中国科学技术大学

# 博士学位论文


## 声学模型区分性训练及其在自动语音识别中的应用

作者姓名：　　鄢 志 杰

学科专业：　　信号与信息处理

导师姓名：　　王 仁 华 教授

完成时间：　　二〇〇八年五月

University of Science and Technology of China

Dissertation for Doctor's Degree


# Discriminative Training of Acoustic Models for Automatic Speech Recognition

Zhi-Jie Yan

Signal & Information Processing

Supervisor: Prof. Ren-Hua Wang

May, 2008

# 中国科学技术大学学位论文原创性和授权使用声明

本人声明所呈交的学位论文,是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

本人授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

保密的学位论文在解密后也遵守此规定。

作者签名：_____

_____ 年 __ 月 __ 日

# 摘 要

声学模型区分性训练已经成为当今主流语音识别系统中最重要的模型训练手段之一。与此同时，对区分性训练准则、模型优化算法以及应用性方法的研究，也日益引起越来越多研究者的重视。在这一背景下，本文围绕声学模型区分性训练及其在自动语音识别中的应用，进行了较系统而深入的研究。并且在准则、优化与应用三个方面都有一定的创新。

首先，本文提出了一种新的区分性训练准则，即最小词分类错误MWCE准则。通过将传统基于句子级的MCE损失代价函数细化到词一级，MWCE准则尝试估计并最小化一个更直接的词级分类错误度量。由于这一词级错误度量更能够匹配大词汇量连续语音识别的目标，即降低词错误率，因此也就能够取得比句子级训练准则更好的识别性能。与其它的一些次句级区分性训练准则(如MWE、MPE)相比，MWCE准则提供了对词级错误的一个全新角度的表达，并在我们的实验中取得了更优的识别性能。这显示从训练准则上继续寻找物理意义更为明确合理的目标来进行优化，仍具有重要的意义。

其次，本文提出了一种新的区分性训练模型参数更新算法，即MMIE准则基于Trust Region的模型参数更新。通过在参数更新过程中引入Trust Region约束，我们使用一种在数学上更为合理、物理意义上更为明确的方式来避免传统EB方法中无界优化问题的一些局限。同时，由于在每次迭代优化中我们都可以得到辅助函数在约束条件下的全局最优解，因此对准则更新的效率也就更高。实验证明，MMIE准则基于Trust Region的模型参数更新在准则优化与识别性能提升两方面都超过了传统的EB模型参数更新方法。

第三，本文提出了对软分类边缘估计SME的一些完善和改进方法。通过将区分性训练领域近年来的一些重要技术引入，我们首次实现了句子级SME估计在大词汇量连续语音识别中的成功应用。接着，我们进一步提出SME估计下的帧级区分性训练方法，通过定义合理的分割度量，在帧尺度上对包含重要区分性信息的训练样本进行筛选。我们在实验中对比了传统MCE准则、句子级SME估计，以及帧级SME估计的性能。结果表明，由于引入了Soft Margin的概念，两种SME估计都能够取得超越MCE准则的性能。而我们提出的帧级SME估计在引入抑制噪声样本的参数后可以取得最好的、明显超过传统MCE准则的识别性能。

最后，本文提出了一种区分性训练的应用性方法，即基于MMIE准则的HMM模型拓扑结构优化方法。我们根据MMIE准则定义出指导模型结构优化的启发性度量，尝试在各个HMM状态间“交换”高斯核以实现各状态混合高斯

成份数目的非均匀分配。此外，还对高斯核交换完成后的特定HMM状态进行时间尺度上的拓扑结构后处理。通过以上这些步骤，我们将模型结构的优化与模型区分性的提高更直接的联系了起来。因此，基于区分性准则的模型拓扑结构优化方法在实验中能够在性能上超过传统的、基于似然度的其它模型结构优化方法。

**关键词：** 区分性训练 声学模型 自动语音识别 最小词分类错误 信任区域软分类边缘估计

# Abstract

Discriminative training of acoustic models has become one of the most important training methods for state-of-the-art speech recognition systems. This topic attracts more and more attentions of researchers, to develop new training criteria, parameter optimization methods, and application techniques. In this context, this thesis focuses on discriminative training of acoustic models and its application in automatic speech recognition. It provides a systematic and in-depth research in this topic, and introduces our innovations in criterion, optimization method, and application of discriminative training.

Firstly, this thesis proposes a novel discriminative training criterion, i.e. Minimum Word Classification Error (MWCE). By localizing conventional string-level MCE loss function to word-level, a more direct measure of word classification error is approximated and minimized. Because the word-level criterion better matches performance evaluation criteria in LVCSR, such as WER, an improved word recognition performance can be achieved. Comparing with other sub-string level criteria (e.g. MWE / MPE), MWCE provides another perspective of word-level classification error, and achieves the best recognition performance in our experiments. This result suggests that it is still meaningful to develop new discriminative training criteria which have explicit physical meaning and more reasonable.

Secondly, the thesis proposes a new parameter optimization method for discriminative training, i.e. trust region based optimization for MMIE criterion. By imposing a trust region constraint into the optimization process, we avoid some disadvantages of the unbounded optimization of conventional EB method. The new optimization method is more reasonable in mathematics, and also physically meaningful. Meanwhile, because we can reach a global optimum in each iteration, the proposed method is more efficient in optimizing criterion. Our experimental results suggest that the trust region based approach outperforms conventional EB method both in optimizing criterion and recognition performance.

Thirdly, this thesis introduces our research to improve the Soft Margin Estimation (SME) method. By imposing some important technologies of discriminative training in recent years, we successfully implement the SME method in LVCSR for the first time. Meanwhile, we propose to use a reasonable frame-level separation measure to

select certain frame samples that contain important discriminative information. We compare conventional MCE, string-level SME, and the proposed frame-level SME in our experiments. The results show that by using the concept of soft margin, both SME methods can achieve a better performance than MCE. And by imposing a factor which removes noisy frames, the frame-level SME achieves the best recognition performance which significantly outperforms MCE.

Lastly, this thesis proposes an application method of discriminative training, i.e. MMIE based HMM topology optimization. We define a heuristic metric according to MMIE criterion, and use it to guild the topology optimization process. The approach tries to "exchange" Gaussian kernels among HMM states so as to allocate model parameters non-uniformly. Besides, a post-process is also carried out to refine model topology in time axis. By doing this, we provide a more direct connection between topology optimization and discrimination. As a result, the discriminative model topology optimization outperforms other conventional, likelihood based optimization methods in our experiments.

**Keywords:** Discriminative Training, Acoustic Model, ASR, MWCE, Trust Region, SME

# 目 录

# 英文缩写及主要符号对照表

| | |
|---|---|
| $a$ | HMM状态转移概率 |
| $b(\cdot)$ | 混合高斯概率密度函数 |
| $c$ | 混合高斯权重 |
| $\mathcal{F}$ | 区分性训练准则 |
| $\mathcal{M}$ | 模型空间 |
| $\mathcal{N}(\cdot)$ | 高斯概率密度函数 |
| $p(\cdot)$ | 概率 |
| $r, R$ | 训练集语料序号及语料总数 |
| $s, S$ | 状态及状态序列 |
| $t, T$ | 时刻及总时间 |
| $w, W$ | 词及词序列 |
| $x, X, o, O$ | 输出 / 观测及输出 / 观测序列 |
| $\gamma$ | 后验概率(占有率) |
| $\lambda, \Lambda$ | 声学模型参数 |
| $\mu$ | 高斯概率密度函数均值及均值向量 |
| $\sigma^2, \mathbf{\Sigma}$ | 高斯概率密度函数方差及协方差矩阵 |

| | |
|---|---|
| AIC | Akaike Information Criterion，Akaike信息准则 |
| AM | Acoustic Model，声学模型 |
| ASR | Automatic Speech Recognition，自动语音识别 |
| BIC | Bayesian Information Criterion，贝叶斯信息准则 |
| CT | Corrective Training，纠正训练(准则) |
| DFE | Discriminative Feature Extraction，区分性特征提取 |
| DP | Dynamic Programming，动态规划 |
| DT | Discriminative Training，区分性训练 |
| DTW | Dynamic Time Warping，动态时间规整 |
| EB | Extended Baum-Welch Algorithm，扩展Baum-Welch算法 |
| EM | Expectation-Maximization Algorithm，期望最大化算法 |
| FB | Forward-Backward Algorithm，前后向算法 |
| FT | Falsifying Training，纠错训练(准则) |
| GD | Gradient Descent，梯度下降算法 |
| GMM | Gaussian Mixture Model，高斯混合模型 |
| GPD | Generalized Probability Descent，广义概率下降 |

| | |
|---|---|
| HLDA | Heteroscedastic Linear Discriminant Analysis，异方差线性判别分析 |
| HMM | Hidden Markov Model，隐马尔科夫模型 |
| HTK | Hidden Markov Toolkit，隐马尔科夫模型工具包 |
| KLD | Kullback-Leibler Divergence，KL距离 |
| LDA | Linear Discriminant Analysis，线性判别分析 |
| LME | Large Margin Estimation，大分类边缘估计 |
| LVCSR | Large Vocabulary Continuous Speech Recognition，大词汇量连续语音识别 |
| MAP | Maximum A-Posteriori，最大后验概率(准则) |
| MCE | Minimum Classification Error，最小分类错误(准则) |
| MDL | Minimum Description Length，最小描述长度准则 |
| MFCC | Mel-Frequency Cepstral Coefficients，梅尔频率倒谱系数 |
| MLE | Maximum Likelihood Estimation，最大似然估计(准则) |
| MLLR | Maximum Likelihood Linear Regression，最大似然线性回归 |
| MMIE | Maximum Mutual Information Estimation，最大互信息量估计(准则) |
| MPE | Minimum Phone Error，最小音素错误(准则) |
| MVE | Minimum Verification Error，最小验证错误准则(准则) |
| MWCE | Minimum Word Classification Error，最小词分类错误(准则) |
| MWE | Minimum Word Error，最小词错误(准则) |
| PLP | Perceptual Linear Prediction，感知线性预测系数 |
| SME | Soft Margin Estimation，软分类边缘估计 |
| SVM | Support Vector Machine，支持向量机 |
| TR | Trust Region，信任区域 |
| UI | User Interface，人机界面 |
| WER | Word Error Rate，词错误率 |
| WPP | Word Posterior Probability，词后验概率 |

# 第 1 章 绪论

## 1.1 语音识别及其简史

语音是人与人之间最自然、最便捷的交流途径之一。如果人与计算机之间也能够通过语音进行交流，那无疑会极大的提高人机界面(User Interface, UI)的易用性。事实上，与计算机进行语音交流一直是人们的梦想，要实现这一愿望则需要依靠语音识别、语言理解、语音合成等多项关键技术。在这些技术中，自动语音识别(Automatic Speech Recognition, ASR)无疑是最重要、最困难的核心技术之一，它的功能就是要把人的语音转换为相应的文本、命令等，以便计算机进行理解和产生相应的操作。

自动语音识别涉及多学科的交叉，它包括声学、生理学、心理学、信号处理、模式识别、人工智能、信息理论、语言学以及计算机科学等很多方面。现代的自动语音识别系统最早可以追朔到20世纪50年代[1]，贝尔实验室的研究者使用模拟元器件，提取语音中元音的共振峰频率变化信息，从而对孤立数字的语音实现了识别。到了20世纪60年代，前苏联有研究者提出动态时间规整(Dynamic Time Warping, DTW)及动态规划(Dynamic Programming, DP)[2]算法，解决了语音模板与语音实例之间的对齐问题，从而为当代的语音识别技术打下了重要的基础。到70年代，随着LPC等语音特征参数的提出[3]，用基音等相关特征进行语音识别的方法被逐渐提出。随着美国国防部高级研究计划署(Defense Advanced Research Projects Agency, DARPA)推动的语音理解研究计划(Speech Understanding Research, SUR)，越来越多学术界和工业界的研究机构纷纷加入到对语音识别的研究中来。这其中，就包括卡耐基梅隆大学(Carnegie Mellon University, CMU)的Harpy[4]、Hearsay系统[5]、BBN的HWIM[6]系统、IBM面向听写机的研究[7]和贝尔实验室面向电信业务的语音识别研究等等。

自动语音识别研究在上世纪八九十年代达到了一个高潮，这与HMM在这一领域的成功应用有非常直接的联系。在HMM引入以后，语音识别终于从简单的基于模板，变为较为完善的利用概率模型体系。在这一时期，有关语音识别中应用HMM的理论和实践逐渐趋于完善[8–10]，而且还出现了一系列在HMM体系下进行自动语音识别的其他关键技术，例如用于HMM模型自适应的最大后验概率准则估计(Maximum A-Posteriori Estimation, MAP Estimation)[11]、最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)[12]，以及用于模型参数绑定的决策树状态聚类[13,14]等。直到现在，这些技术都是语音识别声

学模型方面最为重要的核心。在这一时期，我们可以看到更多更成熟的语音识别系统被逐渐推出，包括CMU的Sphinx系统[15]、BBN的BYBLOS系统[16]，以及SRI的DECIPHER系统[17]等。此外，还可以见到一些面向个人用户的语音识别软件，如微软的Whisper系统[18]、SAPI开发工具，以及IBM的ViaVoice等。在DARPA和美国国家标准和技术研究所(National Institute of Standards and Technology, NIST)的持续推动下，越来越多的语音识别任务被各家研究机构所不断尝试，并努力提高识别性能。这其中最著名的任务包括海军资源管理(Resource Management, RM)、北美商务(North American Business, NAB)、华尔街日报(Wall Street Journal, WSJ)、Switchborad等。这些标准数据库为各家研究机构客观的对比各种技术提供了较为有效的平台。在这个时期，还出现了有别于最大似然估计(Maximum Likelihood Estimation, MLE)准则的声学模型区分性训练(Discriminative Training, DT)[19,20]技术。最后，随着隐马尔科夫模型工具包(Hidden Markov Toolkit, HTK)[21]等软件的推出以及公开化，其他研究机构对自动语音识别研究的门槛大大降低，从而进一步掀起了这一领域研究的热潮。

进入新世纪以后，自动语音识别的研究更是向广度和深度两方面发展。在声学模型方面，区分性训练技术被推上更高的水平[22–24]，一些跳出传统HMM框架的、新的声学模型得到了初步的研究[25]。随着DARPA全球自动语言开发计划(Global Autonomous Language Exploitation, GALE)的提出，业已存在及正在逐渐成熟的各项语音识别技术得到了空前的整合[26,27]。此外，一些基于语音识别的新应用，如语音搜索[28]等，也受到了越来越多的关注。应该说，当今的语音识别系统在很多方面还达不到实用的要求，但这些问题的解决正是研究者孜孜以求的方向。从数十年的历史看来，自动语音识别已经从只能识别孤立音素发展到自然语流下的大词汇量连续语音识别(Large Vocabulary Continuous Speech Recognition, LVCSR)，识别结果的准确性也在逐年提高。随着技术的进展我们有理由相信，在不久的将来，语音识别技术将全面的达到真正实用的水平，从而为更便捷自然的人机语音通信开拓更广泛的前景。

## 1.2 语音识别问题

语音识别就是将观测到的语音样本 $O$ 通过函数 $\mathcal{F}(\cdot)$ 转换为文本 $\mathcal{T} = \mathcal{F}(O)$ 的过程。其核心问题就是寻找适合的转换函数 $\mathcal{F}$，使得转换得到的文本与人工标注一致或尽可能接近。描述自动语音识别输出与人工标注之间相似度的指标通常是错误率(Error Rate)。我们的目标就是要得到一个在操作时能够产生最小错误率的转换函数。

语音识别问题是非常复杂的，我们不可能利用手工制定规则或使用专家知

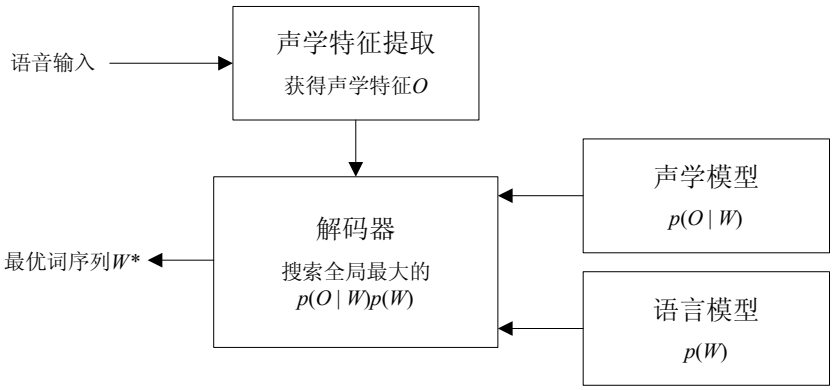


图 1.1 语音识别系统的主要构成

识轻易的得到合适的转换函数。通常的做法是利用统计模型在训练数据中寻找规律，进行统计语音识别(Statistical Speech Recognition)。设某一个词序列表示为 $W = \{w_1, w_2, \ldots, w_N\}$，在统计语音识别框架下评价其与某一段观测 $O$ 匹配程度的度量通常使用如下的贝叶斯公式：

$$p(W \mid O) = \frac{p(O \mid W)p(W)}{p(O)} \tag{1-1}$$

其中， $p(O \mid W)$ 被称为声学模型(Acoustic Model, AM)概率，表示语音本身的声学特征与词串 $W$ 的匹配程度。通常，我们使用隐马尔科夫模型(Hidden Markov Model, HMM)来对声学模型进行建模，其基本理论将在第 2 章作更详细的介绍。$p(W)$ 则是语言模型概率，表示在自然语言中词串 $W$ 本身可能出现的概率。贝叶斯决策理论就是选择使得上式最大的词串 $W^*$ 作为自动语音识别的输出，关于这一理论我们将在第 3 章中进一步说明。

## 1.3 主流语音识别系统及其主要构成

### 1.3.1 声学特征提取

语音识别系统最重要的几个组成部分如图(1.1)所示。语音输入首先被送入声学特征提取模块，从而使得模拟的语音信号数字化，并存储为计算机可以处理的格式。

一个好的特征提取模块应该使得提取出的声学特征能够针对声学模型建模单元具有良好的区分性。同时，特征的体量要较为适中，以便于高效、可靠的估计模型参数。最后，面向语音识别的声学特征应该尽量只含有语音本身的信息，其他诸如说话人、信道、背景环境等干扰性信息应该尽可能避免。一般来说，声学特征提取模块与后续的训练识别过程是独立的，也就是说，我们通常只是根据经验提取合适的声学特征并使用，而不考虑特征与建模单元区分性之间的的关系。但近年来，也出现了一些利用区分性准则对特征进行处理的方法[29,30]，以及

将声学模型的识别结果反馈至特征提取模块的方法[31]。通过这样的手段，特征提取与识别结果的关系将会变得更为紧密，从而提供了进一步提高语音识别系统总体性能的可能。

目前最常用的声学特征一般都是基于傅立叶变换(Fourier Transformation)或线性预测(Linear Prediction)，以及倒谱分析(Cepstral Analysis)等信号处理手段的。其典型代表就是被广泛使用的梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)[32]，以及感知线性预测系数(Perceptual Linear Prediction, PLP)[33]等。此外，还有一些线性变换方法用于特征扩展、特征降维等，以增强声学特征的区分能力或降低维数。这些技术包括使用动态特征[34]、线性判别分析(Linear Discriminant Analysis, LDA)[35]、异方差线性判别分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[36]等。

在本文中，我们主要研究声学特征给定后的声学模型问题。也就是说，无论是在最大似然估计下训练模型，或是在区分性训练准则下对模型参数进行调整，我们都假定使用同样的、不变的声学特征参数。将区分性准则用于特征提取在近年来已经成为一个热点，但这项工作不在本文的范畴之内。

### 1.3.2 声学模型与语言模型

正如前面所提到的，除了声学特征，解码器还需要用到声学模型和语言模型才能进行解码。关于声学模型的相关问题是本文的重点，将会在后续的章节作详细介绍。而目前统计语音识别所用到的语言模型，则绝大部分是n-元(n-gram)统计语言模型。n-gram语言模型的一个基本思想是将词串的生成过程表示为一个词间跳转的 $n-1$ 阶马尔科夫过程(有关马尔科夫过程的简要介绍可见于第 2 章)。那么，词串的先验概率 $p(W)$ 即可表示为：

$$p(W) = p(w_1^K) = \prod_{k=1}^{K} p(w_k \mid w_1^{k-1}) \approx \prod_{k=1}^{K} p(w_k \mid w_{\max[k-n+1,1]}^{k-1}) \tag{1-2}$$

其中，$w_1^0 = \emptyset$。

不难看出，在上式的计算过程中，当前词的概率仅取决于此前 $n-1$ 个词的历史，这也正是n-gram语言模型得名的原因。训练及评价n-gram语言模型的重要指标是语言模型复杂度(Perplexity)[37]，它被定义为词串概率几何平均的倒数，即：

$$PP = [p(w_1^K)]^{-\frac{1}{K}} = \left[ \prod_{k=1}^{K} p(w_k \mid w_{\max[k-n+1,1]}^{k-1}) \right]^{-\frac{1}{K}} \tag{1-3}$$

从原则上讲，语言模型对词串的期望复杂度应该尽可能小，这就表示语言模型在对给定的历史词串进行当前词预测的时候，能够拥有更高的确信度。事

实上，语言模型的训练准则也就是使其对训练集中句子的复杂度最小化。要达到这一目的，就需要统计训练集语料中各词及其组合出现的频率，并以此为基础确定语言模型的参数。给定词表大小 $M$，在n-gram言模型中显然需要统计 $M^N$ 种词语组合出现的频率。而在词汇量较大的情况下，某些组合在训练语料中根本不存在。因此，为了解决这一问题，还必须用到降权(Discounting)[38]及回退(Backing-Off)[39]等方法。降权方法将训练集中能够观察到的词串的个数拿出一部分，通过一定的方式分配给不能观察到的其它词串；而回退方法则是将较短历史信息的词频信息赋予那些观察不到的、需要较长历史信息的词串。在本文中，我们主要关注声学模型，在大多数情况下，我们假设语言模型已知，且其参数不再改变。我们将会在第 3 章中探讨语言模型在区分性训练中的使用问题，但语言模型本身的参数调整则不在本文的关注之内。

### 1.3.3 解码器

要解析出语音输入中所包含的文本信息，需要在解码器(Decoder)中通过搜索算法寻找出最优词串 $W^*$ 作为结果输出。从(1-1)式不难看出：

$$W^* = \arg\max_W p(W \mid O) = \arg\max_W \frac{p(O \mid W)p(W)}{p(O)} = \arg\max_W p(O \mid W)p(W) \quad (1\text{-}4)$$

在语音识别中，解码所必须用到的搜索空间异常巨大。我们可以想象一个不算大的、5000词的词表，即使我们规定每句话最多只能有5个词，那么所有可能的词串数目也将高达 $10^{18}$ 量级( $1+5000+5000^2+\cdots+5000^5 \approx 3.1\times 10^{18}$ )。因此，必须借助一系列方法将如此大规模的搜索问题进行压缩，从而简化到计算机可以处理的程度。

对当今很多主流的解码器而言，压缩搜索空间的最重要近似方法是动态规划思想下的维特比算法(Viterbi Algorithm)[40]。关于这两种方法，我们将在第 2 章作进一步介绍。此外，还有一些解码器[41]使用时间异步的 A 星搜索算法(A*-Search)，又称堆栈解码(Stack-Decoding)，通过引入一些启发性度量来引导搜索过程[42]。

对时间同步的Viterbi算法来说，还需要利用一些方法在解码过程中进行同步的快速概率计算及搜索空间裁剪(Pruning)。这些方法包括快速计算输出概率的一系列方法(如高斯选择算法Guassian Selection[43])，以及Beam裁剪[44]、词法前缀树(Lexical Prefix Tree)[45]、语言模型前看(Language Model Look-Ahead)[46]等。

而对时间异步的其它一些搜索算法而言，通常会使用简单的模型快速得到最可能的一些候选，再在后续过程用更精细的模型对这些候选进行重新打分(Re-Scoring)。第一遍解码所得出的初步候选主要通过n-best列表[47,48]和词图(Word Graph, Lattice)[49]的形式加以存储。由于词图可以用非常简洁的方式表

达较大的声学模型空间，所以在语音识别的很多方面都有应用[31,50–52]。在本文所着重探讨的声学模型区分性训练中，也需要用到词图来保存参考模型空间及竞争模型空间的诸多信息，而这些词图都是由解码器通过搜索得到的。在词图所表达的模型空间中高效的求取统计信息，也是声学模型区分性训练中一项非常重要的工作。

## 1.4 本文的主要内容、创新及组织结构

本文主要探讨声学模型区分性训练及其在自动语音识别中的应用问题。当前对声学模型区分性训练的研究主要集中在训练准则和模型参数优化方法两个方面，而本文在这两方面都有涉及，并有所创新。

首先，在训练准则方面，本文提出了一种全新的最小词分类错误MWCE准则。通过将传统的、基于句子级的MCE损失代价函数细化到词一级，MWCE准则使得模型参数优化的结果与语音识别的最终目标更为匹配，并在实验中取得了超过传统区分性训练准则的识别性能。

其次，在模型参数优化方面，本文提出了基于信任区域Trust Region的HMM模型参数优化算法。通过引入Trust Region，我们使得模型参数优化问题从无界优化变为有界优化，并因此而得以巧妙的利用了待优化函数的诸多数学性质。在我们的实验中，基于Trust Region的模型参数优化算法能够取得超过传统EB优化算法的性能。

接着，在区分性训练模型的推广性能方面，我们研究了软分类边缘估计SME，特别是将已有的句子级SME方法细化到帧一级。通过将区分性训练领域近年来的技术进展融入SME估计，并设计帧一级的分割度量与损失代价函数，SME方法在实验中可以取得超过传统MCE准则的优良识别性能。这也是SME方法在大词汇量连续语音识别任务上首次体现出明显超越传统区分性训练准则的优越性。

最后，在区分性训练的应用方法方面，本文提出了MMIE准则针对HMM模型拓扑结构的优化方法。通过定义与MMIE准则相关的启发性度量，我们打破了传统上均匀分配模型参数的方法的弊端，并通过非均匀分配模型参数的方式提高了声学模型的区分能力。实验证明，在模型参数数目给定的情况下，通过我们提出的区分性模型拓扑结构优化方法，可以使得声学模型的识别性能得到明显的提升。

在本文一开始，我们将会用两章的篇幅介绍语音识别声学模型及其参数训练问题的一些基础。其中，第 2 章主要介绍基于HMM的声学模型以及最大似然准则下的参数估计，第 3 章则主要介绍传统的区分性训练方法，包括训练准则、

参数优化算法以及其他一些问题等。接着在后续章节中，我们将会介绍本文主要的研究工作。包括第 4 章中对新的区分性训练准则MWCE的介绍、第 5 章中对新的模型参数优化方法，即基于Trust Region的模型参数优化的介绍。在接下来的第 6 章中，将会介绍基于SME的区分性训练方法，特别是我们在此基础上所完成的帧级SME区分性训练的工作。在第 7 章中，我们将介绍区分性训练准则MMIE在HMM模型拓扑结构优化上的应用性方法。最后，在第 8 章中，我们将给出全文的总结以及今后可能的一些研究方向。

# 第 2 章 基于HMM的声学模型及其最大似然估计

## 2.1 引言

当今主流的语音识别系统绝大部分采用隐马尔科夫模型HMM作为声学模型的建模基础。可以说，HMM及其相关技术是构成目前语音识别系统的最重要的核心之一。人类语音的特性决定了我们需要一个强大的统计模型来描述一个不断产生观测(语音本身)的离散时间序列(语音中所隐含的信息)，而HMM正好是胜任这一任务的最直接的选择。HMM能够在较少的模型参数下提供较高的精度，而与之相伴而生的动态规划算法则使它具有了对变长时间序列进行分段与分类的能力。HMM的适应能力很强，不论对离散分布还是连续分布，或者是对标量观测以及矢量观测都能够进行建模。HMM的基本假设在于，我们所要描述的时间序列要能够被这样一个参数化的随机过程所描述；而描述此过程的HMM参数又要能够被有效的估计得到。

从Baum提出HMM的理论以来[53]，它在语音及相关领域的应用范围已经变得越来越广泛。除了在语音识别领域扮演核心角色以外，它还被广泛用于语音合成[54]、机器翻译[55]、语言理解[56]等很多方面[10]。在本文中，我们主要关注HMM在语音识别声学模型中的应用，因此我们将首先介绍语音识别背景下HMM模型最核心的一些问题及解法，而这些方法正是HMM占据语音识别声学模型领域支配性地位的关键所在。特别的，我们将在本章中着重介绍声学模型最大似然估计方法，这也是HMM在语音识别领域最基本、最重要的参数估计手段。

本章的后续部分组织如下：(2.2)节介绍HMM的数学定义，(2.3)节介绍HMM经典的评估、解码、训练三问题，并举例说明最大似然准则下HMM参数的估计问题。接着，我们将在(2.4)节说明HMM如何应用于语音识别声学模型建模，最后在(2.5)节给出本章小结。

## 2.2 HMM的数学定义

### 2.2.1 马尔科夫过程及马尔科夫链

马尔科夫过程(Markov Process)是对自然界中一系列现象的数学抽象，它描述了一个最小记忆系统的随机行为。这里的所谓“最小记忆”是指随机过程的当前值仅与其最近的值相关，而与其过去及未来的其它值无关。这实际上是对自然

界规律的一个简化，并被称为马尔科夫假设。从数学上讲，马尔科夫假设也就是假设对连续随机变量 $X$ 及一系列相距足够短的时刻 $t_1 < t_2 < \cdots < t_n$ 来说，有：

$$p(x_{t_n} < \phi \mid x_{t_1}, x_{t_2}, \ldots, x_{t_{n-1}}) = p(x_{t_n} < \phi \mid x_{t_{n-1}}) \tag{2-1}$$

我们可以使用采样将连续时间随机变量转化为离散时间序列，还可以通过量化将随机变量的连续值映射到有穷元素的状态集合中。而这种离散时间、有限状态的马尔科夫过程通常被称为马尔科夫链(Markov Chain)。在马尔科夫链中，设状态集合为 $S = \{1, 2, \ldots, N\}$，随机过程进入当前状态 $s_t$ 的概率仅与上一个状态相关，即：

$$p(s_t \mid s_0, s_1, \ldots, s_{t-1}) = p(s_t \mid s_{t-1}) \tag{2-2}$$

因此，要描述一个马尔科夫链需要如下两个参数：

$$\begin{aligned} a_{ij} &= p(s_t = j \mid s_{t-1} = i) \quad 1 \le i, j \le N \\ \pi_i &= p(s_0 = i) \quad 1 \le i \le N \end{aligned} \tag{2-3}$$

其中，$a_{ij}$ 被称为状态间的转移概率，$\pi_i$ 为状态的初始概率，且它们应满足：

$$\begin{aligned} \sum_{j=1}^{N} a_{ij} &= 1 \quad 1 \le i \le N \\ \sum_{j=1}^{N} \pi_j &= 1 \end{aligned} \tag{2-4}$$

上面的用以描述马尔科夫链的模型有时也被称为“显马尔科夫模型”(Observable Markov Model)，这是一个与隐马尔科夫模型相对的概念。在这种模型所描述的随机过程中，我们可以直接观测到任意时刻随机过程所处的状态，或者可以说，模型的输出与其所处的状态是完全绑定的。我们可以用股票市场中每个交易日的指数变化作一个例子来说明这种模型，如图(2.1)所示[57]。在这里，状态集合中有 3 个状态 1 ~ 3，对应于相对前一交易日上涨、下跌、持平。描述这一随机过程的参数分别为如下的状态转移矩阵 $\mathbf{A}$ 和初始概率矩阵 $\pi$：

$$\begin{aligned} \mathbf{A} = \{a_{ij}\} &= \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \\ \pi = \{\pi_i\} &= \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix} \end{aligned} \tag{2-5}$$

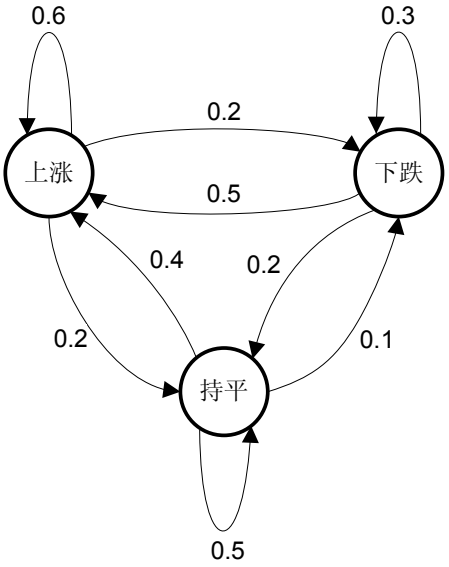


图 2.1　表示股票市场指数涨跌的马尔科夫链模型

不难看出，如果我们要计算连续 5 个交易日上涨的概率，则可表示为：

$$p(1,1,1,1,1) = \pi_1 a_{11} a_{11} a_{11} a_{11} = 0.5 \times (0.6)^4 = 0.0648 \tag{2-6}$$

### 2.2.2　隐马尔科夫模型HMM

上一节所介绍的马尔科夫链的输出与状态是一一对应的，也就是说，状态的输出不是一个随机过程，而是一个确定的事件。很自然的，我们可以将马尔科夫链进行扩展，使得各状态可以以随机的方式对外产生输出，这就是隐马尔科夫模型HMM。在HMM中，实际有两层随机过程：第一层是状态之间的转移，第二层是各状态的输出。前者不被外界所知，是“隐”的，后者则可以被外界所观测到，是“显”的。对于某一特定的状态来说，其产生输出的过程是随机的；而对于某一被观测到的特定输出来说，也只能以概率的形式来推断它所属的状态。HMM的这些特点与马尔科夫链下的情况显然是不同的，但基本可以说，HMM是一个有着状态相关概率输出函数的、扩展的马尔科夫链。

我们将图(2.1)的马尔科夫链拓展到HMM，如图(2.2)所示[57]。可以看到，指数的涨跌与在马尔科夫链中直接与状态绑定不同，变为由概率的形式对外输出，图中各状态旁的输出概率矩阵即表明输出已成为与状态相关的一个随机过程。在上图的例子中，我们将输出概率矩阵定义为：

$$\mathbf{B} = \begin{bmatrix} p(\text{上涨}) \\ p(\text{下跌}) \\ p(\text{持平}) \end{bmatrix} \tag{2-7}$$

由各个状态对指数涨跌的输出概率矩阵我们不难看出，状态 1 对应着“牛市”，因为随机过程在这个状态下出现指数上涨的概率高达0.7。相应的，状态 2

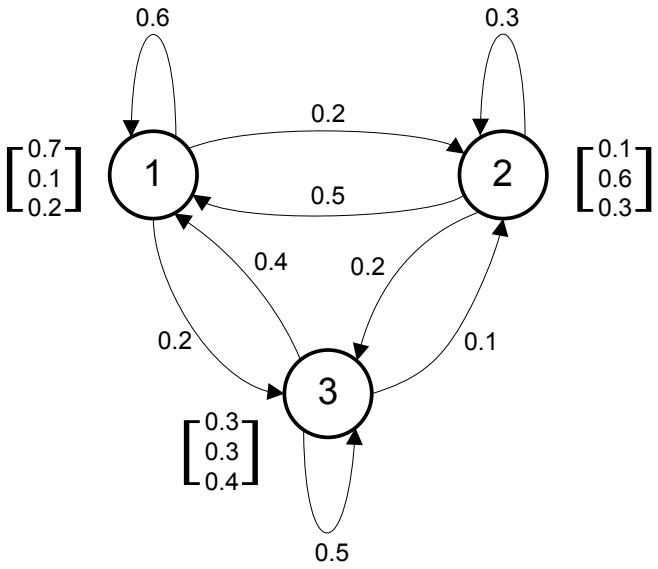


图 2.2　表示股票市场所处趋势的HMM模型

对应着“熊市”，而状态 3 则代表较为稳定的市场。

从数学上讲，定义一个HMM应包含如下一些要素：1、输出(或称观测) $X$；2、状态集合 $\Omega$；3、状态转移矩阵 $\mathbf{A}$；4、状态相关的输出概率 $\mathbf{B}$；5、初始概率矩阵 $\pi$。虽然在上面的举例中我们为了简明而使用了离散的输出和离散的输出概率矩阵，但实际上这两者分别可以是连续随机变量及连续概率密度函数。除此之外，上述参数还应满足：

$$\begin{gathered} a_{ij} \geq 0, \quad b_i(x) \geq 0, \quad \pi_i \geq 0 \\ \sum_{j=1}^{N} a_{ij} = 1 \quad \int b_i(x)\,\mathrm{d}x = 1 \quad \sum_{i=1}^{N} \pi_i = 1 \end{gathered} \tag{2-8}$$

其中，$b_i(x)$ 是状态 $i$ 产生输出 $x$ 的概率。

除了上面已经提到的马尔科夫假设，在HMM中还有一个重要的输出无关假设。也就是当前时刻的输出仅与当前状态相关，而与历史上其他的输出和状态无关：

$$p(x_t \mid x_1^{t-1}, s_1^t) = p(x_t \mid s_t) \tag{2-9}$$

虽然这一假设显然不完全符合自然界的规律，但它的存在却能够使关于HMM的一系列问题变得容易处理得多。在基于HMM的语音识别背景下，这一假设可以极大的降低声学模型参数的数量，从而使得参数估计、解码等问题的复杂度大大的降低。近年来，虽然在语音识别领域也出现了一些打破这一假设的模型[58,59]，但在本文中，我们仍在这个假设的基础上进行后面的推导。

## 2.3 HMM的经典问题

在定义了HMM的基本参数后，要将其真正应用到语音识别领域，还需要解决 3 个最基本的问题。首先，在给定HMM模型 $\Phi$ 以及一串观测序列 $X = \{x_1, x_2, \ldots, x_T\}$ 后，我们需要求得模型 $\Phi$ 产生 $X$ 的似然度 $p(X \mid \Phi)$。这是一个最基本的问题，只有解决它才能够进一步利用概率论中的其它理论来解决更多的问题。其次，在给定 $\Phi$ 以及一串观测序列 $X$ 后，我们要能够搜索出 $\Phi$ 中最可能生成 $X$ 的状态序列 $S = \{s_1, s_2, \ldots, s_T\}$。只有这样，我们才可以对 $X$ 进行解码，找出其下所隐藏的状态转移过程。最后，在给定一串观测序列 $X$ 后，我们还需要通过一定的手段得到模型参数 $\Phi$，并使得 $p(X \mid \Phi)$ 最大化。只有这样，我们才可以在最大似然准则下通过训练的方法求得HMM模型的参数。概括起来，上述三个问题分别可以被称为评估问题、解码问题和训练问题。

### 2.3.1 评估问题

计算 $p(X \mid \Phi)$ 最简单的办法是首先计算某一状态序列 $S$ 产生 $X$ 该概率，再穷举所有可能的状态序列并累加起来，即：

$$p(X \mid \Phi) = \sum_{S} p(X, S \mid \Phi) = \sum_{S} p(S \mid \Phi) p(X \mid S, \Phi) \tag{2-10}$$

根据HMM的数学定义我们不难看出：

$$p(S \mid \Phi) = p(s_1 \mid \Phi) \prod_{t=2}^{T} p(s_t \mid s_{t-1}, \Phi) = \pi_{s_1} a_{s_1 s_2} \ldots a_{s_{T-1} s_T} \tag{2-11}$$

$$p(X \mid S, \Phi) = \prod_{t=1}^{T} b_{s_1}(x_1) b_{s_2}(x_2) \ldots b_{s_T}(x_T) \tag{2-12}$$

因此：

$$p(X \mid \Phi) = \sum_{S} \pi_{s_1} b_{s_1}(x_1) a_{s_1 s_2} b_{s_2}(x_2) \ldots a_{s_{T-1} s_T} b_{s_T}(x_T) \tag{2-13}$$

(2-13)式所表达的物理意义可以形象的解释如下：首先，HMM由起点以 $\pi_{s_1}$ 的概率跳到 $s_1$，并紧接着产生输出概率 $b_{s_1}(x_1)$。而此后，每一次以跳转概率 $a_{s_{t-1} s_t}$ 跳到下一个状态 $s_t$ 时都随之产生输出概率 $b_{s_t}(x_t)$，直至最后一个状态 $s_T$。这样一来，对一个总共有 $N$ 个状态的HMM来说，时长为 $T$ 的所有可能的状态序列多达 $N^T$ 条，要直接计算上式显然需要高达 $O(N^T)$ 的运算量。因此，需要使用所谓前向算法(Forward Algorithm)来归纳计算这一概率。

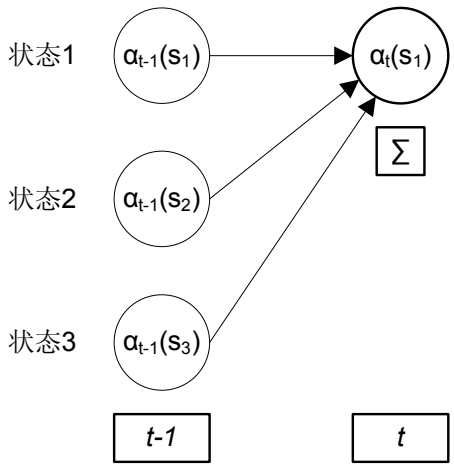


图 2.3　归纳计算前向概率的前向算法示意图

首先，定义前向概率 $\alpha_t(i)$，它表示的是HMM通过一系列状态转移、产生部分的观测 $x_1^t$，并最终于 $t$ 时刻停留于状态 $i$ 的概率：

$$\alpha_t(i) = p(x_1^t, s_t = i \mid \Phi) \tag{2-14}$$

显然，对各状态在时刻1的前向概率我们有 $\alpha_1(i) = \pi_i b_i(x_1)$，而对于其后时刻的前项概率计算则稍微复杂一些。图(2.3)示意了如何在 $t$ 时刻计算状态 1 的前向概率。不难想象，状态转移过程在 $t$ 时刻停留在状态 $i$ 的前向概率，可以通过 $t-1$ 时刻所有状态的前项概率进一步“汇聚”到 $t$ 时刻的状态 $i$ 而得到，即：

$$\alpha_t(i) = \left[ \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} \right] b_i(x_t) \tag{2-15}$$

显然，通过不断的归纳计算，在末时刻 $T$ 累积所有状态的前向概率 $\alpha$，即可得到最终总的输出概率：

$$p(X \mid \Phi) = \sum_{i=1}^{N} \alpha_T(i) \tag{2-16}$$

通过上述的前向算法，我们可以将计算 $p(X \mid \Phi)$ 的复杂度降低至 $O(N^2T)$。

### 2.3.2　解码问题

在更多的时候，我们不仅需要计算评估问题中的 $p(X \mid \Phi)$，更需要知道 $X$ 由 $\Phi$ 中哪条状态转移序列产生的概率最大，即需要搜索出：

$$S^* = \arg\max_{S} p(X, S \mid \Phi) \tag{2-17}$$

通过寻找这样的一条最优状态序列 $S^*$，我们可以通过“显”的 $X$ 来揭示HMM中“隐”的状态信息，而这个问题的实质就是对 $X$ 进行解码。解决这一

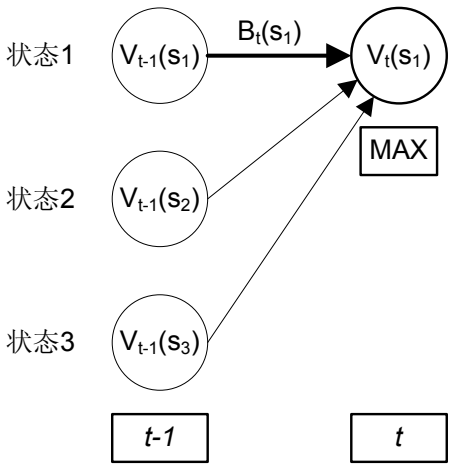


图 2.4　Viterbi算法的示意图

问题的经典方法是Viterbi算法。在Viterbi算法中，我们需要定义状态转移过程在 $t$ 时刻停留于状态 $i$ 的最优状态序列概率 $V_t(i)$，即：

$$V_t(i) = p(x_1^t, s_1^{t-1}, s_t = i \mid \Phi) \tag{2-18}$$

其中，$s_1^{t-1}$ 是从初始时刻到时刻 $t-1$ 并产生 $x_1^{t-1}$ 的最优状态序列。

不难看出，$V_t(i)$ 与前向概率 $\alpha_t(i)$ 有诸多的相似之处。唯一不同的是，$\alpha_t(i)$ 计算的是所有可能的状态序列的概率和，而 $V_t(i)$ 则总是取概率最大的一条状态序列。因此，我们可以用几乎同样的归纳算法来对 $V_t(i)$ 进行计算，而只需要将先前的求和变为取最大值。此外，为了记住概率最大的状态路径以便回溯，我们还需要保存在 $t$ 时刻停留于状态 $i$ 的最优路径信息 $B_t(i)$。总结起来，Viterbi算法的实现可以分为以下几个步骤：首先，初始化 $V_1(i) = \pi_i b_i(x_1)$、$B_1(i) = \emptyset$。然后，仿照前向算法归纳计算每一时刻针对每一状态的 $V$ 和 $B$，即：

$$\begin{aligned} V_t(i) &= \max_{1 \le j \le N} \left[ V_{t-1}(j) a_{ji} \right] b_i(x_t) \\ B_t(i) &= \arg\max_{1 \le j \le N} \left[ V_{t-1}(j) a_{ji} \right] \end{aligned} \tag{2-19}$$

这一归纳过程可以如图(2.4)所示。最后，在终止时刻 $T$，最优状态序列的概率 $p(X, S \mid \Phi) = \max_{1 \le i \le N} V_T(i)$、最优终止状态 $s_T^* = \arg\max_{1 \le i \le N} V_T(i)$。而最优的状态序列则可以从 $s_T^*$ 和 $B_T(s_T^*)$ 开始，经由 $B_t(i)$ 中所保存的回溯信息反向跟踪(Backtracking)，从而得到整条最优状态序列 $S^*$。

Vitebi算法是动态规划算法在HMM中的应用，也可以利用Viterbi算法中计算出的 $p(X, S \mid \Phi)$ 来作为评问题中 $p(X \mid \Phi)$ 的一个近似估计。与计算前向概率时的情形一样，Viterbi算法的计算复杂度同样为 $O(N^2 T)$。

### 2.3.3　训练问题

在解决了HMM的评估问题和解码问题后，对HMM参数的训练问题(或称估计问题)就显得尤其重要了。对HMM参数的估计实际上是评估问题和解码问题的基础，只有得到可靠的HMM参数，才能对观测有效的评估与解码。因此可以说，训练问题是HMM三个经典问题中最基本，同时也是最复杂的一个问题。

目前，对语音识别中HMM的参数估计问题尚没有一个闭式解。通常，我们需要用到迭代的方法来从旧的HMM参数中更新出新参数，从而在最大似然准则下达到 $p(X \mid \Phi)$ 的逐步最大化。要实现上述迭代更新过程，需要用到期望最大化算法(Expectation-Maximization Algorithm, EM)和Baum-Welch算法，又称前后向算法(Forward-Backward Algorithm, FB)。前者用于处理HMM中由状态序列隐变量(Hidden Variable)带来的不完全数据(Incomplete Data)下的训练问题，而后者用于高效的从训练数据中累积统计量(Statistics)，从而有效的抽取模型参数更新所需要的信息。

#### 2.3.3.1　EM算法

EM算法是HMM训练问题的基石，它主要解决了在不完全数据下的最大似然估计问题。从原理上讲，EM算法通过迭代，最大化完全数据(Complete Data)对数似然度的期望，从而间接的最大化对不完全数据的对数似然度。例如，在我们的语音识别训练问题中，我们只能观测到输出序列 $X$，而无法观测到产生这一输出的状态序列 $S$。这里的 $X$ 即为不完全数据，而 $(X, S)$ 的组合对即为完全数据。显然，我们的目的就是要最大化不完全数据下的目标函数 $p(X \mid \Phi)$。在以下的推导中，我们为了更形象的展现语音识别中的问题，将显变量标为 $X$、隐变量标为 $S$。但需要指出的是，EM算法实际应该以更泛化的视角来加以理解：即 $S$ 不应被狭义的理解为HMM中的状态序列，而应被广义的看作不完全数据最大似然估计下的所有隐变量的代表。

由贝叶斯公式，我们有：

$$p(X, S \mid \Phi) = p(S \mid X, \Phi) p(X \mid \Phi) \tag{2-20}$$

因此，

$$\log p(X \mid \Phi) = \log p(X, S \mid \Phi) - \log p(S \mid X, \Phi) \tag{2-21}$$

上式分别在两边针对观测 $X$ 及旧模型参数 $\Phi^{(0)}$ 下的隐变量 $S$ 求期望，有：

$$E[\log p(X \mid \Phi)]_{S \mid X, \Phi^{(0)}} = E[\log p(X, S \mid \Phi)]_{S \mid X, \Phi^{(0)}} - E[\log p(S \mid X, \Phi)]_{S \mid X, \Phi^{(0)}} \tag{2-22}$$

令：

$$Q(\Phi \mid \Phi^{(0)}) = E[\log p(X, S \mid \Phi)]_{S|X,\Phi^{(0)}} = \sum_{S} p(S \mid X, \Phi^{(0)}) \log p(X, S \mid \Phi) \tag{2-23}$$

$$\mathcal{H}(\Phi \mid \Phi^{(0)}) = E[\log p(S \mid X, \Phi)]_{S|X,\Phi^{(0)}} = \sum_{S} p(S \mid X, \Phi^{(0)}) \log p(S \mid X, \Phi) \tag{2-24}$$

显然有：

$$E[\log p(X \mid \Phi)]_{S|X,\Phi^{(0)}} = \log p(X \mid \Phi) = Q(\Phi \mid \Phi^{(0)}) - \mathcal{H}(\Phi \mid \Phi^{(0)}) \tag{2-25}$$

根据简森不等式(Jenson's Inequality)[60]，对于 $a_i > 0$，$\sum_i a_i = 1$ 的一系列系数来说，有：

$$\sum_i a_i \log x_i \le \log \sum_i a_i x_i \tag{2-26}$$

所以，可以推知：

$$\begin{aligned} &\mathcal{H}(\Phi \mid \Phi^{(0)}) - \mathcal{H}(\Phi^{(0)} \mid \Phi^{(0)}) = \sum_{S} p(S \mid X, \Phi^{(0)}) \log \frac{p(S \mid X, \Phi)}{p(S \mid X, \Phi^{(0)})} \\ &\le \log \sum_{S} p(S \mid X, \Phi^{(0)}) \frac{p(S \mid X, \Phi)}{p(S \mid X, \Phi^{(0)})} = \log \sum_{S} p(S \mid X, \Phi) = 0 \end{aligned} \tag{2-27}$$

因此，不难得到：

$$\begin{aligned} &\log p(X \mid \Phi) - \log p(X \mid \Phi^{(0)}) \\ &= [Q(\Phi \mid \Phi^{(0)}) - Q(\Phi^{(0)} \mid \Phi^{(0)})] - [\mathcal{H}(\Phi \mid \Phi^{(0)}) - \mathcal{H}(\Phi^{(0)} \mid \Phi^{(0)})] \\ &\ge Q(\Phi \mid \Phi^{(0)}) - Q(\Phi^{(0)} \mid \Phi^{(0)}) \end{aligned} \tag{2-28}$$

(2-28)式对于迭代优化 $\log p(X \mid \Phi)$ 的意义是显而易见的：在每一步迭代中，我们都可以间接的只对 $Q$ 进行优化，而在优化 $Q$ 的同时，$\log p(X \mid \Phi)$ 的优化幅度将比 $Q$ 的优化幅度来得更大。通常，上面式子中的 $Q(\Phi \mid \Phi^{(0)})$ 被称为 $Q$–函数，或辅助函数(Auxiliary Function)。由此，我们就可以通过在完全数据下优化 $Q$ 来实现对不完全数据下 $\log p(X \mid \Phi)$ 的优化。而最终通过迭代，$\log p(X \mid \Phi)$ 将随着 $Q$ 收敛于某局部最优点。由于在上述过程中我们首先对目标函数取期望(Expectation)，再对取期望后的辅助函数进行实质的最大化(Maximization)，所以整个过程因为这两步而被命名为EM算法。

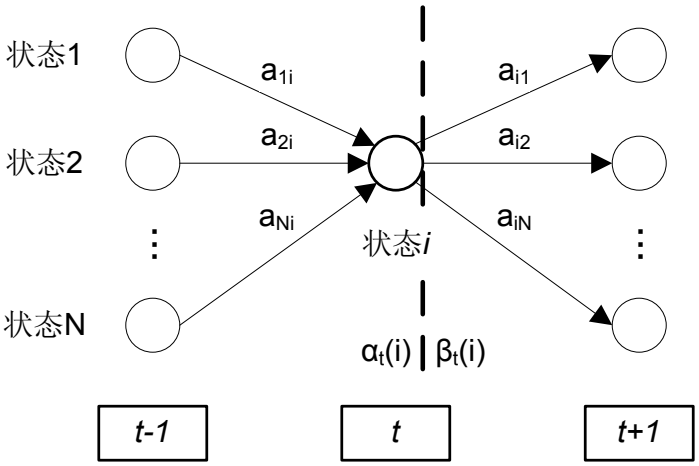


图 2.5　前向概率与后向概率关系的示意图

#### 2.3.3.2　Baum-Welch算法

Baum-Welch算法可以被看作EM算法在HMM背景下的具体应用，在这里，显变量为观测，隐变量为HMM状态序列。Baum-Welch算法又被称为前后向算法，这是因为除了前向概率，它的实现还需要定义如下的后向概率：

$$\beta_t(i) = p(x_{t+1}^T \mid s_t = i, \Phi) \tag{2-29}$$

后向概率的物理意义即是给定HMM状态跳转序列在 $t$ 时刻处于状态 $i$ 的条件下，产生其后的部分观测 $x_{t+1}^T$ 的概率。前向概率 $\alpha_t(i)$ 与后向概率 $\beta_t(i)$ 的关系可以用图(2.5)示意。

与前向概率一样，对后向概率也可以用归纳算法来快速计算。首先，定义 $\beta_T(i) = 1$，其后，从时刻 $T-1$ 开始向前，逐时刻计算：

$$\beta_t(i) = \left[ \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \right] \tag{2-30}$$

除此之外，我们还需要定义跳转占有率 $\gamma_t(i,j)$，它表示在给定HMM模型及观测序列 $X$ 的情况下，状态转移序列在 $t$ 时刻从状态 $i$ 跳转到 $j$ 的概率，即：

$$\begin{aligned} \gamma_t(i,j) &= p(s_{t-1} = i, s_t = j \mid x_1^T, \Phi) \\ &= \frac{p(s_{t-1} = i, s_t = j, x_1^T \mid \Phi)}{p(x_1^T \mid \Phi)} \\ &= \frac{\alpha_{t-1}(i) a_{ij} b_j(x_t) \beta_t(j)}{\sum_{k=1}^{N} \alpha_T(k)} \end{aligned} \tag{2-31}$$

最后，便是利用EM算法的思想及其辅助函数 $Q$，来迭代更新模型参数 $\Phi = \{\pi, \mathbf{A}, \mathbf{B}\}$。首先，根据(2-23)式可得：

$$Q(\Phi \mid \Phi^{(0)}) = \sum_{S} \frac{p(X, S \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \log p(X, S \mid \Phi) \tag{2-32}$$

令 $a_{s_0 s_1} = \pi_{s_1}$，于是有：

$$\begin{aligned} p(X, S \mid \Phi) &= \prod_{t=1}^{T} a_{s_{t-1} s_t} b_{s_t}(x_t) \\ \Rightarrow \log p(X, S \mid \Phi) &= \sum_{t=1}^{T} \log a_{s_{t-1} s_t} + \sum_{t=1}^{T} \log b_{s_t}(x_t) \end{aligned} \tag{2-33}$$

则(2-32)式可以拆分为 $Q(\Phi \mid \Phi^{(0)}) = Q(\mathbf{a} \mid \Phi^{(0)}) + Q(\mathbf{b} \mid \Phi^{(0)})$，其中：

$$Q(\mathbf{a} \mid \Phi^{(0)}) = \sum_{t} \sum_{i} \sum_{j} \frac{p(X, s_{t-1} = i, s_t = j \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \log a_{ij} \tag{2-34}$$

$$Q(\mathbf{b} \mid \Phi^{(0)}) = \sum_{t} \sum_{i} \frac{p(X, s_t = i \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \log b_i(x_t) \tag{2-35}$$

至此，我们可以单独的各自最大化 $Q(\mathbf{a} \mid \Phi^{(0)})$ 和 $Q(\mathbf{b} \mid \Phi^{(0)})$。对于前者，由于 $\sum_{j=1}^{N} a_{ij} = 1$，不难求得：

$$a_{ij} = \frac{\frac{1}{p(X \mid \Phi^{(0)})} \sum_{t} p(X, s_{t-1} = i, s_t = j \mid \Phi^{(0)})}{\frac{1}{p(X \mid \Phi^{(0)})} \sum_{t} p(X, s_{t-1} = i \mid \Phi^{(0)})} = \frac{\sum_{t} \gamma_t(i, j)}{\sum_{t} \sum_{k} \gamma_t(i, k)} \tag{2-36}$$

而对于后者，我们可以用两种情况为例来进行说明。首先，如果 $b_i(x)$ 是离散概率，其码字来自于一个有限长度码本 $O = \{o_1, o_2, \ldots, o_M\}$，那么：

$$Q(\mathbf{b} \mid \Phi^{(0)}) = \sum_{t} \sum_{i} \sum_{k} \frac{p(X, s_t = i \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \log b_i(k) \cdot \delta(x_t, o_k) \tag{2-37}$$

又由 $\sum_{k=1}^{M} b_i(k) = 1$ 可得：

$$b_i(k) = \frac{\frac{1}{p(X \mid \Phi^{(0)})} \sum_{t} p(X, s_t = i \mid \Phi^{(0)}) \cdot \delta(x_t, o_k)}{\frac{1}{p(X \mid \Phi^{(0)})} \sum_{t} p(X, s_t = i \mid \Phi^{(0)})} = \frac{\sum_{t} \sum_{j} \gamma_t(j, i) \cdot \delta(x_t, o_k)}{\sum_{t} \sum_{j} \gamma_t(j, i)} \tag{2-38}$$

而如果 $b_i(x)$ 是一个以单高斯表示的连续概率密度函数，即：

$$b_i(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right] \tag{2-39}$$

由于 $\int b_i(x)\,\mathrm{d}x = 1$ 恒满足，训练问题即转化为对 $b_i(x)$ 中参数 $\mu_i$ 和 $\sigma_i^2$ 的估计问题。由(2-35)式不难得出：

$$\begin{aligned}\frac{\partial Q(\mathbf{b} \mid \Phi^{(0)})}{\partial \mu_i} &= \sum_t \frac{p(X, s_t = i \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \frac{\partial \log b_i(x_t)}{\partial \mu_i} \\ &= \sum_t \frac{p(X, s_t = i \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \frac{x_t - \mu_i}{\sigma_i^2}\end{aligned} \tag{2-40}$$

$$\begin{aligned}\frac{\partial Q(\mathbf{b} \mid \Phi^{(0)})}{\partial \sigma_i^2} &= \sum_t \frac{p(X, s_t = i \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \frac{\partial \log b_i(x_t)}{\partial \sigma_i^2} \\ &= \sum_t \frac{p(X, s_t = i \mid \Phi^{(0)})}{p(X \mid \Phi^{(0)})} \frac{(x_t - \mu_i)^2 - \sigma_i^2}{2\sigma_i^4}\end{aligned} \tag{2-41}$$

令上两式为 0 并求解 $\mu_i$、$\sigma_i^2$，再考察二阶导数的符号可得，使得函数 $Q(\mathbf{b} \mid \Phi^{(0)})$ 最大时的 $\mu_i$ 及 $\sigma_i^2$ 分别为：

$$\mu_i = \frac{\sum_t \sum_j \gamma_t(j,i) x_t}{\sum_t \sum_j \gamma_t(j,i)} \tag{2-42}$$

$$\sigma_i^2 = \frac{\sum_t \sum_j \gamma_t(j,i)(x_t - \mu_i)^2}{\sum_t \sum_j \gamma_t(j,i)} \tag{2-43}$$

至此，我们利用EM算法的思想，使得每次迭代后得到的模型新参数都能够使得目标函数 $p(X \mid \Phi)$ 单调上升。总结起来，Baum-Welch算法的步骤可以分为以下 4 步：1、选择初始模型参数 $\Phi^{(0)}$；2、E-步(E-Step)，构造辅助函数 $Q(\Phi \mid \Phi^{(0)})$；3、M-步(M-Step)，最大化 $Q(\Phi \mid \Phi^{(0)})$，得到新的模型参数 $\Phi$；4、迭代，设置 $\Phi^{(0)} = \Phi$，重复2-4步直至收敛。应用这样的方法，我们就成功的解决了对HMM参数的训练问题。

## 2.4　基于HMM的语音识别声学模型

在解决了HMM的评估、解码及训练这些基础问题后，我们就可以使用HMM对语音识别中的声学模型进行建模。我们需要根据语音的特点选择建模单元，并为不同的建模单元建立不同的HMM。语音识别建模单元的选取应该考虑到一致性、可训练性和可共享性[61]。所谓一致性，是指不同语音实例中相同的语音单元需要具备声学上一致的特征；可训练性是指对于一个建模单元

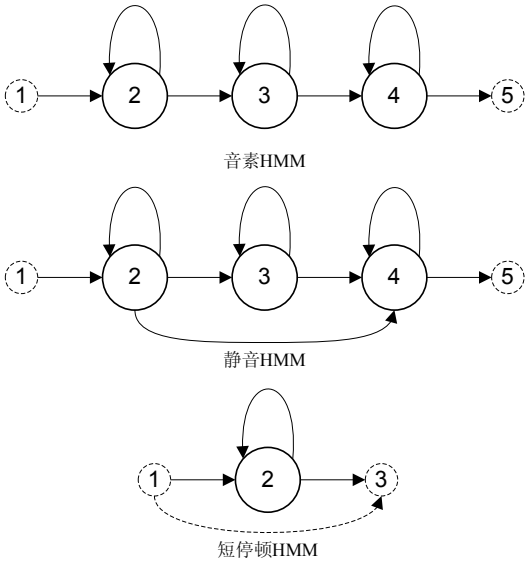


图 2.6　不同建模单元对应的HMM拓扑结构

| 建模单元 | [静音] | 语 | | 音 | | [静音] |
|---|---|---|---|---|---|---|
| mono-phone | sil | y | u | y | in | sil |
| 右相关声韵母 | sil | y+u | u | y+i | in | sil |
| tri-phone | sil | sil-y+u | y-u+y | u-y+in | y-in+sil | sil |

表 2.1　普通话中“语音”一词对应不同建模单元的拆分方法

需要能够得到足够的训练数据来对其参数进行估计；而可共享性则是指是否可以在不同的建模单元之间共享某些具有共性的训练数据。

正是基于上述 3 点，通常的语音识别建模单元一般采用音素(phone)、音节(syllable)以及词(word)等。而在大词汇量连续语音识别中，常用音素作为基本的建模单元。此外，为了考虑连续语音中的协同发音(Coarticulation)现象，一般还采用上下文相关(Context-Dependent)的音素建模，如三元音素(Tri-Phone)建模等。根据普通话语音的特点，还有学者提出采用右相关声韵母来进行建模，即将普通话中的音节拆分为一个右相关声母及独立韵母的组合。我们以普通话中“语音”一词为例，给出其根据一元音素(Mono-Phone)、右相关声韵母及三元音素的建模单元拆分方法，如表(2.1)所示。

在建模单元确定以后，我们就可以根据其特点为各单元分配适当的HMM拓扑结构。一般来说，对普通的音素单元常采用自左向右的无跨越HMM，而对静音模型sil及短停顿模型sp等，则可采用可跨越的3状态及单状态HMM，如图(2.6)所示。

而对于每个HMM状态的输出概率，在大词汇量连续语音识别任务中通常用高斯混合模型(Gaussian Mixture Model, GMM)来表达，即：

$$b_i(x) = \sum_{m=1}^{M} \frac{c_{im}}{\sqrt{(2\pi)^D |\mathbf{\Sigma}_{im}|}} \exp\left[ -\frac{1}{2}(\mathbf{x} - \mu_{im})^\top \mathbf{\Sigma}_{im}^{-1}(\mathbf{x} - \mu_{im})\right] \tag{2-44}$$

其中，$c_{im}$ 是状态 $i$ 中第 $m$ 个混合高斯成份的权重，$D$ 是观测向量 $x$ 的维数，$\mu_{im}$ 和 $\mathbf{\Sigma}_{im}$ 则分别为均值向量及协方差矩阵。在当今典型的语音识别声学模型中，每个状态的混合高斯成份数至少需要 8 到 12 个，因此也就对训练数据量的规模提出了更高的要求。同时应当看到，声学模型建模精度与训练数据量的矛盾在目前仍是极为突出的：为了更精细的建模，我们必须采用上下文相关模型、并使用更多的混合高斯成份数。这样一来，就使得模型参数的数目急剧增加。而与之相对，训练数据相比语音中的复杂现象来说，本来就非常稀疏，要在如此稀疏的数据之上可靠的估计如此之多的模型参数又会成为另一个困难的任务。有鉴于此，声学模型参数绑定技术就显得尤为重要。目前，在大词汇量连续语音识别中最常用的模型参数绑定方法是最大似然准则下基于决策树(Decision Tree)的状态绑定[13,14]。除此之外，我们还可以在模型、状态、混合高斯成份乃至特征等多个层面对模型参数进行绑定与压缩[62–65]。这些技术确保了模型参数在较高的建模精度下仍能得到可靠的估计。

## 2.5 本章小结

在本章中，我们详细介绍了HMM及其在语音识别中的基本应用，包括HMM的数学原理、经典问题，特别是它在最大似然准则下对应于语音识别问题的建模方法和参数估计方法。HMM天生不是针对语音识别问题而提出的，但从目前的情况看来，语音识别却天生就需要HMM这样的统计模型来加以实现。在过去很长一段时间内，HMM都作为最主流的语音识别声学模型建模方法而存在。在HMM基础上所发展的各种配套方法也已十分成熟。虽然在近年来我们可以发现一些其它替代性声学模型的出现，但从近期来看，HMM仍会在语音识别领域占据支配地位。而本文后续章节的工作，也同样将基于HMM声学模型展开。

# 第 3 章　传统区分性训练准则及区分性训练统一准则框架

## 3.1　引言

基于最大似然估计MLE训练的隐马尔科夫模型毫无疑问的已经成为目前自动语音识别系统中声学模型最主流、最常规的标准配置。这与MLE估计的一些优良性质是分不开的。首先，MLE估计提供了一种简单的方法，使得一个较高精度的语音识别声学模型能够迅速的训练得到。其次，由于采用了EM算法、Baum-Welch算法等方法，MLE估计能够放松对参考文本标注精度的要求(例如不要求精细的音段时间标注)，并在每步迭代中确保对目标函数的优化。最后，MLE估计对训练资源的消耗较小，业已存在诸多成熟的算法和工具能够高效的对模型参数进行估计。

但是应该看到，MLE估计在理论上所做的一些假设影响了其在实际条件下训练出最优分类器的能力。具体来说，针对大词汇量连续语音识别这一应用而言，MLE估计训练出最优HMM声学模型需要满足以下几个条件：首先，模型假设要正确，即建模时指定的概率密度函数要能够代表实际语音的“真实”分布；第二，训练数据要趋于无穷，即可以经由无穷多的数据估计出模型的“真实”参数；最后，解码时需要的语言模型要事先已知，且参数要完全“真实”。

不难看出，以上的三点“真实”假设在现实中无一可以真正达到。首先，语音参数分布的“真实”情况是完全不可测得的，更谈不上用通常意义上的指数族函数(在语音识别中通常采用混合高斯函数)来充分模拟。其次，对于语音识别中大量的模型参数而言，训练数据总是稀疏的，实际情况下所能收集到的训练数据量远达不到无穷的要求。最后，解码中语言模型存在的问题与声学模型几乎完全一样，它也同样无法得到所谓“真实”的参数。因此，在现实条件下通过MLE估计得到最优分类器是绝无可能的。

针对这一情况，研究者相继提出了多种区分性训练方法，以期在现实条件下得到较优的分类器。区分性训练方法通常通过定义某一目标函数(Objective Function)，通常称准则(Criterion)，来近似一个与分类代价相关的度量。例如，可以定义一个与分类错误相关的量并最小化它；或是定义一个与识别正确率相关的量，并最大化它。

区分性训练方法在语音识别领域的应用已经超过 20 年。最初，区分性训练通常在小词汇量任务上的能够得到较好的效果，而由于近年来一些关键技术的研究突破，区分性训练已成为大词汇量连续语音识别系统中非常重要的训练手

段。通过区分性训练，我们可以从一定程度上弱化模型假设错误所带来的影响。同时，由于区分性训练致力于优化与识别效果好坏相关的度量，因此也就为提高识别器性能提供了更直接的途径。形象的说，MLE训练告诉模型“这是椅子，那是桌子”，而区分性训练则告诉模型“这是桌子而不是椅子，那是椅子而不是桌子”。MLE训练更重视调整模型参数以反映训练数据的概率分布，而区分性训练则更重视调整模型之间的分类面，以更好的根据设定的准则对训练数据进行分类。

区分性训练研究的重点方向大致有两个，一是定义准则，即表明“需要优化什么”，二是研究优化算法，即如何根据给定的准则有效的优化模型参数。在语音识别领域常用的区分性训练准则主要包括：最大互信息量估计准则(Maximum Mutual Information Estimation, MMIE)、最小分类错误准则(Minimum Classification Error, MCE)，以及最小词 / 音素错误准则(Minimum Word / Phone Error)。而常用的参数优化算法则包括广义概率下降(Generalized Probability Descent, GPD)，以及扩展Baum-Welch(Extened Baum-Welch, EB)算法。在本章中，上述这些方法都将得到一一介绍。

本章的后续部分组织如下：(3.2)节首先介绍贝叶斯决策理论，并解释为何贝叶斯代价在实践中难以真正达到的原因。(3.3)至(3.6)节分别介绍目前最常用的MMIE、MCE、MWE / MPE等几种准则的历史和原理，并在(3.7)节将它们融合进区分性训练统一准则框架中进行说明。接着，在(3.8)节，主要介绍如何根据区分性训练准则对模型参数进行优化。最后，我们会在(3.9)节探讨区分性训练方面目前所存在的一些问题，并在(3.10)节给出本章小结。

## 3.2　贝叶斯决策理论

设有一 $c_1$ 到 $c_M$ 的 $M$ 类分类问题，目标是将任一随即变量 $X$ 分到 $M$ 类中。定义将本属于第 $j$ 类的样本分类为第 $i$ 类的代价为 $e_{ij}$，那么，识别器作出某一决策 $C(x) = c_i$ 的代价函数 $\mathcal{R}$ 即可表示为：

$$\mathcal{R}(c_i \mid x) = \sum_{j=1}^{M} e_{ij} \cdot p(c_j \mid x) \tag{3-1}$$

那么，识别器操作时的总体期望代价即可表示为：

$$\mathcal{L} = \int \mathcal{R}(C(x) \mid x)\, \mathrm{d}p(x) \tag{3-2}$$

通常，分类代价 $e_{ij}$ 被设置为0-1代价，即分类正确的代价为0，分类错误的代价为1：

$$e_{ij} = \begin{cases} 0 & \text{如果 } i = j \\ 1 & \text{如果 } i \neq j \end{cases} \tag{3-3}$$

那么在此情况下，(3-1)式可改写为：

$$\mathcal{R}(c_i \mid x) = 1 - p(c_i \mid x) \tag{3-4}$$

那么，使得(3-2)式最小的分类器行为 $C(x)$ 应满足：

$$C(x) = \arg\min_i \mathcal{R}(c_i \mid x) = \arg\max_i p(c_i \mid x) \tag{3-5}$$

上式就是所谓最大后验概率MAP决策。而由该决策所得到的最小错误代价被称为贝叶斯风险(Bayesian Risk)。当所有后验概率已知时，基于MAP准则的分类器就可以成为贝叶斯决策理论意义下的最优分类器。但显然，在语音识别这类任务中，各类别的后验概率实际无法准确的给出，也正是因为如此，在决策时需要考虑使用其它方法将后验概率的求取转换为先验概率的估计。通常，我们使用如下的贝叶斯公式进行这种转换：

$$p(c_i \mid x) = \frac{p(x \mid c_i) \cdot p(c_i)}{p(x)} \tag{3-6}$$

由于上式中的 $p(x)$ 与决策无关，所以实际上只需获得 $p(x \mid c_i)$ 与 $p(c_i)$ 的先验知识即可。在语音识别中，这两者分别为声学模型概率和语言模型概率，而在这里，我们仅专注于对声学模型 $p(x \mid c_i)$ 的估计。正如本章引言部分所提到的，对 $p(x \mid c_i)$ 采用MLE估计固然可行，但由于实际条件的约束，往往无法得出“真实”的结果。这是因为，首先，我们必须对 $p(x \mid c_i)$ 的概率分布进行建模。然而现实世界的问题往往异常复杂，而我们为了能够处理，又不得不选择一些数学上较为简单的函数以便于操作。这一矛盾使得我们实际上根本无法用简单的数学模型来表达复杂的数据分布，也使得错误的模型假设成为了声学模型建模的根本性问题。其次，即使承认这样错误的模型假设，我们还是无法取得无穷多的数据来对模型参数进行有效的估计。特别是对于语音识别这样的任务来说，相对于现实中大量存在的语音变体(说话人差异、口音差异、环境差异等等)，训练数据总是稀疏的。

基于上述原因，理论意义上最优的MAP决策在语音识别任务中实际不可能作出。而与MAP决策理论密不可分的最大似然估计也就无法得到最优的实际分类效果。也正是因为如此，我们需要使用区分性训练的方法来在现实的诸多限制下提高声学模型的区分能力。可以说，区分性训练的出现正是为了有针对性

的弱化现实世界中上述问题所带来的负面影响的。在接下来的几节中，我们就将介绍当前最常用的几种区分性训练方法。

## 3.3　最大互信息量估计MMIE准则

### 3.3.1　MMIE准则的历史

MMIE准则很早就被应用于语音识别领域。在文献[19]中，Bahl等研究者将MMIE准则应用于基于离散HMM的话者相关2000孤立词识别任务中，并取得了相对于MLE估计 18% 的性能提升。这也是MMIE准则被成功实现的最早的文献之一。在文献[66]中，MMIE准则被推广到连续密度HMM中，并在英文E-Set任务上同样取得了 18% 左右的识别性能提升。在同时期的类似工作中，MMIE准则的研究通常集中在小任务、乃至离散HMM下[67]。相比MLE估计，MMIE准则都能够得到较明显的性能提升。

紧接着，研究者开始试图将MMIE准则应用到小词汇量连续语音识别任务上。在文献[68]中，MMIE被应用到1000词表的连续语音RM任务上，但是识别性能却只有很小的提升。MMIE准则真正被成功应用到基于连续HMM的语音识别任务上，应该首推Normandin在文献[69]中的工作。他在TI Digits连续数字串识别任务上的实验证明，MMIE在连续HMM下可以得到句子级近 50% 的相对错误下降。但当同样的方法应用到大词汇量连续语音识别任务后，MMIE准则对词错误率的降低又不明显了[70]。

在文献[71,72]中，由于词图等相关技术的引入，MMIE准则终于在大词汇量连续语音识别任务上也取得了明显超过MLE估计的性能。在6万多词的WSJ数据库上，MMIE准则能够取得 5% ~ 10% 的相对错误率下降。但在当时的运算资源条件下，要在词图中累积统计量仍是一件相当费时的工作。正是因为如此，在文献[73,74]中，基于整个句子的词图统计量累积工作被放到了帧一级，这就使得运算消耗能够大大降低。而从这些工作所报告的结果来看，这样的简化对识别性能的负面影响实际并不大。当然，随着运算资源在近 10 年来的飞速发展，这样的简化在目前已显得不再需要了。

### 3.3.2　MMIE准则的原理

令 $\mathcal{W}$ 表示语音中所含信息的随机变量(例如音素、孤立词、词串等)，$W$ 为它的实例；再令 $\boldsymbol{O}$ 为表示观测序列的随机变量，$\boldsymbol{O}$ 为它的实例。从信息论的观点出发，我们可以说信息 $W$ 被编码为 $\boldsymbol{O}$。在 $\boldsymbol{O}$ 给定的情况下，描述对 $\mathcal{W}$ 的平均

不确定性的度量是条件熵 $H(\mathcal{W} \mid O)$，它可以被写为：

$$H(\mathcal{W} \mid O) = -\sum_{W,O} p(W,O) \log p(W \mid O) = -E[\log p(W \mid O)] \tag{3-7}$$

从直觉上讲，我们的目标就是要降低这个不确定度，使得我们的解码器在解码时总能做出自认更为“确信”的判决。在实际的语音识别声学模型建模过程中，我们通常使用一个参数化的模型 $\Lambda$ 来近似求得“真实”的后验概率 $p(W \mid O)$。也就是说，我们实际只能得到对 $p(W \mid O)$ 的一个参数化模拟，即 $p_\Lambda(W \mid O)$。进一步，我们有：

$$\begin{aligned}
H_\Lambda(\mathcal{W} \mid O) &= -E[\log p_\Lambda(W \mid O)] \\
&= -\sum_{W,O} p(W,O) \log p_\Lambda(W \mid O) \\
&= -\sum_{W,O} p(W,O) \log \frac{p_\Lambda(W \mid O)}{p(W \mid O)} - \sum_{W,O} p(W,O) \log p(W \mid O) \\
&\geq -\sum_{W,O} p(W,O) \left[ \frac{p_\Lambda(W \mid O)}{p(W \mid O)} - 1 \right] + H(\mathcal{W} \mid O) \quad (\because \log x \leq x - 1) \\
&= H(\mathcal{W} \mid O)
\end{aligned} \tag{3-8}$$

也就是说，由模型得到的条件熵 $H_\Lambda(\mathcal{W} \mid O)$ 是真实条件熵 $H(\mathcal{W} \mid O)$ 的一个上界。最小化 $H_\Lambda(\mathcal{W} \mid O)$ 就能够使其向真实的 $H(\mathcal{W} \mid O)$ 无限趋近，其实质也就是 $p_\Lambda(W \mid O)$ 向 $p(W \mid O)$ 的趋近。当 $H_\Lambda(\mathcal{W} \mid O) = H(\mathcal{W} \mid O)$ 时，我们自然有 $p_\Lambda(W \mid O) \equiv p(W \mid O)$。

在信息论中，我们知道互信息量 $I(\mathcal{W};O)$ 可表示为：

$$I(\mathcal{W};O) = H(\mathcal{W}) - H(\mathcal{W} \mid O) \tag{3-9}$$

假设语言模型熵 $H(\mathcal{W})$ 是固定且已知的，在引入参数化的声学模型近似后，上式即可变为：

$$I_\Lambda(\mathcal{W};O) = H(\mathcal{W}) - H_\Lambda(\mathcal{W} \mid O) \tag{3-10}$$

因此，最小化 $H_\Lambda(\mathcal{W} \mid O)$ 的过程也就是最大化互信息量 $I_\Lambda(\mathcal{W};O)$ 的过程，这也正是MMIE准则得名的原因。在文献[75]中已经证明，在这种情况下的MMIE准则实质上等价于条件最大似然准则(Conditional Maximum Likelihood)[76]。

在实践中，因为训练数据量的约束，我们只能对 $H_\Lambda(\mathcal{W} \mid O)$ 进行估计，并记为 $\hat{H}_\Lambda(\mathcal{W} \mid O)$。通常，我们将期望改为对训练集中所有训练语料的求和，即最小

化：

$$\hat{H}_{\Lambda}(\mathcal{W} \mid \mathcal{O}) = -\frac{1}{R}\sum_{r=1}^{R}\log p_{\Lambda}(W_r \mid O_r) \tag{3-11}$$

在通常的基于HMM语音识别任务中，上式中的求和项就是作为参考的正确模型序列的后验概率。这一后验概率经由计算正确模型序列与所有可能的模型序列的似然度之比而求得，也即：

$$p_{\Lambda}(W_r \mid O_r) = \frac{p_{\Lambda}(O_r \mid W_r)p(W_r)}{\sum_{W' \in \mathcal{M}} p_{\Lambda}(O_r \mid W')p(W')} \tag{3-12}$$

因此，最大互信息量准则还可以看作是对训练集中所有训练语料正确模型序列后验概率的最大化，即最大化：

$$\mathcal{F}_{\mathrm{MMIE}} = \frac{1}{R}\sum_{r=1}^{R}\log p_{\Lambda}(W_r \mid O_r) = \frac{1}{R}\sum_{r=1}^{R}\log \frac{p_{\Lambda}(O_r \mid W_r)p(W_r)}{\sum_{W' \in \mathcal{M}} p_{\Lambda}(O_r \mid W')p(W')} \tag{3-13}$$

这也正好可以成为MMIE准则又一直观解释。

## 3.4　最小分类错误MCE准则

### 3.4.1　MCE准则的历史

作为另一大类区分性训练准则，MCE准则最初由Juang在文献[20]中被应用到语音识别中。顾名思义，MCE准则就是要致力于最小化一个与识别器分类错误相关的度量。我们知道，真正的识别错误度量，如句子错误率等，是关于模型参数的一个离散函数。而为了要成为一个可导并可优化的准则，MCE实际上是对上述错误度量的一个平滑近似。事实上，MCE准则可以看作对训练集上总体经验错误率的平滑逼近，最小化这一准则，即可相应的带来至少是训练集上的错误下降。

一直以来，MCE准则常常在较小规模的语音识别任务上表现出相当良好的性能。在TIMIT连续音素识别任务上，MCE准则可以带来相对MLE 5% 到 14% 的性能提升[77]。同样的，在TI Digits连续数字串识别任务上，MCE则可带来 25% 以上的性能提升[78,79]。在文献[80]中，MCE准则还被应用于中等词汇任务上，并取得了约 10% 的相对错误率下降。

在词图等技术逐渐进入区分性训练之后，MCE准则被继续扩展以适应新条件下的训练任务。在文献[24,81]中，研究者推导了MCE准则在词图环境下的高效实现方法，并在WSJ0及NAB这样的大词汇量连续语音识别任务上进行了实验。实验证明，MCE准则在大任务上仍能带来可靠的性能提升，幅度可达 5% ~ 10% 左右，与MMIE准则相当。

另一个与MCE准则密切相关的准则是最小验证错误准则(Minimum Verification Error, MVE)[82,83]。MVE准则将MCE的分类问题转化为验证问题，而后者的性能在话者识别及置信度判决任务中是至关重要的指标。MVE准则在一些典型任务上能够比MLE准则降低 25% 的虚警、漏警及等错误率。在文献[82]的话者识别任务中，基于MVE的话者自适应相比传统的MAP自适应带来了超过 50% 的性能提升。

在语音识别领域应用MCE准则，一个常常被提及的问题是准则与目标的不匹配。通常的MCE准则降低的是平滑后的句子级分类错误，这与大词汇量连续语音识别的目标，即降低词错误率，是不匹配的。因此，也有工作尝试将句子级的MCE准则细化到词级或音素级[77,84]。总体上来说，这些工作所报告的性能提升相对于传统句子级MCE并不大。而本文也将在第 4 章中对这一问题进行继续的探讨。

### 3.4.2　MCE准则的原理

MCE准则的核心思想，是使用区分函数(Discriminant Function，或称判别函数)设计分类器。设有一 $c_1$ 到 $c_M$ 的 $M$ 类分类问题，在基于区分函数的分类器设计中，我们为每一类定义一个区分函数 $\{g_i(x) \mid i = 1, \ldots, M\}$，并令分类器的分类准则为：

$$C(x) = \arg\max_i g_i(x) \tag{3-14}$$

这种分类器设计方法与MAP准则下的分类器设计有显著的不同。在这里，分类器判决的依据不一定是后验概率，而可以是任意定义的区分函数。这就使得在后验概率难以估计或估计不准时，仍能用区分函数做出合理的判决。更进一步来说，区分函数 $g_i(x)$ 可以不限定在概率体系下，它甚至可以是任意与分类相关的函数，这就为基于区分函数的分类器设计提供了很大的灵活性。最后，由于对区分函数的训练可以在一定程度上弱化对训练数据量的依赖，这就为在有限数据量下获得可靠的分类性能提供了一种可能。

显然，在上述条件下的最优分类器设计问题应该表示成：

$$\{\hat{g}_i(x) \mid i = 1, \ldots, M\} = \arg\min_{g_i(x)} \int \mathcal{R}(C(x) \mid x)\, \mathrm{d}p(x) \tag{3-15}$$

在(3-14)式所定义的分类器行为下，若使用0-1代价，上式中的分类风险可表示为：

$$\mathcal{R}(C(x) \mid x) = \sum_{i=1}^{M} \delta(g_i(x) \neq \max_j g_j(x)) \cdot \delta(x \in c_i) \tag{3-16}$$

将其代入(3-15)式，则需要最小化的总体期望风险即可写为：

$$\mathcal{L} = \int \sum_{i=1}^{M} \delta(g_i(x) \neq \max_j g_j(x)) \cdot \delta(x \in c_i)\, \mathrm{d}p(x) \tag{3-17}$$

不难看出，上式是一个针对区分函数的离散函数。根据它直接对模型参数进行优化是难以进行的。为了构成一个平滑、可优化的准则，MCE设计了如下的误分类度量(Misclassification Measure)：

$$d_i(x) = -g_i(x) + \log\left\{\frac{1}{M-1}\sum_{j,j\neq i}\exp[g_j(x)\cdot\eta]\right\}^{1/\eta} \tag{3-18}$$

以及损失代价函数(Loss Function)：

$$\ell_i(x) = \frac{1}{1+e^{-2\gamma d_i(x)+\xi}} \tag{3-19}$$

相应的，识别器工作时的总体期望代价即可表示为：

$$\widetilde{\mathcal{L}} = \int \sum_{i=1}^{M} \ell_i(x) \cdot \delta(x \in c_i)\, \mathrm{d}p(x) \tag{3-20}$$

对比(3-17)、(3-20)两式不难发现，MCE准则实际是用平滑的损失代价函数 $\ell_i(x)$ 来近似0-1代价中的 $\delta$ 函数。这种近似的逼近程度则是由参数 $\eta$、$\gamma$ 及 $\xi$ 等来控制的。

首先，$\eta$ 控制了对整个竞争空间的利用程度。当 $\eta \to \infty$ 时，(3-18)式实际变为：

$$d_i(x) = -g_i(x) + \max_{j,j\neq i} g_j(x) \tag{3-21}$$

这也就相当于只考虑第 $i$ 类与最强竞争类的对比；而当 $\eta$ 逐步缩小时，更多的竞争类将会被考虑进来，这就使得对模型参数的调整能够兼顾更多的因素。

再看参数 $\gamma$，它实际控制了sigmoid函数的形态。图(3.1)给出了不同 $\gamma$ 下的sigmoid函数形态。可以看到，当 $\gamma$ 较小时，函数较为平缓，显示损失代价随误分类度量相对平滑的变化；而当 $\gamma$ 较大时，函数形态趋于陡峭。在极端的 $\gamma \to \infty$ 情况下，sigmoid损失代价函数则将变为对于误分类度量的阶跃函数。

最后，参数 $\xi$ 可以控制sigmoid函数的平移，这实质上是对不同幅度的误分类度量分配不同的代价。也有研究者将 $\xi$ 与分类边缘(Margin)联系起来，指出这一参数实际可以看作与margin相关的一个度量[85]。

综上所述，当 $\eta \to \infty$、$\gamma \to \infty$、$\xi = 0$ 时，(3-20)式即等价于(3-17)式。而当上述三个参数取现实中的较小值时，我们则可以构造平滑适当的损失代价函数，从而得到可供优化的区分性训练准则。

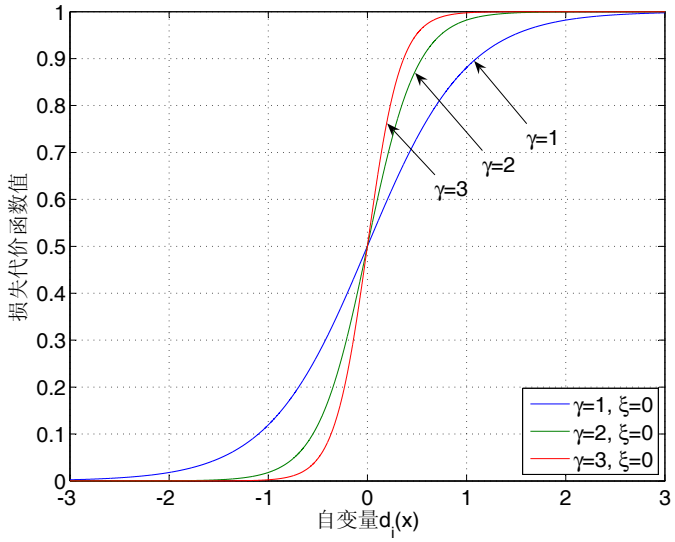


图 3.1　不同 $\gamma$ 参数下的sigmoid函数形态

通常，对(3-20)式的期望代价，我们只能用如下的经验代价来近似：

$$\widetilde{\mathcal{L}}_{\text{emp}} = \sum_{r=1}^{R}\sum_{i=1}^{M} \ell_i(x_r)\cdot\delta(x_r \in c_i) \tag{3-22}$$

其中，$R$ 是训练集的总样本数。不难看出，当 $R\to\infty$ 时，经验代价将收敛于期望代价。在语音识别任务中，为了与解码时的情况相匹配，我们通常以语音单元的生成概率作为区分函数，以训练语料所对应的符号串作为类别。这种方式被称为嵌入式的MCE训练(Embedded MCE Training)，或基于符号模型串的MCE方法(String-Model-Based MCE)，其实质就是在句子级嵌入各个孤立的HMM单元，进行同步优化。例如，对于某一句训练语料 $r$，设其参考正确文本所对应的模型符号串为 $W_r$，则对应正确类别的区分函数可以表示为：

$$g_{W_r}(O_r) = \log p_\Lambda(O_r \mid W_r)p(W_r) \tag{3-23}$$

对应错误类别的区分函数相应的为：

$$g_{\mathcal{M}_r^{\text{MCE}}}(O_r) = \log\left\{\frac{1}{|\mathcal{M}_r^{\text{MCE}}|}\sum_{W'\in\mathcal{M}_r^{\text{MCE}}} p_\Lambda^\eta(O_r \mid W')p^\eta(W')\right\}^{1/\eta} \tag{3-24}$$

其中，$\mathcal{M}_r^{\text{MCE}}$ 为所有可能的、不为 $W_r$ 的模型符号串集合，即 $\mathcal{M}_r^{\text{MCE}} = \mathcal{M}\setminus\{W_r\}$，$|\mathcal{M}_r^{\text{MCE}}|$ 则是 $\mathcal{M}_r^{\text{MCE}}$ 中所含元素的数目。

在实践中，上式的区分函数常常会作一定的简化。例如，可以对 $\eta$ 取无穷，从而使得：

$$g_{\mathcal{M}_r^{\text{MCE}}}(O_r) = \log \max_{W'\in\mathcal{M}_r^{\text{MCE}}} p_\Lambda(O_r \mid W')p(W') \tag{3-25}$$

或者，使得 $\mathcal{M}_r^{\text{MCE}}$ 仅包含数个最大可能的竞争模型串，即所谓基于n-best的MCE；又或者，为了方便在词图中的处理，取 $\eta=1$，并忽略 $|\mathcal{M}_r^{\text{MCE}}|$ 一

项，使得：

$$g_{\mathcal{M}_r^{\mathrm{MCE}}}(O_r) = \log \sum_{W' \in \mathcal{M}_r^{\mathrm{MCE}}} p_\Lambda(O_r \mid W') p(W') \tag{3-26}$$

至此，我们就可以以基于词图的MCE为例，将(3-23)、(3-26)式代入(3-19)式，继而代入(3-22)式，从而重写出如下的MCE准则，即最大化：

$$\mathcal{F}_{\mathrm{MCE}} = \frac{1}{R} \sum_{r=1}^{R} f\left( \log \frac{p_\Lambda(O_r \mid W_r) p(W_r)}{\sum_{W' \in \mathcal{M}_r^{\mathrm{MCE}}} p_\Lambda(O_r \mid W') p(W')} \right) \tag{3-27}$$

其中，平滑函数 $f(z)$ 为：

$$f(z) = -\frac{1}{1 + e^{2\gamma z}} \tag{3-28}$$

注意它与(3-19)式中的sigmoid函数有两点区别：首先，$f$ 函数与sigmoid函数相差一个负号，这是因为我们把MCE的最小化问题统一写成了与MMIE准则一致的最大化问题；其次，sigmoid函数指数项中的负号被乘进了误分类度量，这也是为了将正确模型序列的区分函数写到分子上，使其与MMIE准则在形式上相似。

## 3.5　最小词 / 音素错误MWE / MPE准则

### 3.5.1　MWE / MPE准则的历史

MWE / MPE准则的历史并不长，最早由Povey提出[22]，在最初的文献中，MPE准则在256小时Switchborad任务上相比MLE估计可以取得超过 10% 的相对性能提升，并超过MMIE准则绝对约 1%。紧接着，MWE / MPE在大词汇量连续语音识别任务上显示出了明显超越其它区分性训练准则的性能，因此也就很快在研究和商用的语音识别系统中得到了广泛的应用[26,86,87]。近年来，还存在一些对MWE / MPE准则本身的修改或改进，包括最小分歧准则(Minimum Divergence, MD)[88,89]、最小精确词错误(Minimum Exact Word Error)[90]等。文献[91]对其中的一些改动进行了较详细的对比，从实验结果来看，各种方法之间的性能差异有一些，但并不非常明显。

MWE / MPE准则还被应用到区分性特征提取(Discriminative Feature Extraction, DFE)方面[30,92]。目前，用于特征提取的fMPE已经逐渐开始进入主流系统中[26,87]。此外，MWE / MPE准则还被用于模型自适应方面[93]，在与MAP准则结合后可以取得超过传统MLE-MAP的模型自适应性能。

### 3.5.2　MWE / MPE准则的原理

MWE / MPE准则设计的初衷是寻找一种比MMIE、MCE更为接近大词汇量连续语音识别目标的准则。在这样的任务下，最小化词错误率(或最大化词正确率)比优化句子级的训练准则更为重要。直觉上很容易想到，可以通过最大化期望的词串或音素串的正确率来达到这一目标，即：

$$\Lambda = \arg\max_{\Lambda} E[\mathcal{A}(W, W_r)] = \arg\max_{\Lambda} \sum_{W \in \mathcal{M}} p_{\Lambda}^{\kappa}(W \mid O)\mathcal{A}(W, W_r) \tag{3-29}$$

其中，$\mathcal{A}(W, W_r)$ 是模型串 $W$ 相对正确的参考模型串 $W_r$ 的正确度量(通常为一个近似的词或音素正确个数，后面将会介绍它的通常计算方法)。如果 $\mathcal{A}(W, W_r)$ 定义在词一级，则准则为MWE；如果 $\mathcal{A}(W, W_r)$ 定义在音素一级，则准则为MPE。$p_{\Lambda}^{\kappa}(W \mid O)$ 为缩放后的模型串后验概率，定义为：

$$p_{\Lambda}^{\kappa}(W \mid O) = \frac{p_{\Lambda}^{\kappa}(O \mid W)p^{\kappa}(W)}{\sum_{W' \in \mathcal{M}} p_{\Lambda}^{\kappa}(O \mid W')p^{\kappa}(W')} \tag{3-30}$$

显然，当 $\kappa \to \infty$ 时，(3-29)式变为：

$$\Lambda = \arg\max_{\Lambda} \mathcal{A}\left\{ \arg\max_{W \in \mathcal{M}} [p(O \mid W)p(W)], W_r \right\} \tag{3-31}$$

此时准则完全变成了对解码时将会输出的、有着最大生成概率的模型串的词或音素正确数目的估计。实践中，$\kappa$ 应该取一个较小的有限值。这就使得MWE / MPE准则物理意义成为“使得正确率较高的模型串拥有更高的后验概率”。又由于 $\kappa$ 与调整声学模型与语言模型权重的声学规整(Acoustic Scaling)因子可以结合在一起，有时也就不再将它们作区分而统一写作一个参数。

在用词图表示模型竞争空间的情况下，由于模型串的数目非常庞大，逐一的精确计算 $\mathcal{A}(W, W_r)$ 需要大量的使用动态规划算法，因而极为耗时、难以实现。因此，通常采用近似的方法，将模型串之间的完整比较化整为零，在词图中的词弧(Word Arc)或音素弧(Phone Arc)上分别近似计算各自的正确程度，再使用前后向算法高效的计算出穿过该弧的所有模型串的平均正确程度。设某一词弧或音素弧为 $q$，计算 $q$ 的正确程度的原则是：

$$A(q) = \begin{cases} 1 & \text{如果为正确弧} \\ 0 & \text{如果为替换错误} \\ -1 & \text{如果为插入错误} \end{cases} \tag{3-32}$$

但为了计算方便，通常采用如下的近似将上述硬度量转化为软度量，即：

$$A(q) = \max_{z} \begin{cases} -1 + 2e(q, z) & \text{如果 } q \text{ 和 } z \text{ 标注相同} \\ -1 + e(q, z) & \text{如果 } q \text{ 和 } z \text{ 标注不同} \end{cases} \tag{3-33}$$

其中，$z$ 为参考模型序列中与 $q$ 有时间上交叠的任意弧，而 $e(q,z)$ 为它们交叠部分占 $z$ 总时长的比例[22,94]。

综上，我们将(3-30)式代入(3-29)式、将 $\kappa$ 并入声学规整因子，并在训练集上对所有语料求平均，就可以得到如下的MWE / MPE准则，即最大化：

$$\mathcal{F}_{\text{MWE,MPE}} = \frac{1}{R}\sum_{r=1}^{R}\frac{\sum_{W\in\mathcal{M}} p_{\Lambda}(O_r \mid W)p(W)\mathcal{A}(W,W_r)}{\sum_{W'\in\mathcal{M}} p_{\Lambda}(O_r \mid W')p(W')} \tag{3-34}$$

## 3.6　其他一些区分性训练准则

除了上面这样一些最常用的区分性训练准则以外，还有一些不太常用、或最近才刚刚提出的区分性训练准则，我们在这里也做一个简单的介绍。

首先，对应MMIE准则和MCE准则，分别还有所谓纠正训练(Corrective Training, CT)准则和纠错训练(Falsifying Training, FT)准则。这两种准则与MMIE、MCE的最大区别在于竞争空间选取的不同。在CT准则中，竞争空间选择的是所有可能模型序列中有着最大概率的一条。这样一来，如果整个序列能够被完全正确识别，则模型参数就不会有任何更改；反之，则仅对正确序列与最强的一条竞争序列进行参数更新。而在FT准则中，竞争空间选择的是最具竞争的错误模型序列，因此，训练中也仅调整正确序列与最强竞争序列的模型参数。基本上来说，这两种准则由于考察的竞争空间过于狭小，在大词汇量连续语音识别任务下对训练的推广性考虑不足，因此很少被真正使用过。

除此之外，最小分歧MD准则[88,89]是最近才提出的区分性训练新准则。MD准则可看作是MWE / MPE准则继续向帧一级的细化。同时，MD准则还将MWE / MPE中对标注文本之间的近似距离计算转换为对模型距离的度量。MD准则利用KL距离(Kullback-Leibler Divergence, KLD)[95]进行模型串之间的距离计算，从而避免了难以精确计算的、基于标注文本的距离度量。这样的距离度量不仅更为直接、细化，还有效的避免了在某些训练条件下标注不精确或无法获得人工标注的问题。

最后，目前还有一类基于分类边缘margin的区分性训练准则，如大分类边缘估计(Large Margin Estimation, LME)[96,97]、区分性分类边缘(Discriminative Margin)[85]，以及软分类边缘估计(Soft Margin Estimation, SME)[98–100]等。这些准则的共同特点是在语音识别中引入SVM中margin的概念，从而期望于能够提高区分性训练模型的推广性能。关于这类准则的研究目前刚刚开始变得热门起来，现有的实验也主要集中在原理验证方面。但客观的说，到今天为止，这些方法还有一些关键技术没有得到很好的解决，因此也没有能够在主流的大词汇量连续语音识别任务上报告出明显超越其它传统准则的优越性。

| 准　则 | $f(z)$ | $\mathcal{M}_r$ | $\alpha$ | $\mathcal{G}(W,W_r)$ |
|---|---|---|---|---|
| 最大似然ML | $z$ | $\emptyset$ | - | $\delta(W,W_r)$ |
| 最大互信息量MMIE | | $\mathcal{M}$ | 1 | |
| 纠正训练CT | | $\arg\max_W p(O_r \mid W)p(W)$ | $\infty$ | |
| 最小分类错误MCE | $-\frac{1}{1+e^{2\rho z}}$ | $\mathcal{M}\setminus\{W_r\}$ | - | |
| 纠错训练FT | | $\arg\max_{W,W\neq W_r} p(O_r \mid W)p(W)$ | $\infty$ | |
| 最小词错误MWE | $\exp(z)$ | $\mathcal{M}$ | 1 | $\mathcal{A}(W,W_r)$ |
| 最小音素错误MPE | | | | |
| 最小分歧MD | | | | $-\mathcal{D}_\Lambda(W \parallel W_r)$ |

表 3.1　区分性训练统一准则框架中一组准则的参数选取情况

## 3.7　区分性训练统一准则框架

对比上面提到的MMIE、MCE及MWE / MPE准则不难发现，这三种区分性训练准则之间存在很多共同点。如果能把它们具有共性的部分统一到一起，再分别突出个性，就能够形成一个区分性训练统一准则，从而将它们有机的联系到一起。在文献[81,101]中， Schlüter 就提出了这样的一个区分性训练统一准则框架，并在接下来的工作中得到了充实和完善[24]。

这项工作的意义有以下几个方面：首先，区分性训练统一准则将之前存在的各种区分性训练准则融合到一个框架下，这就为从理论上分析、对比它们之间的区别提供了更直观的角度；其次，由于各种准则被融合在一起，其共性的部分也就可以被统一的程序代码所实现，这就可以从实践上提供对各种准则的相对公平客观的性能对比；最后，新的区分性训练准则和模型参数优化方法都可以在现有框架下进行实现，并与传统方法相对比，这就为区分性训练准则和优化算法方面的研究提供了更好的平台和基础。

区分性训练统一准则框架可表示为最大化：

$$\mathcal{F}_{\text{Unified}} = \frac{1}{R}\sum_{r=1}^{R} f\left(\log\left[\frac{\sum_{W\in\mathcal{M}} p_\Lambda^\alpha(O_r \mid W)\cdot p^\alpha(W)\cdot \mathcal{G}(W,W_r)}{\sum_{W'\in\mathcal{M}_r} p_\Lambda^\alpha(O_r \mid W')\cdot p^\alpha(W')}\right]^{1/\alpha}\right) \tag{3-35}$$

对应各种准则，统一框架中的平滑函数 $f(z)$、竞争空间 $\mathcal{M}_r$、指数因子 $\alpha$，以及增益函数 $\mathcal{G}(W,W_r)$ 的选取列于表(3.1)中。从表中可以看到，各种准则的共性部分都统一体现在同一个准则框架中，而其准则间各自的特性则通过选取对应的参数来得到体现。因此，只要根据区分性训练统一准则对模型参数进行优化，就可以得到所选准则下的模型训练结果。

## 3.8　针对区分性训练准则的模型参数优化算法

在定义了各种区分性训练准则之后，如何根据准则对模型参数进行优化就成为了最重要的问题。一个好的模型参数优化算法必须能够有效的优化准则、

可以高效率的实现，并拥有良好的收敛性能。遗憾的是，到目前为止，对区分性训练准则还没有一种算法能够保证收敛性。这样的局面与最大似然准则下的情况完全不同。因此，研究者只能致力于寻找一些在经验上有着良好优化能力、又有较快收敛性的算法，来对区分性准则进行模型参数优化。常用的模型参数优化方法主要可以分为两类，一是基于梯度下降(Gradient Descent, GD)的算法，二是基于扩展Baum-Welch(EB)方法的算法。

### 3.8.1　基于梯度下降的模型参数优化算法

在早期提出的针对MCE准则的优化方法中，通常采用基于梯度下降的算法，其中最常用的当属广义概率下降GPD算法[20,102,103]。而针对MMIE准则的优化在最初也同样采用了基于梯度的优化算法[19]。除了对一阶梯度的使用，还有一些方法试图通过对二阶导数进行近似来得到更好的优化及收敛效果，包括Quick-Prop、R-Prop等[75,104]。但在这里，我们仅简要介绍MCE准则下经典的GPD优化算法。

对于(3-20)式定义的MCE期望代价，理论上可以证明，如果存在一个有着无穷多元素的随机变量序列 $x_t$，以及一个步长序列 $\epsilon_t$，且满足：

$$\sum_{t=1}^{\infty}\epsilon_t \to \infty, \quad \sum_{t=1}^{\infty}\epsilon_t^2 < \infty \tag{3-36}$$

那么，若使用如下的算法对模型参数进行更新：

$$\Lambda_{t+1} = \Lambda_t - \epsilon_t \mathbf{U} \nabla \ell(x_t;\Lambda)\big|_{\Lambda=\Lambda_t} \tag{3-37}$$

则总体期望代价函数 $\widetilde{\mathcal{L}}$ 将以为 1 的概率收敛到某一局部最小值点(公式中 $\mathbf{U}$ 为一正定矩阵，可以通过其调整训练中模型各参数间的相互权重。在最简单的情况下，$\mathbf{U}$ 可设为单位矩阵)。

上述理论为MCE准则的优化奠定了基础，广义的说，这一理论甚至不限于MCE准则一项，而是对所有其它满足条件的代价函数都可以进行优化。通过这一理论，我们实际不必真正计算总的代价函数，因为总的代价函数是被各个孤立观测的局部代价的梯度 $\nabla\ell(x)$ 所逐步优化的。同时应该看到，在实践中我们不可能拥有无限长的随机变量序列 $x$ (即训练集中的观测样本)，因此上述理论的前提条件实际是并不满足的。但正如经验代价对期望代价的逼近一样，我们仍可以说更多的观测样本可以带来对“真实”局部最优的更好逼近。最后，对步长 $\epsilon$ 的选择也是GPD方法比较微妙和难以操作的部分，实践中常使用经验参数进行设置。

GPD优化方法在最初MCE准则的研究中发挥了很重要的作用，并广泛的应用于连续数字串等小规模的语音识别任务上。但是如果将其扩展到大词汇量连

续语音识别任务上，其可用性就受到了极大的挑战。这主要是因为大规模任务的模型参数量非常庞大，模型之间的竞争关系也非常复杂，使用GPD方法进行优化时的步长选择因此会变得异常困难，也难以把握规律。因此，也就鲜有GPD方法在大词汇量连续语音识别任务上取得较好效果的例子。尤其是在EB方法对区分性训练统一准则的优化提出后，GPD方法已有被逐渐取代的迹象。

### 3.8.2　基于EB的模型参数优化算法

关于EB方法最早的理论基础可见于文献[105]，并在文献[106]中由有理函数进一步推广到基于概率模型的一般目标函数。在离散概率条件下，EB算法的收敛性可以被证明，但收敛所需要的理论上的训练参数 $D$ 在现实中将导致收敛速度极慢。即便如此，我们仍可以通过经验设定参数 $D$，使得EB算法能够在现实中很好的应用于离散概率体系下[105,107]。在离散HMM模型的基础上，Normandin将EB算法进一步推广到连续密度HMM上[69]。虽然在连续HMM下无法证明EB算法的收敛性，但实验证明，通过经验设置训练参数 $D$ 仍能够取得很好的模型参数更新效果。在文献[108]中，EB方法在TIMIT数据库上被证明有超过传统的基于梯度的优化方法的性能。因此，在此后的进一步研究中，EB方法陆续得以在大词汇量连续语音识别任务上进行实验[71,72]，而其更新公式中所用训练参数 $D$ 的经验性设置方法也逐步在实践中得到了探讨和比较[23,69,94,105]。由于EB算法在这类任务上所表现出的良好性能，它已逐渐成为各种区分性训练准则最常用的模型参数更新算法。近年来，也仍然有研究者从不同的角度对它进行解释，并尝试从理论上对训练参数的设置进行探讨[109]。

为了简明的介绍EB算法的原理，我们以区分性训练统一准则框架中的MMIE、MCE准则为例，介绍模型参数是如何依据准则进行优化更新的。此时，(3-35)式中的参数 $\alpha = 1$，且增益函数 $\mathcal{G}(W, W_r) = \delta(W, W_r)$ (关于其它准则在统一框架中的EB优化可以通过类似的方法推导得到)。此时，有：

$$\mathcal{F} = \frac{1}{R}\sum_{r=1}^{R} f\left(\log\left[\frac{p_\Lambda(O_r \mid W_r)\cdot p(W_r)}{\sum_{W'\in\mathcal{M}_r} p_\Lambda(O_r \mid W')\cdot p(W')}\right]\right) \tag{3-38}$$

参照文献[23]中的推导，不难得出：

$$\begin{aligned}\frac{\partial\mathcal{F}}{\partial\lambda_k} &= \frac{1}{R}\sum_{r=1}^{R} f'\left(\log\left[\frac{p_\Lambda(O_r \mid W_r)\cdot p(W_r)}{\sum_{W'\in\mathcal{M}_r} p_\Lambda(O_r \mid W')\cdot p(W')}\right]\right) \\ &\quad\cdot\sum_{t=1}^{T_r}[\gamma_{tr}(k; W_r) - \gamma_{tr}(k)]\cdot\frac{\partial\log p(O_{rt} \mid \lambda_k)}{\partial\lambda_k}\end{aligned} \tag{3-39}$$

其中，$\lambda_k$ 代表模型 $\Lambda$ 中的某一混合高斯成份 $k$ 的所有参数，$\gamma_{tr}(k; W_r)$ 为给定参考模型序列下 $k$ 在 $t$ 时刻的出现后验概率，而 $\gamma_{tr}(k)$ 则为给定竞争空间 $\mathcal{M}_r$

下 $k$ 在 $t$ 时刻出现的后验概率。

显然，由于上式过于复杂，我们很难简单的用设定 $\partial\mathcal{F}/\partial\lambda_k = 0$ 的方式进行优化。因此，一个合理的方案是对目标函数 $\mathcal{F}$ 进行某种近似，找到更易优化的辅助函数 $\mathcal{S}$，从而通过优化 $\mathcal{S}$ 来间接的优化 $\mathcal{F}$。

首先，我们选择如下的辅助函数 $\mathcal{S}$：

$$\begin{aligned}\mathcal{S}(\Lambda,\Lambda^{(0)}) \;=\; & \sum_k \frac{1}{R}\sum_{r=1}^{R} f'\left(\log\left[\frac{p_{\Lambda^{(0)}}(O_r \mid W_r)\cdot p(W_r)}{\sum_{W'\in\mathcal{M}_r} p_{\Lambda^{(0)}}(O_r \mid W')\cdot p(W')}\right]\right) \\ & \cdot \sum_{t=1}^{T_r}[\gamma_{tr}^{(0)}(k;W_r)-\gamma_{tr}^{(0)}(k)]\cdot\log p(O_{rt}\mid\lambda_k)\end{aligned} \tag{3-40}$$

其中，$\Lambda$ 为更新后的模型参数、$\Lambda^{(0)}$ 为更新前的模型参数，而所有标有上标 (0) 的参量均表示采用更新前的模型参数求得。不难看出，$\mathcal{S}$ 与 $\mathcal{F}$ 在更新前的模型参数“原点”上是“相切”的，即：

$$\left.\frac{\partial\mathcal{F}}{\partial\Lambda}\right|_{\Lambda=\Lambda^{(0)}} = \left.\frac{\partial\mathcal{S}}{\partial\Lambda}\right|_{\Lambda=\Lambda^{(0)}} \tag{3-41}$$

这样的辅助函数有时也被称为“弱意义上的”辅助函数(Weak-Sense Auxiliary Function)[94]。需要特别指出的是，在这种类型的函数优化条件下，对辅助函数 $\mathcal{S}$ 的优化其实并不能确保达到对目标函数 $\mathcal{F}$ 的优化。这类辅助函数仅仅能够确保的是：若对辅助函数的优化在某次迭代后没有改变模型参数，即意味着辅助函数与目标函数在模型“原点”的导数均为 0，也即我们已将目标函数优化至某局部最优点。

但是即便使用式(3-40)中的辅助函数，我们仍然无法像EM算法中的那样进行优化。这是因为两个 $\gamma$ 项之差累积求和的结果有可能得到一个负的系数，这就使得 $\mathcal{S}(\Lambda,\Lambda^{(0)})$ 函数中并非所有求和项都是凹(Concave)函数，而这对于一个最大化问题来说是没有有穷解的。因此，还必须引入一个平滑项 $\mathcal{S}^{\text{sm}}(\Lambda,\Lambda^{(0)})$，使得 $\mathcal{S}(\Lambda,\Lambda^{(0)})+\mathcal{S}^{\text{sm}}(\Lambda,\Lambda^{(0)})$ 中所有项都变为凹函数[23,69,94]。实践中，常取如下的平滑项 $\mathcal{S}^{\text{sm}}$，并通过调整训练参数 $D$ 达到所有求和项均为凹函数的要求(设每个混合高斯成份概率均以参数 $\mathcal{N}(\mu_k,\sigma_k^2)$ 表示)：

$$\begin{aligned}&\mathcal{S}^{\text{sm}}(\Lambda,\Lambda^{(0)}) = \\ &\sum_k -\frac{1}{2}\left(D_k\log(2\pi\sigma_k^2)+\frac{D_k(\mu_k^{2(0)}+\sigma_k^{2(0)})-2D_k\mu_k^{(0)}\mu_k+D_k\mu_k^2}{\sigma_k^2}\right)\end{aligned} \tag{3-42}$$

不难看出，$\mathcal{S}^{\text{sm}}(\Lambda,\Lambda^{(0)})$ 满足：

$$\left.\frac{\partial\mathcal{S}^{\text{sm}}}{\partial\Lambda}\right|_{\Lambda=\Lambda^{(0)}} = 0 \tag{3-43}$$

这就使得：

$$\left.\frac{\partial \mathcal{F}}{\partial \Lambda}\right|_{\Lambda=\Lambda^{(0)}} = \left.\frac{\partial(\mathcal{S}+\mathcal{S}^{\text{sm}})}{\partial \Lambda}\right|_{\Lambda=\Lambda^{(0)}} \tag{3-44}$$

即新的辅助函数 $\mathcal{S}+\mathcal{S}^{\text{sm}}$ 仍为 $\mathcal{F}$ 在弱意义上的辅助函数。又由于 $\mathcal{S}+\mathcal{S}^{\text{sm}}$ 项项皆凹，于是就可以按照与EM算法类似的方法直接求得使辅助函数最大的模型参数值。在这里，给出均值和方差的更新公式：

$$\mu_k = \frac{\Gamma_k(O) + D_k \mu_k^{(0)}}{\gamma_k + D_k} \tag{3-45}$$

$$\sigma_k^2 = \frac{\Gamma_k(O^2) + D_k(\mu_k^{2(0)} + \sigma_k^{2(0)})}{\gamma_k + D_k} - \mu_k^2 \tag{3-46}$$

其中，

$$\begin{aligned} \gamma_k \;=\; & \frac{1}{R}\sum_{r=1}^{R} f'\left(\log\left[\frac{p_{\Lambda^{(0)}}(O_r \mid W_r)\cdot p(W_r)}{\sum_{W'\in\mathcal{M}_r} p_{\Lambda^{(0)}}(O_r \mid W')\cdot p(W')}\right]\right) \\ & \cdot \sum_{t=1}^{T_r}[\gamma_{tr}^{(0)}(k;W_r) - \gamma_{tr}^{(0)}(k)] \end{aligned} \tag{3-47}$$

$$\begin{aligned} \Gamma_k(O) \;=\; & \frac{1}{R}\sum_{r=1}^{R} f'\left(\log\left[\frac{p_{\Lambda^{(0)}}(O_r \mid W_r)\cdot p(W_r)}{\sum_{W'\in\mathcal{M}_r} p_{\Lambda^{(0)}}(O_r \mid W')\cdot p(W')}\right]\right) \\ & \cdot \sum_{t=1}^{T_r}[\gamma_{tr}^{(0)}(k;W_r) - \gamma_{tr}^{(0)}(k)]\cdot O_{rt} \end{aligned} \tag{3-48}$$

$$\begin{aligned} \Gamma_k(O^2) \;=\; & \frac{1}{R}\sum_{r=1}^{R} f'\left(\log\left[\frac{p_{\Lambda^{(0)}}(O_r \mid W_r)\cdot p(W_r)}{\sum_{W'\in\mathcal{M}_r} p_{\Lambda^{(0)}}(O_r \mid W')\cdot p(W')}\right]\right) \\ & \cdot \sum_{t=1}^{T_r}[\gamma_{tr}^{(0)}(k;W_r) - \gamma_{tr}^{(0)}(k)]\cdot O_{rt}^2 \end{aligned} \tag{3-49}$$

而至于训练参数 $D$ 的选取，目前仅有一些经验性的方法。相关的讨论及对比可见于文献[23,69,94,105,109]。

## 3.9　区分性训练的其它问题

与基于MLE的模型训练不同，声学模型区分性训练在训练资源消耗上要大很多，这主要表现在对运算资源、存储资源，以及高效算法的迫切需求上。首先，区分性训练需要给出竞争模型空间的信息，这就需要非常高效的搜索算法对训练集中的所有语料进行处理。从CT、FT准则中搜索最具竞争的模型序列，到早

期MCE准则中加入语言模型并搜索最具竞争的n-best序列，一直到目前主流的基于词图的大词汇量解码，区分性训练对竞争空间精度及广度的需求可以说越来越大。但即便是在目前最先进的搜索算法及运算能力下，这样的解码消耗对一个典型的大词汇量连续语音识别任务仍然是惊人的。其次，对于搜索得出的竞争空间，我们还需要有合适的方法加以存储。虽然目前的词图已经可以简约的表征很大的竞争空间，但面对主流规模的训练集语料，其存储消耗仍然极为可观。最后，在训练过程中，区分性训练需要消耗比MLE训练大得多的运算及 I/O 资源，虽然可以藉由某些手段提高效率，但总的资源消耗仍比MLE训练大不少。总之，有效的优化区分性训练流程，提高训练效率，仍是当前可以值得在研究与工程两方面下工夫的方向。

此外，对区分性训练中语言模型的选择甚至优化也是一个有意义的课题。在仅探讨声学模型区分性训练的情况下，目前的实验结果一般表明使用uni-gram语言模型解码出模型竞争空间，可以得到最好的声学模型训练效果。虽然对此现象我们已经有了一些针对结果的解释[23,94]，但目前从理论上还没有一个针对这一问题的令人信服的解答。

第三，在区分性训练中对各模型序列估计后验概率时，常引入声学规整因子来拉近各竞争序列之间的距离，而这实际是对各准则所真正定义的目标函数的一种平滑。实验证明，这样的声学规整对可推广的区分性训练是必要的，但也有观点认为这造成了准则间的区别实际变得非常模糊。更进一步来说，在平滑后的伪目标函数上进行的优化也许甚至根本不能带来对真正目标函数的优化。在这里，理论与实践的不一致问题仍值得寻找更好的解答。

最后，区分性训练在某些时候仍存在较大的推广性问题。某些准则，如MWE / MPE等，尚需要一些辅助性的平滑方法(如i-smoothing[22])来确保推广性能。与基于MLE的训练不同，区分性训练本身就可以说是一个对训练集所谓精细结构(Fine Structure)的学习过程。这样的方式难免会或多或少的陷入过训练的尴尬境地。如何从准则上和优化方法上提高区分性训练的推广性，使之适应测试集与训练集存在一定不匹配的情况，也仍是值得研究的课题。

## 3.10　本章小结

在本章中，我们首先从贝叶斯决策理论出发，揭示了MLE估计在实践中的一些缺陷，并引入了区分性训练的思想。接着，我们一一介绍了一些声学模型区分性训练准则，包括最常用的MMIE、MCE、MWE / MPE等。我们将各种常用的区分性训练准则融入到一个区分性训练统一准则框架中，并演示了如何通过参数的选择在同一个框架中构成不同的准则。最后，我们介绍了目前常用的、针对

区分性训练准则的模型参数优化算法，并在本章末尾结合实践中遇到的问题探讨了区分性训练在当前的一些研究方向。

# 第 4 章 MWCE准则及其在连续语音识别中的应用

## 4.1 引言

近年来，得益于声学模型区分性训练方面的研究进展，区分性训练这一手段已经得到了相当多的重视，逐渐成为声学模型训练的标准方法。区分性训练从十几年前只能应用在连续数字串等小任务上的状况，逐渐发展为今天能够成功的被应用在大词汇量连续语音识别任务上，主要跟其自身训练方法的突破有关。区分性训练之所以能取得这样大的进展，并展示出其巨大的、仍有待发掘的实用潜力，主要应当归结于以下几个方面的技术进展：1、全新的区分性训练准则的提出；2、用以简洁高效的表达竞争空间的词图的使用；3、模型参数更新算法方面的进展。

在本章中，我们主要探讨上述三方面中第一个方面的内容，即提出一种新的区分性训练准则。目前已有的区分性训练准则有很多，但主流的训练准则一般包括传统的最大互信息量估计准则(MMIE)、最小分类错误准则(MCE)，以及近年来提出的最小词 / 音素错误准则(MWE / MPE)等。而其中MWE / MPE准则已经在很多不同的任务上较一致的显示出其超过其他准则的优异性能[94]。

不难看出，MWE / MPE准则实际是从与MMIE准则相似的背景下推导得出的。它们之所以能够得到超过MMIE准则的性能，主要原因应该归结于其利用了次句级(Sub-String Level)的相关信息。所谓次句级信息，主要是指句子级(String Level)以下，包括词(word)、音节(syllable)、音素(phone)等级别的更细致的相关信息。MWE和MPE准则通过利用这些细致信息来最大化一个期望的词或音素正确率。因此，这两种准则一般被视为更加接近语音识别的性能评估准则，如词错误率(Word Error Rate, WER)，因而也顺理成章的能够取得更好的性能。

与MMIE准则相同，传统的用于语音识别声学模型HMM的MCE准则也是定义在句子一级的[78]。句子级的MCE准则通过最小化一个平滑后的句子错误度量，来间接的最小化我们的最终目标，即词错误率WER。但必须注意到，最小化句子错误与最小化词错误之间存在难以忽视的不匹配现象。这两个指标虽然呈正相关，但决不可以被看作是等价的。例如，可以考察如下两种情况：1、某语音识别器每句话必然错且仅错一个词，则整体上句子错误率为 100% ，但词错误率却很低；2、另一语音识别器在每两句话中必有一句完全正确，而另一句却总存在大量(替换、插入及删除)错误，则整体上句子错误率可以有 50% ，而词错误率则可以非常高。从理论上讲， 0% 的句子错误率固然意味着 0% 的词错误率，但

考虑到在现实的大词汇量连续语音识别系统中，我们通常只能对一个有着相当高句子级错误率的基线系统进行优化。因此，句子级与词级错误度量之间的鸿沟是无法简单忽略的。

正是考虑到上述事实，目前国际上已有一些沿着最小分类错误MCE方向的尝试，来发展出一些新的、更能直接体现词一级错误率的区分性训练准则。例如，一般化MCE损失函数(General MCE Loss Function)[110]、基于标注的音素级MCE(Label-Based Phoneme-Level MCE)[77]，以及音素区分性MCE(Phone-Discriminating MCE)[84]等。如果对这些准则进行深入分析便不难看出，上述准则的提出主要还是来源于直观，并缺乏有关准则与真正词一级错误关系的理论推导。更进一步的是，相对于传统的句子级MCE准则，上述文献中所报告的、使用这些新准则后的识别性能提升是不明显的。因此，提出一种基于MCE理论的、与词错误率有更直接关联并能取得较好识别性能的新准则，就显得别具意义。

在本章中，我们便尝试给出这样一个全新的区分性训练准则，即最小词分类错误准则(Minimum Word Classification Error, MWCE)[111]。沿着句子级MCE准则的设计流程，我们力图通过寻找合适的区分函数(Discriminant Function)、误分类度量(Misclassification Measure)，以及损失代价函数(Loss Function)，使得我们能够更为直接的估计出训练语料中词错误的度量。通过最小化这一度量，我们就可以更为明确的最小化词一级的错误。相比句子级MCE准则，新提出的词级MWCE准则能够更直接的解决我们的终极目标，即最小化词错误率WER。因此，可以预见MWCE准则将会带来比传统MCE准则更大的性能提升。

对比之前所提到的那些词一级MCE方法[77,84,110]，在本章中，我们不仅会给出MWCE准则的直观解释，更致力于给出其理论意义上的推导。我们将展示MWCE准则与真正词错误之间的联系，并证明在某种理想条件下，MWCE准则将会变为对训练集中所有被错误识别的词数的估计。当实践中上述理想条件不能完全满足时，MWCE准则仍可被看作是对错误识别词数的平滑近似，从而有效的构成一个可以被优化的区分性训练准则。

我们还成功的将提出的MWCE准则嵌入区分性训练统一准则框架[23,24,81,101]中。由于所有区分性训练准则在这一框架中共享了大部分的实现细节，因此区分性训练统一框架能够公正客观的对不同的准则进行性能对比。我们不仅将MWCE准则与句子级的MMIE、MCE准则进行对比，还将它与词及音素级的MWE、MPE准则相比较。在英文的TIMIT及WSJ0两个识别任务上的实验结果显示，MWCE准则不仅超过了它得以推导的基础准则，即句子级MCE，还超过了其他三种发端于MMIE的准则。

本章的后续部分组织如下：(4.2)节首先简要的回顾传统的句子级MCE准则及其在区分性训练统一准则框架下的实现；然后，(4.3)节介绍如何将传

统MCE准则细化到词一级，从而推导出MWCE准则；接着，在(4.4)节，我们将给出MWCE准则在TIMIT和WSJ0两个数据库上的实验及结果分析；最后，(4.5)节为本章小结。

## 4.2 MCE准则及其与句子分类错误的关系

为了更好的展示传统的MCE准则与MWCE准则的区别，我们首先简要的回顾第 3 章中已经介绍过的句子级MCE准则。

在传统的句子级MCE准则中，对每一句训练语料 $r$，我们通常使用(3-23)和(3-26)式定义如下的误分类度量：

$$\begin{aligned} d_r &= -g_{W_r}(O_r) + g_{\mathcal{M}_r^{\mathrm{MCE}}}(O_r) \\ &= -\log p_\Lambda(O_r \mid W_r)p(W_r) + \log\left\{\frac{1}{|\mathcal{M}_r^{\mathrm{MCE}}|}\sum_{W'\in\mathcal{M}_r^{\mathrm{MCE}}} p_\Lambda^\eta(O_r \mid W')p^\eta(W')\right\}^{1/\eta} \end{aligned} \tag{4-1}$$

而为了近似句子级分类错误，还需要将 $d_r$ 嵌入sigmoid损失代价函数中，即：

$$\mathcal{L}(d_r) = \frac{1}{1+e^{-2\gamma d_r+\xi}} \tag{4-2}$$

不难看出，式(4-2)中的损失代价函数随误分类度量 $d_r$ 的变化会呈现出如下的趋势：在句子完全识别正确的时候，通常有 $d_r < 0$ 或 $d_r \ll 0$，因此，$\mathcal{L}(d_r)$ 也将趋近于 0；而在其他情况下，通常有 $d_r > 0$ 或 $d_r \gg 0$，在这个时候，$\mathcal{L}(d_r)$ 则会趋近于 1。特别是在 $\eta \to \infty$ 且 $\gamma \to \infty$ 的极端情况下，损失代价函数还将直接变为表征句子级分类错误的指示函数。

通过这样的设计，传统MCE准则可以被看作是对训练集上句子级分类错误的一个估计和近似。正是因为如此，在整个训练集上最小化 $\sum_r \mathcal{L}(d_r)$ 就能够最小化句子级分类错误的经验代价，从而也通过这种手段间接的最小化词错误率。

## 4.3 最小词分类错误MWCE准则

虽然通过上面的步骤，传统的句子级MCE准则成功的将解码过程中基于句子的动态规划过程嵌入到训练中来，但这一方法的缺点却可以说是显而易见的。那就是句子级的优化准则(string-level MCE)与词级的评估准则(WER)之间存在相当程度上的不匹配。最小化句子级错误的确可以带来词错误的相应变小，但这样的优化过程显然并不直接。如果能够在此基础上发展出更加关注词级错误的区分性训练准则，则应该能够极大的增强我们优化最终识别目标的效率。

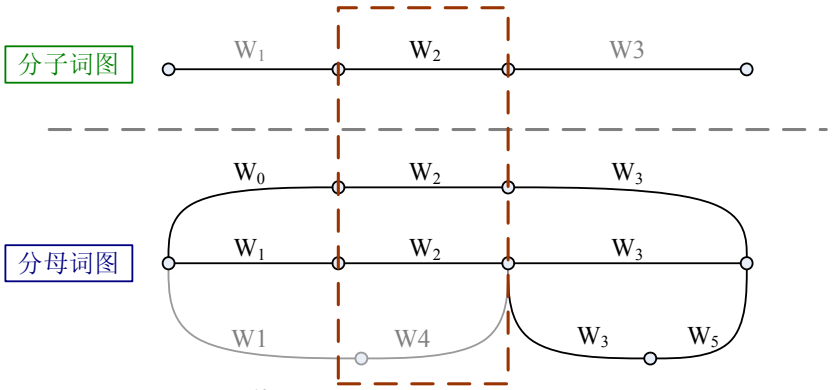


图 4.1　“正确句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 的示意图。包括图中标为深黑色的三条句子，即句子 $W_0W_2W_3$、$W_1W_2W_3$，以及 $W_1W_2W_3W_5$。

在接下来的几节中，我们将会提出词一级的最小词分类错误MWCE准则。我们尝试选择合适的区分函数、误分类度量，以及损失代价函数，使得我们能够更为直接的估计出训练集上的词级分类错误，并对其进行最小化。同时，我们还将从直观解释和理论推导两方面给出MWCE准则与真正词一级错误的关系。

### 4.3.1　MWCE损失代价函数

假设对于第 $r$ 句训练语料的正确参考文本是由 $N_r$ 个词组成的，即 $W_r = \{w_r^1, w_r^2, \ldots, w_r^{N_r}\}$。对每个参照词 $w_r^n$ 来说，我们可以定义一个“正确句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 和一个“错误句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{J}}$，并使之满足：

$$\begin{aligned} &\forall W \in \mathcal{M}_{w_r^n}^{\mathcal{K}}, \exists w \in W, w \equiv w_r^n; \\ &\forall W' \in \mathcal{M}_{w_r^n}^{\mathcal{J}}, \forall w' \in W', w' \neq w_r^n \end{aligned} \tag{4-3}$$

在(4-3)式中，$w \equiv w_r^n$ 表示我们限定词 $w$ 必须与参考词 $w_r^n$ 有着同样的标签和同样的起始时间。因此，从物理意义上讲，“正确句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 将会包括全空间内、在一特定时间段内穿过某一“匹配词” $w$ 的所有句子。相应的，“错误句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ 则包括了全空间内、在指定时间段内不含有任何针对 $w_r^n$ 的“匹配词”的所有句子。这两个句子级和的关系可以如图(4.1)及(4.2)中的例子所示。

显然的，我们有 $\mathcal{M}_{w_r^n}^{\mathcal{K}} \cap \mathcal{M}_{w_r^n}^{\mathcal{J}} = \emptyset$，以及 $\mathcal{M}_{w_r^n}^{\mathcal{K}} \cup \mathcal{M}_{w_r^n}^{\mathcal{J}} = \mathcal{M}$。不难看出，(4-3)式中“匹配词”的定义非常严格，即我们规定了词的标签与时间分割都必须完全匹配。在大词汇量连续语音识别实践中，这样的约束可以适当放松，例如在标签匹配的前提下允许在时间段上存在一定量的差异。

在定义了上述两个句子集合之后，对他们的区分函数可以分别写为：

$$g_{\mathcal{K}}(\Lambda) = \log\left[\frac{1}{|\mathcal{M}_{w_r^n}^{\mathcal{K}}|}\sum_{W \in \mathcal{M}_{w_r^n}^{\mathcal{K}}} p_\Lambda^\eta(O_r \mid W) \cdot p^\eta(W)\right]^{1/\eta} \tag{4-4}$$

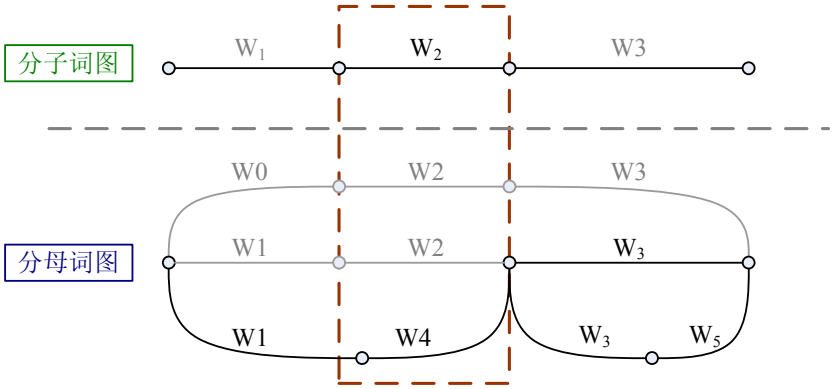


图 4.2　“错误句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ 的示意图。包括图中标为深黑色的两条句子，即句子 $W_1W_4W_3$ 以及 $W_1W_4W_3W_5$。

以及：

$$g_{\mathcal{J}}(\Lambda) = \log\left[\frac{1}{|\mathcal{M}_{w_r^n}^{\mathcal{J}}|}\sum_{W'\in\mathcal{M}_{w_r^n}^{\mathcal{J}}} p_{\Lambda}^{\eta}(O_r \mid W')\cdot p^{\eta}(W')\right]^{1/\eta} \tag{4-5}$$

那么，仿照传统MCE准则，我们设计了如下针对某一参考词 $w_r^n$ 的误分类度量，即：

$$d_{w_r^n} = -g_{\mathcal{K}}(\Lambda) + g_{\mathcal{J}}(\Lambda). \tag{4-6}$$

并相应的把它嵌入sigmoid函数，从而形成如下的损失代价函数：

$$\mathcal{L}(d_{w_r^n}) = \frac{1}{1+e^{-2\gamma(d_{w_r^n}+\xi)}}, \tag{4-7}$$

### 4.3.2　MWCE损失代价函数与词级错误间的关系

首先，让我们试着从直观层面解释式(4-7)中的MWCE损失代价函数与词错误率之间的关系。事实上，这个函数的选取是基于解码时动态规划算法的特性而来的：如果参考词 $w_r^n$ 能够被正确识别，最优的动态规划路径就必须在正确的时间段内穿过这个词；相反的，如果最优的动态规划路径没有能够在这一时间段内包含词 $w_r^n$，我们则可以断言一个词识别错误发生了。相较于传统的句子级MCE，词级的MWCE准则关注且仅关注整个动态规划路径的一段局部。这也就是说，只要某动态规划路径的局部包含了我们所关注的参考词 $w_r^n$，这一路径就会被放进“正确句子集合”。而在包含 $w_r^n$ 的特定局部以外，任何可能的词序列都是允许的。相应的，其他所有未找到“匹配词”的动态规划路径，则会被放入“错误句子集合”中去。

接着，我们尝试从理论角度对MWCE损失代价函数进行分析。仿照传统的句子级MCE准则，我们同样考虑 $\eta\to\infty$ 且 $\gamma\to\infty$ 的极端情况，式(4-6)中的误分

类度量将会转变为：

$$d_{w_r^n} = -\log \max_{W \in \mathcal{M}_{w_r^n}^{\mathcal{K}}} p_\Lambda(O_r \mid W) \cdot p(W) + \log \max_{W' \in \mathcal{M}_{w_r^n}^{\mathcal{J}}} p_\Lambda(O_r \mid W') \cdot p(W') \tag{4-8}$$

而式(4-7)中的损失代价函数则会变为如下的阶梯函数：

$$\mathcal{L}(d_{w_r^n}) = \begin{cases} 0 & \text{如果 } d_{w_r^n} < 0 \\ 1 & \text{如果 } d_{w_r^n} > 0 \end{cases} \tag{4-9}$$

我们知道，语音识别中的动态规划算法会自动的选择有最优得分的句子路径作为识别器的输出。因此，最优路径 $W^*$ 在给定某一参考词 $w_r^n$ 的情况下将会属于且仅属于 $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 和 $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ 中的一个。因此，根据 $d_{w_r^n}$ 的符号，词级别的MWCE损失代价函数可以按如下两种情况讨论：

情况一：$d_{w_r^n} < 0 \Rightarrow \mathcal{L} = 0$

在这种情况下，“正确句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 中的最优路径有着比“错误句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ 中最优路径更高的得分。因此，必有 $W^* \in \mathcal{M}_{w_r^n}^{\mathcal{K}}$。参照式(4-3)中我们对 $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 的选取，$W^*$ 中必然含有正确的参考词 $w_r^n$。在这一情况下，$w_r^n$ 将会被正确的识别出，因此相应的MWCE损失代价函数值也为 0。

情况二：$d_{w_r^n} > 0 \Rightarrow \mathcal{L} = 1$

在这种情况下，“正确句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ 中的最优路径得分低于“错误句子集合” $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ 中最优路径的得分。因此，有 $W^* \in \mathcal{M}_{w_r^n}^{\mathcal{J}}$。与情况一相反，在此情况下 $W^*$ 中必不含有正确的参考词 $w_r^n$。在这一情况下，$w_r^n$ 将不会被正确的识别出，因此相应的MWCE损失代价函数值变为 1。

通过上述两种情况的讨论，不难发现我们实质上是尝试设计MWCE损失代价函数，来对训练集中词错误的数目进行估计和近似。比起传统句子级MCE损失函数，针对这个词级别的损失度量来对模型参数进行优化，能够更符合大词汇量连续语音识别的最终目标，即降低词错误率。还需要指出的是，上述讨论仅仅只是基于 $\eta$ 和 $\gamma$ 趋于无穷的极端情况。在实践中，这两个参数通常还是被设置成有限的、甚至较小的值(无论对传统句子级MCE，还是对其他准则如MMIE、MWE、MPE)。这样做可以考虑进更多的竞争因素，从而被普遍看作能够增强训练后模型的推广性能。

### 4.3.3 区分性训练统一准则框架下的MWCE准则

为了将MWCE准则与其他区分性训练准则进行客观的比较，我们还需要将

其嵌入区分性训练统一准则框架中，而这一过程实质上相当直接。如果我们令 $\alpha=\eta$、$\gamma=\eta\cdot\rho$，并选择合适的 $\xi$ 来约掉式(4-4)和(4-5)中关于句子集合中句子总数的量，就可以通过求和训练集中每个句子平滑后的词错误估计，来得到如下的统一形式准则：

$$\mathcal{F}_{\text{MWCE}}=\sum_{r=1}^{R}\sum_{n=1}^{N_r} f\left(\log\frac{\sum_{W\in\mathcal{M}_{w_r^n}^{\kappa}} p_{\Lambda}^{\alpha}(O_r\mid W)\cdot p^{\alpha}(W)}{\sum_{W'\in\mathcal{M}_{w_r^n}^{\mathcal{J}}} p_{\Lambda}^{\alpha}(O_r\mid W')\cdot p^{\alpha}(W')}\right) \tag{4-10}$$

其中，平滑函数 $f(z)$ 与句子级MCE准则相同，仍对损失代价函数取负号(对统一准则的优化约定为最大化)，即：

$$f(z)=-\frac{1}{1+e^{2\rho z}} \tag{4-11}$$

观察对比(3-27)式与(4-10)式，我们不难发现如下的区别：句子级MCE准则对每句训练语料只需要累积一次统计量(从句首到句末)。而词级的MWCE准则则需要对每句训练语料中的每个参考词都进行一次从句首到句末的统计量累积。这样的方式就带来了运算效率方面的突出问题，并使得MWCE准则比MCE耗时得多。例如，若每个句子包含10-15词，则MWCE准则在训练中所花的时间将会比MCE多一个数量级，在实践中这是难以接受的。

因此，在实现中我们对上述准则进行了一定简化，即对某一参考词 $w_r^n$ 来说，只计算和累积其所覆盖时间范围内的统计量，而抛弃其前后其他时间段内的统计量。这样，对整个句子来说，每帧实际只进行了一次统计量的计算和累积，从而使得简化后MWCE准则的统计量既能够反映针对参考词的正确与错误两方面的竞争情况，又能够达到与句子级MCE准则相当的训练效率。

## 4.4 实验及结果

### 4.4.1 实验细节及参数配置

HTK工具包在最新的3.4版本中提供了一套区分性训练统一准则框架的实现代码，并包括进了传统的MMIE准则，以及新近提出的MWE及MPE准则[21]。从流程上讲，该工具使用一个通过最大似然估计MLE得到的声学模型作为种子，为区分性训练来生成所谓“分子词图”和“分母词图”。其中，“分子词图”即是通常所指的代表参考模型空间的词图，而“分母词图”则是代表整个解码过程中竞争模型空间的词图。HTK工具在统计量计算完成后，采用EB算法来进行声学模型参数更新，因此，该工具能够在多CPU组成的计算集群上进行并行训练。

我们在这一工具的基础上扩展其功能，使之能够支持传统的句子级MCE准则，以及我们新提出的词级MWCE准则。由于各准则在基本算法、参数更新等多

方面都共用相同的代码实现，仅仅只是根据准则在特定的部分有一些差异，所以我们认为这样的方式能够最大限度的保证对比实验的客观和公平。

为了使经过区分性训练的声学模型有更好的推广性能，我们在实验中使用了“i-smoothing”。对于MMIE、MWE和MPE准则，“i-smoothing”参数 $\tau$ 被设置为文档中推荐的值[21]；对于MCE准则，按惯例 $\tau = 0$；而对于MWCE准则，我们按文献[91]中的建议，通过计算取 $\tau = 100$。最后，声学规整因子 $\alpha$ 在后续所有实验中都被设置为 1/15，而 $\rho$ 则被设为 0.04[24]。

### 4.4.2 实验结果

#### 4.4.2.1 TIMIT连续音素识别数据库上的实验

虽然TIMIT连续音素识别任务并非真正的大词汇量连续语音识别任务，但它提供了对纯声学建模的有效评估，且体量较小，因此仍不失为一个很好的用以测试新区分性训练准则的数据库。因此在一开始，我们选择这一数据库进行MWCE准则的代码编写、测试及基本实验。我们首先参照文献[61,112]建立MLE基线系统，使用标准的 3696 句的训练语料和 192 句的核心测试集(Core-Test)进行实验。采用的声学特征为 12 维MFCC+能量，以及它们各自的一阶及二阶差分。我们选择了 48 个音素进行建模，为它们训练tri-phone模型，在计算识别性能时，则又将其归并为 39 个音素[61]。在整个HMM系统中，我们最终共有 990 个绑定状态，每个状态则使用一个含 8 个混合高斯成分的GMM模型表示。由于是单纯的声学层面的识别任务，我们在解码时没有使用任何的语言模型，而是以一个音素循环网络取而代之。对于该MLE基线系统，在测试集上的音素正确率为 62.76%，这与文献[112]中报告的结果亦是可比的。

为区分性训练所生成的“分子词图”是根据专家标注产生的，而“分母词图”则使用了与解码时同样的音素循环网络。值得一提的是，由于TIMIT音素识别任务本身没有“词”的概念，因此MWE准则与MPE准则在此任务上是相同的，而词级的MWCE准则则同样实际完成在音素一级。

通过实验，5 种准则每步迭代的音素识别错误率如图(4.3)所示。除此之外，我们还对比了每种准则能够得到的最低识别错误率，见表(4.1)。从实验结果可以看到，MCE和MWCE两种准则可以在这个数据库上取得较大的性能提升。从此前关于句子级MCE准则的相关文献中，我们也能够看到MCE在这类小任务上往往可以取得很好的效果。相比句子级MCE来说，词(此处实际是音素)级的MWCE准则性能则更胜一筹。相对于MLE基线系统，MWCE可以带来 22.0% 的相对性能提升，是所有准则中性能最优的。

至此，我们完成了在较小任务上实验，基本验证了MWCE准则的可行性，并取得了较好的性能。接下来，将开始真正在大词汇量连续语音识别中的各种实

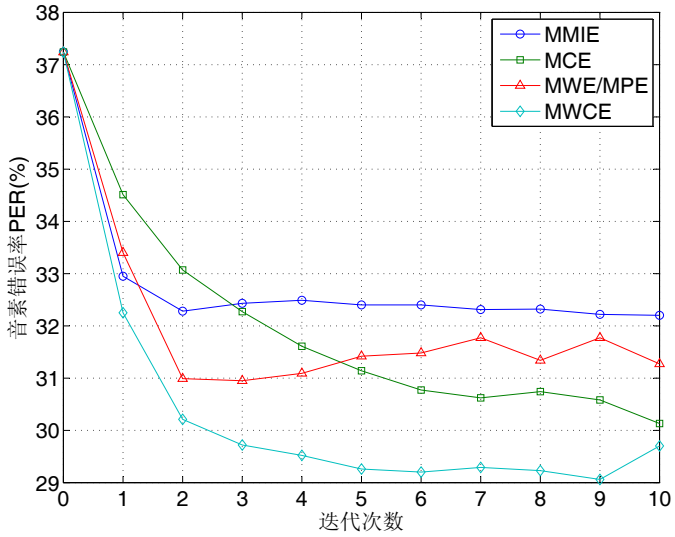


图 4.3　TIMIT数据库上各种准则的音素错误率(Phone Error Rate, PER)

| 准则 | MLE | MMIE | MCE | MW(P)E | MWCE |
|---|---|---|---|---|---|
| PER(%) | 37.24 | 32.20 | 30.13 | 30.95 | 29.06 |
| 相对提升(%) | - | 13.5 | 19.1 | 16.9 | 22.0 |

表 4.1　TIMIT数据库上各种准则的音素错误率(PER)以及对MLE基线系统的相对提升

验。

#### 4.4.2.2　WSJ0数据库上的实验

在WSJ0数据库上的实验采用标准的SI-84训练集，其中共包含 84 个说话人的 7133 句训练语料。测试则采用 1992 年标准 Nov'92 不带标点(non-verbalized)的 5000 词闭集(无集外词)测试集，共包含 8 个说话人的 330 句测试语料。声学特征方面所采用的是CMN(Cepstral Mean Normalization)处理后的 12 维MFCC+能量，以及它们各自的一阶及二阶差分。在MLE基线系统方面，我们基本遵循了文献[113]的训练流程(一个公开的WSJ0训练方案，训练出类似文献[114]中的声学模型系统)。通过训练，我们得到含 2774 个状态的cross-word tri-phone模型，每个状态同样用一个带 8 个高斯混合成分的GMM来表达。分别用bi-gram和tri-gram语言模型对MLE基线系统进行解码，可以分别得到 7.34% 和 4.89% 的词错误率。这个结果与文献[113]中报告的结果亦为可比。

针对区分性训练，我们使用一个弱化的uni-gram语言模型来对“分子词图”进行打分，并生成“分母词图”。使用标准tri-gram语言模型解码后，5种准则的测试集识别性能由图(4.4)和表(4.2)给出。首先，我们可以发现MWE准则在这个任务上能够取得比MPE准则更好的性能，这虽不常见(一般在大词汇量连续语音识别任务上，MPE能够取得比MWE更优的性能)，但与文献[94]中报告的结果是吻合的。再观察句子级MCE准则，它在此任务上仅能取得 8.4% 的性能提升，远低

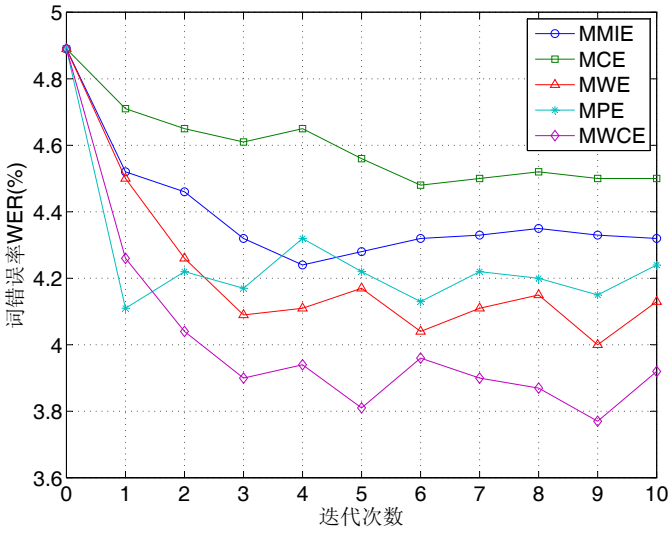


图 4.4　WSJ0 Nov'92 5k测试集上的词错误率(WER)

| 准则 | MLE | MMIE | MCE | MWE | MPE | MWCE |
|---|---|---|---|---|---|---|
| WER(%) | 4.89 | 4.24 | 4.48 | 4.00 | 4.11 | 3.77 |
| 相对提升(%) | - | 13.3 | 8.4 | 18.2 | 16.0 | 22.9 |

表 4.2　WSJ0 Nov'92 5k测试集上各种准则的词错误率(WER)，以及对MLE基线系统的相对提升

于MMIE准则的 13.3%、MWE的 18.2%，以及MPE的 16.0%。这也印证了我们此前关于MCE准则通常在大任务上性能不及其他几种准则的观察。最后，相比其他几种准则，词级的MWCE准则再次取得了最好的识别性能，其相对MLE基线系统的性能提升可以达到 22.9%。

## 4.5　本章小结

在本章中，我们提出了一种全新的区分性训练准则MWCE。通过把传统的、基于句子级的MCE损失代价函数的细化，我们得到了一个更为匹配词一级性能评估准则的优化目标。我们不仅从直观上对MWCE准则进行了解释，还从理论上分析了它与训练集上词分类错误之间的关系。我们证明在某一个特定的极端条件下，MWCE准则实际是对训练集中错误识别词数的估计。我们将MWCE准则嵌入了区分性训练统一准则框架中，并与其他主流准则进行了对比。在TIMIT和WSJ0两个数据库上的实验证明，相对于传统的句子级MCE准则来说，MWCE准则能够取得一致的识别性能提升。MWCE准则甚至在这两个数据库上超过了MMIE，以及其他的次句级准则，如MWE、MPE等。

# 第 5 章 MMIE准则基于Trust Region的HMM模型参数优化

## 5.1 引言

相对于传统的基于梯度下降的GPD等方法而言，EB优化算法由于对步长等参数不十分敏感，又能够较有效的对辅助函数进行优化，因而得到了越来越广泛的应用。尤其是在大词汇量语音识别任务中，由于参数空间巨大，模型之间竞争的相互关系变得非常复杂，EB方法相对于GPD方法在性能和可操控性上的优势更是显露无遗。但虽然如此，我们还是应该看到，EB方法仍有其与生俱来的一些缺陷。

首先，EB优化算法需要在模型更新公式中设置训练参数 $D$ 来使其可以工作[69]。然而，$D$ 的选取存在理论与实际相违背的问题：理论上，参数 $D$ 应该设置在无穷大时，才能保证EB算法在连续概率密度HMM下的收敛性；而实际上，通常的做法却主要是根据经验人为设计一些“公式”[23,69,94]，在较小的 $D$ 下来保证优化算法的可用性和有效性。

其次，EB方法仅要求辅助函数是系数全为正的有理分式，而对辅助函数的一些性质却很少用到。我们知道，在基于HMM的区分性训练中，辅助函数可以转化为一组二次函数(Quadratic Function)形式，如果能对这些二次函数的性质加以充分利用，势必可以更好的对其进行优化。

最后，EB优化算法本质上是一个无界优化(Unbounded Optimization)方法，尚需要一些基于经验观察的方法来提高其推广性，i-smoothing就是一个最突出的例子。正因为EB方法无法控制优化过程中模型参数的变化，使得它必须借助一些辅助的平滑方法来保证优化后模型的性能。这些方法与模型参数变化之间的关系在EB方法下是不明确的，更谈不上具有明确的物理意义。因此，虽然使用这些方法能从一定程度上解决模型的推广性问题，但却很不直接、难以控制。

因此，对于区分性训练中的HMM模型参数优化问题，可以提出另一种完全不同的解决思路。我们知道，需要最大化的目标函数是高度非线性、锯齿状的非凹(Concave)函数。那么首先，我们需要找到一个合适的辅助函数来近似这个目标函数，使得辅助函数既能够反映目标函数的某些特点，还能够拥有一些目标函数所没有的、易于优化的性质。接着，我们必须找到合适的约束条件(Constraint)，使得我们选取的辅助函数能够较好的近似目标函数，即，优化该辅助函数能够有效的带来目标函数的同步优化。最后，我们还需要找到合适的优化算法，充分利用辅助函数所拥有的优良性质，达到对辅助函数本身的最优化。

基于这个思路，我们提出使用基于信任区域Trust Region的优化方法来对最大互信息量估计准则MMIE进行模型参数优化。首先，我们将MMIE目标函数(Objective Function)近似为一个具有诸多易用性质的辅助函数(Auxiliary Function)。接着，我们将会证明，该辅助函数在所谓Trust Region约束下是真正目标函数的合理近似，因此，在Trust Region内对模型参数进行更新将会是“可信”的。最后，由于辅助函数的优良性质，我们还可以证明该函数在Trust Region约束下可以通过相当简洁高效的算法求得其全局最优解。因此，我们可以通过使得辅助函数最优化来同时达到目标函数的最优化。

本章的后续部分组织如下：(5.2)节介绍如何选取合适的辅助函数来近似MMIE目标函数，并解释需要引入适当的约束以保证辅助函数对目标函数的有效近似；(5.3)节引入基于KLD的Trust Region约束，并将这一约束推导为对模型均值和方差变化的具体限制；(5.4)节介绍如何在Trust Region约束下利用数学性质高效的优化我们提出的辅助函数；而(5.5)和(5.6)节则分别是该参数优化方法的实验结果以及本章小结。

## 5.2　MMIE目标函数及基于Trust Region的辅助函数

我们重写式(3-13)中的MMIE目标函数如下：

$$\mathcal{F}_{\text{MMIE}} = \frac{1}{R}\sum_{r}\mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r) = \frac{1}{R}\sum_{r}\Big[\log p(O_r \mid \Lambda; \mathcal{M}_r^{+}) - \log p(O_r \mid \Lambda; \mathcal{M}_r^{-})\Big] \tag{5-1}$$

其中，$\mathcal{M}_r^{+} = \{W_r\}$、$\mathcal{M}_r^{-} = \mathcal{M}$ 分别是参考模型空间及竞争模型空间，且有：

$$\begin{aligned}\log p(O_r \mid \Lambda; \mathcal{M}_r^{+}) &= \log p_{\Lambda}(O_r \mid W_r)\cdot p(W_r)\\ \log p(O_r \mid \Lambda; \mathcal{M}_r^{-}) &= \log \sum_{W'\in\mathcal{M}_r^{-}} p_{\Lambda}(O_r \mid W')\cdot p(W')\end{aligned} \tag{5-2}$$

正如本章引言中所提到过的，首先我们需要寻找合适的辅助函数来近似式(5-1)中的目标函数。在EM算法中我们知道，对于不完全数据的似然度 $\log p(O_r \mid \Lambda)$，均可以拆分为完全数据下的函数 $\mathcal{Q}$ 与 $\mathcal{H}$ 之差，即：

$$\log p(O_r \mid \Lambda) = \mathcal{Q}_r(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r(\Lambda \mid \Lambda^{(0)}) \tag{5-3}$$

其中，

$$\begin{aligned}&\mathcal{Q}_r(\Lambda \mid \Lambda^{(0)}) =\\ &\mathrm{E}_{\mathbf{k}}\Big[\log p(O_r, \mathbf{k} \mid \Lambda)\Big| O_r, \Lambda^{(0)}\Big] = \sum_{\mathbf{k}}\log p(O_r, \mathbf{k} \mid \Lambda)\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)})\end{aligned} \tag{5-4}$$

$$\mathcal{H}_r(\Lambda \mid \Lambda^{(0)}) = \mathrm{E}_{\mathbf{k}}\Big[\log p(\mathbf{k} \mid O_r, \Lambda)\Big| O_r, \Lambda^{(0)}\Big] = \sum_{\mathbf{k}} \log p(\mathbf{k} \mid O_r, \Lambda) \cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}) \tag{5-5}$$

而 $\mathbf{k}$ 代表所有可能的模型高斯核转移序列，$\Lambda^{(0)}$ 与 $\Lambda$ 分别为更新前和更新后的声学模型参数。

使用上述拆分方法，我们可以将参考模型与竞争模型的输出概率表示为：

$$\begin{aligned} \log p(O_r \mid \Lambda; \mathcal{M}_r^{+}) &= \mathcal{Q}_r^{+}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{+}(\Lambda \mid \Lambda^{(0)}) \\ \log p(O_r \mid \Lambda; \mathcal{M}_r^{-}) &= \mathcal{Q}_r^{-}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{-}(\Lambda \mid \Lambda^{(0)}) \end{aligned} \tag{5-6}$$

也即可以将目标函数 $\mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r)$ 表示为：

$$\mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r) = \Big[\mathcal{Q}_r^{+}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{+}(\Lambda \mid \Lambda^{(0)})\Big] - \Big[\mathcal{Q}_r^{-}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{-}(\Lambda \mid \Lambda^{(0)})\Big] \tag{5-7}$$

由此，我们就可以定义如下的辅助函数 $\mathcal{A}$，使得：

$$\mathcal{A}_r(\Lambda; O_r, \mathcal{M}_r) = \Big[\mathcal{Q}_r^{+}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{+}(\Lambda^{(0)} \mid \Lambda^{(0)})\Big] - \Big[\mathcal{Q}_r^{-}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{-}(\Lambda^{(0)} \mid \Lambda^{(0)})\Big] \tag{5-8}$$

不难看出，目标函数 $\mathcal{F}$ 与辅助函数 $\mathcal{A}$ 的主要区别体现在对 $\mathcal{H}$ 函数的参数使用上。在目标函数中，$\mathcal{H}$ 函数的参数需要同时用到更新前后的声学模型，而在辅助函数中，$\mathcal{H}$ 函数的参数却仅使用了更新前的声学模型。虽然目标函数与辅助函数的形式有所区别，但我们仍然可以发现它们之间存在如下关系。首先：

$$\mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r)\Big|_{\Lambda=\Lambda^{(0)}} = \mathcal{A}_r(\Lambda; O_r, \mathcal{M}_r)\Big|_{\Lambda=\Lambda^{(0)}} \tag{5-9}$$

即目标函数与辅助函数在更新前的模型“原点”是重合的，这一性质几乎是显然的。其次，还有：

$$\frac{\partial \mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} = \frac{\partial \mathcal{A}_r(\Lambda; O_r, \mathcal{M}_r)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \tag{5-10}$$

即，目标函数与辅助函数的一阶倒数在模型“原点”是相等的，也就是说目标函数与辅助函数在该“原点”处是相切的。这一性质可以证明如下：

$$\frac{\partial \mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} = \frac{\partial\Big[\mathcal{Q}_r^{+}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{+}(\Lambda \mid \Lambda^{(0)})\Big] - \Big[\mathcal{Q}_r^{-}(\Lambda \mid \Lambda^{(0)}) - \mathcal{H}_r^{-}(\Lambda \mid \Lambda^{(0)})\Big]}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \tag{5-11}$$

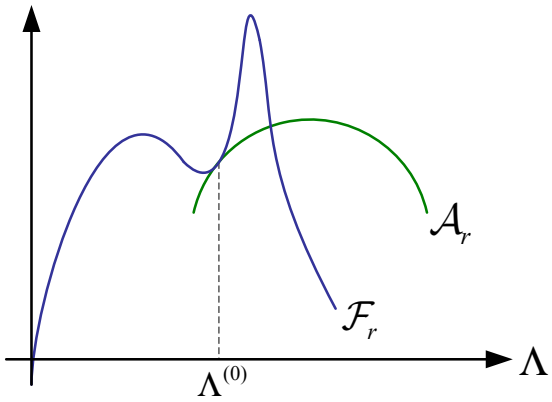


图 5.1　辅助函数 $\mathcal{A}$ 与目标函数 $\mathcal{F}$ 的关系

而：

$$\begin{aligned}
\frac{\partial \mathcal{H}_r(\Lambda \mid \Lambda^{(0)})}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} &= \frac{\partial \sum_{\mathbf{k}} \log p(\mathbf{k} \mid O_r, \Lambda) \cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)})}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \\
&= \sum_{\mathbf{k}} \frac{p(\mathbf{k} \mid O_r, \Lambda^{(0)})}{p(\mathbf{k} \mid O_r, \Lambda)}\Big|_{\Lambda=\Lambda^{(0)}} \cdot \frac{\partial p(\mathbf{k} \mid O_r, \Lambda)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \\
&= \sum_{\mathbf{k}} \frac{\partial p(\mathbf{k} \mid O_r, \Lambda)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \\
&= \frac{\partial \sum_{\mathbf{k}} p(\mathbf{k} \mid O_r, \Lambda)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \\
&= \frac{\partial\, 1}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \\
&= 0
\end{aligned} \tag{5-12}$$

所以：

$$\frac{\partial \mathcal{H}_r^{+}(\Lambda \mid \Lambda^{(0)})}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} = \frac{\partial \mathcal{H}_r^{-}(\Lambda \mid \Lambda^{(0)})}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} = 0 \tag{5-13}$$

因此，

$$\frac{\partial \mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} = \frac{\partial \Big[\mathcal{Q}_r^{+}(\Lambda \mid \Lambda^{(0)}) - \mathcal{Q}_r^{-}(\Lambda \mid \Lambda^{(0)})\Big]}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \tag{5-14}$$

又因为 $\mathcal{H}_r^{+}(\Lambda^{(0)} \mid \Lambda^{(0)})$ 和 $\mathcal{H}_r^{-}(\Lambda^{(0)} \mid \Lambda^{(0)})$ 均为常数，所以：

$$\frac{\partial \mathcal{A}_r(\Lambda; O_r, \mathcal{M}_r)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} = \frac{\partial \Big[\mathcal{Q}_r^{+}(\Lambda \mid \Lambda^{(0)}) - \mathcal{Q}_r^{-}(\Lambda \mid \Lambda^{(0)})\Big]}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(0)}} \tag{5-15}$$

因此，(5-10)式得证。

从式(5-9)、(5-10)可以得出这样结论：我们所设计的辅助函数 $\mathcal{A}$ 在模型参数“原点”附近是对目标函数 $\mathcal{F}$ 的良好近似。$\mathcal{A}$ 与 $\mathcal{F}$ 在该“原点”处不仅相接(函数值相等)而且相切(一阶倒数相等)，如图(5.1)所示。因此，在模型参数“原点”附近对 $\mathcal{A}$ 进行优化，可以间接的达到对 $\mathcal{F}$ 进行优化的目的。

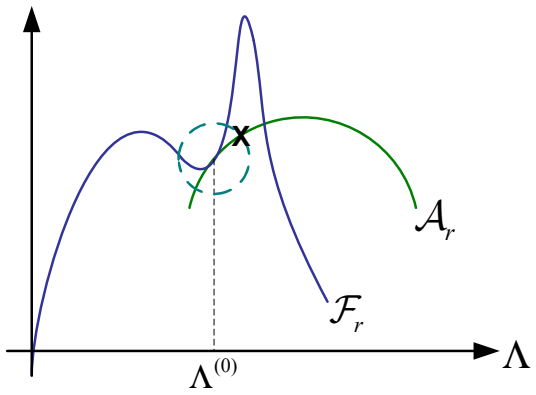


图 5.2　小范围优化时 $\mathcal{A}$ 与 $\mathcal{F}$ 的关系

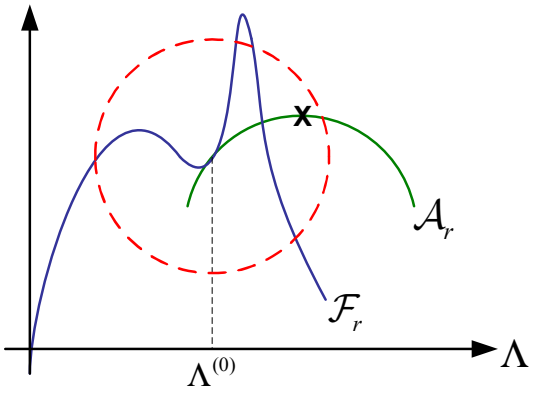


图 5.3　大范围优化时 $\mathcal{A}$ 与 $\mathcal{F}$ 的关系

## 5.3　基于Trust Region的模型参数优化约束

通过以上的步骤，我们可以把对原始目标函数的优化问题转换为对我们定义的、近似的辅助函数的优化问题。但需要注意的是，对辅助函数 $\mathcal{A}$ 的优化并不必然带来对真正目标函数 $\mathcal{F}$ 的优化。这是因为辅助函数仅仅只是对目标函数在声学模型“原点”上的近似，而决不是完全精确的拟合：式(5-9，5-10)仅在模型“原点”处成立，离开了这一“原点”，上述辅助函数的两条性质还能在多大程度上服从于目标函数是难以预见的。

图(5.2)和(5.3)很好的解释了辅助函数优化与目标函数优化之间的不一致性。如果我们从声学模型“原点”出发根据辅助函数 $\mathcal{A}$ 对模型参数进行优化，并不一定会确保目标函数 $\mathcal{F}$ 的优化。在图(5.2)中，模型参数的变动被限制在一个较小的范围内(图中虚线圆圈所示)，因此，将模型参数优化到辅助函数 $\mathcal{A}$ 的最优点(图中 X 符号所示)时，也能同时带来目标函数 $\mathcal{F}$ 的优化；相反，如图(5.3)所示，如果模型参数变动所允许的范围过大，使得辅助函数与目标函数发生了偏移，在将模型参数优化到 $\mathcal{A}$ 的最优点时，$\mathcal{F}$ 却得不到与之一致的优化。

因此，要把对目标函数 $\mathcal{F}$ 的优化转化为对辅助函数 $\mathcal{A}$ 的优化，需要确保如下两点：首先，声学模型必须以其“原点”为基础，逐步迭代优化，直至收敛；其次，每次优化需确保优化后的模型参数不至于偏离“原点”过远，从而确保 $\mathcal{A}$ 与 $\mathcal{F}$ 的一致性关系仍能得到较好保持。

### 5.3.1　基于KLD的Trust Region约束

为了解析的表达上述条件，我们使用如下的KLD距离约束来从数学上控制声学模型参数不致偏离“原点”过远，即定义优化前后模型各高斯核总的KLD距

离变化应小于预设的参数 $\rho^2/2$：

$$\sum_k \mathcal{D}(\lambda_k \parallel \lambda_k^{(0)}) \leq \frac{\rho^2}{2} \tag{5-16}$$

式中，对某一高斯核 $k$，有：

$$\mathcal{D}(\lambda_k \parallel \lambda_k^{(0)}) = \frac{1}{2}\left[(\mu_k - \mu_k^{(0)})^\top \boldsymbol{\Sigma}_k^{(0)-1}(\mu_k - \mu_k^{(0)}) + \mathrm{tr}(\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1(0)}) + \log \frac{|\boldsymbol{\Sigma}_k^{(0)}|}{|\boldsymbol{\Sigma}_k|} - D\right] \tag{5-17}$$

其中，$D$ 为HMM观测向量的维数。

式(5-16)中的KLD距离约束是一个全局的约束，由HMM模型中各个高斯核更新前后的距离累加而成。只要参数 $\rho$ 设置恰当，它就能够确保优化更新后的模型在声学空间上不偏离原模型太远。实践中，应合理设置 $\rho$ 值，使得每次迭代时即使是距离变化量最大的高斯核也仍被约束在合理的范围内。这样，就可以通过不断迭代使模型参数得到逐步优化。

应该看到，式(5-17)中对单个高斯核KLD距离的计算相对比较复杂。如果同时优化模型的均值与方差，将很难直接的应用这一约束。因此，我们需要单独考虑在分别优化模型的均值与方差时，该KLD距离约束将如何应用于具体的模型参数更新。

### 5.3.2 Trust Region约束对HMM模型均值更新的约束

当仅更新声学模型均值时，方差保持定值，即 $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma}_k^{(0)}$。因此，有：

$$\begin{aligned} \mathrm{tr}(\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1(0)}) &= D \\ \log \frac{|\boldsymbol{\Sigma}_k^{(0)}|}{|\boldsymbol{\Sigma}_k|} &= 0 \end{aligned} \tag{5-18}$$

式(5-16)中的全局KLD距离约束变为：

$$\sum_k \left[(\mu_k - \mu_k^{(0)})^\top \boldsymbol{\Sigma}_k^{(0)-1}(\mu_k - \mu_k^{(0)})\right] \leq \rho^2 \tag{5-19}$$

上式还可以进一步写为如下形式：

$$\sum_k \left[(\frac{\tilde{\mu}_k - \tilde{\mu}_k^{(0)}}{\rho})^\top (\frac{\tilde{\mu}_k - \tilde{\mu}_k^{(0)}}{\rho})\right] \leq 1 \tag{5-20}$$

其中，$\tilde{\mu}_k = \boldsymbol{\Sigma}_k^{(0)-\frac{1}{2}}\mu_k$，$\tilde{\mu}_k^{(0)} = \boldsymbol{\Sigma}_k^{(0)-\frac{1}{2}}\mu_k^{(0)}$，分别是均值对原模型标准差的归一化。

令 $\mathbf{x}_k = (\tilde{\mu}_k - \tilde{\mu}_k^{(0)})/\rho$，且定义如下的矩阵：

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_K \end{bmatrix}_{(DK\times 1)} \tag{5-21}$$

式(5-20)即可被简写为 $\mathbf{X}^{\mathrm{T}}\mathbf{X} \le 1$ (其中，$K$ 为HMM模型中总高斯核数)。这一约束条件将会被用于接下来的均值参数优化当中。

### 5.3.3 Trust Region约束对HMM模型方差更新的约束

当仅更新声学模型方差时，均值保持定值，即 $\mu_k \equiv \mu_k^{(0)}$。因此，式(5-16)中的全局KLD距离约束变为：

$$\sum_k \mathrm{tr}(\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1(0)}) - \log\left|\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1(0)}\right| - D \le \rho^2 \tag{5-22}$$

在考虑对角方差矩阵时，可以将上式按观测向量的维数展开，即：

$$\sum_k \sum_d \sigma_{kd}\sigma_{kd}^{-1(0)} - \log(\sigma_{kd}\sigma_{kd}^{-1(0)}) - 1 \le \rho^2 \tag{5-23}$$

令 $\tilde{\sigma}_{kd} = \sigma_{kd}\sigma_{kd}^{-1(0)}$，上式即可改写为：

$$\sum_k \sum_d \tilde{\sigma}_{kd}^2 - \log\tilde{\sigma}_{kd}^2 - 1 \le \rho^2 \tag{5-24}$$

再令 $v_{kd} = \log\tilde{\sigma}_{kd}^2$，式(5-24)可变形为：

$$\sum_k \sum_d \exp(v_{kd}) - v_{kd} - 1 \le \rho^2 \tag{5-25}$$

由于我们约束了每次迭代过程中模型更新不会偏离原模型太远，因此有 $\sigma_{kd} \approx \sigma_{kd}^{(0)}$，所以，$v_{kd} \approx 0$。于是，我们可以把 $\exp(v_{kd})$ 在 $v_{kd} = 0$ 附近进行泰勒展开，有：

$$\exp(v_{kd}) \approx 1 + v_{kd} + v_{kd}^2/2 \tag{5-26}$$

图(5.4)画出了按上式展开 $\exp(v_{kd})$ 的误差情况。从图中可以看到，当更新后方差为更新前方差的约 20 倍时($v_{kd} \approx 1.5$)，两函数之间的误差不超过 1；当更新后方差为更新前方差的约 1/20 时($v_{kd} \approx -1.5$)，两函数之间的误差不超过 0.5。因此，在我们的假设下，式(5-26)所做近似的误差是非常小的。

将(5-26)式代入(5-25)式，约束条件变为：

$$\sum_k \sum_d \frac{v_{kd}^2}{2} \le \rho^2 \tag{5-27}$$

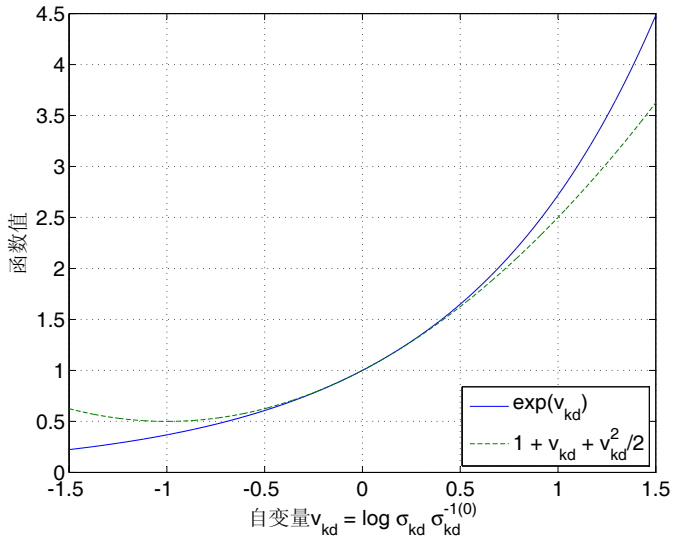


图 5.4　函数 $\exp(v_{kd})$ 在 $v_{kd}=0$ 附近进行泰勒展开的误差情况

再令：

$$\hat{\boldsymbol{\Sigma}}_k = \begin{bmatrix} \log\sigma_{k1}^2 \\ \log\sigma_{k2}^2 \\ \vdots \\ \log\sigma_{kD}^2 \end{bmatrix}_{(D\times 1)} \tag{5-28}$$

$$\hat{\boldsymbol{\Sigma}}_k^{(0)} = \begin{bmatrix} \log\sigma_{k1}^{2(0)} \\ \log\sigma_{k2}^{2(0)} \\ \vdots \\ \log\sigma_{kD}^{2(0)} \end{bmatrix}_{(D\times 1)} \tag{5-29}$$

以及 $\mathbf{y}_k = (\hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Sigma}}_k^{(0)})/\sqrt{2\rho^2}$，

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{bmatrix}_{(DK\times 1)} \tag{5-30}$$

(5-27)式可最终化为：$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} \le 1$。与仅更新均值时的情况一样，这一约束条件也将会被用于接下来的方差参数优化中去。

## 5.4　辅助函数基于Trust Region约束的优化

在上两节中，我们完成了辅助函数的选取，并确定了优化该函数时所应遵循的Trust Region约束。本节将在此基础上探讨如何在Trust Region约束条件下有效

的优化辅助函数 $\mathcal{A}$。

首先，将辅助函数 $\mathcal{A}$ 中的常数项合并为与优化无关的常数 $C$，MMIE准则可改写为：

$$\begin{aligned}
\mathcal{F}_{\text{MMIE}} &= \frac{1}{R}\sum_{r}\mathcal{F}_r(\Lambda; O_r, \mathcal{M}_r) \\
&\approx \frac{1}{R}\sum_{r}\mathcal{A}_r(\Lambda; O_r, \mathcal{M}_r) \\
&= \frac{1}{R}\sum_{r}\left[Q_r^{+}(\Lambda \mid \Lambda^{(0)}) - Q_r^{-}(\Lambda \mid \Lambda^{(0)})\right] + C \\
&= \frac{1}{R}\sum_{r}\sum_{\mathbf{k}} \log p(O_r, \mathbf{k} \mid \Lambda)\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+}) \\
&\qquad - \log p(O_r, \mathbf{k} \mid \Lambda)\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{-})
\end{aligned} \tag{5-31}$$

以式(5-31)求和项中的前半部分为例，把涉及到状态转移概率的模型参数并入常数项 $C$ (我们在这里不对状态转移概率进行优化更新)，则可进一步分解为：

$$\begin{aligned}
&\sum_{r}\sum_{\mathbf{k}} \log p(O_r, \mathbf{k} \mid \Lambda)\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+}) \\
&= \sum_{r}\sum_{\mathbf{k}}\left[\log\prod_{t} p(O_{rt}, \mathbf{k}_t \mid \Lambda)\right]\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+}) + C \\
&= \sum_{r}\sum_{\mathbf{k}}\left[\sum_{t}\log p(O_{rt}, \mathbf{k}_t \mid \Lambda)\right]\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+}) + C \\
&= \sum_{r}\sum_{\mathbf{k}}\left[\sum_{t}\sum_{k}\log p(O_{rt}, k \mid \Lambda)\cdot\delta(k, \mathbf{k}_t)\right]\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+}) + C \\
&= \sum_{r}\sum_{t}\sum_{k}\log p(O_{rt}, k \mid \Lambda)\cdot\left[\sum_{\mathbf{k}}\delta(k, \mathbf{k}_t)\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+})\right] + C
\end{aligned} \tag{5-32}$$

上式中，$\mathbf{k}$ 代表一个合法的高斯核转移序列，$k$ 代表某一特定高斯核，$\mathbf{k}_t$ 则表示 $t$ 时刻时 $\mathbf{k}$ 所处的高斯核。$\delta(k, \mathbf{k}_t)$ 是Kronecker指示函数，即：

$$\delta(k, \mathbf{k}_t) = \begin{cases} 0, & \text{如果 } k \neq \mathbf{k}_t \\ 1, & \text{如果 } k = \mathbf{k}_t \end{cases} \tag{5-33}$$

令：

$$\gamma_{krt}^{+} = \sum_{\mathbf{k}}\delta(k, \mathbf{k}_t)\cdot p(\mathbf{k} \mid O_r, \Lambda^{(0)}; \mathcal{M}_r^{+}) \tag{5-34}$$

不难看出，$\gamma_{krt}^{+}$ 是高斯核 $k$ 在 $t$ 时刻出现于训练句子 $r$ 的正例模型 $\mathcal{M}_r^{+}$ 中的

后验概率。将其代入式(5-32)，则可更简洁的表示为：

$$\begin{aligned}
&\sum_{r}\sum_{\mathbf{k}}\log p(O_r,\mathbf{k}\mid\Lambda)\cdot p(\mathbf{k}\mid O_r,\Lambda^{(0)};\mathcal{M}_r^{+})\\
&=\sum_{r}\Big[\sum_{t}\sum_{k}\log p(O_{rt},k\mid\Lambda)\Big]\cdot\gamma_{krt}^{+}+C\\
&=\sum_{k}\sum_{r}\sum_{t}\gamma_{krt}^{+}\log p(O_{rt},k\mid\Lambda)+C
\end{aligned}\tag{5-35}$$

同理，可得式(5-31)求和项中的后半部分也可写为：

$$\begin{aligned}
&\sum_{r}\sum_{\mathbf{k}}\log p(O_r,\mathbf{k}\mid\Lambda)\cdot p(\mathbf{k}\mid O_r,\Lambda^{(0)};\mathcal{M}_r^{-})\\
&=\sum_{k}\sum_{r}\sum_{t}\gamma_{krt}^{-}\log p(O_{rt},k\mid\Lambda)+C
\end{aligned}\tag{5-36}$$

因此式(5-31)可以表示为正反两个模型的后验概率相减，再乘以对应高斯核输出概率的形式，即：

$$\mathcal{F}_{\mathrm{MMIE}}\approx\frac{1}{R}\sum_{k}\sum_{r}\sum_{t}(\gamma_{krt}^{+}-\gamma_{krt}^{-})\log p(O_{rt},k\mid\Lambda)+C\tag{5-37}$$

又因为：

$$\begin{aligned}
\log p(O_{rt},k\mid\Lambda) &= \log\Big\{\tfrac{1}{\sqrt{(2\pi)^{D}|\boldsymbol{\Sigma}_k|}}\exp\big[-\tfrac{1}{2}(O_{rt}-\mu_k)^{\top}\boldsymbol{\Sigma}_k^{-1}(O_{rt}-\mu_k)\big]\Big\}\\
&= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k|-\frac{1}{2}(O_{rt}-\mu_k)^{\top}\boldsymbol{\Sigma}_k^{-1}(O_{rt}-\mu_k)+C
\end{aligned}\tag{5-38}$$

所以最大化式(5-37)就等价于最小化：

$$\mathcal{F}'=\sum_{k}\sum_{r}\sum_{t}(\gamma_{krt}^{+}-\gamma_{krt}^{-})[\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\frac{1}{2}(O_{rt}-\mu_k)^{\top}\boldsymbol{\Sigma}_k^{-1}(O_{rt}-\mu_k)]\tag{5-39}$$

在接下来的两小节中，我们将分别从均值与方差两方面详细的分析如何在Trust Region约束下优化(最小化)上式。

### 5.4.1　对均值优化问题的数学表达式

去掉与均值优化无关的参数，式(5-39)可重写为：

$$\begin{aligned}
\mathcal{F}'_{\mathrm{mean}} &= \sum_{k}\sum_{r}\sum_{t}(\gamma_{krt}^{+}-\gamma_{krt}^{-})[\frac{1}{2}(O_{rt}-\mu_k)^{\top}\boldsymbol{\Sigma}_k^{-1(0)}(O_{rt}-\mu_k)]+C\\
&= \sum_{k}\sum_{r}\sum_{t}(\gamma_{krt}^{+}-\gamma_{krt}^{-})[\frac{1}{2}\mu_k^{\top}\boldsymbol{\Sigma}_k^{-1(0)}\mu_k-O_{rt}^{\top}\boldsymbol{\Sigma}_k^{-1(0)}\mu_k]+C\\
&= \sum_{k}\sum_{r}\sum_{t}(\gamma_{krt}^{+}-\gamma_{krt}^{-})[\frac{1}{2}\tilde{\mu}_k^{\top}\tilde{\mu}_k-\tilde{O}_{krt}^{\top}\tilde{\mu}_k]+C
\end{aligned}\tag{5-40}$$

其中，$\tilde{O}_{krt} = \boldsymbol{\Sigma}_k^{-\frac{1}{2}(0)} O_{rt}$。在(5.3.2)节中，我们曾令 $\mathbf{x}_k = (\tilde{\mu}_k - \tilde{\mu}_k^{(0)})/\rho$，代入上式，即可等价于最小化：

$$\begin{aligned}\mathcal{F}'_{\text{mean}} &= \sum_k \sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-})[\frac{1}{2}(\rho\mathbf{x}_k + \tilde{\mu}_k^{(0)})^{\top}(\rho\mathbf{x}_k + \tilde{\mu}_k^{(0)}) - \tilde{O}_{krt}^{\top}(\rho\mathbf{x}_k + \tilde{\mu}_k^{(0)})] + C \\ &= \sum_k \sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-})[\frac{1}{2}(\rho^2\mathbf{x}_k^{\top}\mathbf{x}_k) + \rho(\tilde{\mu}_k^{(0)} - \tilde{O}_{krt})^{\top}\mathbf{x}_k] + C\end{aligned} \tag{5-41}$$

再令：

$$\begin{aligned}\Gamma_k^0 &= \textstyle\sum_r \sum_t \rho^2(\gamma_{krt}^{+} - \gamma_{krt}^{-}) \\ \boldsymbol{\Gamma}_k^1 &= \textstyle\sum_r \sum_t \rho(\gamma_{krt}^{+} - \gamma_{krt}^{-})(\tilde{\mu}_k^{(0)} - \tilde{O}_{krt})\end{aligned} \tag{5-42}$$

并定义如下的矩阵形式：

$$\boldsymbol{\Gamma}^0 = \begin{bmatrix} \Gamma_1^0 \cdot I_{D\times D} & & & \\ & \Gamma_2^0 \cdot I_{D\times D} & & \\ & & \ddots & \\ & & & \Gamma_K^0 \cdot I_{D\times D} \end{bmatrix}_{(DK\times DK)} \tag{5-43}$$

$$\boldsymbol{\Gamma}^1 = \begin{bmatrix} \boldsymbol{\Gamma}_1^1 \\ \boldsymbol{\Gamma}_2^1 \\ \vdots \\ \boldsymbol{\Gamma}_K^1 \end{bmatrix}_{(DK\times 1)} \tag{5-44}$$

上述最小化问题可以非常简单明了的表示为( $\mathbf{X}$ 的定义请参见5.3.2节)：

$$\mathcal{F}''_{\text{mean}} = \frac{1}{2}\mathbf{X}^{\top}\boldsymbol{\Gamma}^0\mathbf{X} + \boldsymbol{\Gamma}^{1\top}\mathbf{X} \quad \text{s.t. } \mathbf{X}^{\top}\mathbf{X} \le 1 \tag{5-45}$$

从而，我们将相对复杂的模型优化问题抽象为相对简洁的数学表达式，以利我们从现有数学优化方法中寻求解法。

### 5.4.2 对方差优化问题的数学表达式

当只优化模型方差时，去掉与方差优化无关的参数，式(5-39)又可重写为最小化：

$$\begin{aligned}\mathcal{F}'_{\text{var}} &= \sum_k \sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-})[\frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \frac{1}{2}(O_{rt} - \mu_k^{(0)})^{\top}\boldsymbol{\Sigma}_k^{-1}(O_{rt} - \mu_k^{(0)})] + C \\ &= \sum_k \sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-}) \sum_d [\frac{1}{2}\log\sigma_{kd}^2 + \frac{(O_{rtd} - \mu_{kd}^{(0)})^2}{2\sigma_{kd}^2}] + C\end{aligned} \tag{5-46}$$

同样的，由于我们在(5.3.3)节中曾定义 $\mathbf{y}_k = (\hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Sigma}}_k^{(0)})/\sqrt{2\rho^2}$，那么对于高斯核 $k$ 的第 $d$ 维来说，可知 $\sigma_{kd}^2 = \exp[\sqrt{2\rho^2}y_{kd} + \log\sigma_{kd}^{2(0)}]$。式(5-46)即可变为：

$$\begin{aligned}\mathcal{F}'_{\text{var}} &= \sum_k\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot \\ &\sum_d[\frac{\sqrt{2\rho^2}}{2}y_{kd}+\frac{(O_{rtd}-\mu_{kd}^{(0)})^2}{2\sigma_{kd}^{2(0)}}\cdot\exp(-\sqrt{2\rho^2}y_{kd})]+C\end{aligned} \tag{5-47}$$

因为Trust Region约束要求更新前后的各维方差不能变化过大，因此有 $y_{kd} = (\hat{\sigma}_{kd}^2 - \hat{\sigma}_k^{2(0)})/\sqrt{2\rho^2} \approx 0$。所以，不妨将 $\exp(-\sqrt{2\rho^2}y_{kd})$ 在 $y_{kd} = 0$ 处作泰勒展开，有：

$$\exp(-\sqrt{2\rho^2}y_{kd}) \approx 1 - \sqrt{2\rho^2}y_{kd} + \rho^2 y_{kd}^2 \tag{5-48}$$

将上式代入(5-47)，最小化问题即再变化为：

$$\begin{aligned}\mathcal{F}'_{\text{var}} &= \sum_k\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot \\ &\sum_d\left\{\frac{(O_{rtd}-\mu_{kd}^{(0)})^2\rho^2}{2\sigma_{kd}^{2(0)}}\cdot y_{kd}^2+\frac{\sqrt{2\rho^2}}{2}\cdot\left[1-\left(\frac{O_{rtd}-\mu_{kd}^{(0)}}{\sigma_{kd}^{(0)}}\right)^2\right]\cdot y_{kd}\right\}+C \\ &= \sum_k\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot \\ &\left\{\frac{1}{2}\mathbf{y}_k^{\top}\times\text{diag}\left[\rho^2(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\odot(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\right]\times\mathbf{y}_k\right. \\ &\left.+\frac{\sqrt{2\rho^2}}{2}\cdot\left[1-(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\odot(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\right]^{\top}\times\mathbf{y}_k\right\}+C\end{aligned} \tag{5-49}$$

其中，“$\odot$” 为数组乘法运算符。令：

$$\boldsymbol{\Psi}_k^0 = \sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot\text{diag}\left[\rho^2(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\odot(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\right] \tag{5-50}$$

$$\boldsymbol{\Psi}_k^1 = \frac{\sqrt{2\rho^2}}{2}\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot\left[1-(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\odot(\tilde{O}_{krt}-\tilde{\mu}_k^{(0)})\right] \tag{5-51}$$

并定义如下的矩阵形式：

$$\boldsymbol{\Psi}^0 = \begin{bmatrix}\boldsymbol{\Psi}_1^0 & & & \\ & \boldsymbol{\Psi}_2^0 & & \\ & & \ddots & \\ & & & \boldsymbol{\Psi}_K^0\end{bmatrix}_{(DK\times DK)} \tag{5-52}$$

$$\boldsymbol{\Psi}^1 = \begin{bmatrix} \boldsymbol{\Psi}_1^1 \\ \boldsymbol{\Psi}_2^1 \\ \vdots \\ \boldsymbol{\Psi}_K^1 \end{bmatrix}_{(DK\times 1)} \tag{5-53}$$

上述最小化问题即可以表示为( $\mathbf{Y}$ 的定义请参见5.3.3节)：

$$\mathcal{F}''_{\text{var}} = \frac{1}{2}\mathbf{Y}^\top\boldsymbol{\Psi}^0\mathbf{Y} + \boldsymbol{\Psi}^{1\top}\mathbf{Y} \quad \text{s.t. } \mathbf{Y}^\top\mathbf{Y} \le 1 \tag{5-54}$$

不难看出，(5-54)式与(5-45)式有着相同的形式。

### 5.4.3 二次方程基于Trust Region约束问题的数学解法

事实上，(5-54)式与(5-45)式在数学上早有较为成熟的算法进行优化。为了说明整个优化过程，我们首先给出如下的两条数学定理[115]：

**定理** 5.1： 定义如下的Trust Region问题：

$$\min_{p\in\mathrm{R}^n} m(p) = f + g^\top p + \frac{1}{2}p^\top Bp \qquad \text{s.t. } \| p \| \le \Delta \tag{5-55}$$

那么，矢量 $p^*$ 是上述问题的全局最优解，当且仅当 $p^*$ 满足约束条件，且存在标量 $\lambda \ge 0$，并满足：

$$\begin{gathered} (B+\lambda I)p = -g, \\ \lambda(\Delta - \| p \|) = 0, \\ (B+\lambda I)\text{是半正定矩阵。} \end{gathered} \tag{5-56}$$

**定理** 5.2： 若 $w$ 为一非零矢量， $\lambda_1$、 $\lambda_2$ 为实数，且 $\lambda_0 < \lambda_1 < \lambda_2$。设 $p_1$ 是 $(B+\lambda_1 I)p_1 = w$ 的解，而 $p_2$ 是 $(B+\lambda_2 I)p_2 = w$ 的解，那么有：$\| p_1 \| > \| p_2 \|$。

上述两条定理的证明请参阅相关的数学文献[115]，在这里不再赘述。通过这两条定理，我们可以按如下的情况讨论此一全局最优解 $p^*$ 的解法：

情况一： $B$ 为正定矩阵， $p = -B^{-1}g$，且 $\| p \| \le \Delta$

在这种情况下， $p$ 天然的满足Trust Region约束条件，只需取 $\lambda = 0$ 即可满足定理(5.1)所列出的所有条件。因此，全局最优解为 $p^* = -B^{-1}g$。

情况二： $B$ 为正定矩阵， $p = -B^{-1}g$，但 $\| p \| > \Delta$

在这种情况下， $p$ 不满足Trust Region约束条件。从定理(5.2)可知，需寻找某 $\lambda > \lambda_0 = 0$，计算相应的 $p = -(B+\lambda I)^{-1}g$，并令 $\lambda$ 不断增大，使得 $\| p \|$ 逐渐缩

小，以满足 $\Delta - \| p \| = 0$。此时，定理(5.1)的三个条件都得到满足，因此全局最优解为 $p^* = -(B + \lambda I)^{-1} g$。

情况三：$B$ 为非正定矩阵

在这种情况下，首先应确定某最小的 $\lambda_0 > 0$，使得 $(B + \lambda_0 I)$ 正定、$(B + \lambda_0 I)^{-1}$ 可求。此时，进一步计算 $p_0 = -(B + \lambda_0 I)^{-1} g$，若 $\| p_0 \| < \Delta$，则 $p^* = p_0$；否则，取 $\lambda > \lambda_0$，并逐渐增大 $\lambda$，使得 $\| p \|$ 逐渐缩小，以满足 $\Delta - \| p \| = 0$。此时，定理(5.1)的三个条件都得到满足，因此全局最优解为 $p^* = -(B + \lambda I)^{-1} g$。

通过对以上三种情况的分析，我们可以非常高效的计算每步迭代更新时的全局最优解 $p^*$，并根据 $p^*$ 计算出更新后的模型均值和方差。数学定理(5.1)中的矩阵 $B$ 实际对应式(5-45)中的 $\mathbf{\Gamma}^0$ 与式(5-54)中的 $\mathbf{\Psi}^0$。在使用对角方差矩阵时，它们都是对角阵。因此，$(B + \lambda_0 I)^{-1}$ 的求取过程非常迅速。同时，根据定理(5.2)，$\| p \|$ 实际为 $\lambda$ 的减函数。因此，在调整 $\lambda$ 使得 $\| p \|$ 满足 $\Delta - \| p \| = 0$ 的过程中，我们可以使用双向搜索(Binary Search)大大的提高搜索效率。在实际实验中，我们发现计算 $p^*$ 非常迅速，所花时间相比训练中累积统计量的时间来说，完全可以忽略不计。

## 5.5 实验及结果

我们选取WSJ0数据库进行MMIE准则基于Trust Region的HMM模型参数优化实验，使用的MLE基线系统及区分性训练方面的准备工作与第 4 章中的一致。我们希望对比MMIE准则在EB优化方法下及基于Trust Region约束的优化方法下的识别性能。

由于MMIE准则基于Trust Region的模型优化方法需要确定Trust Region的大小 $\rho$，我们首先希望通过一组实验摸索这一参数设置的合理范围。实验中，我们希望通过穷举一些可能的 $\rho$ 值，来观察这一参数的设置对迭代优化后模型的影响。为了排除其他干扰因素，我们在这里首先仅对模型的均值作更新。具体的实验结果如图(5.5)所示。

实验中，我们尝试了多个不同的Trust Region $\rho$ 值，在这里给出 $\rho$ 分别为 12 至 96 时识别错误率对应迭代步数的变化情况。从实验结果中可以看出，当 $\rho$ 取得比较合适时( $\rho = 12 \sim 24$ )，我们可以观察到较正常和较稳定的识别性能提升。而当 $\rho$ 较大时( $\rho = 48 \sim 96$ )，由于允许模型更新的幅度较大，会出现识别性能在迭代中波动较大，甚至是在几次迭代后出现降低的情况。实际上，我们还测试了更小 $\rho$ 值时的情况( $\rho = 1.5 \sim 6$ )，但由于在这类情况下允许模型更

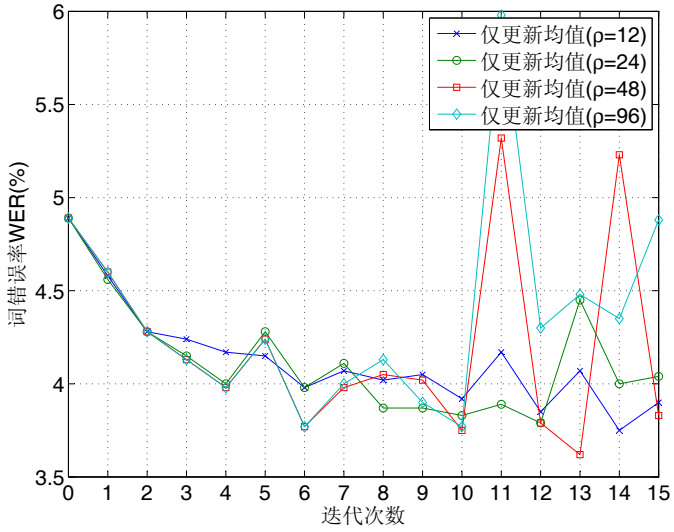


图 5.5　仅更新均值时针对不同 $\rho$ 大小的识别性能

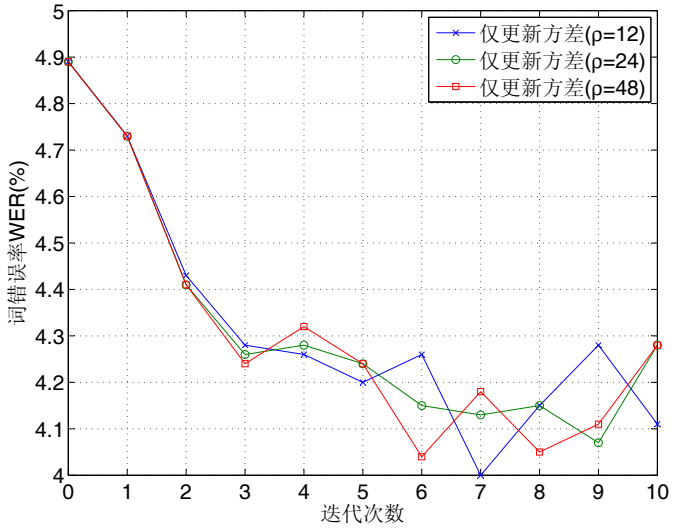


图 5.6　仅更新方差时针对不同 $\rho$ 大小的识别性能

新的幅度过小，造成收敛速度较慢，其结果在这里就不一一列出了。我们可以看到，有效的基于Trust Region的模型参数更新对 $\rho$ 的宽容度还是较大的。虽然 $\rho = 12$ 到 $\rho = 48$ 在距离约束上有高达16倍的差异(我们实际取 $\rho^2/2$ 作为全局约束，参见(5-16)式)，但实际更新后的模型在识别性能上的差异并不大。这显示了基于Trust Region的优化对参数 $\rho$ 并不十分敏感，只需将 $\rho$ 设置在一个较宽的合理范围即可。同时，我们也借由这个实验从经验上获得了 $\rho$ 的有效范围，典型的 $\rho$ 值应正比于模型中的总高斯成份数，并取 $\rho^2 \approx 0.2$ 每高斯核。

我们还实验了仅更新模型方差时的识别率，其结果如图(5.6)所示。从实验结果中可以看出，首先，正如此前的其他文献中所提到的，就更新模型的有效性程度来讲，应该是均值大于方差，再大于转移概率及混合高斯权重。我们的实验证实了这一点，即单纯更新模型方差的识别性能要差于单纯更新模型均值。其次，从识别率数据可以发现，单纯更新方差实际上也能得到相当程度的识别效果改

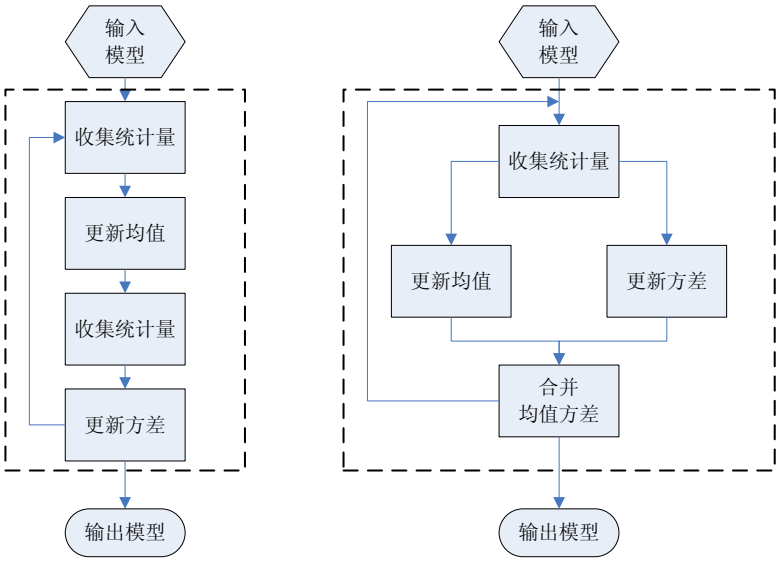


图 5.7　两种更新均值与方差的方案(左：串行更新均值和方差的标准TR方案；右：并行更新均值和方差的TR(ad-hoc)方案)

善。

归根结底，我们还是希望基于Trust Region约束的模型参数更新方法能够同时针对模型的均值与方差进行优化。为此，我们提出了两种实现方案。一种是标准的串行更新方案(以下称TR方案)，一种是取巧的并行更新方案(以下称TR(ad-hoc)方案)。TR方案的主要原则是严格遵循均值方差不能同时优化的原则，这也与我们在进行Trust Region优化公式推导时的约定一致。而TR(ad-hoc)方案则放弃了这一约束，利用同一步迭代得到的统计量分别更新均值与方差，并在最后进行合并。严格的说，TR(ad-hoc)方案无法从理论上保证优化结果的可靠性。但考虑到它实现起来更为快速，因此也有必要考察这种方案对识别性能的具体影响。两种方案的实现流程图如图(5.7)所示。

我们选取了较为折中的Trust Region大小 $\rho = 24$ 来进行实验。主要希望对比EB、TR、TR(ad-hoc)三种方法对MMIE准则优化的效率，以及其所对应的识别性能的提升。

三种方法对MMIE准则的优化效率如图(5.8)所示。从图中我们可以看到，在迭代的开始阶段，两种基于Trust Region的方法在准则优化上的结果不及EB。这是由于从本质上讲EB优化是无界的，而Trust Region约束则使得两种TR方法不会使模型变动得“太远”，因此也就在开始阶段从一定程度上限制了对准则的优化。但随着迭代的继续，由于两种TR方法能够在每一步迭代时都取得全局意义上的“最优点”，因此可以说，每步迭代都是在上一步的最好的基础上进行的。也正是因为这个原因，在两步迭代以后，两种基于Trust Region的优化方法逐渐超过了EB。到迭代结束时，Trust Region方法已远远超过了EB方法。而从两种Trust Region方法内部的对比可以看出，TR方法由于其理论上更为正确，因此在效果

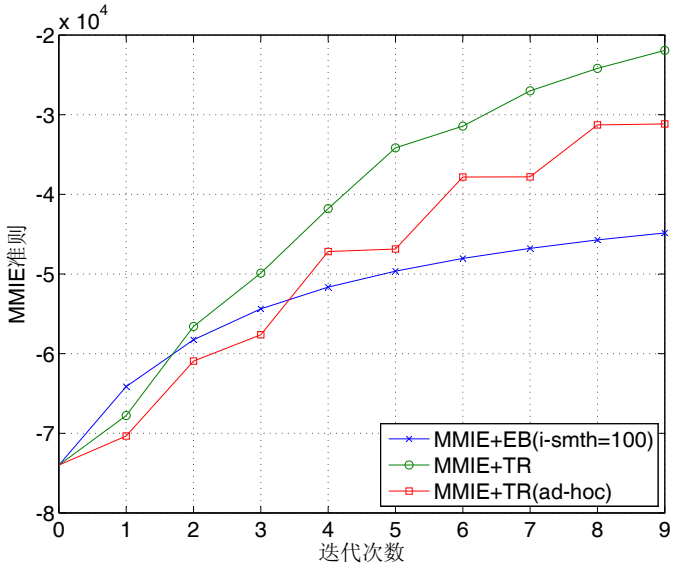


图 5.8　三种方法在 $\rho = 24$ 时对MMIE准则优化的性能对比

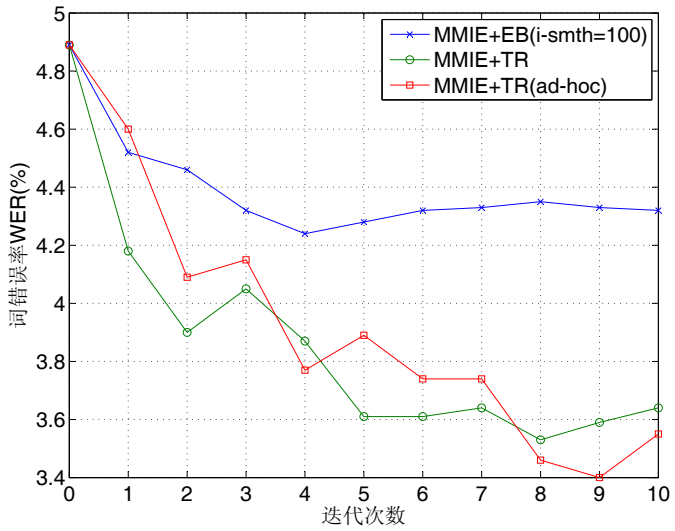


图 5.9　三种方法在 $\rho = 24$ 时的识别性能对比

上要好于TR(ad-hoc)方法。

当然，对准则的良好优化并不必然带来识别性能的提升。我们还需要对比三种方法的识别性能，如图(5.9)所示。从结果可以看出，正如理论所预料的那样，基于Trust Region的两种方法在性能上都优于传统的EB方法。本实验的具体实验数据列于表(5.1)，相对于MMIE+EB 13.3% 的性能提升，两种基于Trust Region的模型参数优化方法能够带来 27.8% 及 30.5% 的提升。这一幅度甚至超过了EB方法在此数据库上最好的MWE+EB的性能(相对提升 18.2% )。而TR与TR(ad-hoc)两种方法的内部对比差别则较不明显。在这样低的错误率下，我们发现两种方法的识别性能实际差别不大。考虑到TR(ad-hoc)方法在训练效率上要优于TR方法，因此虽然前者在理论上存在一些问题，但从实践上看仍不失为一种可以采用的方案。

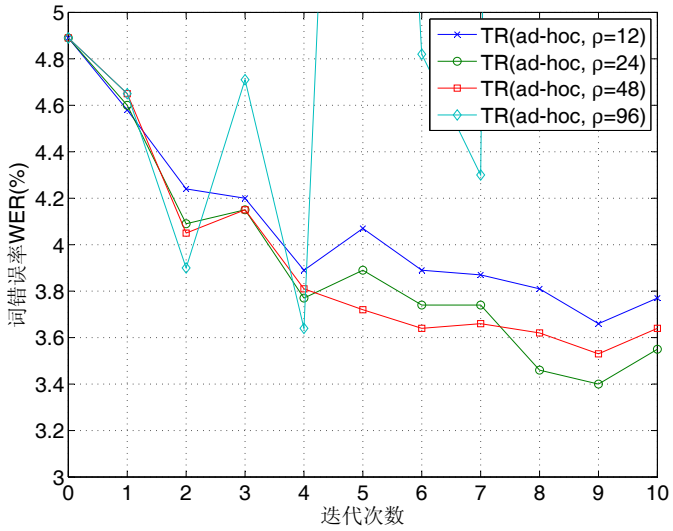


图 5.10　TR(ad-hoc)方法在不同 $\rho$ 值时的识别性能对比

| | MLE | MMIE+EB | MMIE+TR | MMIE+TR (ad-hoc) | MWE+EB |
|---|---|---|---|---|---|
| WER(%) | 4.89 | 4.24 | 3.53 | 3.40 | 4.00 |
| 相对提升(%) | - | 13.3 | 27.8 | 30.5 | 18.2 |

表 5.1　WSJ0 Nov'92 5k任务上的识别性能，以及相对MLE基线系统的相对性能提升

最后，我们选择在此数据库上性能较优的MMIE+TR(ad-hoc)方案，反过去再对控制Trust Region大小的参数 $\rho$ 进行考察，观察之前仅更新均值或方差时所得出的经验性的 $\rho$ 值在同时更新均值方差的情况下是否仍适用，实验结果如图(5.10)所示。可见，先前关于 $\rho$ 值的一些经验在这里仍然适用，只要 $\rho$ 值不被设置得过大，就能够在迭代中取得稳定可靠的性能提升。对于模型更新而言，Trust Region大小的选取主要与模型内部总的高斯成份数有关。一旦定下这个允许的变化量，更新所采取的方式与它的关系就不那么明显了。

## 5.6　本章小结

本章主要从区分性训练中模型参数优化问题的角度出发，提出了一种优于传统EB方法的优化算法，即MMIE准则基于Trust Region的模型参数优化方法。通过对目标函数设计合适的辅助函数，并引入基于KLD的Trust Region约束，我们使得EB中的无界优化问题变为了有界优化，还同时保证了辅助函数对目标函数的有效近似。由于数学文献上已经存在成熟的解决这类优化问题的解法，我们很容易将其应用到声学模型参数更新的具体问题上来。而且，还可以利用具体问题的特殊性质使得优化可以在非常小的运算消耗下完成。实验证明，基于Trust Region约束更新的声学模型不论是在优化MMIE准则，还是在识别正确率上都超过了传统的EB方法。MMIE+TR优化的识别率在WSJ0数据库上甚至还超过了传

统方法中最好的MWE+EB的性能。这表明，在声学模型区分性训练的参数优化方法中引入具有合理物理意义的约束，能够对优化后模型的区分能力起到正面的帮助。这也为我们今后在此方向上的工作提供了有益的参考。

# 第 6 章 SME估计及其帧级区分性训练

## 6.1 引言

在此前的各章中，我们介绍了多种区分性训练准则及优化方法。在训练准则方面，既包括传统的MMIE、MCE、MWE / MPE准则，又介绍了我们提出的MWCE准则；在优化方法方面，则着重介绍了经典的GPD、EB，以及我们提出的基于Trust Region的模型参数优化方法。众多文献以及我们的实验都已经证明，这些方法的使用可以毫无疑问的提高现有大词汇量连续语音识别系统的性能。特别是在测试环境与训练环境较为匹配时，采用区分性训练所得到的性能提升将会尤为明显。

但值得注意的是，测试环境与训练环境的这种“匹配”在有时，甚至是大多数时候，是难以理想达到的。在这种情况下，对声学模型的区分性训练就难以从理论上得到性能保证。这是因为，区分性训练本身是对训练集声学空间精细结构(Fine Structure)的学习。而由于训练集的规模、代表性、覆盖度等问题，会使得区分性训练只能够降低所谓“经验代价”(Empirical Risk)，而非真正的期望代价(Expected Risk)或真实代价(Actual Risk)。这就使区分性训练对训练集上经验代价的降低并不必然带来测试集上的性能提升，也就暴露出了训练环境与测试环境不匹配时的矛盾。

正是基于这样的原因，区分性训练的推广性(Generalization Ability)问题逐渐被研究者所重视，并从多方面试图加以解决。例如，在区分性训练统一准则框架中的声学规整因子的使用[23]、MCE中sigmoid函数的使用[20]、MWE / MPE中i-smoothing的引入[22]，乃至我们提出的基于Trust Region的模型参数优化方法等，都直接或间接的与提高声学模型推广性能有关。但不得不承认的是，虽然我们可以观察到上述这些方法在实践中提高模型推广性的作用，但它们却无一能从理论上给出其与区分性训练模型推广性能的具体联系。

为了解决这一矛盾，近来已有一些研究者尝试将支持向量机(Support Vector Machine, SVM)中有关分类边缘margin的相关理论应用到对声学模型的区分性训练中来。基于margin的分类器在该领域已取得了巨大的成功，其核心思想便是将推广性问题放到统计学习理论[116]中加以看待和解决。具体到语音识别声学模型领域，则主要包括大分类边缘估计(Large Margin Estimation, LME)[96,97]、区分性分类边缘(Discriminative Margin)[85]，以及软分类边缘估计(Soft Margin Estimation, SME)[98,99]等。而在其中，软分类边缘估计SME能够直接的将区分性

训练中决策反馈学习理论与SVM中软分类边缘soft margin的思想相结合，也就因此具有了同时增强模型的区分能力和推广性能的效果。

在早期的对SME方法研究中，已有实验证实它能够在小词汇量语音识别声学模型下取得很好的效果[98]。但当研究工作从小词汇量任务转移到大词汇量连续语音识别任务上之后，SME相对于传统区分性训练准则(如MCE)的性能优势就较不明显了[99]。此外，由于条件的限制，最初实验的设计并没有将区分性训练的一些最新进展引入，因此也就无法将声学模型训练的效果发挥到极致。总结起来，针对此前的SME研究，有以下几点值得加强或改进：1、MLE基础声学模型的性能。此前的研究采用不跨词三音子(within-word tri-phone)模型，而我们知道要将此基线系统性能进一步加强，广泛使用的跨词三音子(cross-word tri-phone)模型是不二的选择；2、累积统计量时对词图的使用。此前的研究从仅使用n-best列表到初步引入了词图，已有了较大进展。但如何充分并高效率的利用词图中的信息，仍有进步的空间；3、参数优化算法。此前的研究仍采用GPD算法进行参数优化，步长选取和调整在大词汇量连续语音识别任务上变得非常困难，因此有必要选择更适合的参数优化算法以提高效率。4、SME级别的细化。此前的SME方法一般定义在句子级(string-level)，在词图引入后，也可以定义在词级(word-level)。从MWE / MPE的实践我们可以看到，有效的利用更精细级别的信息显然对区分性训练的性能大有裨益，因此，仍可考虑更有效的细化SME方法到更精细的级别上。

针对上述 4 个问题，在本章中，我们提出如下的解决方案：首先，我们使用cross-word tri-phone模型建立MLE模型基线系统，将系统的性能提升到一个较高的水平，并以此作为区分性训练的基础；然后，我们采用完善的词图来表达竞争空间，并利用成熟高效的算法在词图中累积区分性训练所需要的统计量；接着，我们取代GPD方法，引入EB方法对模型参数进行优化；最后，我们考虑将问题细化到帧一级，完成SME估计下的帧级区分性训练。通过前几章的内容我们不难发现，上述问题中前3个问题的解决是比较直接的，因此，在本章中，我们着重解决最后一个问题，即分类边缘的进一步细化问题。经过所有以上这些工作，我们可以针对大词汇量连续语音识别任务建立较优的基线系统，并在此之上对比传统MCE准则及我们提出的SME方法的诸多实现方案。

本章的后续部分组织如下：(6.2)节首先介绍软分类边缘估计SME的最初定义及我们对它的发展完善；(6.3)节介绍我们对传统MCE准则及SME估计的诸多方案的对比实验及分析；最后，在(6.4)节将会给出本章小结。

## 6.2 软分类边缘估计SME

在这一节中，我们将会介绍SME估计的最初定义，及我们在此基础上对它的发展。我们将会着重介绍如何设计SME估计以适应大词汇量连续语音识别任务。在以下各节中，我们提出了针对句子级及帧级的SME估计方法，并将其与传统MCE准则在区分性训练统一准则框架下进行对比。与第 4 章一样，由于参与对比的这几种方法共享了大部分的实现细节，因此这样的对比是客观公正的。

### 6.2.1 SME估计基础理论

为了更好的介绍软分类边缘估计SME及其帧级区分性训练，首先简要的回顾SME的基本理论及其最初定义[99]。我们知道，几乎所有的区分性训练准则都可以看作对训练集中某一经过定义的经验代价的最小化。广义的来讲，这些区分性训练准则都可以被表示为最小化：

$$\ell_{\text{emp}}(\Lambda) = \frac{1}{R}\sum_{r}\ell_r(O_r,\Lambda) \tag{6-1}$$

其中，$\Lambda$ 代表声学模型参数，$R$ 为训练集总语料数，而 $\ell_r(O_r,\Lambda)$ 则表示对训练集中第 $r$ 句语料的损失代价函数。以MCE准则为例，损失代价函数定义为：

$$\ell_r(O_r,\Lambda) = \frac{1}{1+e^{-2\gamma d_r(O_r,\Lambda)+\xi}} \tag{6-2}$$

其中，$d_r(O_r,\Lambda)$ 为参考模型指数似然度与竞争模型指数似然度之差，而 $\gamma$、$\xi$ 则为sigmoid函数的参数。

需要特别指出的是，在训练集上最小化上面的经验代价并不必然能够带来测试集上的性能提升。这在统计学习理论中已得到了充分的说明，因为实际上，测试集上的损失代价函数受限于下式[116]：

$$\ell_{\text{test}}(\Lambda) \le \ell_{\text{emp}}(\Lambda) + \sqrt{\frac{1}{R}\left(VC_{\text{dim}}\left(\log\frac{2R}{VC_{\text{dim}}}+1\right)-\log\frac{\delta}{4}\right)} \tag{6-3}$$

即，测试集上的损失代价函数的上界实际可以表示为两式之和，前者是我们熟知的、训练集上的经验代价，而后者则是与所谓VC维( $VC_{\text{dim}}$ )有关的、表征模型推广能力的另一代价。各种通常的区分性训练准则固然能够有效的降低训练集上的经验代价，但它们之间实质上的区别仅在于对经验代价函数的选取所有不同。对于式(6-3)中的推广性项，通常的准则都没有加以认真的考虑。因而，如果我们可以直接的最小化式中的右边部分，也就可以通过结合传统区分性训练准则来最小化测试集上总的损失代价函数。

但是，式(6-3)中的推广性项非常难以求取，这也正是传统区分性训练准则忽略该项的直接原因。通过分析我们发现，在实际情况下( $R$ 为一合理的较大自

然数)，式(6-3)的右边部分是 $VC_{\text{dim}}$ 的单调增函数。而从文献[116]中我们又知，$VC_{\text{dim}}$ 是受限于以分类边缘margin为自变量的某单调减函数的。因此，$VC_{\text{dim}}$ 可以通过增加margin的方法间接的降低，从而带来(6-3)式右半部分推广性项的降低。

至此，我们有了两个需要同时优化的目标函数：一是传统区分性训练时所用到的训练集上的经验代价函数，二是可以带来推广性项优化的margin函数。因此，可以很自然的定义出SME估计的目标函数[98]：

$$L^{\text{SME}}(\Lambda) = \frac{\lambda}{\rho} + \ell_{\text{emp}}(\Lambda) = \frac{\lambda}{\rho} + \frac{1}{R}\sum_{r}\ell_r(O_r, \Lambda) \tag{6-4}$$

其中，$\rho$ 代表软分类边缘soft margin的“大小”，$\lambda$ 是调整前后项权重的参数(较小的 $\lambda$ 意味着经验代价项更重要，较大的 $\lambda$ 意味着意味着推广性项更重要)。从 $\rho$、$VC_{\text{dim}}$ 以及推广性项的关系我们不难看出，$\lambda/\rho$ 与模型推广性有着同向的关系。因此实质上(6-4)式可以看作是对(6-3)式上界的一个模拟。通过这样的定义，SME估计能够直接的最小化模拟的测试集损失代价函数的上界，这也正是其与传统区分性训练准则的最大区别。

### 6.2.2 句子级SME估计

通过上面的原理分析，我们只需要确定一个合适的损失代价函数 $\ell_r(O_r, \Lambda)$，即可代入(6-4)式进行优化。从本质上讲，引入分类边缘的目的在于使在训练集上学习得到的分类面既能有效的对样本进行分类(降低经验代价)，又能保持较高的分类“宽容度”(提高分类边缘margin的大小)。这样一来，在测试集与训练集仅存在较小的偏差的情况下，我们仍可以藉由分类边缘的保护而继续做出正确的分类。因此，产生损失代价的样本应该有两类：1、错误分类的样本——因为其带来较高的经验代价；2、虽然能够正确分类、但却与分类面距离太近的样本——因为其margin较小。

基于这样的原则，在最初的SME估计定义中，损失代价函数被定义在句子一级，并表示为：

$$\ell_r(O_r, \Lambda) = \begin{cases} \rho - d_r(O_r, \Lambda) & \text{如果 } \rho > d_r(O_r, \Lambda) \\ 0 & \text{其他} \end{cases} \tag{6-5}$$

其中，$d_r(O_r, \Lambda)$ 即是传统意义上的误分类度量(Misclassification Measure)，在这里，我们也将其作为句子级SME估计的“分隔度量”(Separation Measure)，并将其与预先设定的margin大小 $\rho$ 作比较。上式在分类面附近的物理意义可以用图(6.1)示意，其中的 $\epsilon_{1\sim6}$ 即代表预设的margin大小 $\rho$ 与分隔度量 $d_r(O_r, \Lambda)$ 之差，即 $\rho - d_r(O_r, \Lambda)$。我们以对方块样本进行分类面学习为例进行说明：首先，所

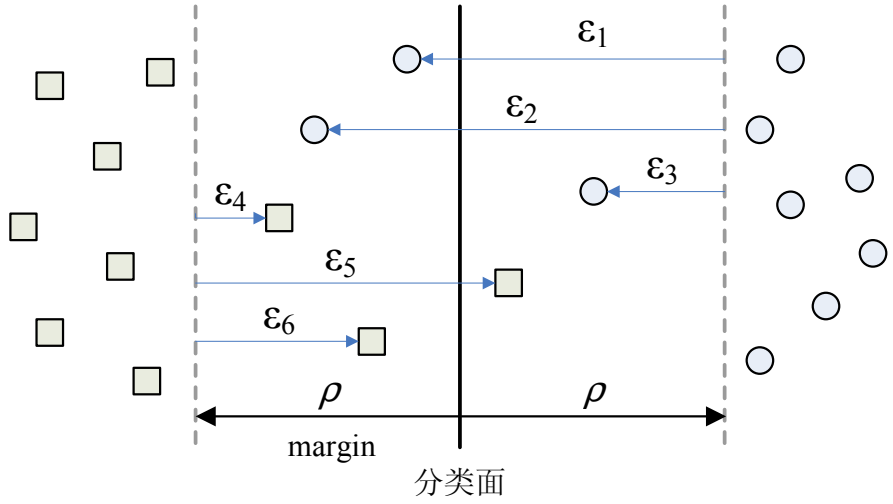


图 6.1　软分类边缘估计SME样本选取的示意图

有在分类边缘margin以外的方块样本(最左边的样本)不产生任何代价，因为它们既能被正确分类，又远离分类面；其次，介于分类边缘和分类面之间的方块样本(此时，分隔度量满足 $0 < d_r(O_r, \Lambda) < \rho$ )产生较小的代价，因为虽然它们能够被正确分类，但却与分类面太近，因此应尽量调整分类面以远离这些样本，以增强推广性能；最后，在分类面右边的方块样本(此时分隔度量满足 $d_r(O_r, \Lambda) < 0$ )产生较大的代价，因为它们尚无法被正确分类，带来较大的经验代价，因此分类面应重点针对这些误分类样本进行相应的调整。

除了(6-5)式所定义的损失代价函数，我们还可以进一步约束 $d_r(O_r, \Lambda)$，而将其修改为如下的函数：

$$\ell_r(O_r, \Lambda) = \begin{cases} \rho - d_r(O_r, \Lambda) & \text{如果 } \rho > d_r(O_r, \Lambda) > \tau \\ 0 & \text{其他} \end{cases} \tag{6-6}$$

不难看出，参数 $\tau$ 的引入将某些过分远离分类面的错误分类样本排除在外。这样设计的目的主要是为了避免在训练过程中因为一些奇异样本点而对正确建模形成干扰。

相对于大分类边缘估计LME[96]来说，软分类边缘估计SME的最大区别在于对误分类样本的使用。更详细的说，LME估计不使用 $d_r(O_r, \Lambda) < 0$、即分类错误的样本进行训练，而只使用分类正确、但却落在分类边缘margin以内的样本( $\rho > d_r(O_r, \Lambda) > 0$ )。这样的样本选择方式在小任务上或许可行，但在大词汇量连续语音识别任务上，句子级 $d_r(O_r, \Lambda) > 0$ 的样本可能非常稀少。而SME估计由于引入了软分类边缘soft margin的思想，使得 $0 > d_r(O_r, \Lambda) > \tau$ 的样本同样可以得到利用，而这些样本对于降低经验代价是有正面作用的。

至此，我们只需要进一步定义误分类度量 $d_r(O_r, \Lambda)$，就可以完成句子级SME估计所有理论方面的设计了。为了与传统的MCE准则可比，我们设计了

与之相同的度量，即：

$$d_r(O_r, \Lambda) = \log \frac{p_\Lambda(O_r \mid W_r)p(W_r)}{\sum_{W \in \mathcal{M}_r} p_\Lambda(O_r \mid W)p(W)} \tag{6-7}$$

其中，$W_r$ 是参考文本词序列，而 $\mathcal{M}_r = \mathcal{M} \setminus \{W_r\}$。

将(6-7)式代入(6-5)或(6-6)式，再代入(6-4)式，我们就完成了句子级SME估计的目标函数。值得一提的是，在传统MCE准则中，$d_r(O_r, \Lambda)$ 是被嵌入到sigmoid函数中而非由(6-5)、(6-6)两式所定义的损失代价函数中。但实质上，MCE准则中所用到的sigmoid函数也可以看作是用来进行训练样本挑选的。只不过挑选的原则和权重与SME中估计完全不同罢了。

### 6.2.3　帧级SME估计

从MWE / MPE相对MMIE准则的改进，到MWCE相对MCE准则的改进中我们可以看到，通过对区分性训练准则从句子级到次句级的细化，我们可以得到更匹配性能评估准则的模型优化途径。而从另一方面来讲，那些因为句子级分隔度量而被排除在训练样本之外的训练语料，其某些局部却仍可能包含至关重要的、可供学习的区分性信息。因此，我们也有必要尝试将最初提出的基于句子级的SME估计细化到更小的级别。

实践中，我们将句子级SME损失代价函数表示为各帧求和的结果，即定义：

$$\ell(O_r, \Lambda) = \sum_t \ell_{rt}(O_{rt}, \Lambda) \tag{6-8}$$

相应的，我们还必须定义帧一级的分隔度量，并令其与预先设置的margin大小一起确定损失代价函数。我们定义如下的帧级分隔度量为 $P_t$，使得：

$$P_t = \sum_{\substack{w, w \in W_r, \\ t_w^{\text{start}} \le t \le t_w^{\text{end}}}} \frac{p_\Lambda(O_r \mid W_r)p(W_r)}{\sum_{W \in \mathcal{M}} p_\Lambda(O_r \mid W)p(W)} \tag{6-9}$$

即，$t$ 时刻的帧级分隔度量 $P_t$ 等于穿过 $t$ 时刻、并属于某正确参考路径的所有词的词后验概率(Word Posterior Probability, WPP)的和。再仿照句子级损失代价函数的定义，可以给出帧级损失代价函数：

$$\ell_{rt}(O_{rt}, \Lambda) = \begin{cases} \rho - d_{rt}(O_{rt}, \Lambda) & \text{如果 } \rho > P_t \\ 0 & \text{其他} \end{cases} \tag{6-10}$$

或：

$$\ell_{rt}(O_{rt},\Lambda)=\begin{cases}\rho-d_{rt}(O_{rt},\Lambda) & \text{如果 } \rho>P_t>\tau\\ 0 & \text{其他}\end{cases} \tag{6-11}$$

其中的帧级误分类度量相应的定义为：

$$d_{rt}(O_{rt},\Lambda)=\log\frac{\sum_{w,w\in W_r,t_w^{\text{start}}\le t\le t_w^{\text{end}}}p_\Lambda(O_r\mid W_r)p(W_r)}{\sum_{w',w'\in W',W'\neq W_r t_{w'}^{\text{start}}\le t\le t_{w'}^{\text{end}}}p_\Lambda(O_r\mid W')p(W')} \tag{6-12}$$

即等于所有在 $t$ 时刻穿过正确参考词 $w$ 的路径的似然度之和，比上所有其他路径的似然度之和再取对数。

不难看出，(6-10)式与(6-11)式着重挑选那些对训练“有益”的帧样本来进行区分性训练。其中，(6-10)式的挑选原则是选取所有的分隔度量小于分类边缘的帧；而(6-11)式则在此基础之上，进一步抛弃了那些可能是噪声样本的、远离分类边界的误分类帧。从后面的实验结果可以看到，(6-11)式中 $\tau$ 的引入对帧级SME估计是至关重要的。

综上，通过将(6-12)式代入(6-10)式或(6-11)式，再代入(6-8)式并进行优化，我们就实现了帧一级的SME估计。帧级SME估计与句子级SME估计的最大区别在于对分隔度量与误分类度量的定义从整句的角度细化到了帧上。这就为充分利用有用的帧级训练样本提供了可行的途径。

## 6.3 实验及结果

为了验证我们提出的句子级及帧级SME估计方法在引入了多种区分性训练方面的最新进展后的性能，我们选择了标准的WSJ0 Nov’92测试任务作为对比各种方法性能的数据库。对于MLE基线系统的训练，我们采用的是标准的、包含7077句训练语料的SI-84训练集。所有MLE相关训练均采用HTK工具进行，最后得到了包含2818个绑定状态的cross-word tri-phone模型系统，而每个状态均使用含有 8 个混合高斯成分的GMM进行建模。声学特征方面，我们采用的是CMN处理后的12维MFCC系数与能量，以及他们的一阶及二阶差分。采用标准的tri-gram语言模型解码时的MLE基线系统的词错误率为 5.06%[①]，这比先前SME估计的实验所用的不跨词种子模型的性能[99]要提高了不少。

针对区分性训练，我们根据正确的参考文本生成了分子词图。对于分母词图，我们首先使用一个bi-gram语言模型解出词图，再使用uni-gram语言模型对其进行重新打分(Re-Scoring)。在训练中，所有声学规整参数均取 1/13，并尽量

① 本章的工作完成于作者在美国佐治亚理工学院(Georgia Institute of Technology)访问研究期间，由于基线系统由对方提供，因此性能指标与第 4 章及第 5 章中有所不同

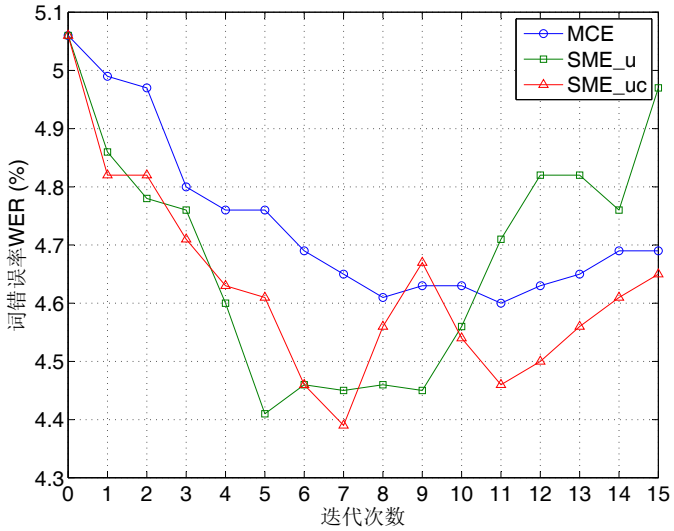


图 6.2　WSJ0 Nov'92 5k测试集上句子级SME估计的词错误率(WER)

| 准则 | MLE | MCE | SME_u | SME_uc |
|---|---|---|---|---|
| WER(%) | 5.06 | 4.60 | 4.41 | 4.39 |
| 相对提升(%) | - | 9.1 | 12.8 | 13.2 |

表 6.1　WSJ0 Nov'92 5k测试集上句子级SME估计的性能对比

使得SME估计与传统MCE准则共享所有能够共享的实现细节。我们抛弃传统的GPD方法，而采用EB方法对模型参数进行优化。此外，由于SME估计本身就致力于提高模型推广性，因此对于那些来源于直觉的平滑方法，如i-smoothing等，我们也弃之不用。

对于传统MCE准则，我们在训练中采用与文献[24]中相同的sigmoid函数参数，即令 $\gamma = 0.04$。通过在区分性训练统一准则框架下的实验，我们得到MCE准则模型的最优词错误率为 4.60%，相对基线系统提高 9.1%。这个提高幅度与文献[24]以及本文第 4 章中的结果是基本可比的，也将作为我们与后续SME估计相对比的参照系统。

### 6.3.1　句子级SME估计

针对SME估计最初提出的两种损失代价函数定义式(6-5)和(6-6)，我们分别将其标为SME_u和SME_uc，并选取参数 $\rho = 2$、$\tau = -30$。从MLE种子模型通过SME估计每步迭代的性能如图(6.2)所示，我们取各种准则所能达到的最优性能作比较，相关的结果如表(6.1)所列。

从实验结果可以看到，三种区分性训练方法均在7-10步迭代时达到最优性能，再往后则不同程度的出现测试集上性能下降的情况。虽然SME估计在很多地方与MCE相同或相似，但通过引入软分类边缘的思想，使得即使是在句子级也能取得超过MCE的性能。相比MCE准则 9.1% 的相对错误率下降，SME_uc方

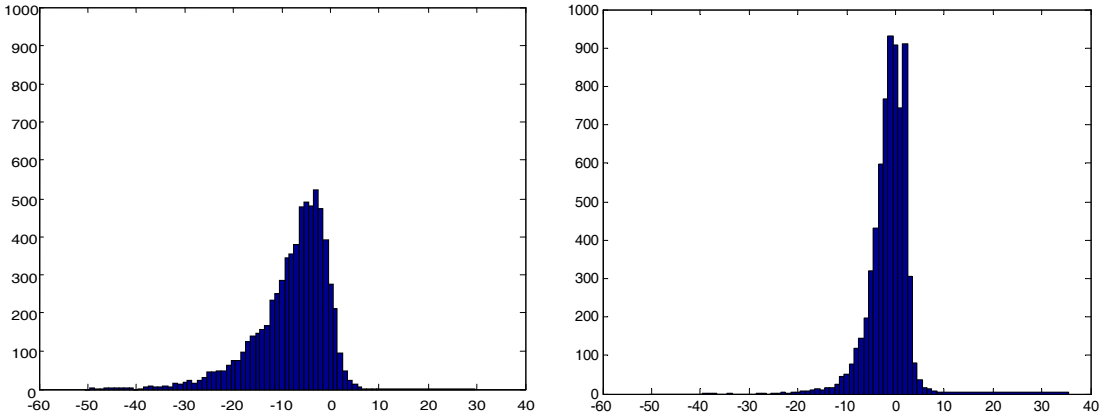


图 6.3　句子级SME估计(SME_u)使用前后的分隔度量 $d_r(O_r, \Lambda)$ 变化情况

法可以取得 13.2% 的更优效果。虽然在这里SME超过MCE的幅度并不算大，但通过诸多区分性训练方面最新技术的引入，我们改变了先前得出的、SME估计只能在小词汇量任务上超过MCE等传统准则的初步结论[99]。同时，通过对两种SME估计的内部性能对比我们可以看到，(6-6)式中 $\tau$ 的引入在句子级SME估计中的效果虽然有一些，但并不显著。在句子级尺度下的奇异样本点一般存在于句子中的某个特定段落，而非整句。因此 $\tau$ 的引入虽然能够抑制该特定段落中的奇异样本，但也同时抑制了段落外的有价值的训练样本。这两方面的矛盾作用抵消之后，造成我们在句子级进行这类操作的性能增益即使有，也很难被明显的观察到。

进一步的，我们分析了句子级SME估计对分隔度量 $d_r(O_r, \Lambda)$ 的优化情况。通过对SME估计前后的分隔度量作直方图，我们得到了如图(6.3)所示的分隔度量优化情况。从图中可以看出，对于初始分隔度量大于 10 的句子，我们几乎观察不到SME_u估计对它们有太明显的变化，因为他们本已远离分类面，且已被正确识别，因而无须过分考虑；但对于初始分隔度量小于 -10 的句子，SME_u估计显然对其显示出了非常强的优化作用。通过优化，大量的句子被调整到了 0 附近，从而使得对它们的区分能力大为增强了。相比之下，SME_uc的优化直方图与SME_u的区别不大，在这里就不再给出。这样的结果也直接造成了在句子级使用 $\tau$ 与不使用 $\tau$ 这两种方式之间的性能差异并不大。

### 6.3.2　帧级SME估计

针对我们提出的帧级SME估计，我们分别将其标为SME_f和SME_fc，对应于式(6-10)和(6-11)。同时，选取参数 $\rho = 0.8$、$\tau = 0.1$。从MLE种子模型通过SME估计每步迭代的性能如图(6.4)所示，相关的结果如表(6.2)所列。

从实验结果中可以看到，SME_f方法可以得到 11.9% 的相对错误下降，而SME_fc方法则可达 18.8%。通过分析我们可以得出以下一些结论：首先，我们所提出的帧级SME估计由于进一步细化了损失代价函数，可以使得性能

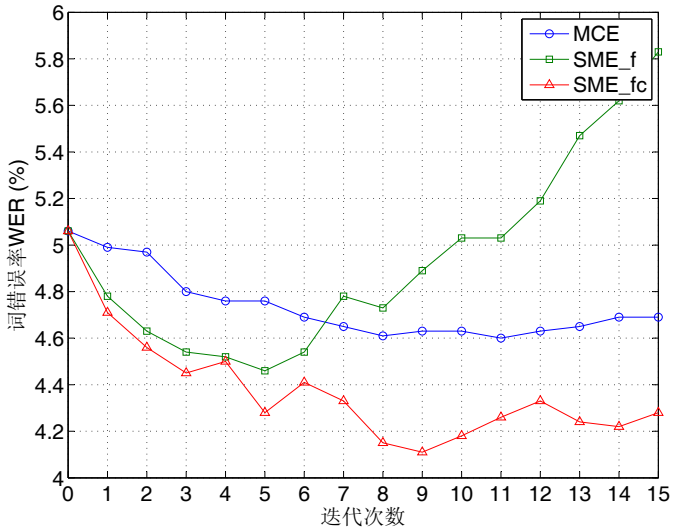


图 6.4　WSJ0 Nov'92 5k测试集上帧级SME估计的词错误率(WER)

| 准则 | MLE | MCE | SME_f | SME_fc |
|---|---|---|---|---|
| WER(%) | 5.06 | 4.60 | 4.46 | 4.11 |
| 相对提升(%) | - | 9.1 | 11.9 | 18.8 |

表 6.2　WSJ0 Nov'92 5k测试集上帧级SME估计的性能对比

相比于句子级SME估计有较大幅度的提高；其次，在帧级SME估计中，引入参数 $\tau$ 以避免噪声帧的影响尤为重要，因为可以看到，单纯的SME_f方法相比SME_u及SME_uc方法并无性能优势，只有引入 $\tau$ 后的SME_fc方法才能较大幅度的超过两种句子级SME方法。我们的实验进一步验证了在区分性训练中细化损失代价函数确实可以带来识别性能上的提升，也验证了我们在帧级分隔度量上引入参数 $\tau$ 的必要性及有效性。

与先前一样，我们分析了帧级SME估计对分隔度量 $P_t$ 的优化情况。通过对SME估计前后的分隔度量作直方图，我们得到了如图(6.5)所示的分隔度量优化情况。从图中可以看出，在使用SME估计之前，分割度量值主要集中在 [0, 0.1] 附近、以及 [0.8, 1.0] 附近。而其它处于 [0.1, 0.9] 区间内的样本帧的分布则比较均匀。仔细分析分隔度量在0附近的那些样本帧可以发现，它们所在句子的分母词图往往被解码得非常混乱，存在对齐错误等情况，因此这些帧更多的是对训练有害的奇异噪声帧。通过(6-11)式中 $\tau$ 的引入，我们可以有效的规避这些帧，从而在优化后形成了图(6.5)中右半幅的情况：在SME_fc训练后，分隔度量呈现出了更加明显的两极分化。其中，[0, 0.1] 段内的样本帧由于 $\tau$ 的关系没有得到太大的优化，几乎保持了原状，从而避免将噪声帧引入而带来的干扰；[0.1, 0.8] 段内的帧得到了重点优化，大量的样本都被优化到了 0.8 这一门限以上，从而提到了区分性；[0.8, 1.0] 段内的帧由于 $\rho$ 的作用基本没有大的变化，且一旦其他段内的样本帧被优化到此段内以后，亦不会再对其进行多余的优化：正如SME的

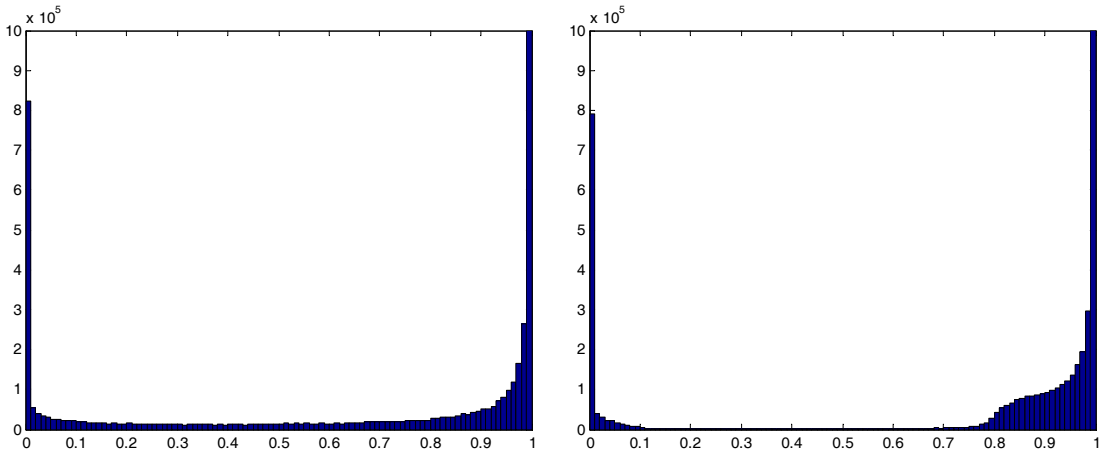


图 6.5　句子级SME估计(SME_fc)使用前后的分隔度量 $P_t$ 变化情况

原理所揭示的那样，对于已经能够较好区分的远离分类面的样本，分类面并不再对其做特别的调整。从图(6.5)中可以充分的看到SME估计的特点与有效性：虽然我们只使用了 [0.1, 0.8] 区间内的样本帧，而且这些样本的数量相对于整个 [0, 1] 区间内的所有样本来讲比例实际并不大，但SME估计却能够取得比使用所有样本更优的识别性能。这显示对于区分性训练中的分类面调整来讲，挑选合适的样本比使用所有样本更为重要，而这也正是基于margin的一类区分性训练方法相对于传统的、不加区别的使用所有样本点的训练准则的最大区别之一。

## 6.4　本章小结

本章主要完成了软分类边缘SME估计方法中两方面的研究工作。第一，对于已经存在的句子级SME估计方法，我们将声学模型区分性训练领域最新技术的进展融入其中，并首次在大词汇量连续语音识别任务上取得了超过传统MCE准则的性能。在此前的SME研究中，该估计方法仅在小词汇量任务上取得了超越传统区分性训练准则的性能，而在大词汇量连续语音识别任务上，则仅能取得超过MLE估计的较小性能提升。通过我们在之前基础上的工作，使得更优的声学模型拓扑结构(cross-word tri-phone)、完善的词图操作、区分性训练统一准则框架、EB模型参数更新方法等技术逐一应用在SME估计中，从而使得句子级SME估计在大词汇量连续语音识别任务上也取得了超过传统MCE准则的优良性能。第二，我们进一步细化了此前句子级SME估计中的损失代价函数的定义，将基于句子一级的函数展开到帧一级。相应的，我们定义了针对帧一级的分隔度量与误分类度量，并引入了控制奇异噪声样本点的参数 $\tau$。通过这些工作，我们合理利用软分类边缘的思想，实现了基于帧级的SME估计。实验结果证明，帧级SME估计由于损失更加细化，可以取得进一步超越句子级SME估计的性能。此外，我们还分析了上述两个级别上SME估计对分隔度量的优化情况，并获得了一些有价值的结论。

# 第 7 章　基于MMIE准则的HMM模型拓扑结构优化

## 7.1　引言

以混合高斯作为概率密度函数的隐马尔科夫模型HMM已经非常成功的应用在了当今主流的自动语音识别系统中。在关于HMM模型的各项理论及实践都已经相当成熟的情况下，以最大似然准则估计MLE训练得到的高斯混合模型GMM通常被视为能最好逼近语音数据“真实”分布的表现形式之一。正是因为如此，以GMM表达的HMM状态，已经毋庸质疑的成为了目前基于HMM的自动语音识别系统中最广泛采用的建模方式。

在通常的、实验性质的语音识别系统中，每个HMM状态中的GMM所包含的高斯核(Gaussian Kernel)数目是固定且相等的。也就是说，给定系统的总高斯核数，通常是以均匀分配的方式将它们平均分给每一个建模单元(如音素、组成音素的各状态等)。在实践中，如果采用这样的高斯核分配方式，有的HMM状态会“过分配”，即他们所得到的高斯核数量超过了实际所需，甚至会因为数据稀疏的原因无法可靠的估计各个混合高斯成分的参数；与此相反，有的HMM状态则会“欠分配”，即它们实际需要更多的高斯核数目，才能确保足够的建模精度以表达数据在这一状态上的真实分布。因此可以说，这种均匀分配高斯核的方式虽然简单易行，但其与生俱来的缺点却限制了它达到最优的建模效果。

也正是因为这样的原因，在实用的或商用的自动语音识别系统中，通常采用非均匀分配的方式将系统中给定的总高斯核数根据一定的准则分配到各个HMM状态上，从而达到优化模型拓扑结构的目的。这些指导高斯核非均匀分配的准则通常同时兼顾信息量与模型参数的数目，并在这两者之间寻找平衡。常用的高斯核非均匀分配准则有Akaike信息准则(Akaike Information Criterion, AIC[117])、贝叶斯信息准则(Bayesian Information Criterion, BIC[118])、最小描述长度准则(Minimum Description Length, MDL[119])等。而其中的BIC / MDL准则因为其描述简单、物理意义明确，又能取得很好的模型选择效果，所以得到了很多实际系统的青睐[120]。

需要指出的是，基于以上各准则的非均匀高斯核分配方法主要是在模型的似然度与复杂度之间寻找平衡，而并不能直接体现非均匀分配模型对各个建模单元的区分度。因此，不难想到通过引入一些区分性准则，如最大互信息量准则MMIE、最小分类错误准则MCE等，来作为非均匀高斯核分配的指导准则。我们希望通过引用这样一些准则，使得高斯核的分配更为直接的指向提高模型的

区分能力，从而也就更为直接的致力于提高系统的识别性能。

在文献[121]中，Normandin定义了基于MMIE准则的启发性度量，并将其应用到HMM模型的连续高斯分裂过程中。Schlüter 在文献[23]中同样采用了类似的方法进行自上而下的高斯分裂。与上述两种方法不同的是，在本章中，指导非均匀高斯核分配的启发性度量是基于一个充分训练的、均匀分配的混合高斯模型来求得的。我们试着在“过分配”与“欠分配”的HMM状态之间“交换”高斯核，从而在保持系统总高斯核数不变的前提下，增强模型的区分性。除此之外，我们还根据这一准则对HMM状态进行时间尺度上的分裂与裁剪。因此对于特定的建模的单元来讲，不仅其分配到的高斯核数目得到了优化，其模型的时间分辨率也得到了优化，所以这种方法可以看作是对整个模型拓扑结构的全面优化。通过使用区分性准则来进行模型拓扑结构优化，我们致力于以更直接的方式来提高模型的识别性能。通过这样的优化，我们可以在保持系统总高斯核数不变的情况下提高系统的识别性能，还可以在保持识别性能不变的情况下压缩模型，从而降低解码阶段的运算消耗。而后者在嵌入式设备等实际应用系统中更是显得尤为重要。我们将所提出的非均匀高斯核分配方法在一个针对嵌入式设备的中文连续数字串系统中进行了验证，并同时对比了传统的均匀分配系统及被广泛采用的、基于BIC / MDL准则的非均匀分配系统。实验结果表明，非均匀分配高斯核的HMM模型在性能上都要优于均匀分配高斯核的模型。而采用我们所提出的基于MMIE准则的高斯核交换方法，能取得最优的识别性能。

本章的后续部分组织如下：首先，我们将在(7.2)节简要回顾传统的基于BIC / MDL准则的非均匀高斯核分配方法；接着在(7.3)节，将重点介绍我们提出的基于MMIE准则的非均匀分配方法；最后，在(7.4)节及(7.5)节，将给出实验结果与本章小结。

## 7.2　基于BIC / MDL准则的非均匀高斯核分配

为了比较我们提出的基于MMIE准则的非均匀高斯核分配方法与传统方法的性能，我们建立了一套基于BIC / MDL准则的非均匀分配HMM模型作为对比系统。为了方便说明准则间的区别，这里也简要的介绍一下BIC / MDL准则的基本原理[119]。对于一个有 $J$ 个状态的HMM系统来说，假设它的各个状态 $j$ 的概率密度函数 $b_j$，都表示为如下的高斯混合密度函数：

$$b_j(O_t) = \sum_{k=1}^{m_j} c_{jk} \mathcal{N}(O_t; \mu_{jk}, \mathbf{\Sigma}_{jk}) \tag{7-1}$$

其中，$m_j$ 是该状态的高斯核成分数目，$c_{jk}$、$\mu_{jk}$ 和 $\mathbf{\Sigma}_{jk}$ 分别是该状态中第 $k$ 个高斯核的权重、均值和协方差矩阵。那么，基于BIC / MDL准则的非均匀高斯

核分配，就是要对状态 $j$ 找到一个最优的高斯核成分数 $m_j^*$，使得该状态的BIC值最大化，即：

$$m_j^* = \arg\max_{m_j} \ \mathrm{BIC}(m_j) \tag{7-2}$$

而BIC值则相应的定义为：

$$\mathrm{BIC}(m_j) = \sum_r \sum_t \gamma_{jrt}^{+} \times \log b_j(O_{rt}) - \frac{1}{2}\lambda_p \times (\#\Phi_{m_j}) \times \log[\sum_r \sum_t \gamma_{jrt}^{+}] \tag{7-3}$$

其中，$\gamma_{jrt}^{+}$ 是状态 $j$ 在训练语料 $r$ 的 $t$ 时刻给定正确参考模型时的状态占有率(后验概率)，$\#\Phi_{m_j}$ 表示在有 $m_j$ 个混合高斯成分时该状态的自由参数个数，而 $\lambda_p$ 则是用来调整模型复杂度惩罚权重的常数。

不难看出，BIC / MDL准则的物理意义即是同时度量模型的似然度以及得到该似然度时模型所占用的参数数目。通过对似然度加上一个与模型参数数目呈正相关的惩罚，就不难找到使得BIC值取得最大的“平衡点”。这个使得BIC值取得最大的 $m_j$ 就将会被作为状态 $j$ 最终的高斯核成分数目。因此，通过对每个状态各自分别进行上面的高斯核数目选择与优化，就能得到使得BIC / MDL准则得以优化的非均匀高斯核分配方案。

## 7.3　基于MMIE准则的非均匀高斯核分配

与前述基于模型似然度及复杂度的分配准则不同，基于区分性准则的非均匀高斯核分配着重优化模型拓扑结构以提高其区分能力。由于区分性准则一直被视为更直接的与语音识别的性能评估准则相关，因而采用区分性准则来指导非均匀高斯核分配将有可能进一步提高模型的识别性能。在本章中，我们使用最大互信息量估计MMIE准则来指导高斯核的非均匀分配。我们尝试在模型的各状态之间“交换”高斯核，从而提高模型的区分能力乃至识别性能。

### 7.3.1　基于MMIE准则的目标函数及启发性度量

首先，我们将(3-13)式中的MMIE准则仿照(5-37)式展开至状态一级，即：

$$\begin{aligned} \mathcal{F}_{\mathrm{MMIE}} &= \frac{1}{R}\sum_j \sum_r \sum_t [\gamma_{jrt}^{+} - \gamma_{jrt}^{-}] \times \log b_j(O_{rt}) + C \\ &= \frac{1}{R}\sum_j \sum_r \sum_t [\gamma_{jrt}^{+} - \gamma_{jrt}^{-}] \times \log[\sum_{k=1}^{m_j} c_{jk}\mathcal{N}(O_{rt};\mu_{jk},\Sigma_{jk})] + C \end{aligned} \tag{7-4}$$

其中，$\gamma_{jrt}^{+}$ 是状态 $j$ 在第 $r$ 句训练语料中给定参考正确模型时，在 $t$ 时刻出现的后验概率，$\gamma_{jrt}^{-}$ 则是其给定竞争模型空间时的后验概率。在(7-4)式中，$\gamma_{jrt}^{+}$ 仅和状态 $j$ 自身的混合高斯成分数及其参数有关，而 $\gamma_{jrt}^{-}$ 则不仅和状态 $j$ 自身

有关，还和其他所有可能与其形成竞争的其他状态有关。对比基于BIC / MDL准则的非均匀高斯核分配，基于MMIE准则进行非均匀高斯核分配的主要区别在于其目标函数(7-4)式无法再看作状态独立、进而分别优化各个状态的高斯核数目。相反，对任意一个状态高斯核成分数目及其参数的改变都会影响整个系统，从而直接使得 $\gamma_{krt}^{-}$ 的值发生改变。也正是因为如此，基于MMIE准则的全局最优分配解很难直接求得，需要另外定义启发性度量来指导高斯核的分配。

在本章中，我们利用目标函数对各高斯核成分权重的导数来做为指导高斯核非均匀分配过程的启发性度量。从导数的物理意义上讲，一个正的导数值表明该高斯核需要增长，而一个负的导数则证明它需要削减，这正好与我们要进行的非均匀高斯核分配的目标相一致。因此，定义状态 $j$ 中高斯核成分 $k$ 的启发性度量为：

$$\begin{aligned} H_{jk} = \frac{\partial \mathcal{F}_{\text{MMIE}}}{\partial c_{jk}} &= \frac{1}{R}\sum_{r}\sum_{t}[\gamma_{jrt}^{+} - \gamma_{jrt}^{-}] \times \frac{\mathcal{N}(O_{rt};\mu_{jk},\boldsymbol{\Sigma}_{jk})}{\sum_{k=1}^{m_j} c_{jk}\mathcal{N}(O_{rt};\mu_{jk},\boldsymbol{\Sigma}_{jk})} \\ &= \frac{1}{R}\sum_{r}\sum_{t}[\gamma_{krt}^{+} - \gamma_{krt}^{-}]/c_{jk} \end{aligned} \tag{7-5}$$

其中，$\gamma_{krt}$ 是对应各高斯核的后验概率，也即是说，该启发性度量使用给定参考模型时该高斯核的后验概率，减去给定竞争模型时的后验概率，并同时以该高斯核的权重加以归整。由于(7-5)式的求和项是由两项相减组成的，那么，针对 $H_{jk}$ 不同的符号和幅度，可以分以下三种情况加以讨论：

情况一：$H_{jk} \approx 0$

这种情况表明，从总的趋势上来讲 $\gamma_{krt}^{+} \approx \gamma_{krt}^{-}$，或者说，高斯核 $k$ 几乎总是在解码中占据支配地位。在这种情况下，由于该高斯核已经被很好的建模，因此没有必要调整它所属状态的混合高斯成分数目。因此，$H_{jk} \approx 0$ 表明了一个分配合理、不需要调整的高斯核；

情况二：$H_{jk} > 0$

这种情况表明，在训练数据中 $\gamma_{krt}^{+} > \gamma_{krt}^{-}$ 的情况占多数，或者说，该高斯核及其所属的状态作为参考文本中的正确模型，却没有能够在解码中竞争过其它的错误状态及其其中的高斯核。在这种情况下，从总的趋势上讲，就会最终出现 $H_{jk} > 0$ 的情况。因此，启发性度量 $H_{jk} > 0$ 表明了一个欠分配的、需要增加其状态成分数目的高斯核。同时，除了符号以外，$H_{jk}$ 的幅度还表示了该高斯核建模的精度不够的严重程度；

情况三：$H_{jk} < 0$

这种情况表明，在训练数据中 $\gamma_{krt}^{+} < \gamma_{krt}^{-}$ 的情况占多数，也就是说，该高斯核本身不存在于参考文本中，但在解码时却取得了较大的优势。在这种情况下，由于 $\gamma_{krt}^{+}$ 为 0，而 $\gamma_{krt}^{-}$ 为正值甚至较大，从总的趋势上讲，就会最终出现 $H_{jk} < 0$ 的情况。显然，在这种情况发生时，该高斯核更多的是以干扰的形式出现在解码过程中，这种干扰与正确的高斯核形成了错误的竞争。因此，该高斯核应当被削弱以减少这种竞争。在这里，$H_{jk}$ 的幅度同时还表示了该高斯核干扰正确解码的严重程度。

综上，由启发性度量 $H_{jk}$ 的上述性质，基于MMIE准则的非均匀高斯核分配可以按以下 3 步进行：1、对系统中所有的高斯核计算 $H_{jk}$，并按此进行排序；2、将排在最末尾的 $N$ 个高斯核( $H_{jk}$ 符号为负、幅度最大)进行裁减；3、对排在最靠前的 $N$ 个高斯核( $H_{jk}$ 符号为正、幅度最大)进行分裂。不难看出，在上述过程中，由于裁减和分裂的高斯核在数量上是相等的，这实际上是一个高斯核的“交换”过程。也正是因为这样，系统中总的高斯核数目在整个模型结构优化过程中是保持不变的。

### 7.3.2　模型拓扑结构后处理

经过上述高斯核调整过程，一部分HMM状态的高斯核数目可能会被减少到 0，而另一些状态的高斯核则可能会增长过多，这两种情况都会加剧各状态之间在概率密度上不可比的问题，从而不利于建模。因此，还需要对模型进行时间尺度上的拓扑结构后处理来进一步调整优化参数分配。

情况一：裁减高斯核数目被减至 0 的状态

假设经过上述高斯核交换过程，某状态的高斯核数目已被减至 0，这实际表明该状态在时间尺度上对区分性已无贡献，因此，有必要对其进行裁减。设被裁减的状态为 $s$，而所有进入 $s$ 的状态集合为 $P$、所有跳出 $s$ 的状态集合为 $Q$，那么，裁减 $s$ 后的状态转移概率将会相应变化为：

$$\begin{aligned} &a_{ps} = 0, \forall p \in P; \quad a_{sq} = 0, \forall q \in Q \\ &a_{pq} = a'_{pq} + a'_{ps} \cdot \frac{1}{1 - a'_{ss}} \cdot a'_{sq}, \forall p \in P, q \in Q \end{aligned} \tag{7-6}$$

其中，$a$ 为各状态在状态 $s$ 被裁减后的转移概率，而 $a'$ 则为状态 $s$ 被裁减前的转移概率。整个状态裁减和转移概率调整的过程可以如图(7.1)所示。

情况二：分裂高斯核数目增加过多的状态

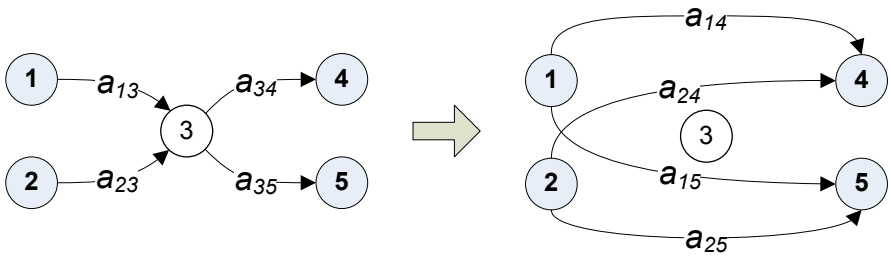


图 7.1　裁减高斯核数目被减至 0 的状态并重新计算转移概率

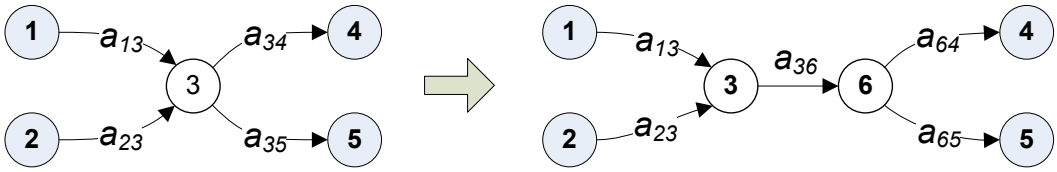


图 7.2　分裂高斯核数目过多的状态并加入转移概率

假设经过上述高斯核交换过程，某状态的高斯核数目被增加至超过了按经验预设的门限。此时，就需要相应的对该状态进行时间尺度上的分裂，使得该状态中过多的高斯核平均分配到时间上相连的两个状态中去。相比上面的状态裁减过程，这种状态分裂较为简单，如图(7.2)所示，只需要将状态 $s$ 的所有高斯核平均分配到两个新的状态 $s_1$、$s_2$ 中，并加入 $s_1$ 到 $s_2$ 的跳转概率即可。在实验中，$s_1$ 与 $s_2$ 之间的转移概率可由经验设置，一般可设为 0.5。总结起来，上述基于最大互信息量准则的非均匀高斯核交换及模型拓扑结构后处理过程，可以用如表(7.1)所示的伪代码来表达。

## 7.4　实验及结果

### 7.4.1　实验配置

我们在一个面向嵌入式设备的中文连续数字串语音识别任务上验证了上述基于MMIE准则的模型拓扑结构优化方法。由于嵌入式设备的运算资源有限，在模型参数较少时尽可能提高识别性能，就显得格外重要。而高斯核非均匀分配正好可以满足这样的需求。

我们所采用数据库的语音采自办公室环境，共包含 100 个发音人( 50 男 50 女)。对每一个发音人，我们都录制了约 120 句由“零”到“九”及“幺”所组成的、长度分别为 1 至 12 的连续数字串。在数据库设计时，我们保证了每个数字的样本数分布是均匀的。我们挑选了其中的 50 人( 6209 句)作为训练集，另外的 50 人( 6199 句)作为测试集。

我们使用HTK进行模型训练及识别实验。首先，我们训练了传统的基于均匀高斯核分配的基线系统模型。对于该模型，我们采用了 10 状态的自左向右HMM作为11个数字的基本建模结构。此外，一个 3 状态的静音silence模型及

| | | | |
|---|---|---|---|
| 初始化： | | | |
| | 训练一套基于MLE估计的、高斯核均匀分配的模型 | | |
| 基于MMIE准则的非均匀高斯核交换： | | | |
| | 从训练集中统计启发性度量 $H_{jk}$ | | |
| | 排序 $H_{jk}$，生成排序列表 | | |
| | 高斯核交换： | | |
| | | 设已交换高斯核数 $n=0$、总高斯核交换数为 $N$<br>当 $n<N$ 时： | |
| | | | 定位 $H_{jk}$ 为负、幅度最大的高斯核：$i=\arg\min H_{jk}$<br>裁减高斯核 $i$，将其所对应的 $H_{jk}$ 删除出排序列表 |
| | | | 定位 $H_{jk}$ 为正、幅度最大的高斯核：$i=\arg\max H_{jk}$<br>分裂高斯核 $i$，将其所对应的 $H_{jk}$ 删除出排序列表 |
| | | | $n=n+1$ |
| | 模型拓扑结构后处理 | | |
| | 基于MLE的模型参数重估 | | |

表 7.1　基于MMIE准则的非均匀高斯核交换及模型拓扑结构后处理

一个单状态的短停顿short pause (sp)模型也被引入进系统中以对静寂段进行建模。我们分别训练了每状态高斯核数分别为 2、4、8、12、16 的均匀分配模型，作为我们可供对比的基线系统。

在基于传统BIC准则的非均匀高斯核分配模型方面，我们使用一个均匀分配的、每状态 32 高斯的模型作为初始种子，并利用BIC准则压缩至平均每状态 2、4、8、12 及 16 混合高斯。压缩完成后，我们还进行了 4 步MLE迭代以优化模型的均值方差等参数。

最后，对于我们所提出的基于最大互信息量准则的高斯核交换方法，则直接采用了 5 个基线系统模型做初始种子，分别训练平均每状态 2、4、8、12 及 16 混合高斯的非均匀分配模型。在实际的高斯核交换过程中，设定每一次交换模型中 10% 的高斯核，而整个过程被迭代执行了 5 次。每次模型拓扑结构发生改变后，也同样进行了 4 步MLE迭代以更新模型参数。

### 7.4.2　实验结果

均匀分配的基线系统、传统的基于BIC准则的非均匀分配方法，以及我们所提出的基于MMIE准则的非均匀分配方法在我们的中文连续数字串任务上的识别性能如图(7.3)和(7.4)所示。

从实验结果我们可以看出：1、两种非均匀高斯核分配模型的识别性能都一致性的优于作为基线系统的、均匀分配的模型；2、对比两种不同准则的非均匀分配模型，可以看到基于MMIE准则的非均匀高斯核分配模型的识别性能又要一致的优于基于BIC准则的非均匀分配模型。

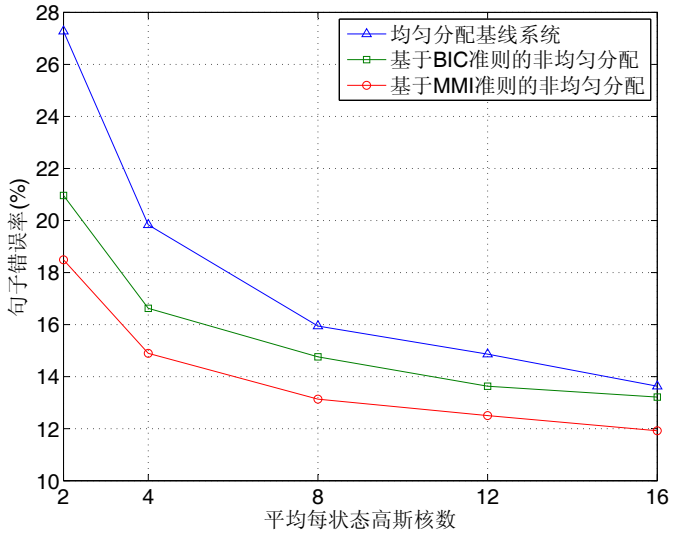


图 7.3　三种模型的句子错误率随状态平均高斯核数的变化

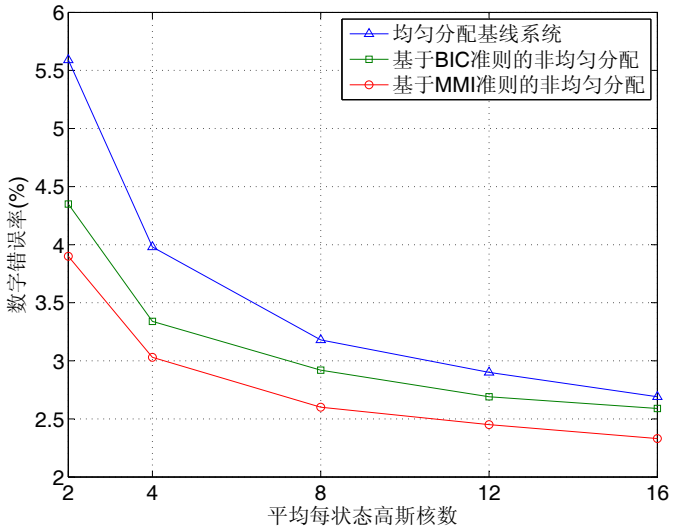


图 7.4　三种模型的数字错误率随状态平均高斯核数的变化

从图中不难看出，纵向的从提高识别率的角度来看，非均匀分配模型在总高斯核数相等的情况下总能取得比均匀分配模型更好的识别性能；横向的从模型压缩的角度来看，当识别率基本相当时，非均匀分配模型所需要的高斯核数目只需要均匀分配模型的 2/3 甚至 1/2 (例如均匀分配的 12 混合高斯模型与MMIE非均匀分配的平均 4 ~ 8 混合高斯模型性能相当)。这种现象在连续数字串识别这类小任务上的趋势显得更加的明显。

我们不妨以平均每状态 4 高斯的三种模型为例来分析不同的准则对模型拓扑结构优化有着什么样的影响。表(7.2)给出了均匀分配模型、BIC非均匀分配模型、MMIE非均匀分配模型在训练集上针对各个数字高斯核分配情况和相应的训练集识别错误率。不难看出，对均匀分配高斯核时错误最多的三个数字“五”、“一”和“二”来说，基于BIC准则和MMIE准则的两种分配方法有着同样的分配修正方向(一致的增多或减少高斯核数目)；同时，基于区分性

| | 一 | 二 | 三 | 四 | 五 | 六 | 七 | 八 | 九 | 零 | 幺 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 均匀分配 | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 |
| 错误率(%) | 11.17 | 6.75 | 0.15 | 0.66 | 13.69 | 1.48 | 0.81 | 0.75 | 0.92 | 3.64 | 1.11 |
| BIC非均匀 | 42 | 39 | 49 | 44 | 30 | 33 | 39 | 42 | 45 | 32 | 45 |
| 错误率(%) | 11.17 | 5.68 | 0.12 | 0.59 | 10.25 | 1.84 | 0.66 | 0.31 | 0.67 | 3.83 | 1.50 |
| MMIE非均匀 | 56 | 36 | 40 | 39 | 28 | 40 | 39 | 38 | 43 | 39 | 42 |
| 错误率(%) | 10.17 | 4.77 | 0.09 | 0.69 | 8.58 | 1.36 | 0.50 | 0.50 | 0.83 | 3.12 | 1.20 |

表 7.2　三种模型在平均每状态 4 高斯核时的高斯核分配情况及对应的训练集错误率

| | 均匀分配模型 | BIC非均匀分配 | MMIE非均匀分配 |
|---|---|---|---|
| 句子错误率(%) | 19.83 | 16.62 | 14.89 |
| 数字错误率(%) | 3.98 | 3.34 | 3.03 |

表 7.3　三种模型在平均每状态 4 高斯核时的识别性能对比

的MMIE准则相比基于似然度的BIC准则，对模型参数分配的调整幅度更大。与此相反，对于错误最少的几个数字，如“三”、“四”等，基于似然度的BIC准则可能仍会对它们进行一定幅度的调整以提高整体似然度；但基于区分性的MMIE准则却几乎不对这些数字进行太大的变动。这种现象的产生，与我们所采用的模型拓扑结构优化准则的物理意义是直接相关的：基于BIC准则的优化方法更着眼于提高模型对训练集的似然度(在我们的实验中，均匀分配模型的帧平均对数似然度为 -51.82，BIC模型为 -51.03，MMIE模型为 -51.91 )，而基于MMIE准则的优化方法则更注重调整在训练集中识别率较低的混淆模型，从而直接提高这些模型的识别率。从总的效果来看，基于MMIE准则优化的模型在三种模型中似然度最低而识别率最高，这正好从一个侧面证明了似然度与识别率虽有联系，但却并非是一种紧密的直接联系。将三种模型应用于测试集进行测试，其识别性能如表(7.3)所示。相对于基线系统，基于BIC准则的非均匀分配把句子及数字两个级别的错误率分别相对降低了 16.2% 和 16.1%；而基于MMIE准则的非均匀分配通过对模型结构的进一步优化，使得句子及数字相对错误率分别降低了 24.9% 和 23.9%。这个结果表明，根据训练数据进行的模型拓扑结构优化，能够较一致的反映到测试集上。

## 7.5　本章小结

本章提出了一种基于区分性的、MMIE准则的声学模型拓扑结构优化方法，并将其与传统的基于模型似然度和复杂度的BIC / MDL优化方法进行了对比。通过定义合理的启发性度量，我们尝试在一个训练充分的多混合高斯均匀分配模型的各个状态之间“交换”高斯核，从而使得MMIE准则得以优化。实验结果表

明，由于基于区分性准则的优化更为直接的将模型拓扑结构的调整与模型的区分能力联系起来，因此，也就能够取得比单纯基于似然度的方法更好的识别性能。本章的工作主要在中文连续数字串这一较小的任务下进行，相关方法在应用到更大规模任务下后有可能遇到的运算量及推广性方面的问题，而这个课题可以作为今后研究的方向。

# 第 8 章 总结

## 8.1 本文的主要工作

本文是对作者在攻读博士学位期间，在语音识别声学模型方面所做研究工作的一个汇总。这些工作主要围绕对声学模型的区分性训练展开，在区分性训练准则、模型参数优化方法，以及区分性训练的应用性方法等方面进行了较详细的研究和探索。其主要创新包括：1、提出了一种新的区分性训练准则MWCE，通过准则的细化来匹配大词汇量连续语音识别的目标，从而直接作用于降低词错误率；2、提出了区分性训练中基于Trust Region的HMM参数更新方法，改进了传统优化方法的一些缺陷，更有效的在MMIE准则下实现了对目标准则的优化和对识别性能的提升；3、进一步完善了传统句子级SME估计的现有实现，并提出SME估计的帧级区分性训练方法。通过引入Soft Margin的思想，实现了对训练样本的挑选，提高了区分性训练的可推广性；4、提出了区分性训练准则在应用背景下对声学模型拓扑结构优化的方法，通过在MMIE准则下定义启发性度量，指导模型拓扑结构在空间和时间两方面的优化，从而提高了声学模型在同等参数数目下的识别性能。

具体来说，首先，本文介绍了我们所提出一种新的区分性训练准则MWCE。由于区分性训练准则主要解决“优化什么”的问题，准则之间的性能差异往往体现在对需要优化的具体问题的定义上。MWCE准则所定义的优化目标可以看作是对传统的、基于句子级的MCE准则的细化。通过这样的细化，我们可以使得对模型参数的优化更直接的匹配大词汇量连续语音识别的最终目标，即降低词错误率WER。因此，MWCE准则也就很自然的能够取得比传统句子级准则更优异的识别性能。相比其他的次句级准则，如词级的MWE、音素级的MPE来说，MWCE准则提供了对词级错误的一个不同角度的表达。它主要沿着区分函数、误分类度量及损失代价函数这条线索来近似估计词分类错误，并对其进行优化。我们在区分性训练统一准则框架下实现了各种准则的客观对比，在WSJ0和TIMIT数据库上，MWCE准则都取得了最好的识别效果。这项工作进一步证明，细化区分性训练准则以使之更贴近识别目标，通常都会取得比传统的句子级准则更好的性能。而由于目前仍无法得到对于词或音素级错误的精确估计，寻找更贴切、更合理的准则并进行优化仍具有一定的空间。

其次，我们针对区分性训练的模型参数优化问题，提出了MMIE准则基于Trust Region的模型参数优化方法。这一方法将传统EB方法的无界优化问题

加以约束，从而在一个可靠的信任区域内对模型参数进行逐步调整。通过使用这一方法，我们可以用一种数学上更为合理、物理意义上更为明确的方式来避免EB算法中经验性的求取训练参数 $D$ 的问题。同时，由于在每步迭代中我们都可以得到辅助函数在约束条件下的全局最优解，因此参数更新的效率也就更高。我们在WSJ0数据库上对这一参数更新算法进行了诸多实验，给出了几种迭代更新方式及参考的Trust Region大小。我们对比了基于Trust Region的方法与传统EB方法在优化准则与降低错误率两方面的性能。从实验结果可以发现，新方法在准则优化与提高识别率两方面都有着较好的性能。

第三，我们针对声学模型区分性训练准则方面已经有了初步研究的SME估计，提出了一系列补充和改进方法。首先，我们将已经存在的句子级SME准则加以完善，将区分性训练领域近年来的一些重要技术引入，实现了句子级SME准则在大词汇量连续语音识别中的成功应用。其次，我们提出了更细化的帧级SME区分性训练，通过引入SVM中软分类边缘soft margin的概念，在帧尺度上对正确参考与错误竞争进行筛选，寻找对模型训练有益的样本。我们在WSJ0数据库上详细分析了句子及帧两种尺度上SME准则的优化行为，并对比了它们之间、以及它们与传统MCE准则之间的识别性能。实验结果表明，两种SME准则都能取得超过MCE准则的性能。而在引入排除干扰样本帧的参数 $\tau$ 后，帧级的SME方法可以取得最优的识别性能，并明显超过传统MCE。这也是SME方法在大词汇量连续语音识别上第一次取得明显超过传统方法的性能。

最后，作为区分性训练准则的一种应用，我们还提出了一种基于MMIE准则的HMM模型拓扑结构优化方法。这种方法通过区分性准则定义出一个启发性度量，并通过它在HMM声学模型的各个状态之间交换高斯核，从而实现参数分配的优化配置。除此之外，我们还提出在时间尺度上对HMM建模单元的状态拓扑结构进行后处理，进一步优化其在这一尺度上的区分能力。我们在一个面向嵌入式应用的中文连续数字串语音识别的任务上对我们提出的方案进行了实验，并分析比较了传统准则及区分性准则下模型拓扑结构优化行为的异同。实验结果表明，基于MMIE准则的模型拓扑结构优化能够更为直接的将模型结构与模型的区分能力联系起来。因此，也就能够取得比单纯基于似然度的其它传统准则更优的识别效果。

## 8.2 进一步的研究方向

对本文所介绍的工作来说，仍有如下一些可以继续深入的方向：首先，对MWCE准则的实现还需要一些关键的近似，我们实际上仍不能得出对真正词分类错误的精确估计；其次，对基于Trust Region的模型参数优化方法而言，我们

还可以继续探讨它对除MMIE以外的准则的优化方法，并寻找更合适的、自适应的，乃至自动的Trust Region大小定义方法；第三，我们还可以对SME方法进一步展开，将其思想应用到其它区分性训练方法乃至区分性训练统一准则框架下，探讨更好的方法来解决区分性训练模型的推广性问题[①]。最后，我们的HMM模型拓扑结构优化方法由于运算量的关系，目前尚只能应用在较小规模的任务上。如何在大规模连续语音识别任务上进行这样的模型结构优化，并继而将区分性准则推广到对模型单元决策树绑定的优化上，都是可以继续研究的方向。

从更广的范围来讲，声学模型区分性训练及其应用在近年来已经取得了长足的进步，并逐渐成为大词汇量连续语音识别中的标准训练手段。但同样应该看到，虽然在性能上区分性训练已经能够较稳定的取得超过传统最大似然估计的性能，但区分性训练自身仍有一些待解决的问题值得继续研究。

大体上讲，进一步研究的方向仍主要集中在准则与优化两个方面。应该说，目前的区分性训练准则已经相当多，但各自在不同任务上的表现却不尽相同。我们尚没有一种准则能较稳定的在大多数任务上都取得较好的性能，这显示出我们对要“优化什么”这一根本性问题的认识还有待进一步加强。除此之外，在林林总总的准则下，我们都会采用一些近似(如声学规整因子等)来达到表面上更优的测试集性能。但实际上，亦有研究者指出，这些近似实际模糊了各准则之间的区别，并使得我们进行参数优化后的实际结果偏离了此前准则的理论定义。如何有效的跨越准则理论与实际实现之间的鸿沟，也可以作为一个思考的方向。

在模型参数优化方面，目前常用的优化方法仍只是在经验和实践上能够取得一定的效果。而在数学理论上，这些方法却远称不上完美。面对如此复杂的区分性训练目标函数，在数学上寻找更好的优化方法，也是非常值得研究的课题。

声学模型区分性训练的推广性问题也是对其更广泛应用的潜在威胁。区分性训练从本质上讲仍是对训练集所呈现出的语音空间精细结构的学习。当训练与测试环境不匹配时，区分性训练准则所表现出的敏感性常常超过最大似然估计准则。要解决这一问题，从根本上讲仍需要从准则和优化两方面加以考虑。近年来所出现的一些基于margin的方法可能是解决这一问题的有效途径之一。但这类方法还需要进一步的工作来证明其在大词汇量连续语音识别中的可行性。

最后，区分性准则在其它方面的应用，如特征提取、模型自适应乃至置信度判决等方面，也是非常有意义的新课题。我们可以很清楚的看到近年来的一些研究工作在这些方面的进展。事实上，区分性训练的一些思想可以应用在语音识别、乃至更广泛领域的方方面面。对声学模型的区分性训练致力于从区分性

① 在Georgia Institute of Technology的李锦宇尚未发表的论文中，已将帧级SME方法与MWCE准则相结合，并在WSJ0数据库上取得了超过所有传统准则的识别性能。其论文Soft Margin Estimation with Various Separation Levels for LVCSR (Jinyu Li, Zhi-Jie Yan, Chin-Hui Lee and Ren-Hua Wang)已投稿至ICSLP2008

的角度看待语音识别问题，追求用更好的训练准则和优化方法来进行更正确的识别，而这实质上也应该是语音识别全过程中所有方法的最终追求之所在。

# 插图索引

# 表格索引

# 参考文献


[1] Davis K H, Biddulph R, Balashek S. Automatic Recognition of Spoken Digits. The Journal of the Acoustical Society of America, 1952, 24(6):637–642.

[2] Vintsyuk T K. Speech Discrimination by Dynamic Programming. Cybernetics and Systems Analysis, 1968, 4(1):81–88.

[3] Itakura F, Saito S. A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies. Electronics and Communications in Japan, 1970, 53(A):36–43.

[4] Lowerre B. The Harpy Speech Understanding System. Morgan Kaufmann Publishers Inc., 1990: 576–586.

[5] Erman L D. Overview of the Hearsay Speech Understanding Research. ACM SIGART Bulletin, 1976, (56):9–16.

[6] Klatt D H. Review of the ARPA Speech Understanding Project. The Journal of the Acoustical Society of America, 1977, 62(6):1345–1366.

[7] Jelinek F, Bahl L, Mercer R. Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech. IEEE Trans. on Information Theory, 1975, 21(3):250–256.

[8] Schwartz R, Chow Y, Kimball O, et al. Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech. Proceedings of ICASSP1985, 1985. Vol. 10, 1205-1208.

[9] Juang B H, Levinson S, Sondhi M. Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains. IEEE Trans. on Information Theory, 1986, 32(2):307–309.

[10] Rabiner L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989, 77(3):257–286.

[11] Gauvain J L, Lee C H. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Trans. on Speech and Audio Processing, 1994, 2(2):291–298.

[12] Leggetter C, Woodland P. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language, 1995, 9(2):171–185.

[13] Young S, Odell J, Woodland P. Tree-Based State Tying for High Accuracy Acoustic Modelling. Proceedings of ARPA Workshop on Human Language Technology, 1994. 307-312.

[14] Odell J. The Use of Context in Large Vocabulary Speech Recognition[D]. Cambridge University, 1995.

[15] Lee K F. Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The Sphinx System[D]. Carnegie Mellon University, 1988.

[16] Chow Y, Dunham M, Kimball O, et al. BYBLOS: The BBN Continuous Speech Recognition System. Proceedings of ICASSP1987, 1987. Vol. 12, 89-92.

[17] Murveit H, Cohen M, Price P, et al. SRI's DECIPHER System. Proceedings of Workshop on Speech and Natural Language, 1989. 238-242.

[18] Huang X, Acero A, Alleva F, et al. Microsoft Windows Highly Intelligent Speech Recognizer: Whisper. Proceedings of ICASSP1995, 1995. Vol. 1, 93-96.

[19] Bahl L R, Brown P F, Souza P V, et al. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. Proceedings of ICASSP1986, 1986. 49-52.

[20] Juang B H, Katagiri S. Discriminative Learning for Minimum Error Classification. IEEE Trans. on Signal Processing, 1992, 40(12):3043–3054.

[21] Young S, et al . The HTK Book (Revised for HTK version 3.4). 2006.

[22] Povey D, Woodland P. Minimum Phone Error and I-Smoothing for Improved Discriminative Training. Proceedings of ICASSP2002, 2002. Vol. 1, 105-108.

[23] Schlüter R. Investigations on Discriminative Training Criteria[D]. RWTH Aachen University, 2000.

[24] Macherey W, Haferkamp L, Schlüter R, et al. Investigations on Error Minimizing Training Criteria for Discriminative Training in Automatic Speech Recognition. Proceedings of EuroSpeech2005, 2005. 2133-2136.

[25] Hasegawa-Johnson M, Baker J, Borys S, et al. Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop. Proceedings of ICASSP2005, 2005. Vol. 1, 213-216.

[26] Soltau H, Kingsbury B, Mangu L, et al. The IBM 2004 Conversational Telephony System for Rich Transcription. Proceedings of ICASSP2005, 2005. Vol. 1, 205-208.

[27] Soltau H, Saon G, Kingsbury B, et al. The IBM 2006 GALE Arabic ASR System. Proceedings of ICASSP2007, 2007. Vol. 4, 349-353.

[28] Seide F, Yu P, Ma C, et al. Vocabulary-Independent Search in Spontaneous Speech. Proceedings of ICASSP2004, 2004. Vol. 1, 253-256.

[29] Biem A, Katagiri S, Juang B H. Pattern Recognition Using Discriminative Feature Extraction. IEEE Trans. on Signal Processing, 1997, 45(2):500–504.

[30] Povey D, Kingsbury B, Mangu L, et al. fMPE: Discriminatively Trained Features for Speech Recognition. Proceedings of ICASSP2005, 2005. Vol. 1, 961-964.

[31] Yan Z J, Soong F K, Wang R H. Word Graph Based Feature Enhancement for Noisy Speech Recognition. Proceedings of ICASSP2007, 2007. Vol. 4, 373-376.

[32] Davis S, Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1980, 28(4):357–366.

[33] Hermansky H. Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustical Society of America, 1990, 87(4):1738–1752.

[34] Picone J. Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE, 1993, 81(9):1215–1247.

[35] Hunt M, Lefèbvre C. A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech. Proceedings of ICASSP1989, 1989. Vol. 1, 262-265.

[36] Kumar N. Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition[D]. Johns Hopkins University, 1997.

[37] Bahl L R, Jelinek F, Mercer R L. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1983, 5:179–190.

[38] Katz S. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1987, 35(3):400–401.

[39] Ney H, Essen U, Kneser R. On Structuring Probabilistic Dependences in Stochastic Language Modelling. Computer Speech and Language, 1994, 8(1):1–38.

[40] Viterbi A. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Trans. on Information Theory, 1967, 13(2):260–269.

[41] Lee A, Kawahara T, Doshita S. An Efficient Two-Pass Search Algorithm Using Word Trellis Index. Proceedings of ICSLP1998, 1998. 1831-1834.

[42] Jelinek F. Fast Sequential Decoding Algorithm Using a Stack. IBM Journal of Research and Development, 1969, 13(6):675–685.

[43] Bocchieri E. Vector Quantization for the Efficient Computation of Continuous Density Likelihoods. Proceedings of ICASSP1993, 1993. Vol. 2, 692-695.

[44] Ney H, Mergel D, Noll A, et al. A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition. Proceedings of ICASSP1987, 1987. Vol. 12, 833-836.

[45] Ney H, Haeb-Umbach R, Tran B H, et al. Improvements in Beam Search for 10000-Word Continuous Speech Recognition. Proceedings of ICASSP1992, 1992. Vol. 1, 9-12.

[46] Steinbiss V, Tran B H, Ney H. Improvements in Beam Search. Proceedings of ICSLP1994, 1994. 2143-2146.

[47] Schwartz R, Chow Y L. The N-Best Algorithms: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses. Proceedings of ICASSP1990, 1990. Vol. 1, 81-84.

[48] Soong F, Huang E F. A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypotheses in Continuous Speech Recognition. Proceedings of ICASSP1991, 1991. Vol. 1, 705-708.

[49] Ney H, Aubert X. A Word Graph Algorithm for Large Vocabulary, Continuous Speech Recognition . Proceedings of ICSLP1994, 1994. 1355-1358.

[50] Mangu L. Finding Consensus in Speech Recognition[D]. Johns Hopkins University, 2000.

[51] Lo W K, Soong F. Generalized Posterior Probability for Minimum Error Verification of Recognized Sentences. Proceedings of ICASSP2005, 2005. Vol. 1, 85-88.

[52] Qian Y. Use of Tone Information in Cantonese LVCSR Based on Generalized Character Posterior Probability Decoding[D]. The Chinese University of Hong Kong, 2005.

[53] Baum L, Eagon J. An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology. Bulletin of American Mathematical Society, 1967, 73:360–363.

[54] Tokuda K, Yoshimura T, Masuko T, et al. Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. Proceedings of ICASSP2000, 2000. Vol. 3, 1315-1318.

[55] Brown P F, Pietra S D, Pietra V J D, et al. The Mathematic of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 1993, 19(2):263–311.

[56] DeRose S J. Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics, 1988, 14(1):31–39.

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

[58] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of ICML2001, 2001. 282-289.

[59] Gunawardana A, Mahajan M, Acero A, et al. Hidden Conditional Random Fields for Phone Classification. Proceedings of Eurospeech2005, 2005. 1117-1120.

[60] Jensen J L W V. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica, 1906, 30(1):175–193.

[61] Lee K F, Hon H W. Speaker-Independent Phone Recognition Using Hidden Markov Models. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1989, 37(11):1641–1648.

[62] 李小兵. 高效简约的语音识别声学模型[博士学位论文]. 中国科学技术大学, 2006.

[63] Yan Z J, Liu P, Du J, et al. Training Discriminative HMM by Optimal Allocation of Gaussian Kernels. Proceedings of ISCSLP2006, 2006. 289-298.

[64] 鄢志杰, 胡郁, 王仁华. 一种基于区分性准则的模型结构优化方法. 中文信息学报, 2008, 22(2):99–105.

[65] 雷雄国, 鄢志杰, 王智国, 等. 非均匀高斯绑定技术的研究. 第九届全国人机语音通讯学术会议, 2007.

[66] Brown P. The Acoustic-Modeling Problem in Automatic Speech Recognition[D]. Carnegie Mellon University, 1987.

[67] Merialdo B. Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information. Proceedings of ICASSP1988, 1988. Vol. 1, 111-114.

[68] Chow Y L. Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition Using the N-Best Algorithm. Proceedings of ICASSP1990, 1990. Vol. 2, 701-704.

[69] Normandin Y. Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem[D]. McGill University, 1991.

[70] Normandin Y. MMIE Training for Large Vocabulary Continuous Speech Recognition. Proceedings of ICSLP1994, 1994. Vol. 3, 1367-1370.

[71] Valtchev V, Odell J, Woodland P, et al. Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition. Proceedings of ICASSP1996, 1996. Vol. 2, 605-608.

[72] Valtchev V, Odell J J, Woodland P C, et al. MMIE Training of Large Vocabulary Recognition Systems. Speech Communication, 1997, 22(4):303–314.

[73] Bahl L, Padmanabhan M, Nahamoo D, et al. Discriminative Training of Gaussian Mixture Models for Large Vocabulary Speech Recognition Systems. Proceedings of ICASSP1996, 1996. Vol. 2, 613-616.

[74] Povey D, Woodland P. Frame Discrimination Training of HMMs for Large Vocabulary Speech Recognition. Proceedings of ICASSP1999, 1999. Vol. 1, 333-336.

[75] Valtchev V. Discriminative Methods in HMM-Based Speech Recognition[D]. Cambridge University, 1995.

[76] Nádas A, Nahamoo D, Picheny M. On A Model-Robust Training Method for Speech Recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1988, 36(9):1432–1436.

[77] McDermott E, Katagiri S. String-Level MCE for Continuous Phoneme Recognition. Proceedings of EuroSpeech1997, 1997. 123-126.

[78] Chou W, Lee C H, Juang B H. Minimum Error Rate Training Based on N-Best String Models. Proceedings of ICASSP1993, 1993. Vol. 2, 652-655.

[79] Chou W, Lee C H, Juang B H. Minimum Error Rate Training of Inter-Word Context Dependent Acoustic Model Units in Speech Recognition. Proceedings of ICSLP1994, 1994. Vol. 2, 439-442.

[80] Saul L, Rahim M. Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition. IEEE Trans. on Speech and Audio Processing, 2000, 8(2):115–125.

[81] Schlüter R, Macherey W, Müller B, et al. Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition. Speech Communication, 2001, 34:287–310.

[82] Korkmazskiy F, Juang B H. Discriminative Adaptation for Speaker Verification. Proceedings of ICSLP1996, 1996. Vol. 3, 1744-1747.

[83] Sukkar R, Lee C H. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition. IEEE Trans. on Speech and Audio Processing, 1996, 4(6):420–429.

[84] Fu Q, He X, Deng L. Phone-Discriminating Minimum Classification Error (P-MCE) Training for Phonetic Recognition. Proceedings of InterSpeech2007, 2007. 2073-2076.

[85] Yu D, Deng L, He X, et al. Use of Incrementally Regulated Discriminative Margins in MCE Training for Speech Recognition. Proceedings of InterSpeech2006, 2006. 2418-2421.

[86] Hain T, Woodland P, Evermann G, et al. Automatic Transcription of Conversational Telephone Speech. IEEE Trans. on Speech and Audio Processing, 2005, 13(6):1173–1185.

[87] Chen S, Kingsbury B, Mangu L, et al. Advances in Speech Transcription at IBM under the DARPA EARS Program. IEEE Trans. on Audio, Speech and Language Processing, 2006, 14(5):1596–1608.

[88] Du J, Liu P, Soong F, et al. Minimum Divergence Based Discriminative Training. Proceedings of ICSLP2006, 2006. 2410-2413.

[89] Du J, Liu P, Jiang H, et al. A New Minimum Divergence Approach to Discriminative Training. Proceedings of ICASSP2007, 2007. Vol. 4, 677-680.

[90] Heigold G, Macherey W, Schlüter R, et al. Minimum Exact Word Error Training. Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2005. 186-190.

[91] Povey D, Kingsbury B. Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training. Proceedings of ICASSP2007, 2007. Vol. 4, 321-324.

[92] Povey D. Improvements to fMPE for Discriminative Training of Features. Proceedings of Eurospeech2005, 2005. 2977-2980.

[93] Povey D, Gales M, Kim D, et al. MMI-MAP and MPE-MAP for Acoustic Model Adaptation. Proceedings of Eurospeech2003, 2003. 1981-1984.

[94] Povey D. Discriminative Training for Large Vocabulary Speech Recognition[D]. Cambridge University, 2004.

[95] Kullback S, Leibler R A. On Information and Sufficiency. The Annals of Mathematical Statistics, 1951, 22(1):79–86.

[96] Li X, Jiang H, Liu C. Large Margin HMMs for Speech Recognition. Proceedings of ICASSP2005, 2005. Vol. 5, 513-516.

[97] Sha F, Saul L K. Large Margin Hidden Markov Models for Automatic Speech Recognition. In: Schölkopf B, Platt J, Hoffman T, (eds.). Proceedings of Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press, 2007: 1249–1256.

[98] Li J, Yuan M, Lee C H. Soft Margin Estimation of Hidden Markov Model Parameters. Proceedings of InterSpeech2006, 2006. 2422-2425.

[99] Li J, Siniscalchi S, Lee C H. Approximate Test Risk Minimization Through Soft Margin Estimation. Proceedings of ICASSP2007, 2007. Vol. 4, 653-656.

[100] Li J, Yan Z J, Lee C H, et al. A Study on Soft Margin Estimation for LVCSR. Proceedings of ASRU2007, 2007. 268-271.

[101] Schlüter R, Macherey W. Comparison of Discriminative Training Criteria. Proceedings of ICASSP1998, 1998. Vol. 1, 493-496.

[102] Chou W, Juang B, Lee C. Segmental GPD Training of HMM Based Speech Recognizer. Proceedings of ICASSP1992, 1992. Vol. 1, 473-476.

[103] Katagiri S, Juang B H, Lee C H. Pattern Recognition Using a Family of Design Algorithms Based Upon the Generalized Probabilistic Descent Method. Proceedings of the IEEE, 1998, 86(11):2345–2373.

[104] McDermott E. Discriminative Training for Speech Recognition[D]. Waseda University, 1997.

[105] Gopalakrishnan P, Kanevsky D, Nádas A, et al. An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems. IEEE Trans. on Information Theory, 1991, 37(1):107–113.

[106] Kanevsky D. A Generalization of the Baum Algorithm to Functions on Non-Linear Manifolds. Proceedings of ICASSP1995, 1995. Vol. 1, 473-476.

[107] Cardin R, Normandin Y, De Mori R. High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation. Proceedings of ICASSP1991, 1991. Vol. 1, 533-536.

[108] Kapadia S, Valtchev V, Young S. MMI Training for Continuous Phoneme Recognition on the TIMIT Database. Proceedings of ICASSP1993, 1993. Vol. 2, 491-494.

[109] Afify M. Extended Baum-Welch Reestimation of Gaussian Mixture Models Based on Reverse Jensen Inequality. Proceedings of Eurospeech2005, 2005. 1113-1116.

[110] McDermott E, Katagiri S. Minimum Error Training for Speech Recognition. Proceedings of IEEE Workshop on Neural Networks for Signal Processing, 1994. 259-268.

[111] Yan Z J, Zhu B, Hu Y, et al. Minimum Word Classification Error Training of HMMs for Automatic Speech Recognition. Proceedings of ICASSP2008, 2008. 4521-4524.

[112] Young S. The General Use of Tying in Phoneme-Based HMM Speech Recognisers. Proceedings of ICASSP1992, 1992. Vol. 1, 569-572.

[113] Vertanen K. HTK Wall Street Journal (WSJ) Training Recipe. http://www.inference.phy.cam.ac.uk/kv227/htk.

[114] Woodland P, Odell J, Valtchev V, et al. Large Vocabulary Continuous Speech Recognition Using HTK. Proceedings of ICASSP1994, 1994. Vol. 2, 125-128.

[115] Nocedal J, Wright S J. Numerical Optimization. 2nd ed., Springer, 2006.

[116] Vapnik V N. The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[117] Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. Proceedings of 2nd International Symposium on Information Theory, 1973. 267-281.

[118] Schwarz G. Estimating the Dimension of a Model. Annals of Statistics, 1978, 6(2):461–464.

[119] Rissanen J. Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Company, 1989.

[120] Chen S, Eide E, Gales M, et al. Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News. Proceedings of ICASSP1999, 1999. 37-40.

[121] Normandin Y. Optimal Splitting of HMM Gaussian Mixture Components with MMIE Training. Proceedings of ICASSP1995, 1995. 449-452.

# 致　谢

本文的完成首先要感谢我的导师王仁华教授。王老师身上有着老一辈知识分子特有的风范，对学生关怀备至，又严格要求。王老师对语音领域的战略眼光，对科学研究的严谨态度，对日常工作的勤奋精神，都深深的影响着我和每一位同学。而王老师在学术上“走出去、请进来”的开放态度更是让我们能够从更多的角度观察和思考研究中的问题。师从王老师是我的幸运，更是我的荣耀和骄傲。在过去的五年里，王老师对我的培养已经让我受益匪浅。我相信王老师的教诲也必将在我今后的人生道路上给我更多的帮助。

感谢微软亚洲研究院的宋謌平教授。在博士学位研究期间，我作为访问学生有幸到微软亚洲研究院语音组学习，并得到了宋老师诸多悉心的指导。宋老师渊博的知识、认真的态度，以及对学生的关爱、对工作的热情，都给我留下了非常深刻的印象。特别是宋老师对研究问题深入浅出的分析和一针见血的见解，使我从做研究的方法到具体的研究问题上都受益良多。

感谢实验室的戴礼荣教授为我的课程学习和研究工作提出了很多中肯而宝贵的意见和建议。感谢刘必成老师为我及其他同学的工作提供了良好的硬件保障。感谢实验室已经毕业的李小兵博士、刘波硕士、郭罡硕士、范明硕士，以及仍在实验室工作的其他师兄弟：胡郁、凌震华、魏思、杜俊、刘聪、竺博等等。我们一起在实验室形成了良好的研究氛围，并互相学习、互相帮助，本文中的很多想法也都是在与他们的共同讨论和工作中逐渐产生的。

感谢加拿大 York University 的江辉教授，本文第 5 章中的研究工作是江教授到讯飞语音实验室访问期间，在他的指导下完成的。感谢美国 Georgia Institute of Technology 的李锦辉教授和李锦宇师兄，本文第 6 章中的研究工作是我赴美访问学习期间在李教授的指导下与锦宇师兄合作完成的。上述三位都是国际上非常活跃的语音领域的专家和研究者，与他们的合作极大的开阔了我在研究上的视野。

最后，要特别感谢我的父母和家人。父母抚养我成人，并教会我很多做人的道理，我顺利的完成学业与他们的辛勤付出是分不开的。由衷的感谢我的未婚妻汪磊小姐，她一直在我身边陪伴和支持着我，我们之间相互的关心和鼓励让生活总是充满着希望。

# 个人简历及在读期间发表的学术论文

鄢志杰，男，1982年3月出生于重庆市。主要研究方向为语音识别，包括语音识别前端鲁棒性方法、语音增强、语音识别声学模型建模、声学模型区分性训练，以及语音识别应用系统等。

## 主要学习经历

| 时　间 | 学　业 | 就读院系 |
|---|---|---|
| 1999 ~ 2003 | 本　科 | 中国科学技术大学电子工程与信息科学系 |
| 2003 ~ 2005 | 硕士研究生 | 中国科学技术大学电子工程与信息科学系 |
| 2005 ~ 2008 | 博士研究生 | 中国科学技术大学电子工程与信息科学系 |

## 主要研究经历

| 时　间 | 研究机构、身份及研究内容 |
|---|---|
| 2003 ~ 2005 | 中国科学技术大学讯飞语音实验室　语音识别组成员 |
| | 研究噪声鲁棒性语音识别前端方法、说话人验证、采用多特征(基频段长等)融合的语音识别等内容。 |
| 2005 ~ 2006 | 微软亚洲研究院　语音组访问学生 |
| | 研究噪声语音增强及噪声鲁棒性语音识别、区分性训练及区分性准则下的模型拓扑结构优化。提出综合利用信号处理及模型补偿的、基于词图的特征增强方法。 |
| 2006 ~ 2008 | 中国科学技术大学讯飞语音实验室　语音识别组组长 |
| | 研究声学模型区分性训练准则及参数优化方法。参与组织、设计及搭建完成面向实用的电话语音识别研究原型系统。 |
| 2007 | 美国Georgia Institute of Technology　访问学者 |
| | 研究区分性训练中Soft Margin Estimation方法在大词汇量连续语音识别中的准则尺度细化及其应用。 |

## 近年所获奖项

| 时　间 | 所获奖项 |
|---|---|
| 2007 | ICASSP 2007 Student Paper Contest Winner (优秀学生论文奖) |
| 2006 | Microsoft Fellowship 2006 (微软学者奖) |

# 在读期间发表的学术论文

[1] **Yan Z.-J.**, Zhu B., Hu Y. Wang R.-H., Minimum Word Classification Error Training of HMMs for Automatic Speech Recognition, Proceedings of ICASSP2008, 2008. 4521-4524.

[2] **鄢志杰**，胡郁，王仁华，一种基于区分性准则的模型结构优化方法，中文信息学报，2008，22(2):99-105.

[3] **Yan Z.-J.**, Soong F. K., Wang R.-H., Word Graph Based Feature Enhancement for Noisy Speech Recognition, Proceedings of ICASSP2007, 2007. Vol. 4, 373-376.

[4] Li J., **Yan Z.-J.**, Lee C.-H., Wang R.-H., A Study on Soft Margin Estimation for LVCSR, Proceedings of ASRU2007, 2007. 268-271.

[5] 竺博，**鄢志杰**，胡郁，王仁华，区分性参数重分配在HMM模型结构优化中的应用，第九届全国人机语音通讯学术会议，2007.

[6] 雷雄国，**鄢志杰**，王智国，吴及，非均匀高斯绑定技术的研究，第九届全国人机语音通讯学术会议，2007.

[7] **Yan Z.-J.**, Zhou J.-L., Soong F. K., Wang R.-H., Signal Trajectory Based Noise Compensation for Robust Speech Recognition, Lecture Notes in Computer Science: Chinese Spoken Language Processing, Springer, 2006, 4274:335-345.

[8] **Yan Z.-J.**, Liu P., Du J., Soong F. K., Wang R.-H., Training Discriminative HMM by Optimal Allocation of Gaussian Kernels, Proceedings of ISCSLP2006, 2006. 289-298.

[9] Liu C., **Yan Z.-J.**, Hu Y., Wang R.-H., A Comparative Study on Confidence Measure in Mandarin Command Word Recognition, Proceedings of ISCSLP2006, 2006. 355-366.