature-space transforms it is necessary to specify the parameters $\eta = \{\Lambda, \Psi, \Phi\}$ for the prior densities. Following the SMAPLR approach [119], it is possible to structure the prior densities for the transforms hierarchically in a tree structure (see Figure 6.1).

The tree can be constructed using any appropriate clustering method. One possibility is to cluster the Gaussian components is the HTK regression class tree approach [136]. This implementation uses a centroid-splitting algorithm in which similar Gaussian components are determined via an Euclidian distance measure between distributions. As an alternative, Watanabe et al. [132] use the Kullback-Leibler distance function as a distance measure between the state-dependent Gaussian components.

Under this structural approach, the prior distribution $p(T_k)$ at tree layer k is specified by parameters estimated under the posterior distribution $p(T_{k-1}|\hat{o}_{k-1}, \hat{w}_{k-1})$ of the corresponding parent node at layer $k-1$. Specifically, the mode Λ_k of the prior is set to the mode of the posterior distribution $P(T_{k-1}|\hat{o}_{k-1}, \hat{w}_{k-1})$. Let j be a given node and i be its parent node. The mode of the posterior in node i is the MAP estimate \bar{T}_i of the transformation matrix T_i , as given by equation (6.18). Hence, in accordance to equation (6.5), we set $\Lambda_j = \bar{T}_i$. This approximation provides a

transformation estimate \bar{T}_j , associated with node j , that is in some sense 'close to' the MAP estimate \bar{T}_i obtained in its parent node i , through the prior constraint.

The scale of the prior distribution $g(T_j; \eta)$ is controlled by the two hyperparameters $\{\Phi_j, \Psi_j\}$. By introducing some additional assumptions for the structure of the scale parameters $\{\Phi_j, \Psi_j\}$, it becomes possible to simplify the estimation of the transforms. Following Siohan et al. [119] we can fix Ψ to the identity matrix, $\Psi = I$, and set Φ to a scaled identity matrix, $\Phi = c \cdot I$. Given these approximations, the prior distribution of (6.3) reduces to equation (6.6). Furthermore, \tilde{G}_i and \tilde{k}_i can be rewritten as

$$\begin{aligned}\tilde{G}_i &= G_i + \frac{1}{2}c^{-1}(I^{-1} + I^{-T}) \\ &= G_i + c^{-1} \cdot I\end{aligned}\tag{6.20}$$

and

$$\begin{aligned}\tilde{k}_i &= k_i + \frac{1}{2}[(cI)^{-1}\Lambda I^{-1}]_i + \frac{1}{2}[(cI)^{-T}\Lambda I^{-T}]_i \\ &= k_i + c^{-1}[M]_i\end{aligned}\tag{6.21}$$

since the off-diagonal terms $(\phi^{-1})_{i,j}, \forall j \neq i$ of Φ^{-1} are zero.

By inspecting equations (6.20) and (6.21), we can see that these simplifications allow to control the scaling of the prior distribution only with a scalar coefficient c . As c decreases, the influence of the terms related to the prior increases. On the other hand, as c increases, the influence of the prior decreases and the MAP estimate of the transform relies mainly on the training data. At the limit, i.e. setting c to infinity, the prior information vanishes and equations (6.20) and (6.21) reduce to equations (6.11), (6.12) correspondingly. Hence, as $c \rightarrow \infty$, the MAP reestimation formula (6.18) approaches the ML one (6.19), exhibiting the asymptotical similarity of the two estimates.

In summary, at a given node j first the location parameter Λ_j of the prior is determined by setting $\Lambda_j = \bar{T}_i$, where T_i is the MAP estimate of the parent node i . Then, a judicious choice of the scale coefficient c ensures robustness by keeping \bar{T}_j close to \bar{T}_i . One possibility is to adjust the scaling factor c based on the depth of

the tree. Based on the above discussion, for the scaling of the prior by c , we could set a large c at the root node and decreasing it for the nodes close to the leaves of the tree.

The proposed structural MAP feature-space estimation algorithm is summarized in the following steps:

1. Start at the root node (indexed by 1) with an initial prior distribution $p(T_1)$ from the matrix variate normal density centered at an identity transformation at the root node. That is, set $\Lambda_1 = I$.
2. Derive the MAP estimate \bar{T}_1 given the training data (\hat{o}_1, \hat{w}_1) associated with the root node using equation (6.18).
3. Explore the tree in a breadth-first manner. For each node j of parent i :
 - (a) Set the hyperparameter Λ_j of the prior distribution to the MAP estimate \bar{T}_i of the parent node i . That is set $\Lambda_j = \bar{T}_i$.
 - (b) Derive the MAP estimate \bar{T}_j given the training data (\hat{o}_j, \hat{w}_j) associated with node j using equation (6.18).

6.6 Discussion

In Sections 6.4 and 6.5 we introduced a novel structural maximum a posteriori (MAP) estimation framework for feature-space transforms inspired by the structural MAP estimation treatment of model-space transforms presented by Siohan et al. [119]. A Bayesian counterpart of the maximum likelihood linear transforms (MLLT) was formulated based on MAP estimation. Prior density estimation issues were addressed by the use of a hierarchical tree structure in the transform parameter space. Under this structural framework, the posterior distributions for the transforms at one level were used to specify the priors for the transforms at adjacent levels. This could prevent overtraining even for transforms derived close to leaf nodes since the prior information has been obtained from the top nodes. The proposed structural MAP estimation framework for feature-space transforms is particularly useful when training data are sparse for which the ML approach, i.e. MLLT, may give inaccurate estimates.

The prior distribution of the transformation parameters T was selected from the family of elliptically symmetric distributions. In general, transformation-based adaptation schemes under elliptically symmetric priors have close-form solutions. The same prior distribution was used by Chou [20] and Siohan et al. [119] for maximum a posteriori linear regression (MAPLR) based model adaptation.

In a different approach, Wang and Zhao [131] used as a prior the generalized Gaussian density (GGD), also known as power exponential distribution. In the GGD model a shape parameter describes the exponential rate of decay and, in general, the shape or skewness of the distribution. Smaller values of the shape parameter correspond to heavier tails and therefore to more peaked distributions. The authors proposed a recursive Bayesian estimation approach for transformation-based adaptation using an EM variant with a second order iterative M step. Thus, the parameter estimation was based on a gradient ascent rather than closed-form solution. This recursive learning technique allowed the use of heavy-tailed priors, i.e. the GGD model, rather than conjugate or elliptically symmetric priors. The authors found that heavy tailed a priori density functions gave better recognition performance. However, the use of such heavy tailed priors came at the expense of a closed-form solution.

Chapter 7

Cross-Corpus Normalization Of Diverse Acoustic Data

The previous chapters discussed estimation procedures for feature-space linear transforms under three widely-used estimation criteria: Maximum Likelihood, Maximum A Posteriori and Maximum Mutual Information. In Chapter 4 we showed how the MMI criterion can be used for feature normalization in HMM training. A fully discriminative training procedure was obtained that estimates both the linear transform and Gaussian parameters under the MMI criterion. In Chapter 6 we described a structural MAP estimation framework for feature-space transforms. A Bayesian counterpart of the maximum likelihood linear transforms (MLLT) was formulated based on the maximum a posteriori estimation. This structural MAP estimation procedure for feature-space transforms provides a framework to overcome estimation problems posed by sparse training data for which the ML approach, i.e. MLLT, may give inaccurate estimates.

In this chapter we investigate the use of heterogeneous data sources for acoustic training. One of the prominent problems in modelling the process of speech communication is that of *variability*. This problem is further exaggerated by the use of diverse acoustic data. To tackle variability across acoustic sources we will employ the feature normalization techniques discussed in earlier chapters. We will describe an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data. A larger in-domain training set is created by effectively transforming the out-of-domain data before incorporation in training.

7.1 Acoustic Training from Heterogeneous Data Sources

The common refrain in automatic speech recognition (ASR) system development is that when it comes to acoustic training, there's no data like more data [91]. At the same time, data added should somehow be similar to the existing data set which in itself should be closely related to the final task to which the ASR system will be applied. Any number of contributing factors, such as language, dialect, acoustic channel, sampling rate, domain or topic, speaking style, speaker age and education, enter into the characterization of an acoustic training set.

Typically data are available from a single source, such as a single controlled data collection effort that gathers speech from a known population under somewhat controlled circumstances. This yields a relatively homogeneous collection of speech, and models trained on such a collection will perform well when incorporated into an ASR system evaluated on new speech of a similar nature. However, if a second collection of speech differs in any of these or other dimensions, for instance if the acoustic channel varies, simply adding the second collection to the first collection to create a larger acoustic training set may in fact lead to degradation in recognition performance. Loosely speaking, if the model is not able to account for the added variability, the acoustic model training process will be disrupted. This chapter focuses on simple acoustic normalization techniques that we show make it possible to augment an acoustic training set with diverse speech data that would otherwise lead to performance degradation.

Many approaches have been proposed to model unwanted variations in sampled speech and language. In Chapter 2 we reviewed several techniques applied in acoustic modeling based on linear transformations. These techniques reduce non-informative variability between speech frames or adjust the acoustic models to better match a desired condition. Language modeling can also benefit from techniques to incorporate out-of-domain data from the same language [64, 63] or in-domain data from other languages [69].

Here, we describe an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data. The approach is a straightforward application of model-based acoustic normalization techniques used to map the out-of-domain feature space onto the in-domain data. In this way, a larger in-

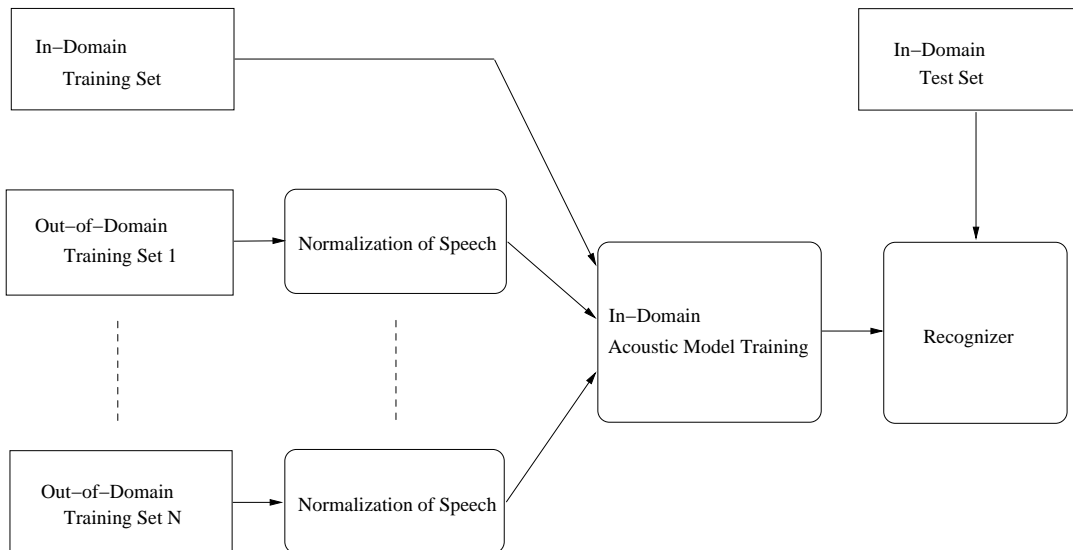


Figure 7.1: Schematic diagram of the cross-corpus acoustic normalization. Each out-of-domain feature space is transformed before being used in training.

domain training set is created by effectively transforming the out-of-domain data before incorporation in training. A schematic diagram of the proposed acoustic normalization procedure is illustrated in Figure 7.1. This acoustic normalization procedure yields a relatively homogeneous in-domain collection of speech, and models trained on such a collection will perform well when incorporated into an ASR system evaluated on new speech of a similar in-domain nature. Moreover, any appropriate adaptation scheme can be used to fine-tune the model parameters to new speech.

7.2 Cross-Corpus Normalization

We start with a collection of C training sets ($c = 1, \dots, C$), where $c = 1$ denotes the in-domain data set. Assuming that the in-domain training data are sufficiently representative of the type of speech expected to be recognized, our modelling technique transforms the out-of-domain feature space to match the space of the in-domain training population. The transformed acoustic feature vector o is found as $Ao + b$, where A is a nonsingular matrix and b is a vector. It is these transforms $[b \ A]$ that will be estimated over the out-of-domain training sets. Although this modeling approach is quite general and could be extended to a variety of normalization techniques and estimation criteria, we study only transform-based acoustic

normalization in HMMs under the maximum likelihood (ML) estimation criterion.

Following equation (1.33), the emission density of state s is reparametrized as

$$q(\zeta|s, c; \theta) = \frac{|A_{\mathcal{R}(s)}^{(c)}|}{\sqrt{(2\pi)^m |\Sigma_s|}} e^{-\frac{1}{2}(T_{\mathcal{R}(s)}^{(c)}\zeta - \mu_s)^T \Sigma_s^{-1} (T_{\mathcal{R}(s)}^{(c)}\zeta - \mu_s)}. \quad (7.1)$$

Note the dependence on c ; the observation distribution depends on the training set to which it is applied. Here $T_r^{(c)}$ denotes the extended source dependent transformation matrix $[b_r^{(c)} \ A_r^{(c)}]$ associated with states $S_r = \{s | \mathcal{R}(s) = r\}$ for classes $r = 1, \dots, R$; ζ is the extended observation vector $[1 \ o^T]^T$; and μ_s and Σ_s are the mean and variance for the observation distribution of state s . The Σ_s are constrained to be diagonal covariance matrices. We assume that the in-domain data do not need to be normalized at the corpus level, and this is in fact a key step in the modeling approach. To this end, we simply set $A_r^{(t)} = I$ and $b_r^{(t)} = \mathbf{0} \ \forall r$. The entire parameter set is specified as $\theta = (T_r^{(c)}, \mu_s, \Sigma_s)$.

Our goal is to estimate the transforms and the HMM parameters under the ML criterion. The estimation is based on the observed random process $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$ that consists of an \hat{n} -length word sequence $\hat{w}_1^{\hat{n}}$ and an \hat{l} -length sequence of m -dimensional acoustic vectors $\hat{o}_1^{\hat{l}}$. To incorporate information about the source identity into the statistical framework, we modify the observed random process to include a sequence $\hat{c}_1^{\hat{l}}$ that labels each observation vector by the source that produced it: $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}}, \hat{c}_1^{\hat{l}})$. The training objective therefore becomes the maximization of $p(\hat{o}_1^{\hat{l}} | M_{\hat{w}_1^{\hat{n}}}, \hat{c}_1^{\hat{l}}; \theta)$. This estimation is performed as a two-stage iterative procedure. At each iteration, we first maximize the ML criterion with respect to the affine transforms while keeping the Gaussian parameters fixed, and then reestimate the Gaussian parameters using the updated values of the normalizing transforms.

Maximum likelihood reestimation of the parameters is performed using the Expectation-Maximization (EM) [29] algorithm. The EM auxiliary function, as given by equation (1.15), can be modified to include the source identity as follows

$$Q(\bar{\theta}, \theta) = \sum_{r,c} \sum_{s \in S_r} \sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \theta) \log q(\hat{\zeta}_\tau | s, c; \bar{\theta}) + C. \quad (7.2)$$

Here, C is a constant independent of θ coefficients and $\gamma_s(\tau; \theta) = q_{s_\tau}(s | M_{\hat{w}_1^{\hat{n}}}, \hat{o}_1^{\hat{l}}, \hat{c}_1^{\hat{l}}; \theta)$

is the conditional occupancy probability of state s at time τ given the training acoustics, transcription and source identity.

Using calculus to do the maximization yields the following update rule to be satisfied by the parameter estimation procedures: given a parameter estimate θ , a new estimate $\bar{\theta}$ is found so as to satisfy

$$\bar{\theta} : \sum_{r,c} \sum_{s \in S_r} \sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \theta) \nabla_{\theta} \log q(\hat{\zeta}_\tau | s, c; \bar{\theta}) = 0. \quad (7.3)$$

We will show in the subsequent sections how this estimation criterion can be used for cross-corpus normalization of heterogeneous data sources in HMM training.

7.2.1 Corpus-Normalizing Transform Estimation

In the first part of the two-stage estimation procedure we fix the HMM means and variances and maximize the ML criterion with respect to the affine transforms. The presentation incorporates Gales' [46] treatment of MLLT. The reestimation formula for the transforms is derived from the update relationship of equation (7.3). The optimization is performed on a row by row basis, and therefore, the derivation involves the differentiation of the reparameterized emission density q (equation (7.1)) with respect to the i^{th} row of $T_r^{(c)}$, denoted by $[T_r^{(c)}]_i$.

Substituting the expression for the reparameterized emission density (equation (7.1)) into the update relationship of equation (7.3) and ignoring all terms independent of $[T_r^{(c)}]_i$ yields

$$[\bar{T}_r^{(c)}]_i : \sum_{s \in S_r} \sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \theta) \nabla_{[T_r^{(c)}]_i} \left(\log |A_r^{(c)}| - \frac{1}{2} \left(T_r^{(c)} \zeta_\tau - \mu_s \right)^T \Sigma_s^{-1} \left(T_r^{(c)} \zeta_\tau - \mu_s \right) \right) = 0. \quad (7.4)$$

Given that the covariance matrices Σ_s are constrained to be diagonal it is possible to rewrite the update relationship of equation (7.4) as

$$\begin{aligned}
[\bar{T}_r^{(c)}]_i : \beta \nabla_{[T_r^{(c)}]_i} \log \left(p_i [T_r^{(c)}]_i \right) \\
- \frac{1}{2} \nabla_{[T_r^{(c)}]_i} \sum_{i=1}^m \left([T_r^{(c)}]_i G_i [T_r^{(c)T}]_i - 2 [T_r^{(c)}]_i k_i^T \right) = 0 \quad (7.5)
\end{aligned}$$

where p_i is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, ($c_{ij} = \text{cof}(A_{ij})$) and

$$\begin{aligned}
G_i &= \sum_{s \in S_r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T \right) \\
k_i &= \sum_{s \in S_r} \frac{\mu_{s,i}}{\sigma_{s,i}^2} \left(\sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \theta) \hat{\zeta}_\tau^T \right) \\
\beta &= \sum_{s \in S_r} \left(\sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \theta) \right)
\end{aligned}$$

Calculating the gradient $\nabla_{[T_r^{(c)}]_i}$ in the update relationship of equation (7.5) yields:

$$\beta \frac{p_i}{p_i [\bar{T}_r^{(c)T}]_i} = [\bar{T}_r^{(c)}]_i G_i - k_i. \quad (7.6)$$

An iterative solution to the optimization of equation (7.6) is described by Gales [46], where each row of $T_r^{(c)}$ is optimized given the current value of all the other rows. It can be shown that the i^{th} row of the transformation matrix is found by

$$[\bar{T}_r^{(c)}]_i = (\alpha p_i + k_i) G_i^{-1} \quad (7.7)$$

where α satisfies a quadratic expression (Equation B1.8, [46]).

7.2.2 Gaussian Parameters Estimation

This section describes the estimation scheme for the state dependent Gaussian means and variances under the ML criterion. With the transforms estimated as

described in Section 7.2.1, we denote the parameter set as $\tilde{\theta} = (\bar{T}_r^{(c)}, \mu_s, \Sigma_s)$. The parameter update relationship of equation (7.3) can be simplified by noticing that each of the states is uniquely assigned to one of R disjoint transform classes S_r , according to the relation $\mathcal{R}(s) = r$

$$(\bar{\mu}_s, \bar{\Sigma}_s) : \sum_c \sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \tilde{\theta}) \nabla_{\mu_s, \Sigma_s} \log q(T_{\mathcal{R}(s)}^{(c)} \hat{\zeta}_\tau | s, c; \tilde{\theta}) = 0. \quad (7.8)$$

Here the posteriors $\gamma_s(\tau; \tilde{\theta})$ are estimated for each state using the new transform estimates and old Gaussian model parameters. To simultaneously update the Gaussian means and variances in the same pass we will take the derivative of the state dependent emission density with respect to μ_s and Σ_s .

Mean estimation

The gradient of $\log q(\zeta | s, c; \tilde{\theta})$ with respect to the parameter component μ_s is given by

$$\begin{aligned} \nabla_{\mu_s} \log q(\zeta | s, c; \tilde{\theta}) &= \nabla_{\mu_s} \left(-\frac{1}{2} \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right)^T \Sigma_s^{-1} \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right) \right) \\ &= \Sigma_s^{-1} \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right) \end{aligned}$$

Substituting into equation (7.8) and rearranging gives the update equation for μ_s

$$\bar{\mu}_s = \frac{\sum_c \sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \tilde{\theta}) \bar{T}_{\mathcal{R}(s)}^{(c)} \hat{\zeta}_\tau}{\sum_c \sum_{\tau: \hat{c}_\tau = c} \gamma_s(\tau; \tilde{\theta})} \quad (7.9)$$

Variance estimation

The gradient of $\log q(\zeta | s, c; \tilde{\theta})$ with respect to Σ_s^{-1} is given by

$$\begin{aligned} \nabla_{\Sigma_s^{-1}} \log q(\zeta | s, c; \tilde{\theta}) &= \nabla_{\Sigma_s^{-1}} \left(\log |\Sigma_s| - \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right)^T \Sigma_s^{-1} \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right) \right) \\ &= \Sigma_s - \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right) \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \zeta - \mu_s \right)^T \end{aligned}$$

Substituting the gradient into equation (7.8) yields

$$\sum_{\tau:\hat{c}_\tau=c} \gamma_s(\tau; \tilde{\theta}) \left(\bar{\Sigma}_s - \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \hat{\zeta}_\tau - \bar{\mu}_s \right) \left(\bar{T}_{\mathcal{R}(s)}^{(c)} \hat{\zeta}_\tau - \bar{\mu}_s \right)^T \right) = 0$$

Using the fact that $\bar{\mu}_s$ is given by equation (7.9) we can obtain the reestimation formula for the new estimate of $\bar{\Sigma}_s$ as

$$\bar{\Sigma}_s = \frac{\sum_c \sum_{\tau:\hat{c}_\tau=c} \gamma_s(\tau; \tilde{\theta}) \bar{T}_{\mathcal{R}(s)}^{(c)} \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}_{\mathcal{R}(s)}^{(c)T}}{\sum_c \sum_{\tau:\hat{c}_\tau=c} \gamma_s(\tau; \tilde{\theta})} - \bar{\mu}_s \bar{\mu}_s^T. \quad (7.10)$$

This concludes the estimation procedure for the parameters of the cross-corpus normalization model. We have presented a two-step, iterative procedure. The transforms are estimated via equation (7.7) which is iterated until the $[T_r^{(c)}]_i$ parameters converge. After this the Gaussian parameter estimates are found via equations (7.9) and (7.10).

7.3 Modelling Speaker Variation within Cross-Corpus Normalization

The proposed cross-corpus acoustic normalization attempts to reduce corpus-wide variations. However, a training corpus usually consists of speech collected from a population of speakers. Due to the wide range of speakers in a typical training corpus, the corpus-normalizing transforms also have to cope with inter-speaker variability. The factors which contribute to speaker variation can be split into two categories [97]: *acoustic* such as realisational, duration, intonation and physiological factors, and *phonological* such as lexical and stress differences factors. Acoustic variations can be modelled by feature-normalizing transforms while the phonological differences can be handled by grammar or pronunciation models [105, 110]. In the following sections we will discuss some extensions of the cross-corpus normalization procedure that attempt to minimize speaker-to-corpus variations.

7.3.1 Maximum Likelihood Speaker-to-Corpus Normalization

A straightforward extension of the cross-corpus normalization technique is obtained by using a divide-and-conquer procedure. Rather than applying a common normalizing transform for all the data in a out-of-domain corpus, we divide the corpus into homogeneous clusters of speakers, or even by each speaker separately. The cross-corpus normalization procedure of Section 7.2 can be easily reformulated to account for speaker-dependent normalizing transforms. To incorporate information about the speaker identities into the acoustic normalization framework, we modify the observed random processes to include a sequence \hat{k}_1^l that labels each observation vector by the speaker who uttered it: $(\hat{o}_1^l, \hat{k}_1^l, \hat{w}_1^{\hat{n}})$. The training objective therefore becomes the maximization of $p(\hat{o}_1^l | M_{\hat{w}_1^{\hat{n}}}, \hat{k}_1^l; \theta)$ where $M_{\hat{w}_1^{\hat{n}}}$ is the composite HMM model corresponding to the word sequence $\hat{w}_1^{\hat{n}}$. Similar to the cross-corpus normalization procedure, we assume that the in-domain speakers does not need to be normalized, and therefore we simply set $A_r^{(k)} = I$ and $b_r^{(k)} = \mathbf{0}$ for all speakers k in the in-domain set. The re-estimation formulae for the model parameters can be readily obtained by following the derivation of Section 7.2. As a result this modelling technique transforms the feature space of each out-of-domain speaker to match the space of the in-domain training population.

The above ML speaker-to corpus normalization scheme is in fact a special case of the feature-space SAT [46] (see also Section 2.2). Our scheme differs from the feature-space SAT on the treatment of the in-domain speakers. The feature-space SAT applies normalizing transforms to both in-domain and out-of-domain speakers, while our scheme normalizes the out-of-domain speakers only. In this way, the ML speaker-to corpus normalization scheme is aimed at reducing non-linguistic differences between each out-of-domain speaker and the in-domain speaker population.

However, the amount of training data for each speaker in a training collection varies significantly and there might be speakers with limited training data. As discussed in Chapter 6, small amounts of training data may not be representative of the statistical properties of the speaker acoustics and may lead to the well known overtraining property of the maximum likelihood estimator. It is therefore desirable to constrain the possible values of the transformation parameters to avoid getting biased and inaccurate estimates.

One possible solution to the overtraining problem is to group the out-of-domain training speakers into several clusters and estimate a common transform for all the speakers in a cluster. While this approach can give well-trained transforms in the expense of modelling resolution, it does not utilize the cross-corpus normalizing transforms. A better solution could be obtained if we use the cross-corpus transforms as prior knowledge for the estimation of speaker-dependent transforms. This approach is discussed in the next section.

7.3.2 Structural MAP Speaker-to-Corpus Normalization

Our goal is to obtain as accurate speaker-dependent transforms as possible without sacrificing model robustness. Thus, rather than tying transforms across groups of speakers to avoid overfitting, we adopt the structural MAP estimation procedure proposed in Section 6.5. This choice is motivated by three observations.

First, the MAP criterion provides a way of incorporating the cross-corpus transforms as prior knowledge about the possible values of the speaker-dependent transform parameters. As we mentioned in Section 1.6.3, the use of the prior adds a regularization term to the reestimation equation, and penalizes estimates that deviate from the prior. Hence, the MAP criterion can reduce the risk of overfitting for speakers with limited amount of data.

Second, the tree structure may provide a better use of the data since transformations are hierarchically derived; the cross-corpus transforms are being used to constrain the estimation of the more local transforms.

Third, using the MAP criterion we take advantage of the asymptotic property of MAP estimation for large sizes of training data: as long as the initial MAP and ML estimates are identical, the MAP estimate converges to the same point as ML when the amount of training data approaches infinity. Therefore, for speakers with sufficient amount of data, the MAP derived normalizing transforms will be 'similar' to the corresponding well-trained ML derived transforms.

Therefore, using hierarchically structured prior densities we obtain an automatic and dynamically controlled procedure for the estimation of the speaker-dependent transforms.

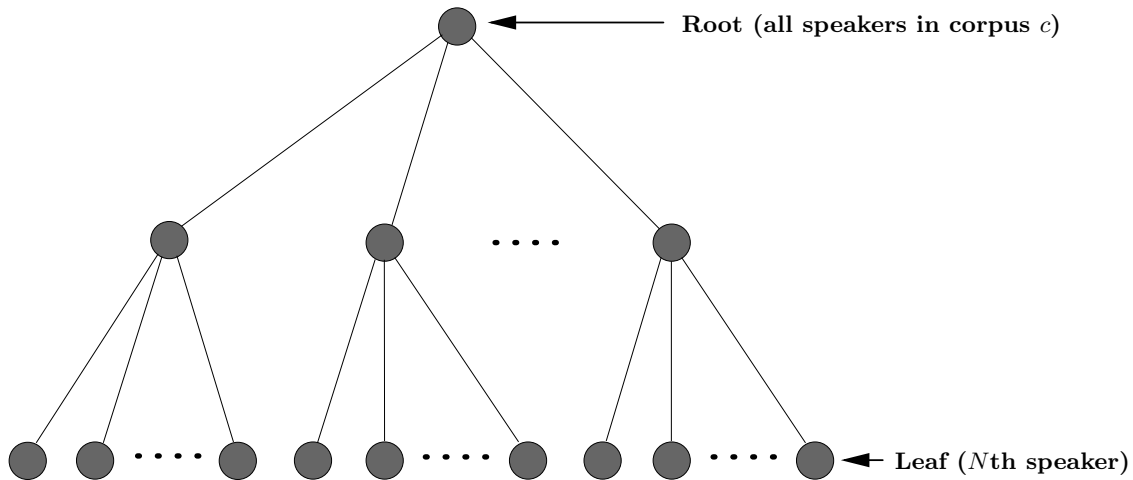


Figure 7.2: Tree structure for training speakers in out-of-domain corpus. The root node contains all the speakers in the out-of-domain corpus, and the leaf nodes contain each distinct speaker.

Hierarchical tree structure

The structural MAP estimation procedure of Section 6.5 requires the use of a hierarchical tree structure in the transform parameter space. Here we consider a tree structure for each out-of-domain corpus that hierarchically groups the speakers in each corpus, as illustrated in Figure 7.2. The root node contains all the speakers in the out-of-domain corpus, and the leaf nodes contain each distinct speaker. The tree can be constructed using any appropriate speaker clustering method [67, 70, 101, 107]. Given a hierarchical tree structure, the estimation of feature-space transforms is carried out following the procedure described in Section 6.5.

7.4 Summary

In this chapter we proposed the use of heterogeneous data sources for acoustic training. The proposed cross-corpus acoustic normalization procedure should make it possible to enlarge the acoustic training set with out-of-domain acoustic data that otherwise could possibly lead to performance degradation. To account for the variability across acoustic sources, we employed the feature normalization techniques developed in previous chapters. A larger in-domain training set could be created by effectively transforming the out-of-domain feature space to match the in-domain training population.

This acoustic normalization procedure was initially developed to reduce corpus-wide variations. However, a training corpus usually consists of speech collected from a population of speakers. Therefore, we then extended the cross-corpus normalization procedure to account for inter-speaker variations too. As a result, the feature space of each out-of-domain speaker is transformed to match the space of the in-domain train population.

The extension from coarse transforms (on the corpus level) to fine transforms (on the speaker level) is an attempt to balance the fundamental conflict between model resolution and generalizability. Coarse transforms are trained from sufficient amounts of data but can only capture corpus-wide variations. On the other hand, finer transforms describe training data better, but are estimated from significantly less data and consequently can be less general. To avoid overtraining while keeping transforms as 'sharp' as possible, we adopted the structural MAP estimation procedure of Section 6.5. This MAP estimation procedure reduces the risk of overfitting for speakers with limited amount of data, while still preserving good asymptotic properties as the size of training data increases. Moreover, it provides a way of incorporating the cross-corpus transforms as prior knowledge about the possible values of speaker-dependent transform parameters.

It should be noted that a similar feature normalization approach has already been used for unsupervised speaker adaptation by Padmanabhan et al. [101]. This speaker adaptation strategy is based on finding a subset of training speakers in the training corpus who are acoustically close to the test speaker, and then computing a set of linear transforms for each of the selected training speakers that maps the training speaker's data to the acoustic space of the test speaker. Then, the parameters of an initial speaker-independent system are reestimated using the transformed data from the selected training speakers.

This scheme provides a fair amount of gain over other adaptation schemes, such as MLLR, at the cost of higher computational complexity. The reason is that for the purpose of speaker clustering it is necessary to obtain an acoustic characterization for each training speaker. This is achieved by estimating a speaker-dependent model for each training speaker via MAP adaptation with the priors selected from the speaker-independent model. Then, the acoustic likelihood of each test speaker's data is computed using each training speaker's model. The top N speakers, in order of this likelihood, are selected. Therefore, this computational cost is a major

constraint for real-time applications. In contrast, our scheme normalizes the data during acoustic model training, and therefore, the computation load due to the normalization procedure lies in training rather than in recognition.

Chapter 8

Cross-Corpus Normalization of Mandarin Speech Corpora

The goal of the cross-corpus normalization procedure presented in Chapter 7 is to enlarge an ASR acoustic training set with out-of-domain acoustic data. This approach is a straightforward application of model-based acoustic normalization techniques to map the out-of-domain feature spaces onto the in-domain data. A larger in-domain training set can be created by effectively transforming the out-of-domain data before incorporation in training. This chapter will put the cross-corpus normalization procedure in practice by investigating the use of diverse Mandarin speech corpora for building a Mandarin Conversational Telephone Speech ASR system.

8.1 Mandarin Speech Corpora Description

Our task is to build a Mandarin Conversational Telephone Speech ASR system for the CallFriend (CF) [39] domain using data not only from the closely matched CF training corpus but also with data added from two out-of-domain sources. The out-of-domain corpora consist of the CallHome (CH) [1, 26] and Flight (FL) corpus [141]. The following list summarizes the top-level characteristics of the data sources used in acoustic and language model training:

- CallFriend corpus: Both parties located in the continental United States and Canada.
- CallHome corpus: Calls originated in the US with the other speaker(s) in

locations overseas. Due to signal transmission over international connections, channel noise and distortions occur in CallHome, especially on the overseas end [80].

- Flight corpus: Based entirely in China.
- Speakers in CallFriend and CallHome took advantage of a free phone call.
- Conversations in CallFriend and CallHome were mostly between family members and friends.
- The Flight corpus consists of telephone conversations between travel agents and customers calling to ask about flights and to make reservations.

A more detailed analysis on the Flight corpus transcription comes to the following summaries [141]:

1. Heavy background noises mainly at the operators end, as well as telephone channel noises;
2. Comparative low-volume or sometimes even unclear speech at the customers end;
3. Serious phoneme deletion and co-articulation; and
4. So severe spontaneous linguistic phenomena that sometime a long sentence with several spoken language phenomena is very difficult even for the transcribers to understand.

We now compare the three Mandarin corpora based on the amount of data available from each speaker in each Mandarin data source. Figure 8.1 consists of three subgraphs; one for each Mandarin data source. Each graph shows the histogram of the amount of data from each speaker from the corresponding data source used in acoustic model training. We can see that the amount of training data for each speaker in all three collections varies significantly and there are many speakers with limited training data. Moreover, the majority of the speakers in the Flight corpus have limited training data relative to the speakers in the CallFriend and even in the CallHome corpus.

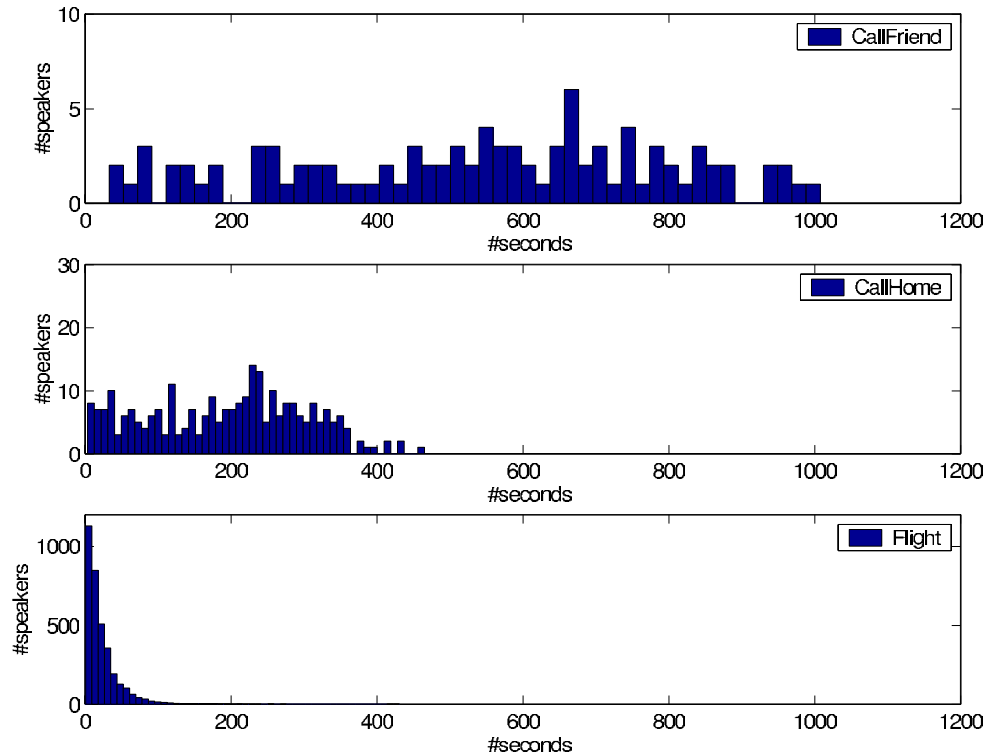


Figure 8.1: Histogram of the amount of data for each speaker in each Mandarin data source used in acoustic model training.

The three Mandarin data sources used in acoustic and language model training exhibit several differences that can potentially affect performance. First, the acoustic signal is transmitted through domestic channels (CallFriend and Flight) and international channels (CallHome). Second, even though all speakers in the three collections are native Mandarin speakers, they differ in their geographical location. This can potentially lead to pronunciation variabilities among the speakers. Third, there is a significant amount of speakers with limited training data, particularly in the Flight corpus. Finally, there is extreme variety in topics. While the topic was restricted to travel arrangements in the Flight corpus, the speakers in the CallFriend and CallHome dialogs could chose their topics freely. Thus, the vocabulary used in the Flight corpus is much more restricted than the vocabulary used in the CallFriend and CallHome corpora.

Despite these differences, since these three databases contain conversational Mandarin collected over the telephone, it is reasonable to investigate whether the CallHome and Flight data can be helpful in building ASR systems for the CallFriend domain.

8.2 ASR System Description

	Data Sources in Training	#Hours	#Conversations
In-domain	CallFriend (CF)	14	42
Out-of-domain	CallHome (CH)	14	100
	Flight (FL)	22	1790

Table 8.1: Mandarin data sources used in acoustic and language model training.

The testbed used for this research was the 1 hour CallFriend development set defined by BBN [135]. As we mentioned, the training data come from three different Chinese corpora. These are (see also Table 8.1): a 14 hour, 42-conversation CallFriend (CF) corpus; a 14 hour, 100-conversation CallHome (CH) corpus; and a 22 hour, 1790-conversation Chinese spontaneous telephone speech corpus in the flight enquiry and reservation domain (FL) [141]. Both the CF and CH collection are part of the training set defined for the EARS RT-03 evaluation.

The baseline acoustic models were built using HTK [136]. The system is a speaker independent continuous mixture density, tied state, cross-word, gender-independent, context-dependent Initial-Final (I/F), HMM system. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration components [59]. Cepstral mean and variance normalization was performed over each conversation side. The acoustic models used cross-word I/F with decision tree clustered states [136], where questions about phonetic context as well as word boundaries were used for clustering. Details of the ASR system design are available [140].

Corpus	Weight
CallFriend	0.77
CallHome	0.21
Flight	0.02

Table 8.2: Weights used for each linearly interpolated language model. The interpolation weights were chosen so as to minimize the perplexity on held-out CallFriend transcriptions.

Decoding experiments were performed using the AT&T Large Vocabulary Decoder [92], using a word bigram language model constructed as follows. Three word bigram language models were trained over each set of transcriptions and were linearly interpolated [120] to form a single bigram language model. This bigram, which con-

Data Sources in Training				CER	
CF	CH	FL	Total Hours	SI	SI+MLLR
✓			14	60.8	58.7
	✓		14	62.2	59.8
		✓	22	69.2	65.8
✓	✓		28	57.9	55.9
✓		✓	36	60.8	58.7
✓	✓	✓	50	59.3	56.6

Table 8.3: Character Error Rate (%) of baseline systems trained from various corpus combinations as evaluated on the CF test set. Results are reported with and without unsupervised MLLR speaker adaptation.

tained 116766 bigrams and 50629 unigrams, was used for all decoding experiments. The interpolation weights which are tabulated in Table 8.2, were chosen so as to minimize the perplexity on held-out CF transcriptions. The weight (0.02) on the FL language model confirms our prior observation made in Section 8.1, namely, how different the FL transcriptions are from the CF transcriptions.

8.3 Unnormalized Out-of-Domain Acoustic Data

Initial baseline experiments were performed to measure the performance of models trained using each of the three training sources. Various training sets were created through combinations of the sources without cross-corpus normalization. Table 8.3 summarizes the performance of ASR systems estimated over these training sets. Results on the CF test set, both with and without unsupervised MLLR speaker adaptation [74], are given.

We first conducted baseline experiments to quantify the mismatch between each of the three training corpora and the test corpus in terms of recognition performance. Not surprisingly, acoustic models trained only with CF gave the best performance on the CF test set (CER 60.8%/58.7%). The CH trained acoustic models had poorer but comparable performance (CER 62.2%/59.8%). On the other hand, the FL-based acoustic models were significantly worse than either CF or CH models (CER 69.2%/65.8%).

We then investigated the combination of each of the out-of-domain corpora and the in-domain corpus in training acoustic models for the CF task. The acoustic models were obtained by pooling the in-domain training data with each out-of-

Data Sources & Normalization				CER	
CF	CH	FL	#transforms	SI	SI+MLLR
I	T		1 per corpus	57.6	55.8
I	I	T	1 per corpus	58.1	55.7
I	T	T	1 per corpus	57.8	55.5

Table 8.4: Character Error Rate (%) of systems by normalizing out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that a source was included in training with / without normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation.

domain training data and estimating the HMM parameters in the standard ML fashion, i.e. without the cross-corpus normalizing transforms. It was found that a simple merging of the CF and CH data yielded an improvement relative to using either corpus alone. However, adding the FL set to the CF data gave absolutely no improvement relative to using the CF data alone. Moreover, adding the FL set to the CF and CH sets degrades performance relative to training with CF and CH alone.

The results of this section show that the performance of acoustic models trained from a combination of in-domain and out-of-domain data depends on the similarity of each training set to the test set. Simply adding out-of-domain data can actually degrade performance.

8.4 Cross-Corpus Normalized Out-of-Domain Acoustic Data

We next conducted a series of experiments to assess the effectiveness of cross-corpus acoustic normalization as proposed in Section 7.2. This procedure does need a starting point from which the initial set of transforms can be estimated. All normalization experiments are seeded by the CF+CH system of Section 8.3, which was trained over the combined CF and CH training sets and was best of the unnormalized systems (CER 57.9%/55.9%). A single transform was estimated for each out-of-domain corpus in these preliminary normalization experiments. The cross-corpus normalization experiments are reported in Table 8.4.

We first investigated the combination of the CF and CH corpora. Applying the cross-corpus normalizing transform to the CH data gave a modest 0.3% improvement

Data Sources & Normalization				Estimation Criterion	CER	
CF	CH	FL	#transforms		SI	SI+MLLR
I	T	T	1 per speaker	ML	57.7	55.2
I	T	T	1 per speaker	MAP	57.6	54.9

Table 8.5: Character Error Rate (%) of systems by normalizing on the speaker level out-of-domain acoustic training data relative to in-domain data. In the first system the transforms were estimated under the ML criterion; in the second under the MAP criterion. An ‘T’ / ‘T’ indicates that each speaker in the source was included in training with / without normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation.

relative to the unnormalized CF+CH system when no MLLR speaker adaptation was used during decoding. However, this improvement effectively diminishes with the presence of speaker adaptation on the test side.

We then added the FL set to the CF and CH sets. We initially treated the CF and CH corpora as in-domain data sources and the FL corpus as out-of-domain source. Under this scenario, only the FL data were transformed. The normalization of the FL corpus gave an improvement (CER 58.1%/55.7%) relative to the unnormalized CF+CH+FL system. Then, we applied the cross-corpus normalization to both the CH and FL corpora. Normalizing both out-of-domain data sources yielded a slightly better result (CER 57.8%/55.5%) relative to normalizing the FL corpus alone. In conclusion, the cross-corpus normalization makes it possible to improve performance by adding a severely mismatched corpus.

8.5 Speaker-to-Corpus Normalized Out-of-Domain Acoustic Data

In the previous section we applied the cross-corpus acoustic normalization procedure, as proposed in Section 7.2, to reduce corpus-wide variations between each out-of-domain data source and the in-domain CallFriend corpus. We saw that by first removing non-linguistic differences between collections, on the corpus-level, we were able to enlarge the CallFriend ASR acoustic training set and improve performance.

We next study the extensions of the cross-corpus normalization procedure, as proposed in Sections 7.3.1 and 7.3.2, that attempt to minimize speaker-to-corpus

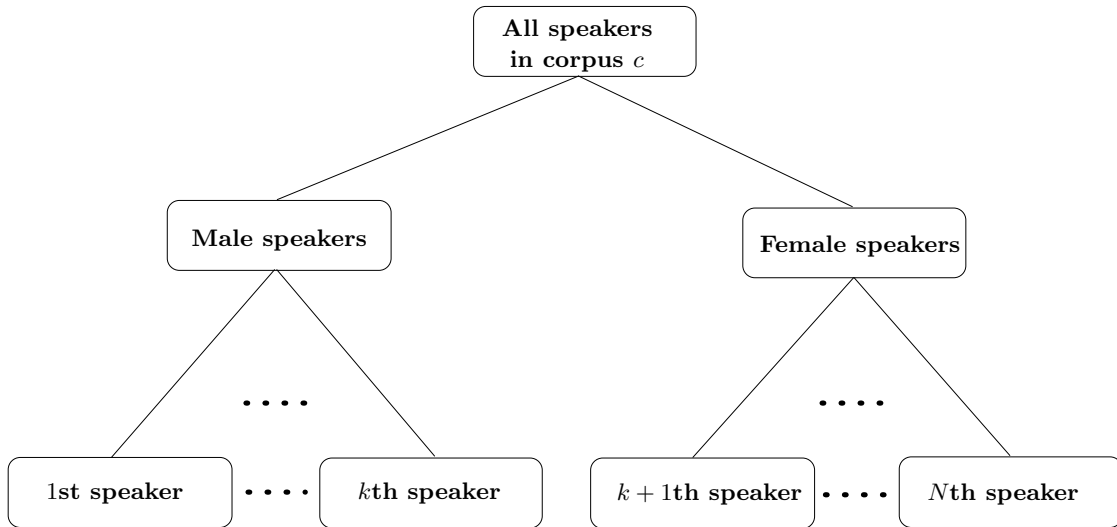


Figure 8.2: Three-level tree structure for training speakers in out-of-domain corpus c . The root node contains all the speakers in the out-of-domain corpus, the second level divides the speakers by their gender and the leaf nodes contain each distinct speaker.

variations. All normalization experiments were seeded by the CF+CH system of Section 8.3. A single transform was estimated for each out-of-domain speaker in these normalization experiments. The speaker-to-corpus normalization experiments are reported in Table 8.5.

We first applied the Maximum Likelihood speaker-to-corpus normalization scheme, as proposed in Section 7.3.1, to both the speakers in the CH and FL corpora. Normalizing on the speaker level gave not only an improvement (CER 57.7%/55.2%) relative to the unnormalized CF+CH+FL system but also an improvement relative to normalization on the corpus level CF+CH+FL system (CER 57.8%/55.5% - See Table 8.4).

However, applying the acoustic normalization technique on the speaker level rather than the corpus level implies that the transformation parameters are estimated from much smaller amounts of data. In fact, as we discussed in Section 8.1 (see also Figure 8.1), there is a significant number of speakers in the out-of-domain corpora that have limited amounts of training data. As we mentioned in Chapter 6, small amounts of training data may not be representative of the statistical properties of the speaker acoustics and may lead to inaccurate estimates.

To overcome possible overtraining, we applied the structural MAP speaker-to-corpus normalization scheme, as proposed in Section 7.3.2, again to both the speakers

in the CH and FL corpora. The structural MAP estimation procedure requires the use of a hierarchical tree structure in the transform parameter space. Here, we used a simple three-level tree, as shown in Figure 8.2, for each out-of-domain corpus. The root node (first level) contains all the speakers in the out-of-domain corpus, the second level divides the speakers by their gender and the third level (leaf nodes) contain each distinct speaker. The estimation of the transforms was based on the algorithm described in Section 6.5. The initial prior distribution at the root node of each tree was centered at the corresponding cross-corpus transformation estimate. That is, we chosen $p(T_1^{(CH)})$ and $p(T_1^{(FL)})$ such that their mode is $M_1^{(CH)} = \bar{T}^{(CH)}$ and $M_1^{(FL)} = \bar{T}^{(FL)}$, respectively. Under this structural framework, we were able to incorporate the cross-corpus transforms as prior knowledge about the possible values of the speaker-dependent transform parameters. Normalizing on the speaker level using the MAP derived transforms yielded a better result relative to normalizing on the speaker level using the ML derived transforms (CER 57.6%/54.9% vs 57.7%/55.2%).

8.5.1 Distance Measures Between Model Sets

A probabilistic distance measure, i.e. the Kullback-Leibler (KL) divergence, has been proposed to measure the dissimilarity between pairs of HMMs [68]. Using the KL divergence, we will investigate in this section how 'close' are the MAP derived transforms $T_{(MAP)}$ to the ML derived transforms $T_{(ML)}$. As we discussed in Section 1.7, the Gaussian emission densities are reparameterized by the parameters of the transforms. Hence, the usual set of HMM parameters θ (see Chapter 1) is augmented by the transforms T . Let η denote the entire HMM parameter set as $\eta = (\theta, T)$.

We fix the parameters θ to the corresponding parameters of the baseline CF+CH system of Section 8.3, which seeded the estimation of the MAP and ML derived transforms. The only parameters being modified are the parameters of the transforms. Hence, we have two sets of parameters $\eta_{(MAP)} = (\theta, T_{(MAP)})$ and $\eta_{(ML)} = (\theta, T_{(ML)})$. Rather than consider the KL divergence between pairs of complete HMM models, we use a measure that depends only on the parameters being modified, that is the parameters of the transforms. To this end, we use an average KL divergence between pairs of Gaussians on a per mixture component level, defined as

$$D(T_{(MAP)}, T_{(ML)}) = \frac{1}{\sum_{s=1}^S M(s)} \sum_{s=1}^S \sum_{m=1}^{M(s)} D_{KL} (q(\zeta|s, m; T_{(MAP)}), q(\zeta|s, m; T_{(ML)})) \quad (8.1)$$

Here, $M(s)$ is the number of mixture components in state s ; $q(\zeta|s, m; T_{(MAP)})$ and $q(\zeta|s, m; T_{(ML)})$ are the m th Gaussian mixture component of state s , which are reparameterized by $T_{(MAP)}$ and $T_{(ML)}$, respectively.

The KL divergence D_{KL} in equation (8.1) is defined as [24]

$$D_{KL} (q(\zeta|s, m; T_{(MAP)}), q(\zeta|s, m; T_{(ML)})) = \int_{\zeta} q(\zeta|s, m; T_{(MAP)}) \log \frac{q(\zeta|s, m; T_{(MAP)})}{q(\zeta|s, m; T_{(ML)})} d\zeta.$$

As shown in Appendix A, the above divergence can be expressed in a closed form as

$$\begin{aligned} D_{KL} (q(\zeta|s, m; T_{(MAP)}), q(\zeta|s, m; T_{(ML)})) &= \log \frac{|A_{(MAP)}|}{|A_{(ML)}|} \\ &\quad - \frac{1}{2} \text{tr} (J(T_{(MAP)}^T \Sigma_{m,s}^{-1} T_{(MAP)} - T_{(ML)}^T \Sigma_{m,s}^{-1} T_{(ML)})) \\ &\quad + [J]_1 (T_{(MAP)} - T_{(ML)})^T \Sigma_{m,s}^{-1} \mu_{m,s} \end{aligned}$$

where J is defined as the matrix

$$\begin{bmatrix} 1 & [A_{(MAP)}^{-1} (\mu_{m,s} - b_{(MAP)})]^T \\ A_{(MAP)}^{-1} (\mu_{m,s} - b_{(MAP)}) & A_{(MAP)}^{-1} [\Sigma_{m,s} + (\mu_{m,s} - b_{(MAP)}) (\mu_{m,s} - b_{(MAP)})^T] A_{(MAP)}^{-T} \end{bmatrix}$$

and $[J]_1$ denotes the 1st row vector of J .

In Figure 8.3 we plot the average KL divergence $D(T_{(MAP)}^k, T_{(ML)}^k)$, as defined in equation (8.1), for every pair of transforms $(T_{(MAP)}^k, T_{(ML)}^k)$ that corresponds to each speaker k in the out-of-domain corpora, against the amount of data available for the estimation of the speaker-dependent transforms. Note that for presentation purposes a logarithmic scale was used for both axes. The following observations can be made from Figure 8.3:

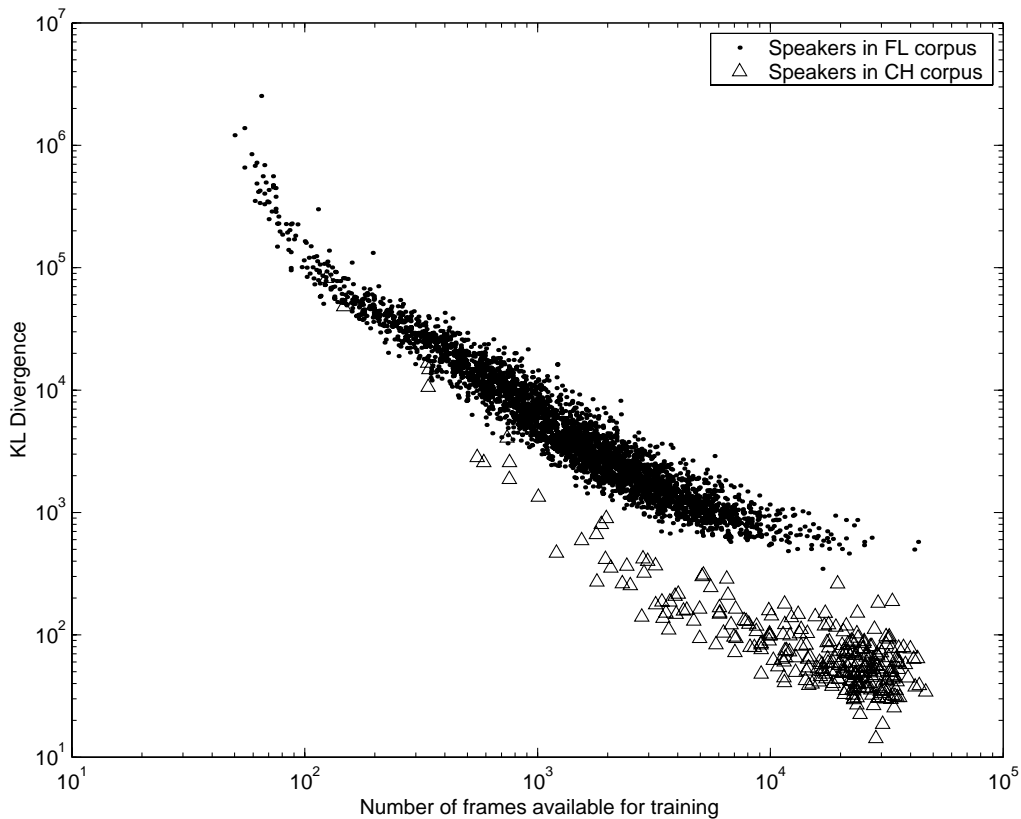


Figure 8.3: Average Kullback-Leibler (KL) divergence $D(T_{(MAP)}^k, T_{(ML)}^k)$, as defined in equation (8.1), of every pair of transforms $(T_{(MAP)}^k, T_{(ML)}^k)$ that corresponds to each speaker k in the out-of-domain corpora, plotted against the amount of data available for the estimation of the speaker-dependent transforms $T_k^{(MAP)}$ and $T_k^{(ML)}$. For presentation purposes a logarithmic scale was used for both axes.

- The KL divergence $D(T_{(MAP)}^k, T_{(ML)}^k)$ is 'on-the-average' inversely proportional to the amount of the training data. This behavior, along with the results of Table 8.5, comes into accordance with the prior theoretical discussion of Section 7.3.2. Namely, the MAP estimation procedure reduced the risk of overfitting for speakers with limited amount of data, while preserving asymptotic properties as the size of training data increased.
- For a given amount of training data the KL divergence for the speakers in the FL corpus was greater than the KL divergence for the speakers in the CH corpus. This maybe explained in part by the inclusion of the CH speakers in training the baseline system against which both transforms are estimated.

Data Sources & Normalization				CER	
CF	CH	FL	#transforms	SAT	SAT+MLLR
I	I	I		59.4	55.6
I	T	T	1 per corpus	58.0	54.6
I	T	T	1 per speaker	-	54.4

Table 8.6: Character Error Rate (%) of SAT derived systems from unnormalized and normalized out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that a source was included in speaker adaptive training with / without cross-corpus normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation.

8.6 Speaker Adaptive Training on Normalized Out-of-Domain Acoustic Data

A commonly used approach for improving ASR performance is speaker adaptive training (SAT) [2] in which speaker dependent transforms are used to reduce speaker-specific variations in the speech signal. Our training set is a collection of heterogeneous corpora, and we investigate whether cross-corpus normalization procedures can be used jointly with speaker adaptive training to improve recognition performance.

Table 8.6 compares the performance of SAT acoustic models trained over unnormalized acoustic data to SAT acoustic models trained over an in-domain training set created by transforming the out-of-domain corpora prior to speaker adaptive training. Throughout these SAT experiments we used a fixed set of two regression classes for the speaker dependent transforms - one class for speech states and one class for silence states. The first SAT system was seeded by the unnormalized CF+CH+FL system of Section 8.3 (CER 59.3%/56.6%) and subsequently trained over the unnormalized CF, CH and CH training sets. The second SAT system was seeded by the models of Section 8.4 (CER 57.8%/55.5%) which were trained over the in-domain CF data and the normalized out-of-domain CH and CF training sets. SAT training was performed as usual, but over the cross-domain normalized data.

In the following we focus on the recognition performance incorporating unsupervised speaker adaptation over the test set. Applying SAT in the standard fashion, i.e. without cross-corpus normalization, yields 1.0% absolute gain over the unnormalized CF+CH+FL system (CER 55.6% vs. 56.6% - see Table 8.3). This is comparable to the gains from cross-corpus normalization alone: in Section 8.4 we found that apply-

ing the cross-corpus normalizing transforms to both the out-of-domain corpora gave 1.1% absolute gain over the same unnormalized CF+CH+FL system (CER 55.5% vs. 56.6%). Results in Table 8.6 show that SAT can be further improved by 1.0% (CER 54.6% vs. 55.6%) if we first compensate for the cross-corpus differences across the training sets. When we consider the combined gains from SAT and cross-corpus normalization against the ML baseline system, (CER 54.6% vs. 56.6%) the total gain is 2.0%, an indication that the cross-corpus normalization and SAT procedures yield additive improvement, and are thus capturing complementary influences, as desired.

Finally, the third SAT system in Table 8.6 was seeded by the model of Section 8.5 (CER 57.6%/54.9% - See Table 8.5) which was trained over the in-domain CF data and the normalized by the MAP derived speaker-dependent transforms CH and FL training sets. Applying SAT in the standard fashion, i.e. relaxing the constraint that the transform be identity on the in-domain training corpus, gave an additional 0.5% gain (CER 54.9% vs. 54.4%). Furthermore, if we compare the third SAT system to the second SAT system (CER 54.6% vs. 54.4%) which was seeded by the models of Section 8.4 (CER 57.8%/55.5% - See Table 8.4) we observe a slightly better result. This can be attributed to the better initialization point of the third SAT system relative to the second SAT system (CER 57.6%/54.9% vs. 57.8%/55.5%).

8.7 Summary

In this chapter we investigated the use of heterogeneous data sources for acoustic training. Three diverse Mandarin speech corpora, the CallFriend, Callhome and Flight corpus were employed to build a CallFriend ASR system. Table 8.7 summarizes the main experimental results of this chapter.

Baseline experimental results (result (b) vs. result (c) in Table 8.7) showed that simply adding out-of-domain data, i.e. Flight corpus, actually degraded performance. Then, the proposed acoustic normalization procedure, which is a straightforward application of model-based acoustic normalization techniques to map the out-of-domain feature spaces onto the in-domain data, was put into practice. We applied the normalization procedure both on the corpus-level, as proposed in Section 7.2, to reduce corpus-wide variations between each out-of-domain data source and the in-domain CallFriend corpus, and on the speaker-level, as proposed in Sec-

System	Data Sources & Normalization				Estimation Criterion	CER	
	CF	CH	FL	#transforms		SI	SI+MLLR
(a)	I					60.8	58.7
(b)	I	I				57.9	55.9
(c)	I	I	I			59.3	56.6
(d)	I	T	T	1 per corpus	ML	57.8	55.5
(e)	I	T	T	1 per speaker	ML	57.7	55.2
(f)	I	T	T	1 per speaker	MAP	57.6	54.9

Table 8.7: Summary of Character Error Rate (%) of systems by normalizing out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that each speaker in the source was included in training with / without normalization, respectively. Results are reported with and without unsupervised MLLR speaker adaptation. Systems (a)-(c) are described in Section 8.3, system (d) is described in Section 8.4, and systems (e)-(f) in Section 8.5.

tions 7.3.1 and 7.3.2, to minimize speaker-to-corpus variations. All normalization experiments were seeded by the CF+CH system (result (b) in Table 8.7).

We first studied the use of single cross-corpus transforms; a single transform was estimated for each out-of-domain corpus. The cross-corpus normalization made it possible to improve performance by adding a severely mismatched corpus (result (d) in Table 8.7). Then, we applied the normalization procedure on the speaker-level. We estimated a single transform for each speaker in the out-of-domain corpora under the ML and MAP criterion. Both normalization schemes gave an improvement in performance relative to the normalized on the corpus level system (results (e) & (f) vs. result (d) in Table 8.7). However, the system with MAP derived transforms outperformed the system with ML derived transforms.

The superiority of the normalization procedure via the MAP derived transforms over the ML transforms is due to the data sparseness issues encounter in a big portion of the speaker population (see also Figure 8.1). The use of the corpus-level normalizing transforms as prior information in the structural MAP estimation procedure reduced the overfitting of the speaker-level transforms. We are able to obtain as accurate speaker-dependent transforms as possible without sacrificing model robustness. Thus, the structural MAP normalization procedure provides a robust method of normalizing out-of-domain data based on very little data while still preserving good asymptotic properties as the size of training data increases (see also Figure 8.3). Furthermore, it provides a hierarchical estimation procedure: the cross-corpus (corpus-dependent) transforms are being used to constrain the estimation of

the more local (speaker-dependent) transforms.

We have also found that cross-domain normalization can also improve Speaker Adaptive Training. Experimental results show that performing SAT over cross-corpus normalized data effectively doubles the gains obtained from SAT alone on this corpus. Interestingly, the gains from SAT and cross-corpus normalization are almost exactly additive, which is strong evidence that they are capturing different phenomena. In this, we emphasize that we are carefully employing existing modeling algorithms. What we have shown is that careful initialization and application of transform-based modeling techniques can be used to capture different effects in heterogeneous data.

There are many interesting modeling issues involved in sharing speech, especially across languages [115, 129], such as the varying effect of phonetic context, the presence or absence of phones, and other issues such as the role of prosodic features such as pitch and pause duration. However, none of those detailed issues can be studied unless it is possible to work with multiple data sources without degrading the baseline performance. The cross-corpus acoustic normalization framework is meant to be a basis to enable further studies of the more subtle issues in combining multiple data sources.

Chapter 9

Minimum Risk Acoustic Clustering for Acoustic Model Combination

In Chapter 7 and 8 we investigated the use of heterogeneous data sources for acoustic training. Our goal was to build a single set of acoustic models using data from different acoustic sources. We described an acoustic normalization procedure, i.e. cross-corpus acoustic normalization, for enlarging an ASR acoustic training set with out-of-domain acoustic data. A larger in-domain training set was created by effectively transforming the out-of-domain data before incorporation in training.

In this chapter we describe procedures for combining multiple acoustic models, obtained using training corpora from different languages, in order to improve ASR performance in languages for which large amounts of training data are not available. We treat these models as multiple sources of information whose scores are combined in a log-linear model to compute the hypothesis likelihood. The model combination can either be performed in a static way, with constant combination weights, or in a dynamic way, with parameters that can vary for different segments of a hypothesis. The aim is to optimize the parameters so as to achieve minimum word error rate. In order to achieve robust parameter estimation in the dynamic combination case, the parameters are defined to be piecewise constant on different phonetic classes that form a partition of the space of hypothesis segments. The partition is defined, using phonological knowledge, on segments that correspond to hypothesized phones. We examine different ways to define such a partition, including an automatic approach

that gives a binary tree structured partition which tries to achieve the minimum word error rate with the minimum number of classes.

9.1 Multilingual Acoustic Modeling

Multilingual acoustic modeling is motivated by the need for speech recognizers in languages and dialects for which acoustic training data is not available in large quantities. The goal of multilingual acoustic modeling is to improve ASR performance in a target language by borrowing models and data from other languages.

Previous work has largely focused on the task of building a single set of target language acoustic models using data from different source languages. For example, cross-lingual phonetic mappings between the source and target languages can be created so that a pool of multilingual data can be used to train a single set of multilingual acoustic models [112]. Alternatively, well-trained acoustic models from a source language can be adapted to the target language using standard acoustic adaptation techniques [44, 96].

An alternative methodology has recently been presented [17, 72] that extends the above procedures by using discriminative model combination techniques (DMC) [14]. Rather than merging source language acoustic data to train a single system, this technique produces a likelihood score for a recognition hypothesis by combining the scores produced for the hypothesis segments by multiple, independent, source language ASR systems. First, mappings are derived from the phones of the source languages to those of the target language. This allows mapping source language acoustic models onto target language speech. The mapped source models are then adapted by MLLR/MAP on a small amount of transcribed target language acoustic data. The best possible monolingual target language system is trained using the available data in the target language. Using this monolingual system, N-Best lists are produced for the test set. Finally, these N-Best lists are rescored by the adapted source language ASR systems and these scores are combined to produce a new likelihood for each hypothesis. The new best hypothesis is chosen based on this rescored.

Here we discuss refinements of this last, crucial step, i.e. the optimal combination of the available acoustic models from different languages. We treat these source language acoustic models as independent information sources that provide separate,

independent scores for each hypothesis, which are combined in a log-linear (GLM) model. The weights of this log-linear combination can either be *static* or *dynamic*. Static weights are optimally determined for each source language system (on held out data) and are held constant for all hypothesized segments. Dynamic weights allow the different language systems to contribute variably, as a function of the hypothesis. We compare the static and the dynamic combination approaches and discuss dynamic weighting algorithms in detail.

9.2 Log-linear Combination of Multiple Information Sources

The DMC approach [14] was used in previous work [17, 72] to combine the multiple monolingual ASR systems. DMC aims at an optimal integration of all possible information sources (in our case the acoustic models from multiple languages and a language model for the target language) into one log-linear posterior probability distribution. The weights of the scores for each information source can either be constant (static combination) or variable across the hypothesis (dynamic combination).

9.2.1 Static combination

Assume we have available m knowledge sources from which we can obtain scores for a hypothesis W . We define the probability $P(W|\mathcal{I})$, where \mathcal{I} is the information available to the sources, which includes but is not limited to the acoustic vector O , to be an exponential combination of the scores $S_i(W|\mathcal{I})$ obtained using each of the information sources. In the following we will denote these scores just with $S_i(W)$:

$$P(W|\mathcal{I}) = \frac{1}{Z(\Lambda, \mathcal{I})} \prod_{i=1}^m S_i(W)^{\lambda_i} \quad (9.1)$$

where $Z(\Lambda, \mathcal{I})$ is a normalization factor so that the probabilities for all $W \in \mathcal{H}$ add to one. Restricting the space of hypotheses in a finite set \mathcal{H} is equivalent with doing N-Best [113] or word lattice rescoring [133], where the set \mathcal{H} is the set of the highest scored hypotheses.

9.2.2 Dynamic combination

Here we combine the scores from the available information sources dynamically, within the simple form of an exponential model, by weighting each of the scores with different exponents, for different segments of a hypothesis. This allows the source language models to contribute variably depending on the hypothesis. For instance, one language may approximate a set of target language phones particularly well, while other target language phones are not modelled well at all. Thus we want the weights with which the scores from each language are combined to depend on the identity of the phones in the different hypotheses.

Each source model can have a different time alignment for the same hypothesis and thus the segments might correspond to different acoustic intervals. We denote as w_{ij} the j th segment for hypothesis W corresponding to the segmentation associated with the i th source. Then we can define:

$$P(W|\mathcal{I}) = \frac{1}{Z(\Lambda, \mathcal{I})} \prod_{i=1}^m \prod_{j=1}^{K_i(W)} S_i(w_{ij})^{\lambda_i(w_{ij})} \quad (9.2)$$

where the exponent for the score of each segment is a function of the segment. The robust optimization of the parameters $\lambda(\cdot)$ is the focus of the remainder of this chapter.

Since we would like to have a small number of parameters $\lambda(\cdot)$ to optimize, we define a function for each source language $F_i : (\mathcal{W}) \rightarrow \{1, \dots, C_i\}$ that maps the space of (w_{ij}) into a small number of discrete classes. Then we can define for a hypothesis W the score under the i th independent source:

$$S_{ic}(W) = \prod_{w_{ij} \in W: F_i(w_{ij}, \mathcal{I})=c} S_i(w_{ij}) \quad (9.3)$$

We can rewrite (9.2) as:

$$P(W|\mathcal{I}) = \frac{1}{Z(\Lambda, \mathcal{I})} \prod_{i=1}^m \prod_{c=1}^{C_i} S_{ic}(W)^{\lambda_i(c)} \quad (9.4)$$

where we have grouped the scores for the segments of the hypothesis according to the class of each segment.

9.2.3 Optimization issues

The above defined model of equation (9.4) is used to rescore the N-Best lists and choose the MAP candidate. We train the parameters $\lambda(\cdot)$ in formulas (9.1) and (9.4) so that the empirical word error count induced by the model is minimized. Since the objective function is not smooth, gradient descend techniques are not appropriate for estimation. We use the simplex downhill method known as amoeba search [95] to minimize the word errors on a held out set [128]. We also consider the approach of Beyerlein [14] which minimizes a smooth function that approximates the expected error and has a closed form solution for the parameters of an exponential model.

In the case of dynamic combination, apart from the problem of finding the optimal parameters $\lambda(\cdot)$, we face another optimization issue: that of finding the optimal functions $F_i(\cdot)$ for each information source. Ideally we would like to jointly optimize the function and the parameters associated with it, i.e. find the partition of the space whose optimal parameters achieve the minimum number of word errors. In Section 9.3.3 we present an algorithm that approximates this search.

9.3 Multilingual Acoustic Model Combination

First we describe our experimental setup and give the results for some simple, knowledge based partitions. Then we describe the automatic approach that constructs a partition in an attempt to achieve the lowest word error rate (WER) with the minimum number of parameters $\lambda(\cdot)$.

9.3.1 Database description

As our source-language acoustic training data we used English Broadcast News obtained from LDC. The target language was Czech for which we used 1.0 hour of the Charles University Corpus of Financial News (CUCFN). This is read speech and was used in the 1999 Summer Workshop at JHU [72] to train the baseline acoustic models for Czech. We also obtained from the same corpus an extra 1.0 hour of Czech data which we use to train the combination parameters of the log-linear model. The test set was 1.0 hour of Czech Voice of America broadcasts.

The acoustic models were trained from mel-frequency, cepstral data using HTK [136]. We used state-clustered cross-word triphone HMMs with 3 states each and 6

	$\#\lambda_{AC}$		WER (%)
BASELINES			
(a)	1	A_{cz}	29.1
(b)	1	A_{en}	28.4
STATIC COMBINATION			
(c)	2	$A_{cz} + A_{en}$	27.8
DYNAMIC COMBINATION (knowledge based partition)			
(d)	6	$(V+C+S)_{cz} + (V+C+S)_{en}$	27.5
(e)	71	$(12V+27C+1S)_{cz} + (10V+20C+1S)_{en}$	26.8
(f)	28	$(12V+1C+1S)_{cz} + (10V+1C+1S)_{en}$	26.9
(g)	51	$(1V+27C+1S)_{cz} + (1V+20C+1S)_{en}$	27.2
(tree partition)			
(h)	8	tree leaves in Figure 9.1	27.0

Table 9.1: Combination of English and Czech acoustic models using different acoustic classification schemes.

gaussian mixtures per state. There were a total of 6040 shared states in the system.

1000-Best hypotheses were obtained for the training and test data using the 1.0 hour baseline Czech language triphone acoustic model.

In all the experiments described in the following sections a bigram language model [18] and word insertion penalty were included. Their weights were optimized along with the acoustic model weights $\lambda_{AC}(\cdot)$.

We can see the baseline WER result using only the one hour trained Czech acoustic models (A_{cz}) in line (a) of Table 9.1. We compare this with the WER when we used only the English acoustic models adapted on the one hour of Czech data A_{en} (b). We notice that the WER is lower with the English models. This is due to the fact that we evaluate the English model by rescoreing the N-Best obtained with the original Czech models. Decoding with the English models and the Czech language model is worse than the 1.0 hour Czech system.

9.3.2 Knowledge based partition

The first system we evaluated utilized both the English and Czech acoustic models score in a static combination (system $A_{cz} + A_{en}$ in Table 9.1, line (c)). Here the two models were combined as in formula (9.1) using two weights: λ_{cz} and λ_{en} . Thus using only one extra weight yields a significant improvement over the monolingual

systems.

Next we explored dynamic combination by examining different knowledge based partitions for the hypothesized phone segments for each information source used. We consider partitions only for the acoustic information sources; the language model receives only one weight.

The first partition simply clusters together the vowel, consonant and silence models for the English and Czech language $((V+C+S)_{cz/en})$ and assigns one weight for the phone models in each cluster. This model has 6 acoustic weights to be trained (one for each of the classes for each of the languages). The accuracy improves slightly over the static combination (line (d) in Table 9.1).

We find that by allowing each phone model to have its own weight we achieve a further improvement (Table 9.1-line (e)). In that system there are 40 weights for the Czech phone models (12V+27C+1S) and 31 weights for the English (10V+20C+1S). But we found that we could reduce the number of parameters significantly by allowing separate weights for only the vowels for each language and tying the weights for the consonants to one parameter for each model, without significantly changing WER (result (f) in Table 9.1). On the other hand when only the consonants were allowed to have separate weights the result deteriorated (result (g) in Table 9.1).

The above results suggest that we should aim to obtain the optimal partition: the minimum number of tied parameters that achieve the lowest WER.

9.3.3 Searching for optimal partition of the parameter space

The goal of this experiment is to obtain the accuracy improvements achieved in the previous section with parsimonious acoustic clustering. We use an automatic approach to find an optimal partitioning of model scores into classes. The algorithm iteratively builds a binary partition of the hypothesized phone segments using a top-down hierarchical procedure. It starts with all phones (from all acoustic models) in one class-node. Phonological questions are used repeatedly to split each available class into two sub-classes. For every node, all questions are examined: the weights $\lambda_i(c)$ are optimized for the resulted partitions so that the empirical word error count induced by the model, as given by formula (9.4), is minimized on the held out set. The optimization of the weights is described in Section 9.2.3. For each possible partition we have now trained the weights and computed the corresponding WER

on the held out set. Then, the tree grows by choosing to split the node using the question that results in classes with the best WER improvement.

We also investigated building the tree with splitting criterion the improvement in the smooth approximation of WER [14]. This objective function has a closed form solution so the tree building algorithm was much faster. However the resulting WER was higher than found with the direct minimization of WER.

The questions asked in the tree are allowed to separate classes of phones from the pool of phone-models available, and assign a separate weight to each of the classes. These questions involve either general classes of models (from both languages), such as “Is this a liquid phone?”, or “Is it in the class of phone A” (this is the class of phones (A, AA, AH, AE, AX), or they can separate a specific phone, for example “Is this the ‘EH’ Czech phone”.

In Figure 9.1 we see the binary tree partition chosen by the algorithm after finding 8 classes. We notice that the algorithm produces a right branching tree. That means that each question chosen identifies small classes (most of them consist of one phone), while the majority of phone models are still tied with one common weight. This suggests that only a few phones contribute most of the improvement in discrimination. Another observation we make from the figure is that the majority of questions involve vowels. This means that putting individual weights for some vowels is more beneficial than other classes, which is consistent with the results (f)-(g) of Table 9.1 that suggests vowels with individual weights are more beneficial than consonants.

The results in Table 9.1 show that we obtain gains comparable to the knowledge based approach with far fewer parameters. We were not successful in finding a partition that achieved a higher accuracy than the sparse partition with one phone per class. However, this technique should prove robust as more information sources are incorporated into the model.

In Figure 9.2 we plot the accuracy achieved by the system as a function of the number of classes chosen by the automatic algorithm (on both the training and the test data); the accuracy achieved by the system with the sparse phone partition (knowledge based partition) is also plotted. We notice that after 8 classes the change in WER from adding extra classes is negligible.

9.4 Discussion

In this chapter, we have presented a new approach for sub-word multilingual acoustic model combination. This approach aimed at an optimal integration of all possible information sources (in our case the acoustic models from multiple languages and a language model for the target language) into one log-linear posterior probability distribution. The weights of this log-linear combination can either be *static* or *dynamic*. Static weights were optimally determined for each source language system (on held out data) and were held constant for all hypothesized segments. Dynamic weights allowed the different language systems to contribute variably, as a function of the hypothesis.

In the case of dynamic combination, apart from the problem of finding the optimal parameters $\lambda(\cdot)$, we faced another optimization issue: that of finding the optimal functions $F_i(\cdot)$ for each information source. In Section 9.3.3 we described an automatic approach that constructs such partition in an attempt to achieve the lowest word error rate (WER) with the minimum number of parameters $\lambda(\cdot)$.

Experimental results compared the static and the dynamic combination approaches and showed that dynamic combination of multilingual acoustic phonetic classes is superior to static combination of multilingual acoustic scores. Previous attempts at multilingual acoustic clustering have been mainly employed in maximum likelihood modeling [23]. Here, we have shown that minimum risk acoustic clustering is effective in finding acoustic classes that directly minimize the word error rate using MAP decision rules.

We note in closing that the cross-corpus acoustic normalization framework of Chapter 7 can be potentially useful for multilingual ASR systems. As we mentioned, most of the aforementioned training approaches are based on multilingual speech data pooling followed by acoustic model adaptation to fit the characteristics of a target language. However, as in the experiments described earlier in Section 8.3, it often happens that simply combining data from multiple languages actually hurts ASR performance in a single language. Therefore, it can be advantageous to study multilingual ASR by first removing non-linguistic differences between collections.

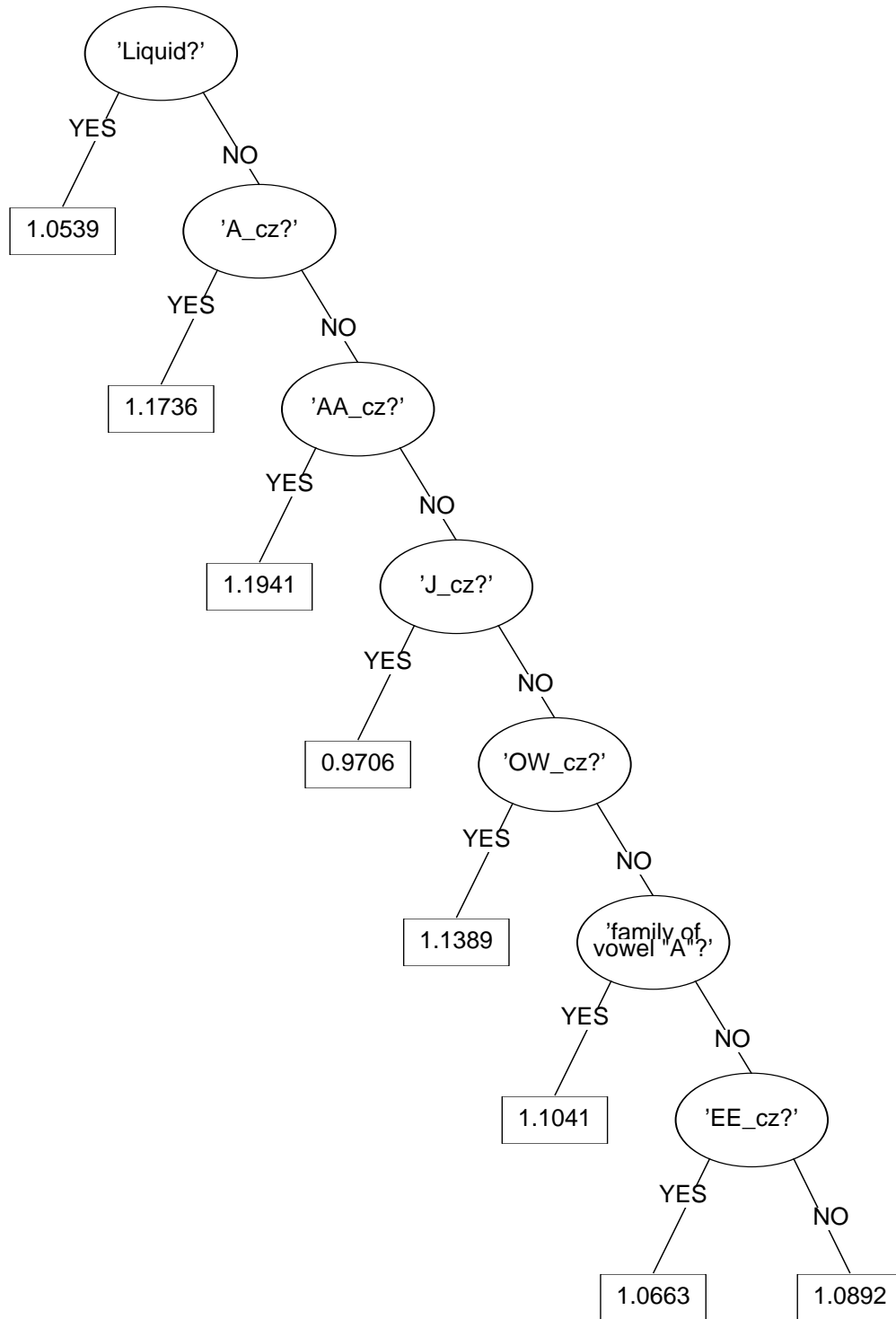


Figure 9.1: The binary tree partition constructed by the automatic partition algorithm. The class weights are shown in each leaf.

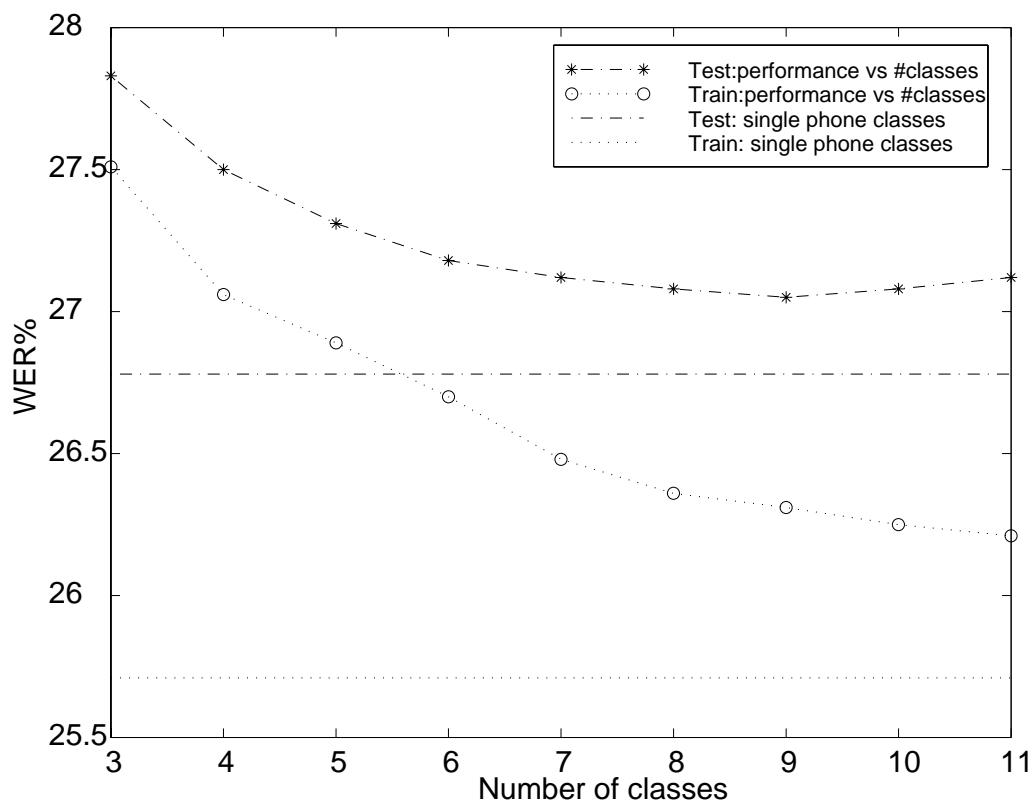


Figure 9.2: Word Error Rate (%) of systems derived with the knowledge based partition and the automatic partition algorithm of Section 9.3.3, and tested on both the training and test data as a function of the number of classes. For the knowledge based partition system each phone model has its own weight (71 classes in total).

Chapter 10

Conclusions and Future Work

10.1 Thesis Summary

As we discussed in Chapter 2, linear transforms have been used extensively in hidden Markov model (HMM) based Automatic Speech Recognition (ASR) systems for feature normalization and modelling of correlations between spectral parameters. Until recently, these transformation techniques have been based on the maximum likelihood (ML) parameter estimation framework. In this thesis, we proposed discriminative and Bayesian training approaches, as an alternative to the ML criterion, for the parameters of the linear transforms. Discriminative training, under the Maximum Mutual Information (MMI) criterion, is attractive because it directly attempts to optimize performance on the acoustic training. On the other hand, Bayesian learning provides a formulation to combine prior information for the parameters of the linear transforms, with acoustic information, as provided by the training data. In this way, Bayesian learning approaches prevent the training algorithm from finding solutions that deviate significantly from what is expected based on prior knowledge. Thus, Bayesian learning is particularly useful for dealing with problems posed by sparse training data.

In summary, this thesis placed emphasis on:

- Novel estimation procedures for the linear transforms based on different estimation criteria, i.e. the Maximum Mutual Information and the maximum a posteriori criterion, with the aim of improving the overall recognition performance of ASR systems.

- The use of linear transforms in acoustic normalization procedures for enlarging an ASR acoustic training set with out-of-domain acoustic data, in an effort to combine heterogeneous data sources for acoustic training.
- A new approach for sub-word multilingual acoustic model combination. Specifically, procedures for combining multiple acoustic models, obtained using training corpora from different languages, in order to improve ASR performance in languages for which large amounts of training data are not available.

We began this thesis with a brief overview of ASR systems and a review of the use of linear transforms along with their elementary properties. Then, in Chapter 2, we presented prior research into the application of linear transforms in ASR systems. In Chapter 3, we described the objectives of this work.

Then, in Chapter 4, we described the integration of Discriminative Linear Transforms into Maximum Mutual Information (MMI) estimation for Large Vocabulary Speech Recognition. We developed novel estimation procedures that find Discriminative Linear Transforms jointly with MMI for feature normalization and we presented reestimation formulae for this training scenario. These fully discriminative procedures were derived by maximizing the Conditional Maximum Likelihood (CML) auxiliary function. The transforms obtained under the CML criterion are termed Discriminative Likelihood Linear Transforms (DLLT).

Then, in Chapter 5, we validated DLLT as an estimation procedure and compared the proposed discriminative training technique to its Maximum Likelihood counterpart for Large Vocabulary Conversational Speech Recognition (LVCSR) tasks. DLLT was evaluated on the SWITCHBOARD corpus and gave approximately 0.8% absolute Word Error Rate improvement over the ML estimation procedure (see Table 5.2 and Table 5.6). In conclusion, DLLT provides a discriminative estimation framework for feature normalization in HMM training for LVCSR tasks that outperforms in recognition performance its Maximum Likelihood counterpart.

We also introduced in Sections 6.4 and 6.5, a novel structural maximum a posteriori (MAP) estimation framework for feature-space transforms inspired by the structural MAP estimation treatment of model-space transforms presented by Siohan et al. [119]. A Bayesian counterpart of the maximum likelihood linear transforms (MLLT) was formulated based on MAP estimation. The prior distribution of the transformation parameters was modelled by the family of matrix variate normal dis-

tributions. The proposed structural MAP estimation framework for feature-space transforms is particularly useful for dealing with problems posed by sparse training data for which the ML approach, i.e. MLLT, can give inaccurate estimates.

Then, in Chapter 7, we described an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data (see Figure 7.1). The approach was a straightforward application of model-based acoustic normalization techniques to map the out-of-domain feature space onto the in-domain data. Under this procedure, a larger in-domain training set can be created by transforming the out-of-domain feature space to match the in-domain training population.

In Chapter 8, we put the cross-corpus normalization procedure into practice by investigating the use of diverse Mandarin speech corpora for building a Mandarin Conversational Telephone Speech ASR system. Baseline experimental results (see Table 8.7) showed that simply adding out-of-domain data, actually degraded performance. We then applied the normalization procedure both on the corpus-level, as proposed in Section 7.2, to reduce corpus-wide variabilities between each out-of-domain data source and the in-domain corpus, and on the speaker-level, as proposed in Sections 7.3.1 and 7.3.2, to minimize speaker-to-corpus variabilities. The cross-corpus normalization made it possible to improve performance by adding a severely mismatched corpus (see Table 8.7) that otherwise leads to performance degradation. Also, experimental results showed that MAP estimation of linear transforms provides improved performance relative to ML estimation procedures when data sparseness is encountered. Finally, we found that cross-domain normalization can also improve Speaker Adaptive Training (SAT). The experimental results of Section 8.6 showed that performing SAT over cross-corpus normalized data effectively doubled the gains obtained from SAT alone on this corpus.

Finally, in Chapter 9, we presented a new approach for sub-word multilingual acoustic model combination. This approach aimed at an optimal integration of all possible information sources into one log-linear posterior probability distribution. The weights of this log-linear combination can either be *static* or *dynamic*. Static weights were optimally determined for each source language system (on held out data) and were held constant for all hypothesized segments. Dynamic weights allowed the different language systems to contribute variably, as a function of the hypothesis. We developed an automatic approach to find the optimal partitioning of model scores into classes. This automatic approach which gives a binary tree

structured partition tries to achieve the minimum word error rate with the minimum number of classes. The optimization of the weights was based on the simplex downhill method known as amoeba search.

Experimental results compared the static and the dynamic multilingual acoustic model combination approaches. We used a combination of English and Czech acoustic models trained from English Broadcast News and Czech from the Charles University Corpus of Financial News (CUCFN), respectively. The experimental results showed that dynamic combination of multilingual acoustic phonetic classes is superior to static combination of multilingual acoustic scores (see Table 9.1). In conclusion, we have shown that minimum risk acoustic clustering is effective in finding acoustic classes that directly minimize the word error rate using MAP decision rules.

10.2 Suggestions For Future Work

We now outline some directions for future research:

Discriminative cross-corpus acoustic normalization We have found that discriminative training under the Maximum Mutual Information criterion of linear transforms and HMM parameters for feature normalization outperforms ML training. The cross-corpus normalization procedure, which incorporates linear transforms into acoustic normalization, was based on the ML criterion. Therefore, the acoustic normalization technique can be transformed into a discriminative modeling technique, in a manner analogous to the approach we have taken in Chapter 4, in order to improve the effectiveness of the normalization procedure.

Combination of discriminative training within a Bayesian formalism We have found that MAP estimation of linear transforms provides improved performance relative to ML estimation procedures when data sparseness is encountered. Hence, it is also possible to improve MMI estimates, based on insufficient data, by incorporating prior information about the transformation parameters into the discriminative training framework.

Controlling the contribution of out-of-domain data The linear transforms, employed in the cross-corpus normalization procedure, attempt to capture non-linguistic acoustic variations between heterogeneous collections. However, in many cases the mismatch between acoustic sources is nonlinear and the functional form is unknown [131]. As an alternative to relying entirely on the modeling accuracy of the linear transforms, a weighted version of the classical ML estimation framework can be used [41, 61, 83]. A closely related approach to the weighted maximum likelihood applied to ASR systems is the *selective training* procedure by Arslan and Hansen [4] which controls the influence of outliers in the training data. We expect the primary challenge to be the identification of the out-of-domain data that needs to be downweighted and the estimation of the corresponding weights. One possibility is to use a distance that measures the deviation of the normalized observation from the current estimated model and to downweight the corresponding likelihood function according to this distance [41]. The overall goal is to control the contribution of the out-of-domain data in such a way that the accuracy of the resulting model for a specific task is maximized.

Appendix A

Measuring the Distance Between Gaussian Densities Based on Kullback-Leibler Divergence

In this Appendix we provide a detailed derivation of the average Kullback-Leibler (KL) divergence between pairs of Gaussian densities on a per mixture component level, defined as

$$D(T_1, T_2) = \frac{1}{\sum_{s=1}^S M(s)} \sum_{s=1}^S \sum_{m=1}^{M(s)} D_{KL}(q(\zeta|s, m; T_1), q(\zeta|s, m; T_2)) \quad (\text{A.1})$$

where $M(s)$ is the number of mixture components in state s ; $q(\zeta|s, m; T_1)$ and $q(\zeta|s, m; T_2)$ are the m th Gaussian mixture component of state s , which are reparameterized by T_1 and T_2 , respectively.

The KL divergence D_{KL} in equation (A.1) is defined as [24]

$$D_{KL}(q(\zeta|s, m; T_1), q(\zeta|s, m; T_2)) = E_{q_{\zeta; T_1}} \left[\log \frac{q_{\zeta; T_1}}{q_{\zeta; T_2}} \right] \quad (\text{A.2})$$

We then substitute in equation (A.2) the expression for the reparameterized Gaussian densities, as given by equation (1.33), and calculate the logarithm in equation (A.2) as

$$\begin{aligned}
\log \frac{q_{\zeta;T_1}}{q_{\zeta;T_2}} &= \log \frac{\frac{|A_1|}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2}(T_1\zeta - \mu)^T \Sigma^{-1}(T_1\zeta - \mu)\right)}{\frac{|A_2|}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2}(T_2\zeta - \mu)^T \Sigma^{-1}(T_2\zeta - \mu)\right)} \\
&= \log \frac{|A_1|}{|A_2|} - \frac{1}{2} \left((T_1\zeta - \mu)^T \Sigma^{-1}(T_1\zeta - \mu) \right) + \frac{1}{2} \left((T_2\zeta - \mu)^T \Sigma^{-1}(T_2\zeta - \mu) \right) \\
&= \log \frac{|A_1|}{|A_2|} - \frac{1}{2} \zeta^T (T_1^T \Sigma^{-1} T_1 - T_2^T \Sigma^{-1} T_2) \zeta + \zeta^T (T_1^T - T_2^T) \Sigma^{-1} \mu \\
&= \log \frac{|A_1|}{|A_2|} - \frac{1}{2} \text{tr} \left(\zeta^T \zeta (T_1^T \Sigma^{-1} T_1 - T_2^T \Sigma^{-1} T_2) \right) + \zeta^T (T_1^T - T_2^T) \Sigma^{-1} \mu
\end{aligned}$$

Given the above expression for the logarithm, the KL divergence D_{KL} of equation (A.2) is given as

$$\begin{aligned}
E_{q_{\zeta;T_1}} \left[\log \frac{q_{\zeta;T_1}}{q_{\zeta;T_2}} \right] &= E_{q_{\zeta;T_1}} \left[\log \frac{|A_1|}{|A_2|} - \frac{1}{2} \text{tr} \left(\zeta^T \zeta (T_1^T \Sigma^{-1} T_1 - T_2^T \Sigma^{-1} T_2) \right) \right. \\
&\quad \left. + \zeta^T (T_1^T - T_2^T) \Sigma^{-1} \mu \right] \\
&= \log \frac{|A_1|}{|A_2|} - \frac{1}{2} E_{q_{\zeta;T_1}} \left[\text{tr} \left(\zeta^T \zeta (T_1^T \Sigma^{-1} T_1 - T_2^T \Sigma^{-1} T_2) \right) \right] \\
&\quad + E_{q_{\zeta;T_1}} \left[\zeta^T (T_1^T - T_2^T) \Sigma^{-1} \mu \right] \\
&= \log \frac{|A_1|}{|A_2|} - \frac{1}{2} \text{tr} \left(E_{q_{\zeta;T_1}} \left[\zeta^T \zeta \right] (T_1^T \Sigma^{-1} T_1 - T_2^T \Sigma^{-1} T_2) \right) \\
&\quad + E_{q_{\zeta;T_1}} \left[\zeta^T \right] (T_1^T - T_2^T) \Sigma^{-1} \mu \\
&= \log \frac{|A_1|}{|A_2|} - \frac{1}{2} \text{tr} \left(J (T_1^T \Sigma^{-1} T_1 - T_2^T \Sigma^{-1} T_2) \right) + [J]_1 (T_1^T - T_2^T) \Sigma^{-1} \mu
\end{aligned}$$

where J is defined as the matrix

$$\begin{bmatrix} 1 & [A_1^{-1}(\mu - b_1)]^T \\ A_1^{-1}(\mu - b_1) & A_1^{-1}[\Sigma + (\mu - b_1)(\mu - b_1)^T]A_1^{-T} \end{bmatrix}. \quad (\text{A.3})$$

Bibliography

- [1] R. Agarwal, B. Wheatley, Y. Muthusamy, and T. Staples. Diagnostic profiling for speech technology development: Call Home analysis. *Proceedings of Speech Research Symposium*, pages 131–137, June 1995.
- [2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *International Conference on Spoken Language Processing*, pages 1137–1140, October 1996.
- [3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Mathematical Statistics. J. Wiley & Sons, New York, 1984.
- [4] L.M. Arslan and J.H.L. Hansen. Selective training for hidden Markov models with applications to speech classification. *IEEE Transactions on Speech and Audio Processing*, 7(1):46–54, January 1999.
- [5] B. Atal. Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55:1304–1312, June 1974.
- [6] S. Axelrod, R. Gopinath, and P. Olsen. Modeling with a subspace constraint on inverse covariance matrices. In *International Conference on Spoken Language Processing*, pages 2177–2180, September 2002.
- [7] L. R. Bahl, P.F. Brown, P. V. de Souza, and R.L. Mercer. Maximum mutual information estimation of hidden markov models parameters for speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 49–52. IEEE, April 1986.

- [8] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.
- [9] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, and M.A. Picheny. Context dependent vector quantization for continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 632–635. IEEE, April 1993.
- [10] J.K. Baker. The DRAGON system - an overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):24–29, February 1975.
- [11] R. Bakis. Continuous speech recognition via centisecond acoustic states. In *91st Meeting of the Acoustical Society of America*, April 1976.
- [12] L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73:360–363, 1967.
- [13] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563, December 1966.
- [14] P. Beyerlein. Discriminative model combination. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 481–484. IEEE, May 1998.
- [15] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day Series in Probability and Statistics. Holden-Day, Inc., Oakland, California, 1977.
- [16] W. Byrne. The JHU March 2001 Hub-5 Conversational Speech Transcription System. In *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.
- [17] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang. Towards language independent acoustic modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1029–1032. IEEE, June 2000.

- [18] W. Byrne, J. Hajic, P. Ircing, F. Jelinek, S. Khudanpur, J. McDonough, N. Paterrek, and J. Psutka. Large vocabulary speech recognition for read and broadcast czech. In *Workshop on Text, Speech and Dialog*, Marianske Lanze, Czech Republic, 1999.
- [19] J.-T. Chien. Online hierarchical transformation of hidden markov models for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7(6):656–667, November 1999.
- [20] W. Chou. Maximum a posterior linear regression with elliptically symmetric matrix variate priors. In *European Conference on Speech Communication and Technology*, volume 1, pages 1–4, September 1999.
- [21] W. Chou and X. He. Maximum a posterior linear regression based variance adaptation of continuous density HMMs. Technical Report ALR-2002-045, Avaya Labs Research, 2002.
- [22] Y.-L. Chow. Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 701–704. IEEE, April 1990.
- [23] P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward. Towards a universal speech recognizer for multiple languages. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 591 – 598. IEEE, December 1997.
- [24] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
- [25] S. J. Cox and J. S. Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 294–297. IEEE, May 1989.
- [26] T. Crystal, M. Cowing, A. Martin, and D. Pallett. LVCSR. *Proceedings of Speech Research Symposium*, pages 21–39, June 1995.
- [27] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE*

- Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(4):357–366, August 1980.
- [28] M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.
- [29] A. P. Dempster, A. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
- [30] V. Digalakis. Online adaptation of hidden Markov models using incremental estimation algorithms. *IEEE Transactions on Speech and Audio Processing*, 7(3):253–261, May 1999.
- [31] V. Digalakis, L.G. Neumeyer, and M. Perakakis. Quantization of cepstral parameters for speech recognition over the world wide web. *IEEE Journal on Selected Areas in Communications*, 17(1):82–90, January 1999.
- [32] V. Digalakis, S. Tsakalidis, C. Harizakis, and L. Neumeyer. Efficient speech recognition using subvector quantization and discrete-mixture HMMs. *Computer Speech and Language*, 14(1):33–46, January 2000.
- [33] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366, September 1995.
- [34] W. R. Dilon and M. Goldstein. *Multivariate Analysis*. John Wiley and Sons, 1984.
- [35] V. Doumptotis. *Discriminative Training for Speaker Adaptation and Minimum Bayes Risk Estimation in Large Vocabulary Speech Recognition*. PhD thesis, The Johns Hopkins University, 2004.
- [36] V. Doumptotis and W. Byrne. Pinched Lattice Minimum Bayes Risk discriminative training for large vocabulary continuous speech recognition. In *International Conference on Spoken Language Processing*, pages 1717–1720, October 2004.
- [37] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.

- [38] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 346–248. IEEE, May 1996.
- [39] RT-03 Spring Evaluation, 2003. [Online]. Available: <http://www.nist.gov/speech/tests/rt/rt2003/spring/>.
- [40] F. Liu F, R. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *Proceedings of ARPA Human Language Technology Workshop*, March 1993.
- [41] C. Field and B. Smith. Robust estimation - a weighted maximum likelihood approach. *International Statistical Review*, 62(3):405–424, 1994.
- [42] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–768, 1922.
- [43] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
- [44] P. Fung, C.Y. Ma, and W.K. Liu. Map-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese. In *European Conference on Speech Communication and Technology*, pages 871–874, September 1999.
- [45] M. Gales and P. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, 10(4):249–264, October 1996.
- [46] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98, April 1998.
- [47] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3):272–281, May 1999.
- [48] M. J. F. Gales. Adaptive training for robust ASR. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 15–20, December 2001.
- [49] J.-L. Gauvain and C.-H. Lee. Maximum *a posteriori* estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.

- [50] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520. IEEE, March 1992.
- [51] V. Goel, S. Axelrod, R. Gopinath, P. Olsen, and K. Visweswariah. Discriminative estimation of subspace precision and mean (spam) models. In *European Conference on Speech Communication and Technology*, pages 2617–2620, September 2003.
- [52] P. S. Gopalakrishnan, Dimitri Kanevsky, Arthur Nádás, and David Nahamoo. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory*, 37(1):107–113, January 1991.
- [53] R. A. Gopinath. Maximum likelihood modeling with Gaussian distributions for classification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 661–664. IEEE, May 1998.
- [54] Y. Grenier. Speaker adaptation through canonical correlation analysis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 888–891. IEEE, April 1980.
- [55] Y. Grenier, L. Miclet, J. C. Maurin, and H. Michel. Speaker adaptation for phoneme recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1273–1275. IEEE, April 1981.
- [56] A. Gunawardana. Maximum mutual information estimation of acoustic HMM emission densities. Technical Report CLSP Reasearch Note No. 40, CLSP, The Johns Hopkins Univerisity, 3400 N. Charles St., Baltimore, MD 21218, USA, 2001.
- [57] A. Gunawardana and W. Byrne. Discriminative speaker adaptation with conditional maximum likelihood linear regression. In *European Conference on Speech Communication and Technology*, pages 1203–1206, September 2001.
- [58] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Monographs and surveys in pure and applied mathematics. Chapman & Hall/CRC, 1999.

- [59] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.
- [60] A. J. Hewett. *Training and Speaker Adaptation in Template-based Speech Recognition*. PhD thesis, Cambridge University, 1989.
- [61] F. Hu and V. Zidek. The weighted likelihood. *The Canadian Journal of Statistics*, 30(3):347–371, 2002.
- [62] M.J. Hunt and C. Lefèbvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 262–265. IEEE, May 1989.
- [63] R. Iyer and M. Ostendorf. Transforming out-of-domain estimates to improve in-domain language models. *European Conference on Speech Communication and Technology*, pages 1975–1978, September 1997.
- [64] R. Iyer, M. Ostendorf, and H. Gish. Using out-of-domain data to improve in-domain language models. *IEEE Signal Processing Letters*, 4(8):221–223, August 1997.
- [65] P. Jain and H. Hermansky. Improved mean and variance normalization for robust speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, May 2001.
- [66] J. Jaschul. Speaker adaptation by a linear transformation with optimized parameters. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1657–1670. IEEE, May 1982.
- [67] H. Jin, F. Kubala, and R. Schwartz. Automatic speaker clustering. In *Darpa Speech Recognition Workshop*, pages 108–111, 1997.
- [68] B.-H. Juang and L. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64(2):391–408, February 1985.
- [69] S. Khudanpur and W. Kim. Contemporaneous text as side-information in statistical language modeling. *Computer Speech and Language*, 18(2):143–162, April 2004.

- [70] T. Kosaka and S. Sagayama. Tree-structured speaker clustering for fast speaker adaptation. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 19–22. IEEE, April 1994.
- [71] N. Kumar. *Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition*. PhD thesis, The Johns Hopkins University, 1997.
- [72] Towards language independent acoustic modeling: Final report, 1999. [Online]. Available: <http://www.clsp.jhu.edu/ws99/projects/asr>.
- [73] C. J. Leggetter. *Improved acoustic Modeling for HMMs using Linear Transformations*. PhD thesis, Cambridge University, 1995.
- [74] C. J. Leggetter and P. C. Woodland. Speaker adaptation of continuous density HMMs using multivariate linear regression. *International Conference on Spoken Language Processing*, pages 451–454, September 1994.
- [75] C. J. Leggetter and P. C. Woodland. Speaker adaptation using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University, June 1994.
- [76] E. L. Lehmann. Efficient likelihood estimators. *The American Statistician*, 34(4):233–235, November 1980.
- [77] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics - Doklady*, 10(10):707–710, 1966.
- [78] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi. An introduction to the application of the theory of probabilistic functions on a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, 1983.
- [79] D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Point of View*. Cambridge University Press, London, 1965.
- [80] F.-H. Liu, M. Picheny, P. Srinivasa, M. Monkowski, and J. Chen. Speech recognition on Mandarin CallHome: a large-vocabulary, conversational, and telephone speech corpus. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 157–160. IEEE, May 1996.

- [81] A. Ljolje. The importance of cepstral parameter correlations in speech recognition. *Computer Speech and Language*, 8(3):223–232, July 1994.
- [82] A. Ljolje. The AT&T LVCSR-2001 system. In *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.
- [83] M. Markatou. Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56(2):483–486, 1999.
- [84] A. Martin, J. Fiscus, M. Przybocki, and B. Fisher. The evaluation: Word error rates and confidence analysis. In *Hub-5 Workshop*, Linthicum Heights, Maryland, 1998. NIST. [Online]. Available: http://www.nist.gov/speech/tests/ctr/hub5e_98/hub5e_98.htm.
- [85] A. Martin, M. Przybocki, J. Fiscus, and D. Pallett. The 2000 NIST evaluation for recognition of conversational speech over the telephone. In *Proceedings of the Speech Transcription Workshop*. NIST, 2000.
- [86] J. McDonough. *Speaker Compensation with All-Pass Transforms*. PhD thesis, The Johns Hopkins University, 2000.
- [87] J. McDonough, T. Schaaf, and A. Waibel. On maximum mutual information speaker-adapted training. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 601–604. IEEE, May 2002.
- [88] J. McDonough and A. Waibel. Maximum mutual information speaker adapted training with semi-tied covariance matrices. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 128–131, April 2003.
- [89] J. McDonough, G. Zavaliagos, and H. Gish. An approach to speaker adaptation based on analytic functions. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 721–724. IEEE, May 1996.
- [90] W. Meeker and L. Escobar. *Statistical methods for reliability data*. Wiley Series in Probability and Statistics. J. Wiley & Sons, New York, 1998.
- [91] R. L. Mercer. Language modelling for speech recognition. In *IEEE Workshop on Speech Recognition*, Arden House, Harriman, N.Y, U.S.A, May 1988.

- [92] M. Mohri and M. Riley. Integrated context-dependent networks in very large vocabulary speech recognition. In *European Conference on Speech Communication and Technology*, pages 811–814, September 1999.
- [93] A. Nádas. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional maximum likelihood. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-31(4):814–817, August 1983.
- [94] A. Nádas, D. Nahamoo, and M. A. Picheny. On a model-robust training method for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(9):1432–1436, September 1988.
- [95] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7(4):308–313, January 1965.
- [96] C. Nieuwoudt and E. C. Botha. Cross-language use of acoustic information for automatic speech recognition. *Speech Communication*, 38(1):101–113, September 2002.
- [97] F. Nolan. *The Phonetic Bases of Speech Recognition*. Cambridge University Press, 1983.
- [98] Y. Normandin. *Hidden Markov Models, Maximum Mutual Information, and the Speech Recognition Problem*. PhD thesis, McGill University, Montreal, 1991.
- [99] Y. Normandin. Maximum mutual information estimation of hidden Markov models. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*, chapter 3, pages 57–81. Kluwer, 1996.
- [100] P. Olsen and R. Gopinath. Modeling inverse covariance matrices by basis expansion. *IEEE Transactions on Speech and Audio Processing*, 12(1):37–46, January 2004.
- [101] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny. Speaker clustering and transformation for speaker adaptation in speech recognition

- systems. *IEEE Transactions on Speech and Audio Processing*, 6(1):71–77, January 1998.
- [102] M. Padmanabhan and S. Dharanipragada. Maximum-likelihood nonlinear transformation for acoustic adaptation. *IEEE Transactions on Speech and Audio Processing*, 12(6):572–578, November 2004.
- [103] JHU RT-02 Workshop presentation, 2002. [Online]. Available: <http://www.clsp.jhu.edu/research/rteval/rt-02/jhu-rt02-pres.pdf>.
- [104] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
- [105] M. Riley and A. Ljolje. Automatic generation of detailed pronunciation lexicons. In *Automatic Speech and Speaker Recognition : Advanced Topics*, page 285302. Kluwer Academic Press, 1995.
- [106] H. Robbins. The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 35:1–20, 1964.
- [107] A. Sankar, F. Beaufays, and V. Digalakis. Training data clustering for improved speech recognition. In *European Conference on Speech Communication and Technology*, pages 502–505, 1995.
- [108] A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(3):190–202, May 1996.
- [109] D. Sankoff and J. Kruskal (Eds). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Welsey, 1983.
- [110] M. Saraclar. *Pronunciation Modeling for Conversational Speech Recognition*. PhD thesis, The Johns Hopkins University, June 2000.
- [111] R. Schlüter. *Investigations on Discriminative Training Criteria*. PhD thesis, RWTH Aachen - University of Technology, 2000.

- [112] T. Schultz and A. Waibel. Fast bootstrapping of lvcsr systems with multilingual phoneme sets. In *European Conference on Speech Communication and Technology*, pages 371–374, September 1997.
- [113] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, and G. Zavaliagkos. New uses for the n-best sentence hypotheses within the Byblos speech recognition system. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1–4. IEEE, March 1992.
- [114] R. Schwartz, Y-L. Chow, and F. Kubala. Rapid speaker adaptation using a probabilistic spectral mapping. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 633–636. IEEE, April 1987.
- [115] Special Session. Multilinguality in Speech Processing. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, May 2004.
- [116] E. Shriberg. Disfluencies in SWITCHBOARD. In *International Conference on Spoken Language Processing*, pages 11–14, October 1996.
- [117] E. Shriberg. Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetics Sciences*, pages 619–622, August 1999.
- [118] O. Siohan, C. Chesta, and C.-H. Lee. Hidden Markov model adaptation using maximum a posteriori linear regression. In *Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.
- [119] O. Siohan, T. Myrvoll, and C.-H. Lee. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer Speech and Language*, 16(1):5–24, January 2002.
- [120] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *International Conference on Spoken Language Processing*, pages 901–904, September 2002.
- [121] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng. The SRI march 2000 Hub-5 conversational speech transcription system. In *Proceedings of the Speech Transcription Workshop*. NIST, 2000.

- [122] S. Tsakalidis, V. Doumptotis, and W. Byrne. Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation. In *International Conference on Spoken Language Processing*, pages 2585–2588, September 2002.
- [123] S. Tsakalidis, V. Doumptotis, and W. Byrne. Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation. *IEEE Transactions on Speech and Audio Processing*, 13(3):367–376, May 2005.
- [124] L. Uebel and P. Woodland. An investigation into vocal tract length normalization. In *European Conference on Speech Communication and Technology*, pages 2519–2522, 1999.
- [125] L. F. Uebel and P. C. Woodland. Discriminative linear transforms for speaker adaptation. In *Proceedings of the Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, pages 61–64. ISCA, 2001.
- [126] L. F. Uebel and P. C. Woodland. Improvements in linear transforms based speaker adaptation. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 49–52. IEEE, May 2001.
- [127] V. Valtchev, J.J. Odell, P.C. Woodland, and S.J. Young. MMIE training of large vocabulary speech recognition systems. *Speech Communication*, 22(4):303–314, 1997.
- [128] D. Vergyri. Use of word level side information to improve speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1823 – 1826. IEEE, June 2000.
- [129] D. Vergyri, S. Tsakalidis, and W. Byrne. Minimum Risk Acoustic Clustering for Multilingual Acoustic Model Combination. In *International Conference on Spoken Language Processing*, pages 873–876, October 2000.
- [130] A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–601, 1949.
- [131] S. Wang and Y. Zhao. On-line Bayesian tree-structured transformation of HMMs with optimal model selection for speaker adaptation. *IEEE Transactions on Speech and Audio Processing*, 9(6):663–677, September 2001.

- [132] T. Watanabe, K. Shinoda, K. Takagi, and E. Yamada. Speech recognition using tree-structured probability density function. In *International Conference on Spoken Language Processing*, pages 223–226, 1994.
- [133] P. C. Woodland, C. J. Leggetter, J. J. Odell, S. J. Young, and V. Valtchev. The 1994 HTK large vocabulary speech recognition system. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 73–76. IEEE, May 1995.
- [134] P. C. Woodland and D. Povey. Large scale discriminative training for speech recognition. In *Proceedings of the Tutorial and Research Workshop on Automatic Speech Recognition*, pages 7–16. ISCA, 2000.
- [135] RT-03 Spring Workshop, May 2003. [Online]. Available: <http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/cts-combined-sm-ok-v14.pdf>.
- [136] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book, Version 3.0*, July 2000.
- [137] G. Zavaliagkos, R. Schwartz, and J. Makhoul. Batch, incremental and instantaneous adaptation techniques for speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 676–679. IEEE, May 1995.
- [138] Y. Zhao. An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(3):380–394, July 1994.
- [139] F. Zheng. A syllable-synchronous network search algorithm for word decoding in chinese speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 601–604. IEEE, March 1999.
- [140] F. Zheng, Z. Song, P. Fung, and W. Byrne. Mandarin pronunciation modeling based on the CASS corpus. *J. Comp. Sci. Tech. (Science Press, Beijing, China)*, 17(3), May 2002.

- [141] T. F. Zheng, P. Yan, H. Sun, M. Xu, and W. Wu. Collection of a chinese spontaneous telephone speech corpus and proposal of robust rules for robust natural language parsing. In *SNLP-O-COCOSDA*, pages 60–67, May 2002.

Vita

Stavros Tsakalidis received the Bachelor of Science degree in electrical engineering from the Technical University of Crete, Greece, in 1998 and the M.S. and Ph.D. degree in electrical engineering from The Johns Hopkins University, in 2000 and 2005 respectively. During the summer of 2000 he worked at Nuance Communications as a research scientist. He is the recipient of the 1998 Ericsson Award of Excellence in Telecommunications for his Bachelor thesis and has one U.S. patent in the field of data processing. His research focus is on acoustic modeling for large vocabulary speech recognition.