摘　要

尽管非特定人的语音识别系统已经达到了令人鼓舞的性能，但是在实际应用时由于说话人和环境的改变通常会使得系统性能显著下降。当遇到特殊口音的说话人，或者环境有一定的噪音时，系统的误识率甚至有可能增加原来的 5 倍。语音识别要走向实用，就必须克服这个鲁棒性问题，因此语音自适应技术的意义非常重要。

本文从说话人自适应技术入手讨论了语音自适应的各种方法。通过对说话人引起的声学差异的讨论，我们分析和实现了两种常用的说话人自适应方法：最大后验概率（MAP）方法和最大似然线性回归（MLLR）方法。实验证明这两种自适应方法对说话人自适应和环境自适应都有良好效果。

在此基础上，本文提出一种适合于强健语音识别的快速综合渐进自适应方法。通过在渐进的 MAP 方法中引入一个简化的 MLLR 模块，这种新方法成功地结合了 MAP 和 MLLR 两种方法的优点：MLLR 方法比较简单，而且在自适应数据很少时自适应速度比较快；而 MAP 方法给出了结合先验知识和自适应数据的最优解，有良好的渐进性。在新方法中，简化的 MLLR 模块使用了一个全局的转移矩阵，用来对付环境和说话人生理引起的差异，为 MAP 模块提供了更加精确的初始模型。而渐进的 MAP 模块则主要用来精细的刻画基于音素层次的差异，同时也确保了整个方法的渐进性。另外，针对综合渐进自适应方法中 MAP 和 MLLR 模块的特点，我们同时采用了一种新的渐进使用自适应数据的策略。在论文最后的实验中，这种综合方法即使在自适应数据比较少的情况下也可以取得好的效果。在无噪音和有噪音的环境中分别可以降低 23.03%和 29.69%的识别字错误率。实验证明，这种新方法能够有效的克服说话人差异和环境差异对识别系统的影响，适合强健语音识别系统的要求。

关键词：语音识别，说话人自适应，环境自适应，MAP，MLLR

ABSTRACT

Speaker independent speech recognition systems have achieved great progress in recent years. However the recognition performance degrades rapidly when there is a mismatch between the testing and the training conditions, e.g. an outlier speaker, the presence of low-level background noise, and so on. Therefore, adaptive techniques are critically important to overcome these mismatches and to propel speech recognition to practical applications.

This paper discussed various algorithms of adaptive techniques, especially in the field of speaker adaptation. By analyzing the acoustical variations between speakers, the paper discussed and implemented two classical speaker adaptation methods: *Maximum a Posteriori* (MAP) and *Maximum Likelihood Linear Regression* (MLLR). At the end of the paper, experimental results show that these two methods work well in both speaker adaptation and environment adaptation.

Then, a new approach for rapid, incremental adaptation is presented in this paper. While MLLR method can achieve a fast adaptation rate when only few data is available, MAP has desirable asymptotic properties. By introducing a simplified MLLR module to the incremental MAP processing, the new approach integrates these two methods to make use of their advantages and offset their disadvantages. In the new approach, the simplified MLLR module uses a single globe regression class to minimize the mismatches caused by the environment differences and speaker anatomical differences, and provides a more accurate initial model to the MAP processing. The incremental MAP module is used for a further subtle removal of phoneme-level variations, and to insure the asymptotic properties of the whole approach. Based on the characteristics of MAP and MLLR, a new incremental adaptation scheme of using the adaptation data is also adopted. In our experiments, the new approach performances well even when adaptation is conducted with only a few short utterances. The new one improves the *Word Error Rate* (WER) by 23.03% in a quiet environment and by 29.69% in a noisy environment respectively. These results demonstrate that this new approach can effectively deal with the speaker variations and environment variations, and is well suited for the robust speech recognition.

Keywords: speech recognition, speaker adaptation, environment adaptation, MAP, MLLR

目　录

第一章 引 言

处于信息革命浪潮时代的今天，人们对于各种各样信息的需求与日俱增，从而人们急切需要更好的信息处理方式。语音，作为人类信息交流的最自然、最有效、最灵活而又最为广泛使用的途径，越来越引起研究者的关注。

1.1 语音识别

语音识别（Speech Recognition）是指采用计算机从人的语音信号中自动提取最有意义的信息，从而确定语音信号的语言含义的过程。作为一个科学研究领域，它与声学、语音学、语言学、脑科学、生理学、心理学、人工智能、数字信号处理理论、模式识别理论、统计信息理论、最优化理论、计算机科学等众多学科紧密相连。

1.1.1 语音识别的意义

随着人们对语音识别认识的深入，人们对语音识别也提出了越来越高的目标。语音识别的最终目的就是象人与人之间谈话交流信息一样，实现人－机自由对话，也就是赋予机器以听觉，使机器能听懂人的语言，辨明话音的内容或说话人，将人的语音正确地转化为书面语言或有意义的符号，或者进一步使机器能够按照人的意志进行操作，把人类从繁重或危险的劳动中解脱出来。据预测,语音识别将成为继键盘和鼠标器之后,人机交互界面革命中的下一次飞跃。正如IDC的PC分析员Richard Zwetchkenbaum所说:“语言是最自然的界面”。

语音识别具有很大的实际应用价值，其发展、成熟和实用化将推动许多产业的迅速发展，其中包括计算机、办公室自动化、通信、国防、机器人等等。目前可以想象的语音识别主要应用有：语音输入系统，作为一种最自然的文字

输入方法，用口述代替键盘向计算机输入文字，这将给办公室自动化和出版界带来革命性的变化；语音控制系统，为人们在手动控制以外又提供了一种更安全、更方便的控制方法，特别是当系统工作在一些特定的环境（如黑暗场所或手脚已被占用来进行其它动作的环境）或一些特殊的用户（如残疾人）时；基于对话系统的数据库查询系统，为用户提供了更为自然、友好和便捷的数据库检索或查询，可以广泛运用在银行、交易所、民航等机构；除此之外，语音识别还可以用于口语翻译系统、计算机辅助教学、自动身份确认等很多领域。

1.1.2 语音识别的历史与现状

自动语音识别（ASR, Automatic Speech Recognition）研究开始于五十年代初。当时电子信号频谱分析仪器开始被用于从语音信号中识别简单、少量的音节和音素。其中有代表性的是 1952 年美国 Bell Laboratories 研制的 Audry 系统[1]和 1956 年 RCA Laboratories 的单音节词识别系统[2]。

六十年代，数字计算机的迅速发展使人们对语音信号的研究由对模拟信号的分析转向数字技术。在这一时期 Fant[3, 4]和 Flanagan[5]对语音产生的研究使人们对语音产生的机理有了一个较系统的了解。人们还对人类听觉的生理和心理进行了研究，发现了人耳对声音中的不同频率成分有不同的分辨力的反应力，提出了临界频带理论。这一时期，在语音识别的算法方面尚未找到合适计算机分析的模型和算法。但人们研究了分段（Segmentation）、分类（Classification）和模式匹配（Pattern Matching）等问题。与此同时，自然语言领域的一些基础性研究也在进行。六十年代多方面的基础性研究为七十年代语音识别的迅速发展打下了基础。

七十年代，语音识别无论在理论上，还是在系统实现上，都有了迅速的发展。1975 年 Itacura[6]发现基于线性预测编码（LPC, Linear Predictive Coding）的谱系数是识别器很好的特征，不但识别效果大有提高，计算复杂度也比较小。同一时期，六十年代 Vintsyuk 所提出的动态时间规整（DTW, Dynamic Time Warping）算法[7]也成功的应用于语音识别中。从此基于 LPC 分析和 DTW 算法的识别系统纷纷建立起来。七十年代另一个重大的里程碑，就是 CMU 的 Baker[8]和 IBM 的 Jelinek[9]意识到可以将马尔可夫模型（HMM, Hidden Markov Model）

应用于语音识别。七十年代出现了许多成功的孤立词识别系统，如：CMU 的 Hearsay-II[10]、IBM 的大词汇量自动语音听写系统[11]、Bell Labs 用于通迅的与话者无关的语音识别系统[12]。

到了八十年代，语音识别技术有了新的综合性的发展。矢量量化（VQ, Vector Quantization）[13]和隐马尔可夫模型（HMM, Hidden Markov Models）[14, 15]在语音识别中获得了广泛的应用，从而产生了象 CMU 的 SPHINX[16, 17]这样的成功的非特定人连续语音识别系统。另外，八十年代人工神经网络的研究热潮也波及语音领域，出现了基于人工神经网络（ANN, Artificial Neural Networks）[18]或者人工神经网络和隐马尔可夫模型的混合模型[19, 20, 21]的识别系统。

进入九十年代，随着信号处理、声学模型、语言模型、解码搜索算法等理论日益成熟，计算机软硬件系统性能不断提高，出现了一些大词汇量连续语音识别系统，如 IBM 的 ViaVoice[22]，Microsoft 的 Whisper[23]，CMU 的 SPHINX-II[24]等等。这些系统大体上采用了相似的技术，不仅有基于隐马尔可夫模型的声学模型，而且包含了较复杂的语言模型以及先进的解码算法。有的系统还加入了自然语言理解部分，使系统性能进一步提高。

目前已有不少语音识别系统进入实用化阶段，走上了市场，这里列出最近《个人电脑》杂志报道的世界主要语音识别软件的评比结果[25]。这个评比结果表明现在实用的语音识别系统已经发展到了非特定人、超大规模词汇量和连续语音识别阶段，并具有大约 93%的初始识别正确率。

表 1-1 Dragon NaturallySpeaking、FreeSpeech 2000、L&H Voice Xpress Professional 和 ViaVoice Pro Millennium Edition 四种语音识别系统的性能比较

	Dragon Naturally-Speaking	FreeSpeech 2000	L&H Voice Xpress Professional	ViaVoice Pro Millennium Edition
初始识别率	95%	91%	93%	95%
是否支持多用户	支持	支持	支持	支持
基本活动词汇表	160,000	60,000	34,000	64,000
最大活动词汇表	250,000	670,000	64,000	2,000,000

1.1.3 语音识别系统的框架

虽然目前的实用的语音识别系统使用各种不同的模型和解码方法，但图 1-1 给出了语音识别系统的一般性的框架结构。语音信号通过信号处理模块生成识别器使用的一系列特征向量；识别器再利用语言模型和声学模型得到对应输入特征向量有最大概率的词序列；同时提供给自适应模块有用的信息用来对语言模型和声学模型进行修改。

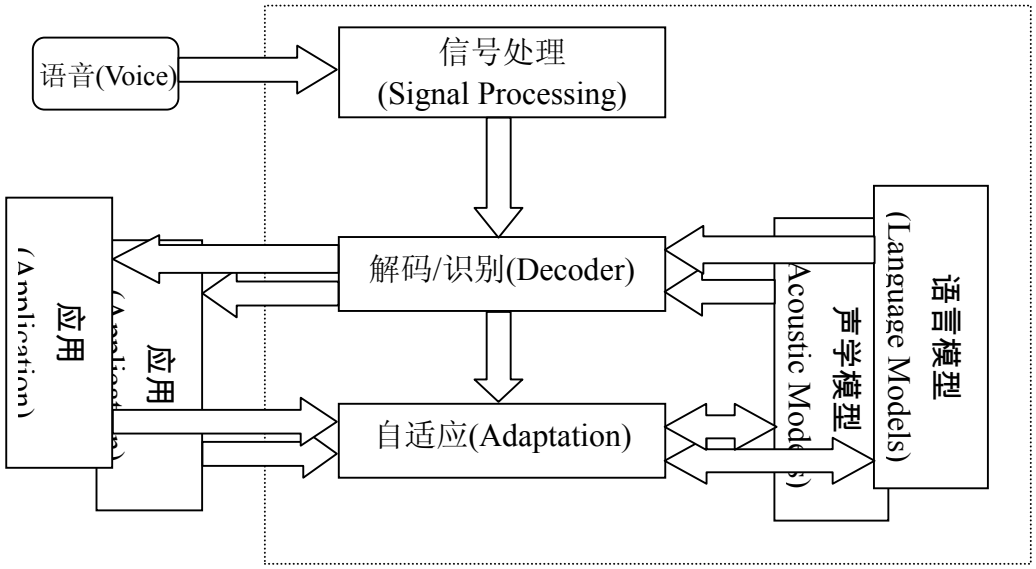


图 1-1 语音识别系统的框架

1.2 语音自适应技术

如图1-1中所示，目前的大多数使用语音识别系统中都包含了一个非常重要的模块：自适应模块。它的作用主要是用各种自适应技术来调整声学模型和语言模型，使系统适应新的应用状况。虽然一个训练好的系统可以适应很多不同的情况，但模型和实际操作状况间总存在一定的差异。所以使语音识别系统可以通过少量的矫正数据尽量减小这种差异是十分重要的。自适应技术就是这样

一种技术，它对系统参数进行调整，从而使系统更好的匹配由于麦克风、传输通道、环境噪音、说话人、文体和应用的上下文等引起的差异。

1.2.1 说话人自适应

目前语音识别技术在中小词汇量的非特定人（Speaker Independent, SI）识别系统中已经可以达到很高的识别正确率。比如基于RM1[1]的非特定人测试数据集平均的词错误率达到了3%左右[26]。尽管平均错误率很低，但有一些说话人的错误明显高与其他人。由于使用很广泛的说话人来训练非特定人的系统模型，使得说话人之间的差异被当作说话人内部的差异处理了。这样使得每一个声学模型包含了大量的差异，有可能降低对于单个的说话人的精确建模。这一点可以由对同一个说话人的语音对比非特定人系统和特定人系统（Speaker Dependent, SD）证实。假设有充足的数据训练这两个系统，那么特定人系统性能要比非特定人系统好2到3倍。如文献[26]给出的比较结果，见表1-2。

表 1-2　非特定人与特定人系统性能比较：基于同一说话人的语音（RM1 测试集中），比较非特定人和特定人系统。其中 SI 由 3990 个 SI 句子训练得到，SD 在 SI 基础上再训练 600 个 SD 句子。

说话人	SI错误率(%)	SD错误率(%)
bef0_3	3.2	2.3
cmr0_2	7.4	1.6
das1_2	1.8	0.9
dms1_2	3.2	1.0
dtb0_4	3.3	1.2
dtd0_5	4.6	2.3
ers0_7	3.5	2.6
hxs0_6	6.5	1.5
jws0_4	4.5	1.8
pgh0_1	2.6	2.1
rkm0_5	8.3	2.6
tab0_7	2.2	1.8
平均	4.3	1.8

[1] The ARPA Resource Management database (RM1) [94]

特定人系统需针对一个单独说话者进行训练，一般而言所需的语音数据量至少应达600句话（大词汇系统，词汇量在5000以上）[27]。输入如此大量的语音数据对于每个使用者是一项沉重的负担，而且处理这些数据所需的CPU时间也要若干小时，这都使特定人系统的实用性受到很大限制。

为了解决这个问题，开始引入说话人自适应技术，即在一个已经训练好的初始模型系统上，用一定的新说话人的语音数据（我们叫做自适应数据，Adaptation Data）来试图提高模型系统对这个说话人的建模精度。另一种理解可以是对一个针对老说话人充分训练的特定人系统和一个用少量新说话人的训练语音数据训练的非特定人系统的结合，从而使系统的识别率接近于对新说话人经过充分训练的特定人系统的水平。

除了解决非特定人系统存在的问题之外，说话人自适应技术也可以用来增强识别系统对环境的自适应能力，特别是提高对环境噪音或麦克风差异的适应能力。

本文研究工作主要集中在说话人自适应技术的研究上。

1.2.2 其他自适应技术

除了说话人自适应(对说话人的声音特点的自适应能力)外，自适应技术还包括下面几个方面：

对环境的自适应能力，特别是提高对环境噪音或麦克风差异的适应能力。对环境噪音的适应可以有两种简单的办法：一种是去掉输入语音中的噪音，使得语音变得纯净。此时，系统对噪音的自适应能力就体现在如何根据不同的环境噪音，采取不同的去噪方法，尽量减少噪音对后续操作的影响。这需要对各种噪音进行分析，以制定相应的处理方法。还有一种是直接用含有噪音的语音来训练模型，使得噪音成为模型的一个固有部分。此时，系统对噪音的自适应能力体现在噪音模型如何反映测试使用时的噪音环境。当环境噪音与训练噪音不一致时，系统必须对含噪模型进行调整，从而排除掉噪音对系统识别性能的影响。

对说话人的语言特点的自适应能力。对语言特点的自适应，主要是对输入的文体格式和语体格式的自适应能力。系统应该根据输入语音流的特点进行一定的调整，使得系统的各种参数对特定的语音更具针对性。在语体格式上，主要有口语体和书面体两种。口语语言存在着大量的省略、迟疑、停顿、临时插入、重复强调、自我纠错以及非法语法结构和无意义语音等现象[28]。

对说话人的语种特点的自适应能力。这一点是未来多语种复合系统的要求，即未来的语音识别系统可以自动识别多种语种。

1.2.3 国内外发展动态

语音信号处理自适应技术的研究是伴随着语音识别技术的飞速发展而产生并发展起来的。目前自适应技术已经成为了语音识别技术中的一个不可缺少的重要部分，并且开始应用在大多数实用语音产品和研究平台中，如 IBM 的 ViaVioce[?]，Microsoft 公司的 Whisper[?]。

从整个语音识别研究的发展前景上看，语音识别系统的顽健性（Robust）将是未来几年的研究重点之一[29，30]。因为这是语音识别系统由实验转为实用过程中的一个最为迫切最关键的问题。而说话人自适应技术是其中不容忽视的一个重点和难点。这项技术已经引起了越来越多的研究机构的关注和兴趣，几乎所有从事语音识别方面研究的科研与企事业单位都开始投入专门的精力从事自适应技术的研究。各种语音研究的国际学术会议（如 ICASSP，EuroSpeech，ICSLP 等）也开始把说话人自适应作为单项专题进行讨论。

目前，国际上说话人自适应的主要方法可以大致分为下面几种：

说话人正规化（Speaker Normalization）[37-42，67]，其目的是建立一个正规化的说话人空间，使得任何人的语音都可以映射其中。这样可以把说话人间的差异降到最低。正规化的方法很多，其中使用比较多的有：声道长度归一（VTN）和倒谱均值规正（CMN）。

说话人聚类（Speaker Clustering）[17，43，45，65]，通过一定的聚类或者分类算法，对不同说话人的模型进行聚类分组。识别时直接选取与目标说话人最接近的模型组进行识别。这种方法是十分简单有效的方法，被许多系统广泛使用。

谱变换（Spectral Transformation）[47-53，26，91，92]，是通过使用线性或非线性的变换把一个说话人的语音谱空间映射到另一个人的谱空间上，从而实现自适应。需要指出的是，这种变换即可以适用于特征空间也可以在 VQ 码本或 HMM 参数上进行。

参数调整（Model Parameter Adaptation）[54-63，36，79，82，85]，是把原有 SI 系统的 HMM 参数作为先验知识（a Prior information），遵照 Bayes 估计准则求出达到最大后验概率（Maximum a Posterior，MAP）时系统采用的最佳参数。

各种方法的详细介绍和讲解请参见本论文的第二章。

我国的语音识别研究起步比较晚，但由于汉语语音识别的重要性日益突出，最近十年的发展十分迅速。所以相应从九十年代开始的说话人自适应的研究我国基本可以和国外同步。目前国内从事这方面研究的机构主要有：清华大学、中国科学院声学研究所、中国科学院自动化研究所、香港大学、中国科学技术大学、国防科技大学、北京邮电大学等等[31，36-38，40-42，57-61，69，73，74，82]。

1.3 论文组成

论文的工作是进行语音识别中的自适应技术的研究，主要内容是说话人自适应技术的实现及研究。

1.3.1 论文工作

主要做了如下的工作：(1) 实现基于最大后验概率（MAP）方法的说话人自适应。(2) 实现基于最大似然线性回归（MLLR）方法的说话人自适应。(3) 提出一个综合的渐进自适应方法。(4) 使用说话人自适应技术对环境和噪音进行自适应。

1.3.2 论文安排

本文内容安排如下：

第一章 概括描述语音识别、语音自适应、以及本文主要研究工作；

第二章 分析了造成说话人差异的原因，概要介绍说话人自适应的基本概念和原理，详细介绍了各种常见的说话人自适应方法;

第三章 详细给出了基于最大后验概率（MAP）和基于最大似然线性回归（MLLR）的两种自适应方法的原理和实现方法;

第四章 详细描述了我们提出的综合渐进自适应方法;

第五章 给出了实验和结果的分析；

第六章 对本文进行总结。

第二章 说话人自适应技术

本章详细给出了说话人自适应技术的产生背景、基本原理、分类、以及一些常见的方法。

2.1 特定人系统与非特定人系统

目前语音识别系统如果从对说话人的依赖程度上分类，可以分为特定人系统（Speaker Dependent，SD）和非特定人系统（Speaker Independent，SI）。

顾名思义，特定人的语音识别系统只适用于某个特定的用户，并要求该使用者预先提供足够多的个人语音数据以训练系统。这种特定性使系统不包含模糊的“平均”或“标准”信息，因而具有语言无关性，无论口音如何，只要使用者能在训练及识别过程中始终保持一致就可得到良好的识别效果。现有特定人系统的识别率已普遍达到 95％以上[31]。然而单用户使用的局限性大大束缚了特定人系统的进一步推广与应用，一旦有新的用户加人，系统则要求进行重新训练，否则识别率将急剧下降。一般而言，训练所需要的语音数据量应达到几百句话（大词汇量），以每句话占用 2－3 秒计，录入训练语音将占用 20 分钟以上。而几乎在所有情况下，输人如此大量的语音数据会令每个使用者望而生畏；处理这些数据所需的 CPU 时间也无异于雪上加霜，我们就更无须论及是否可以提供足够好的客观环境与时间来进行训练了。

然而在现实生活中，有许多情况要求频繁地替换使用人员，例如办公环境下的口述录音，数据库（交通报告，航空时间表等）的信息检索。此时，非特定人的语音识别系统表现出巨大优势。这种非特定人系统能够在免除对每一用户进行大量训练的前提下，为相当广泛的用户群提供良好的识别效果。理想情况就是对任一说话人，无论口音、谈话风格如何，都能给予满意的识别效果。

这显然与人们的最初设想相吻合。遗憾的是正如第一章中表 1-2 所示，现有非特定人系统的识别精度还无法满足实际使用的要求，其错误率可相当于对应的特定人系统的两至三倍，在某些情况下甚至能高达 5 倍。此外，即使是一个工作良好的非特定人系统在遇到特殊的说话人（与一般人说话有明显差别，我们称为 outlier）时，识别率也会显著下降[32]。

特定人系统和非特定人系统性能的差距的原因是很明显的。非特定人系统使用很广泛的说话人语音来训练识别系统的模型，虽然能够保证有足够的数据来精确刻画语音单元的各种复杂的时变特性、协同发音等，同时却使得说话人之间的差异被忽略，从而降低了系统对于单个的说话人建模的精度。下面我们具体分析一下说话人差异的产生原因和分类。

2.2 说话人差异（Speaker Differences）

影响识别系统的识别效果的因素有很多，不过他们可以一般划分成两类：说话人之间的（Inter-Speaker）和说话人内部的（Intra-Speaker）。

2.2.1 说话人之间的差异（Inter-Speaker Differences）

每一个人的说话都有自己的特点。当一个人说话时，他所发出的语音受到很多因素的影响，比如：他的声道的长度、宽度和物理形状，年龄，性别，健康状况，文化程度，个人的发音习惯等。这些差异使得一个人的语音可能和另一个人完全不一样。这一点我们可以从图 2-1 中看得很清楚。说话人之间的差异主要两个方面：生理差异和说话习惯差异[33]。

生理差异主要是缘于每个人的发声器官的形状、大小和动态特性都不同。这种生理差异对语音的基频有显著的影响，使得不同人产生不同的声学特征（这也是造成男女之间差异的主要因素）。这种情况的一个极端的例子是说话人性别对语音频谱参数的影响。如果建立一个便于分析的语音产生模型，可以发现，基音频率 f 值取决于声带的尺寸和特性，以及声带所受的张力。一般而言，男性说话者的 f 值大致分布在 60～200HZ 范围内，女性说话者和小孩的 f 值在

200～450HZ 之间[?]。目前研究表明[34]，男性和女性在发元音时有着明显不同的共振峰频率，男性发的元音基频更低，共振峰带宽更窄，并且频谱更平缓。这也是为什么用男性语音训练的特定人系统在女性测试或双性测试时有很差效果的原因。

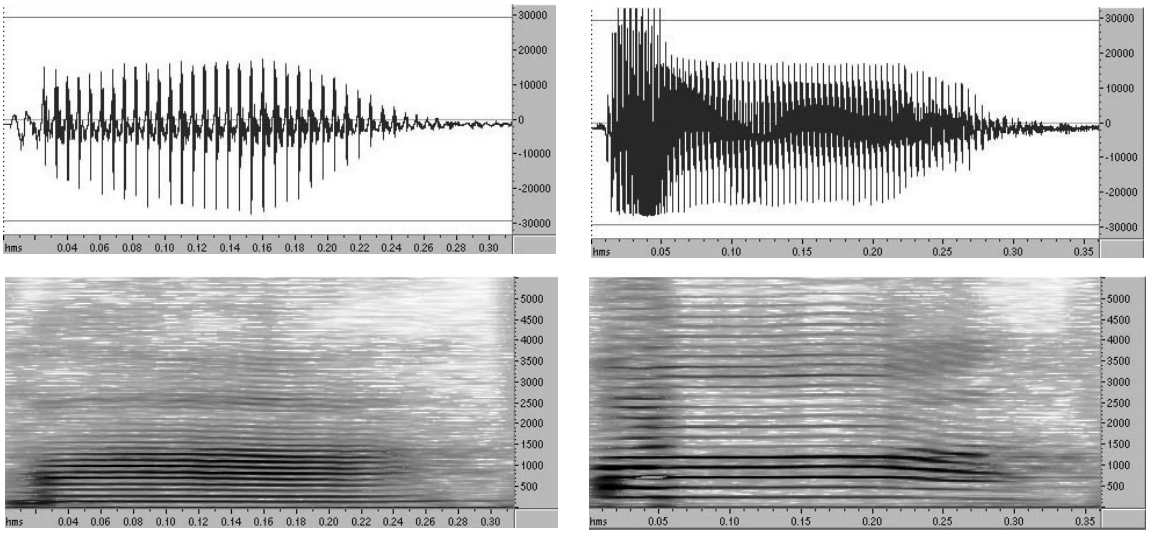


图 2-1　两个不同说话人发数字“8”语音的时频波形图和语谱图。我们可以清楚看出不同说话人之间的差异。

说话习惯则和说话人学习说话的过程有关，这种习惯直接影响发声的清晰效果和共振峰传输率的不同。具体的比如说话人的语速和口音，这两点即使在人的听力理解中也十分重要。各人不同的说话习惯，包括每个人的教育和文化背景的不同，所用方言的不同，所属的社会层次和集团不同以及个人的经历、气质的差异。国外许多专家专门针对影响发音的各种社会因素，包括地域环境，宗教信仰，文化背景等作了深入的研究，并著书阐述它们对各种口音的影响[35]。F.Nolan 在文献[?]中指出，音节之间的协同发音模型也会随着口音的改变而变化。文献[?]中指出，事实上口音的影响大约可以使得识别系统的错误率增加 2~3 倍。

由于发声的原理是十分复杂的，所以这种说话人之间的差异用简单的分类方法来解决是很困难的。

2.2.2 说话人内部的差异（Intra-Speaker Differences）

即使我们忽略说话人之间的差异，对于同一个说话人，在不同的时间、不同的心理和生理状态下，讲述同一内容的话语也会有相当大的差异。这是因为每次发音之间存在着声道形状和语速的差异。而当一个人由于感情的变化大声或小声说话时这种差异就更加明显。这种一个人自己的发音差异我们称之为说话人内部的差异。它主要包括语速、感情语气和健康状况等因素的影响[?]。这些因素中的一个有变化，就可能使这个说话者训练好的识别系统的性能有很大的退化。

总体上看造成不同说话人声学变化（说话人之间的差异）的因素细微而广泛，要比某个具体说话者的语声变化因素（说话人内部的差异）大得多也更难以捕捉和描述。在一些识别系统中，需要区分说话人之间的差异和说话人内部的差异。比如从许多人的语音中识别某个人的语音，就要考虑说话人之间的差异，减轻说话人内部的差异。不过，对于非特定人的语音识别系统，不管是说话人的改变，还是发音条件的变化，两方面都要考虑。遗憾的是，迄今为止，人们还没有能够建立一套比较精确的模型对此进行描述，因此只有求助于统计的方法，通过大量的训练获取某种“平均意义”上的信息，从而减少个人特性的参与。但也正由于个人信息的刻意削弱，造成了系统对某个特定人识别效果的下降。为了解决这个问题，说话人自适应技术应运而生。

2.3 说话人自适应

为了解决第一章第二节和本章第一节里我们提到的特定人系统和非特定人系统中训练数据量和说话人差异这一对矛盾，人们提出了建立一种过渡性模型，由新的说话人提供较少量的数据样本，系统通过提取其中的有用信息并按照一定的算法对原有的非特定人模型进行修正，最终得到更适合该话者的模型。这种技术即被称为说话人自适应（Speaker Adaptation，SA），相应的系统有人称为 SA 系统。其中原有的说话人常称为参考说话人（Reference Speaker），新的说话人为目标说话人（Target Speaker）。说话人自适应也可以理解成是试图使用比特定人系统训练所需更少的数据来提高对特定人的建模精度的方法，如图 2-2 所示。

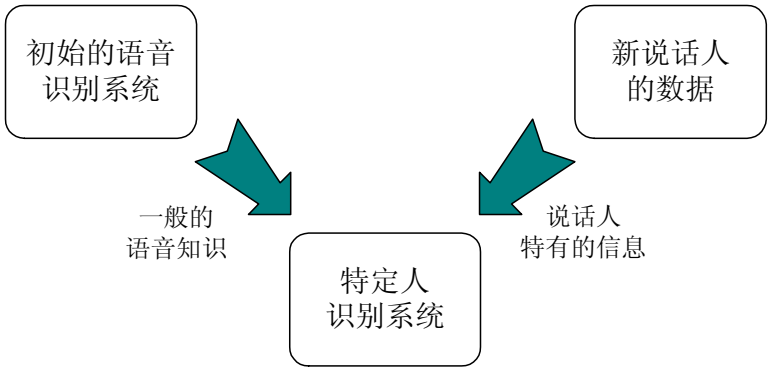


图 2-2　说话人自适应技术：一个特定人识别系统可以通过把从充分训练的模型中得到的通用语音知识和从新说话人的数据中得到的该说话人特有的信息结合起来实现。

对说话人自适应算法的研究，在最早的语音识别系统建立后就开始了。随着语音识别技术的飞速发展，说话人自适应越发得到大家的重视。

2.3.1 说话人自适应的分类

说话人自适应技术，按在什么时候，以什么方式进行自适应和怎样使用自适应数据可以分为下面几种：

有监督自适应（Supervised），即在某特定人使用识别系统之前，提供由系统规定的语音输入，然后系统做自适应过程。目标说话人所说的训练语音预先规定好，所训练的单字、单词或句子是系统已知的。

无监督自适应（Unsupervised），即目标人说话人只需提供少量标注数据或不提供自适应数据，由系统以渐进方式逐渐调整系统参数，以适应于此目标说话人。系统不知道目标说话人所说的语音内容，模型或参数的修正通过识别系统的反馈来实现的。

静态的自适应（Batch/Static），即识别系统一次性使用所有自适应数据进行自适应，生成新的识别模型。

渐进的自适应（Incremental/Dynamic），即识别系统是在运行过程中逐渐调整到最佳状态的，不断使用新的数据来自适应。调整的过程一般不为使用者所知。这种方式也叫做在线自适应（Online Adaptation）。

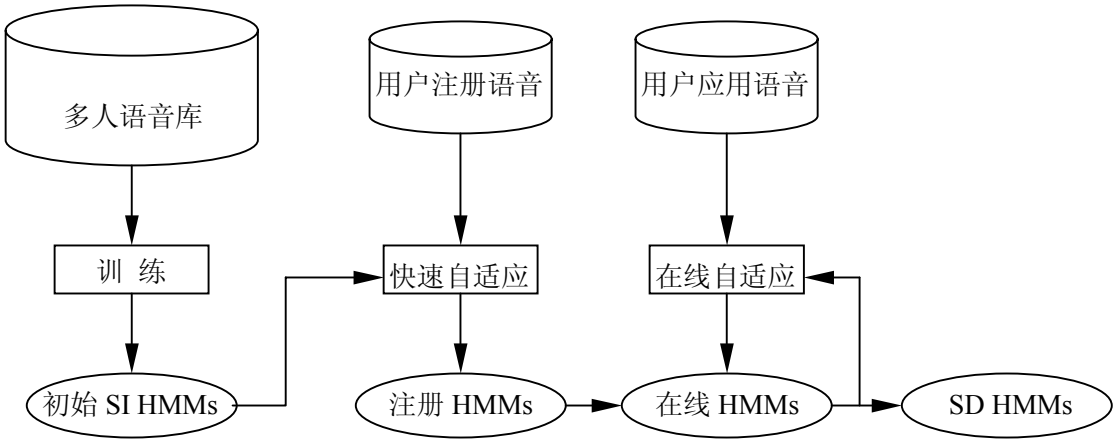


图 2-3　一个基于 HMMs 的说话人自适应系统

通常使用的是静态有监督的自适应和渐进无监督的自适应，前者自适应数据来自注册输入，后者自适应数据来自识别的前向反馈。图 2-3 就给出了这样一个实际系统的例子[36]。当自适应数据的数量无法预知或系统可以提供进一步的自适应数据时，应该使用渐进的自适应。这种情况下，有无监督的方法都可以使用，不过无监督的更适合自然条件。

2.3.2 说话人自适应的主要方法

说话人自适应的方法有很多，本文把他们大致分为以下四类：说话人聚类（Speaker Clustering）、说话人正规化（Speaker Normalization）、谱变换（Spectral Transformation, Spectral Mapping）、模型参数调整（Model Parameter Adaptation）。不过值得指出的是说话人自适应方法的分类并不唯一，有些方法之间没有特别清晰的界限。如一些文献[37]中，说话人正规化算做一种在特征空间上的谱变换方法。另外在实际的应用中，大多数系统往往综合使用多种自适应技术，本节

最后会给出几个实例。下面就按本文的分类方法具体介绍这几种说话人自适应方法：

2.3.2.1 说话人正规化（Speaker Normalization）

尽管象前面所说的说话人的差异很大，但人仍然可以很轻松的识别理解不同口音和性别的各种人的语音。这说明人的大脑可能可以进行一些正规化过程，去除语音个性化的特征。这样在识别系统中说话人之间的差异就可以解决了。说话人正规化技术的思想就是来源于人的识别过程。

说话人正规化的目的是建立一个正规化的说话人空间，使得任何人的语音可以映射其中。这样可以把说话人之间的差异降到最低（最好声学特性不变）。也可以理解成说话人正规化是试图把新说话人的语音特点转化成参考说话人的，这样可以使用已有的参考说话人的特定人识别系统来识别新说话人的语音。图 2-4 是一幅说话人正规化的示意图。

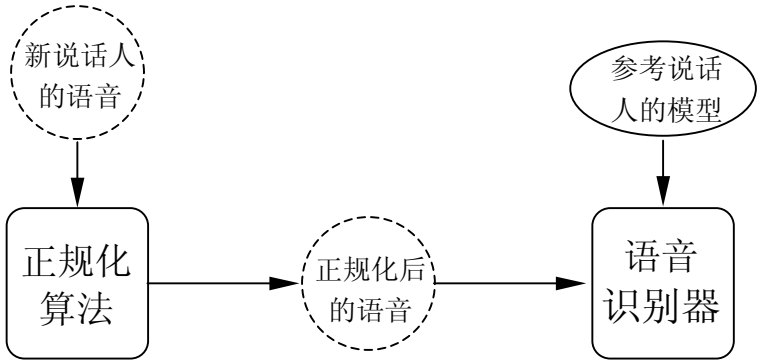


图 2-4　说话人正规化过程示意图

说话人正规化技术的问题在于语音的多样性。要想找到一种一般的技术能够很好的映射到正规化空间是很困难的。比较常用的有两种方法：

倒谱均值规正（Cepstrum mean normalization，CMN）[?, 38]

研究表明，长时倒谱均值可以用来刻画说话人和信道的特征。事实上，CMN

也是诸多用来补偿说话人和信道影响的成功例子。CMN 有两个处理步骤：首先，用一个基于能量的有声/无声检测算法遍历整个语流，计算有声帧的倒谱均值；然后，将该语流的所有帧参数都减去倒谱均值，得到新的特征参数。由于强制训练和测试的所有语流的倒谱均值为零，CMN 可以补偿由于训练和测试时说话人和信道差异可能带来的卷积畸变。

声道长度归一（Vocal tract length normalization，VTN）[?，39，67]

声道长度归一通过补偿声道长度的差异来规正不同说话人的差异。具体的实现方法很多，大致可以分为两类：1）通过对语音频率特性（最常用的如共振峰频率）的估计，直接估计声道长度因子；2）利用最大似然准则来估计声道长度因子。两种方法的后续处理基本一样，都是利用转移（Spectral Shift）算法，如频率弯曲（Frequency Warping）来对语音频谱进行归正，从而校正声道长度不同的影响。

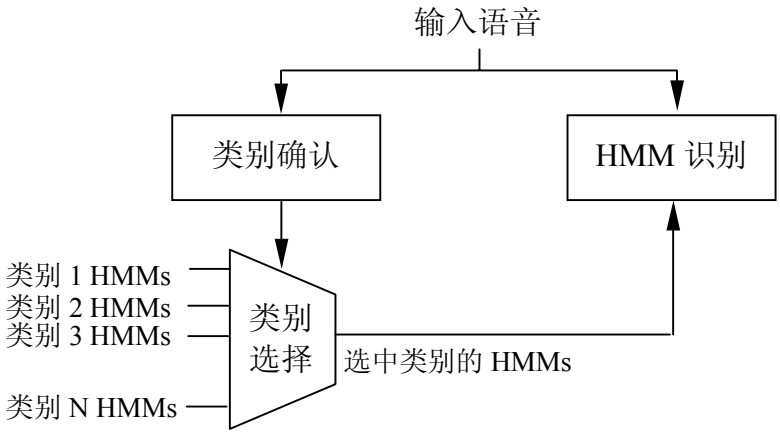
除了上面对特征参数或语音进行变化的方法外，说话人正规化的另一个思路寻找新的特征参数。不同的特征参数，对于不同的说话人的鲁棒性有一定的差异。有文献表明，在常见的特征参数中，MFCC 比其他的如 LPCC 等，对说话人的适应性要好[?]。而寻找更好的、高鲁棒性的特征参数，则是长期以来研究工作的目标之一。比如文献[40，41，42]中提出一种基于 Mellin 变换的语音新特征的自适应方法。由于 Mellin 变换的尺度不变性，这种新的特征对说话人的声道长度是不敏感的，因此可以大大减小由于声道长度而引起的说话人之间的差异。在作者的实验中，新特征的识别效果比 Mel 倒谱好许多，甚至比基于最大似然的声道归正自适应技术还好。

由于说话人正规化需要复杂的映射，并且对所有语音信息进行同样处理，忽视了语音事件的内容，所以单独使用效率不是十分理想。

2.3.2.2 说话人聚类（Speaker Clustering）

说话人聚类（Clustering）或分类（Classification）是实现说话人自适应的一个比较简单的方法。它设想对应不同的说话人有与其对应的许多模型，自适应方法就是找出对应于目标说话人的模型。实际上由于要事先训练且每一个不同

的说话人需要大量的数据，所以产生大量的训练好的模型是不现实的。一般采用的是使用相对少量的模型，每个模型来代表一些相近的说话人。说话人聚类的算法把相近的说话人编成组，用每组的人的语音一起训练出一套模型。这样每个说话人的语音需求就减少了，且能够得到比较少的模型。如图 2-5 所示，对于识别过程中的自适应就是选取与目标说话人最相近的组的模型，并用它来



识别。

图 2-5　说话人聚类自适应过程示意图

由于说话人聚类明显比一般的非特定人系统要好，所以许多人使用它做自适应。Lee[?]在 CMU 的 SPHINXS 系统中曾采用聚类算法实现了这个想法，但自适应与 SI 系统原有结果大体相当，无明显改进。文献[?]中总结的原因是分类后每个类使用的训练数据和非特定人系统使用的相比太少了，从而识别参数的精确度降低。Imamura[43]使用基于距离的交叉熵的方法对已训练好的模型进行聚类，这种方法使得识别性能有所提高。Mathan[44]使用了一种分层的树型结构来聚类，二叉树的每个叶子节点对应一个说话人的 HMM 网络，新的说话人的语音自动向下一直走到叶子节点选取 HMM 网络。后来 Kosaka[45]又提出了一种改进的树型结构的聚类算法，每一步把有最大 Bhattacharyya 距离的一簇分开，聚类过程由阈值控制。说话人群体按照有顶向下逐层细化的分级结构形成树状分类，最上面的层主要包含全局信息成分，层次越靠下，蕴含的个人信息比重越大。选择类别时可以到达的最大层次深度被认为是关键的参数之一，可以根据

最大似然准则（ML）自动确定。训练人个数为 170 时，选择类别后的识别率由原有 SI 系统的 74.2％提高到 76.4％，平均误识率相对降低了 8.5％。

所有这些算法在原理上是相近的。这些算法的优点是只需极少量的训练语音就可以判断目标说话人的类别，因此使用非常快捷方便。但是也同样有一些缺点，如：系统对训练量要求苛刻，参与训练的说话人群体必须尽可能的大并覆盖得足够广泛；当类别数量较大时，对于大词汇表系统存储多套参数所需的存储量过于巨大；此外对于与所有训练说话人发音显著不同的说话人，选择类别的方法几乎不会引起任何改善。不过对于这些问题根据数据应用的目的不同可以做相应的处理。如采取平滑（Smooth1ng）或内插（Interpolation）技术在各类别中加入整体信息。另外说话人聚类的方法现在更多的是和其他自适应方法相结合使用，为其他的方法提供初始或参考模型。

性别依赖的模型（Gender-dependent Models）可以看作是说话人聚类的一种特例应用。它仅根据说话人的性别把说话人分成两类，在训练时同性的说话人归为一类，并对每类分别识别。它和一般说话人聚类技术的不同是，它可以不需要对新的说话人进行类别检测。实践证明[?]，性别依赖的模型可以减少词的识别错误率约 10%。而且，虽然更复杂的说话人聚类模型可以进一步降低识别错误率，但除非有大量的类别否则效果并不比性别依赖的模型好很多。所以目前很多系统使用的是性别依赖的模型。

2.3.2.3 谱变换（Spectral Transformation）

谱变换，又被称做谱映射（Spectral Mapping），这几年开始广泛的使用在说话人自适应中。其基本思路是不同说话人之间差异有很大一部分表现在他们在说同一话语时其语音短时谱的差异上，而这种差异又是由于每个人的发声器官的个体差异造成的，因此有可能通过一种线性变换，从一种谱映射为另一种谱，从而实现自适应。这种映射既适用于信号空间和特征空间（Feature Space），也可以在 VQ 码本或者 HMM 参数上进行，如图 2-6 所示[46]。

实现时，首先利用新说话人的少量语音找出新老说话人语音短时谱间的对

应关系，从而可以用某种变换将前者映射为后者，或者反之，或者将二者同时映射到第三空间。然后用已有的经过充分训练的 HMM 参数对新说话人语音进行识别。这两种方法的示意图如图 2-7 和图 2-8。

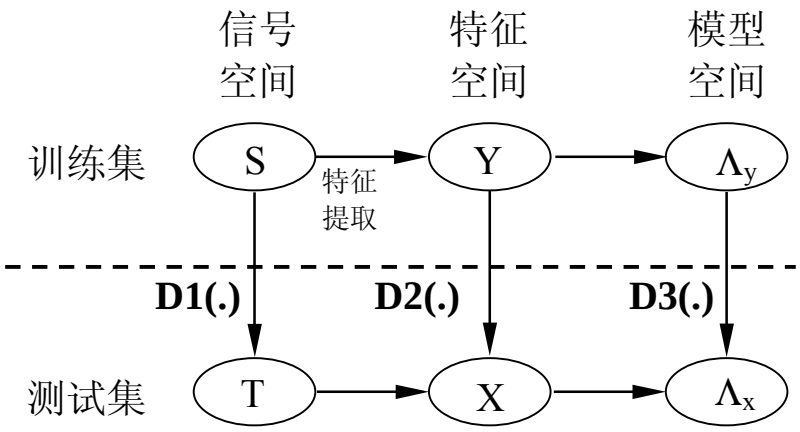


图 2-6　信号空间、特征空间和模型空间中的谱转换

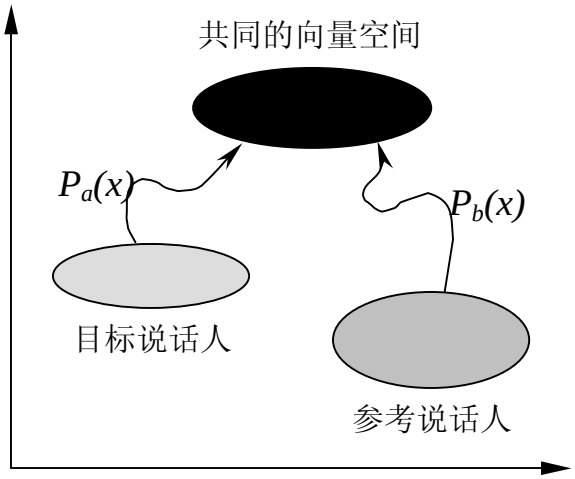


图 2-7　所有说话人的语音表示在一个公用的向量空间

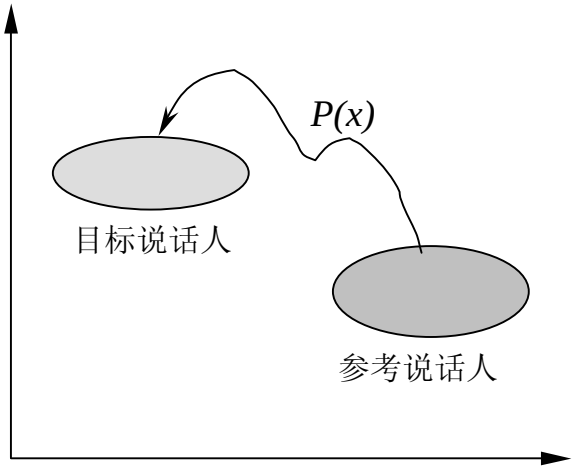


图 2-8　单一的变换从参考说话人到目标说话人

第一种方法是对参考说话人的语音建立参考模板，并产生使目标说话人和参考说话人之间距离最小的谱转换。如图 2-7 所示，该方法把目标说话人的语音和参考模板都映射到一个共同的向量空间。这需要两个转换函数：为目标说

话人的（P_a）和为参考模板的（P_b）。

$$y_s = P_a(x_s) \tag{2-1}$$

$$y_r = P_b(x_r) \tag{2-2}$$

其中 x_s 和 y_s 分别是目标说话人的原始和转换后的观察向量，x_r 和 y_r 分别是参考说话人的原始和转换后的模板。识别的距离衡量就是在 y_s 和 y_r 之间进行的。

另一种简化而又有效的谱变换方法是类频率转移法：

$$y = P(x) = Tx + a \tag{2-3}$$

其中 T 是转移矩阵，a 是个向量。x 是输入语音，y 是转换后的语音（见图 2-8）。这种转换即可以对所有的输入语音帧一起转换和模板比较，也可以分别对不同的语音段用不同的转换。虽然后者的计算量更大且需要的数据也多，但效果比前者要好。

上面的两种方法不仅可以直接使用在信号空间和特征空间，类似的方法还使用在离散 HMM（discrete HMM，DHMM）识别系统中，对 DHMM 的 VQ 码本进行在概率域上的转换。比如美国 BBN 公司开发的 BYBLOS 系统[47]就采取了谱变换的方案。原理如图 2-9。系统采用 DHMM 识别框架，先建立一个特定人的 SD 系统。再要求一个新的说话人只说一短段语音（例如 2 分钟），利用这一短段语音找到新老二者的对应关系，以便将充分训练的老说话人的全套识别参数用于新说话人。在谱转换的自适应算法中保持状态转移概率不变，只对输出概率进行自适应处理，处理步骤如下：

1）让新说话人说 2 分种规定话语，老说话人的训练语音中也包括这 2 分钟规定话语。对新、老说话人分别形成这 2 分钟话语的特征矢量序列。用 DTW 算法将这两个序列对齐，对齐的准则是使二者的欧式距离为最小。

2）用老说话人的 VQ 码本对这两个序列进行编码，分别得到新、老说话人的 VQ 码字标号序列。统计出码字的概率依赖关系映射矩阵。

3）用概率映射矩阵改善老的说话人的识别参数。

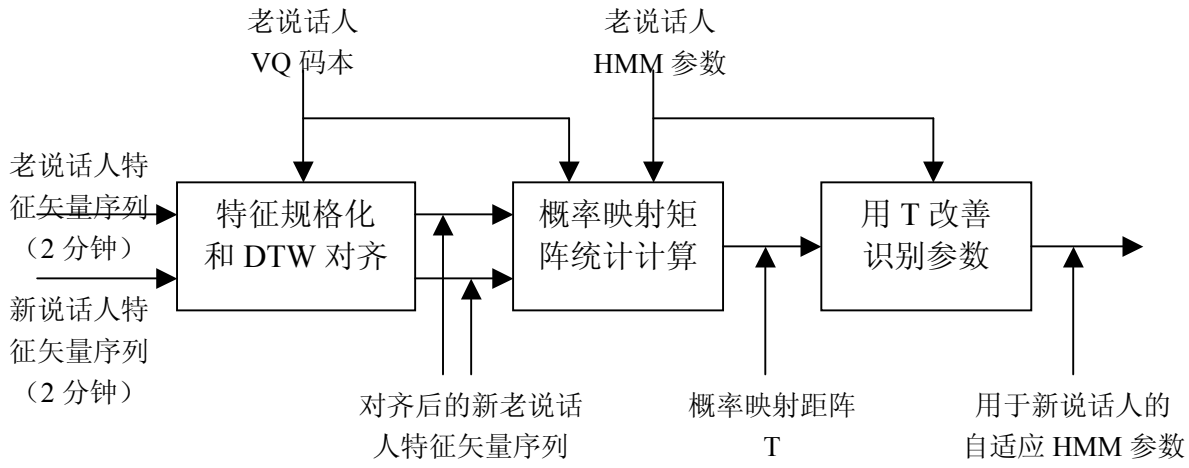


图 2-9　一个采用特征规格化和概率谱映射自适应系统

在 IBM 公司的 5000 词 TANGORA 系统中[48]，采用了另外一种称之为“说话人马尔可夫模型”的谱映射自适应算法并且获得较佳的效果。与前者不同的是，这里是将老说话人的谱空间映射到新的说话人上；这样可以充分利用已有的训练数据。

目前很多基于连续 HMM（continuous density HMM，CDHMM）的系统采用一种叫最大似然线性回归（MLLR）的自适应方法[?, 49, 50, 51, 52, 53]。这种方法可以看作是谱变换在 CDHMM 参数空间上的应用，本文将在第三章对它做详细的介绍。

除了线性的谱转换外，还有一些非线性的映射方法，不过其效果并不比线性的好，所以实际使用中用的比较少。

2.3.2.4 模型参数调整（Model Parameter Adaptation）

最近模型参数调整的自适应方法在说话人自适应中运用很广泛。对于提高非特定人系统的性能，谱变换和模型选择（即说话人聚类）的方法有一定的局限性。尽管一些谱变换技术对模型参数也进行修改，但它们更偏重于试图实现

目标说话人与参考说话人的匹配，而不是去尽量提高模型对目标说话人的精确建模。

模型参数调整主要是针对连续 HMM（Continuous Density HMM，CDHMM）模型参数的转换。不过经过扩展也可以使用在离散 HMM（Discrete HMM，DHMM）和半连续 HMM（Semi-continuous HMM，SCHMM）系统中。对于离散 HMM 系统自适应，主要是对 VQ 码本和 HMM 参数转换；而连续分布 HMM 系统则只需要对 HMM 参数（状态输出概率函数和状态转移概率）转换。

模型参数调整方法使用的是最大后验概率（Maximum a Posteriori，MAP）的重估方法[54, 55, 56, 57, ?]。系统把原有 SI 系统的 HMM 参数作为先验知识，遵照 Bayes 估计准则求出达到最大后验概率时系统采用的最佳参数。所以这种方法也被称为贝叶斯自适应（Bayesian Adaptation）。MAP 方法和经典的 Baum-Welch 最大似然（Maximum Likelihood，ML）重估算法比较，需要的训练数据很少，解决了自适应数据不足的问题。而且理论上当训练数据趋于无穷时，用 MAP 法所得的模型与用充分语料做最大似然（ML）训练所得的模型相等价。有关 MAP 方法的具体原理和实现方法，本文在第三章中将做详尽阐述和讨论。

最近 Qiang Huo[58, 59, 60, 61]提出一种基于准贝叶斯（Quasi-Bayes）学习的渐进式 HMM 自适应框架。其核心思想是同时渐进的更新近似后验分布参数和 HMM 参数。把近似后验分布参数在下一步 MAP 估算中作为先验分布。尽管这种方法的收敛性没有理论证明，但实验表明这种基于 QB 的渐进方法几乎可以达到 Batch 的方法的效果。

MAP 方法要求对所有模型的参数进行估算，但由于自适应数据的有限，有些模型没有对应的数据来自适应。这种情况在大词汇的语音识别系统中更明显。有的方法对尚无训练数据的模型就不做变换，不过大多数系统都通过模型的共享和平滑技术来解决[62, 63]。在第三章中我们会详细介绍一种向量域平滑（Vector Field Smoothing，VFS）的方法。

2.3.2.5 综合方法

在前面的方法介绍中，各个方法是分开介绍的，不过在实际的应用中，决大多数系统是综合使用多种自适应技术的。比如谱转换和模型参数转换就常一

起使用；文献[?]中结合了 MAP 和 VFS 的自适应方法，在系统渐进性和快速性两方面都取得较好的效果。文献[?]给出了一个很综合的自适应系统。如图 2-10 所示，文中使用了 MAP-VFS 的方法，并利用说话人聚类来选择产生初始模型。该系统用 6.3s 的自适应数据，使系统的错误率由 22.0%降到 17.7%。

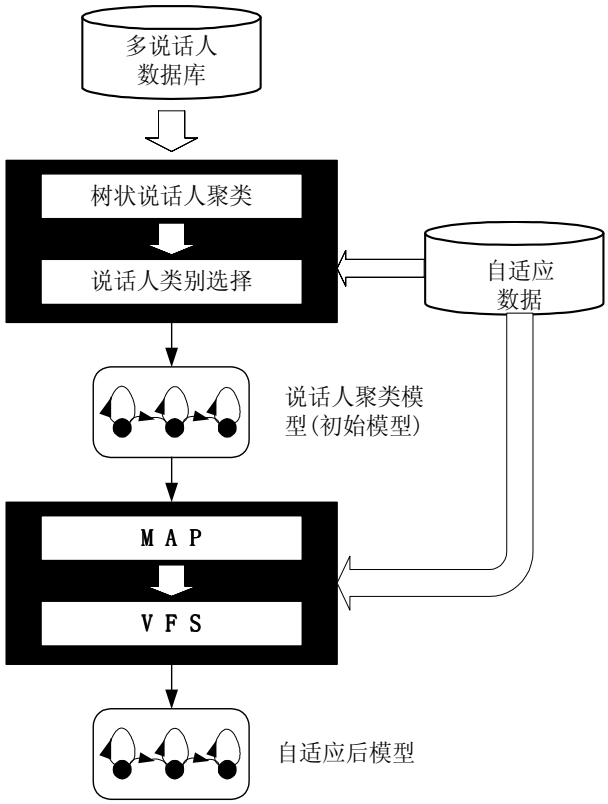


图 2-10 说话人自适应系统实例[50]

2.4 综述

说话人自适应技术是运用少量的自适应数据来训练提高非特定人系统对特定人的建模精度。它可以有效的克服说话人之间的差异给非特定人系统带来的难题。目前大多数的语音识别系统都采用了一种或几种的说话人自适应技术。不过由于人们对识别系统的完健性要求的提高和语音识别系统的不断发展，说话人自适应技术还有很多课题值得发展和改进。

第三章 MAP 与 MLLR

本章将详细给出了我们实现的两种自适应方法：基于最大后验概率（MAP）方法和基于最大似然线性回归（MLLR）方法的原理和实现。

3.1 基于 HMM 模型参数转换的自适应方法

本章介绍的两个自适应方法都可以归为模型参数转换（Model Parameter Transformation）的自适应方法，它们主要是针对语音识别系统中的 HMM 模型参数做自适应转换，如图 3-1 所示。这里 HMM 模型既可以是离散 HMM（DHMM）模型也可以是连续分布 HMM（CDHMM）模型。由于在识别过程中 HMM 的状态输出概率函数起主要的作用，所以本论文中只讨论对 CDHMM 的状态输出概率函数参数（主要是均值）进行自适应。

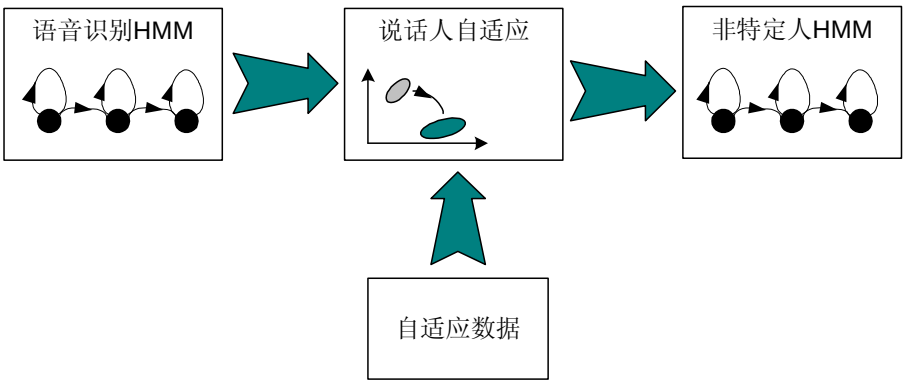


图 3-1　模型参数转换的自适应方法：通过少量自适应数据对原有识别系统的 HMM 模型参数进行转换，得到相当于特定人识别系统的 HMM 模型。

3.2 最大后验概率（Maximum a Posteriori，MAP）

MAP 是目前模型参数调整即 Bayesian Adaptation 的一个主要方法之一。它通过引入了先验（a priori）知识来求最大的后验（a posteriori）概率，从而提高自适应效果。所以和最大似然（Maximum Likelihood）重估方法相对应，这个方法被叫做最大后验概率（Maximum a Posteriori）重估方法。

最早引入 MAP 重估方法来自适应，是针对离散 HMM 模型参数对新说话人进行自适应[64]。后来扩展到对连续 HMM 模型参数的自适应[?]。再后来，Gauvain 和 Lee 又给出了转换带有混合高斯输出分布的 HMM 模型参数的办法[?]。目前各种基于 HMM 的识别系统，特别是规模比较大的识别系统，大多数都采用 MAP 方法做自适应。近年来，许多人也围绕 MAP 方法进行一些研究，提出了很多改进的 MAP 方法，不过究其原理都基本一致。

下面我们从 MAP 和 ML 的区别入手，看看 MAP 的原理和实现。

3.2.1 MAP 和 ML

由于说话人自适应是希望使用少量的新说话人的自适应数据来对系统进行自适应，所以常遇到训练数据稀疏（sparse）的问题。而目前 HMM 模型参数训练一般使用的是经典的 Baum-Welch 最大似然（Maximum Likelihood，ML）重估算法，这种 ML 算法只有在大量而充分的语料训练的前提下才能达到最优。所以，标准的 HMM 模型参数的 ML 训练方法在数据稀疏的情况下效果不好（参见第五章的实验结果）。于是为了解决自适应数据不足的问题，人们把 HMM 模型的先验信息引入了模型训练过程，发展出了最大后验概率方法（MAP）。

MAP 重估和 ML 重估的原理很相似，其根本区别在于是否在重估过程中使用参数的先验分布。假设$O = \{o_1, o_2, ..., o_T\}$是概率密度函数（Probability Density Function，p.d.f.）为 $p(O)$ 的一系列观察值，λ是定义分布的参数集合。重估问题可以看作是给定训练数据序列O，估算λ的过程。这个过程我们可以通过求下式实现：

$$\lambda_{estimate} = \arg\max_{\lambda} p(\lambda \mid O) \tag{3-1}$$

应用贝叶斯准则（Bayes Rule），其中 $p(\lambda)$ 是 HMM 参数的先验分布[2]：

$$p(\lambda \mid O) = \frac{p(O \mid \lambda) p(\lambda)}{p(O)} \tag{3-2}$$

得到：

$$\lambda_{estimate} = \arg\max_{\lambda} \frac{p(O \mid \lambda) p(\lambda)}{p(O)} \tag{3-3}$$

在传统的 ML 的估计公式中，认为模型参数λ虽然是不知道的但却是固定的，同时概率密度函数 $p(O)$ 与模型参数无关，所以忽略了分式中的分母 $p(O)$ 和分子中的 $p(\lambda)$，只是使 $p(O \mid \lambda)$ 达到最大。即，

$$\lambda_{ML} = \arg\max_{\lambda} p(O \mid \lambda) \tag{3-4}$$

而 MAP 重估方法最大的特点是引入了对 HMM 参数的先验分布 $p(\lambda)$ 的考虑，即认为模型参数λ是符合先验分布 $p(\lambda)$ 的随机变量。所以在重估过程中只忽略公式中的分母部分。即，

$$\lambda_{MAP} = \arg\max_{\lambda} p(\lambda \mid O) = \arg\max_{\lambda} p(O \mid \lambda) p(\lambda) \tag{3-5}$$

对比公式（3-4）和公式（3-5），我们可以认为 ML 重估是 MAP 重估在忽略参数先验分布假设情况下的特例，而 MAP 则是在 ML 重估中引入了对 HMM 参数λ的先验分布 $p(\lambda)$。

值得注意的是先验知识（如待估参数的先验分布）的使用是 MAP 成功的关键。而对 MAP 重估的简化起重要作用的是一个叫分布共轭家族（Conjugate Families of Distribution）的概念。一个随机向量的共轭先验（Conjugate Prior）定义为这个向量的概率密度函数（p.d.f.）参数的先验分布，这样后验分布 $p(\lambda \mid O)$ 和先验分布 $p(\lambda)$ 都属于同一个分布函数家族。例如，一个高斯（Gaussian）概率密度函数的均值的共轭先验也是一个高斯密度。

[2]对于具有高斯混合密度的 CDHMM 指所有参数的先验知识，包括初始概率、转移概率和混合系数的分布，还有均值和方差参数的联合分布等。由于在 CDHMM 参数中，尤以均值向量对识别结果的影响最大，所以一般的 MAP 只对均值重估。

3.2.2 MAP 重估

下面我们从公式(2-5)推导出 MAP 的均值重估公式。假设观察值 $x_1, x_2, ..., x_n$ 是符合均值 ϕ 未知，方差 σ^2 已知的高斯分布。那么似然函数 $p(x|\phi)$ 可以表示为：

$$p(x|\phi) = \frac{1}{(2\pi)^{n/2}\sigma^2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i-\phi}{\sigma})^2\right] \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i-\phi}{\sigma})^2\right] \quad (3\text{-}6)$$

为了简化公式（3-6），我们用下式：

$$\sum_{i=1}^{n}(x_i-\phi)^2 = n(\phi-\bar{x}_n)^2 + \sum_{i=1}^{n}(x_i-\bar{x}_n)^2 \quad (3\text{-}7)$$

其中 $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n}x_i$，即 $x = \{x_1, x_2, ..., x_n\}$ 的样本均值。

我们改写公式（3-6）里的 $p(x|\phi)$ 为公式（3-8）：

$$p(x|\phi) \propto \exp\left[-\frac{n}{2\sigma^2}(\phi-\bar{x}_n)^2\right]\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right] \quad (3\text{-}8)$$

同时我们假设 ϕ 的共轭先验分布也是高斯分布，并且均值是 μ，方差是 ν^2。于是先验分布 $p(\phi)$ 可以表示为下式：

$$p(\phi) = \frac{1}{(2\pi)^{1/2}\nu}\exp\left[-\frac{1}{2}(\frac{\phi-\mu}{\nu})^2\right] \propto \exp\left[-\frac{1}{2}(\frac{\phi-\mu}{\nu})^2\right] \quad (3\text{-}9)$$

MAP 重估如公式（3-10）所示试图找到 ϕ_{MAP} 使得后验概率最大：

$$\phi_{MAP} = \arg\max_{\phi} p(\phi|x) = \arg\max_{\phi} p(x|\phi)p(\phi) \quad (3\text{-}10)$$

由于对数函数是增函数，所以 ϕ_{MAP} 也可以表示为下式：

$$\phi_{MAP} = \arg\max_{\phi}\left(\log p(x|\phi) + \log p(\phi)\right) \quad (3\text{-}11)$$

所以 ϕ_{MAP} 可以通过解下面的微分等式得到：

$$\frac{\partial \log p(x|\phi)}{\partial\phi} + \frac{\partial \log p(\phi)}{\partial\phi} = 0 \quad (3\text{-}12)$$

把公式（3-8）和公式（3-9）带入公式（3-12）得：

$$\frac{\partial\left[-\frac{1}{2}(\frac{\phi-\mu}{\nu})^2\right]}{\partial\phi}+\frac{\partial\left[-\frac{n}{2\sigma^2}(\phi-\bar{x}_n)^2\right]}{\partial\phi}=0 \tag{3-13}$$

化简上式，可以解得：

$$\phi_{MAP}=\frac{\sigma^2\mu+n\nu^2\bar{x}_n}{\sigma^2+n\nu^2} \tag{3-14}$$

我们把公式（3-14）里的符号改变成我们习惯的，并做一下变量替换，可以得到对于高斯分布均值的 MAP 重估的一般公式：

$$\hat{\mu}_k=\frac{\tau_k\mu_k+n_k m_k}{\tau_k+n_k} \tag{3-15}$$

式中 μ_k 与 $\hat{\mu}_k$ 分别表示自适应前后的均值向量，n_k 表示对应第 k 个高斯分布的训练样本总数，m_k 表示用 ML 估算出的对应该高斯的样本均值向量。τ_k 是模型先验分布的一个重要参数，它控制着自适应对先验信息 μ_k 的依赖程度，可以预先设定，也可用在训练过程中由数据估计，我们将在下一小节中专门介绍它的求法。

从公式（3-15）我们可以看出，自适应后的均值向量实际是初始均值与 ML 估算的均值的一种线性加权之和。对于自适应数据比较少的情况，n_k 比较小，于是计算均值时先验的知识占优。当自适应数据增加时，n_k 相应变大，于是 ML 重估对 MAP 的结果的影响增大。这也说明了 MAP 自适应方法的一个重要特性：随着自适应数据数量的增大，MAP 重估逐渐逼近 ML 重估。而 ML 是模型通常的训练方法，所以 MAP 的结果是可以达到充分训练过的特定人识别的效果。并且在理论上当训练数据趋于无穷时，用 MAP 法所得的模型与用充分语料做最大似然（ML）训练所得的模型相等价。

对于公式（3-15）我们也可以转换改写成下面的形式，从而用类 Baum-Welch 的方法具体实现。

$$\hat{\mu}_k=\frac{\tau_k\cdot\mu_k+\sum_{t=1}^{T}c_{kt}\cdot o_t}{\tau_k+\sum_{t=1}^{T}c_{kt}} \tag{3-16}$$

其中

$$c_{kt} = \frac{\omega_k \cdot N(o_t / \mu_k, \sum_k)}{\sum_k \omega_k \cdot N(o_t / \mu_k, \sum_k)} \tag{3-17}$$

式中 o_t 表示对应的训练数据。 $N(o_t / \mu_k, \sum_k)$ 是一个高斯分布， $\sum_k$ 是高斯分布的协方差矩阵， ω_k 是混合系数。这里 τ_k 和公式（3-15）里的差一个系数。从公式（3-16）中可以看出自适应调整后的均值向量实际是初始均值与相应各训练数据的线性加权之和。

上面给出的只是对均值重估的推导。如果对均值和协方差矩阵都要进行重估推导就比较复杂了，可以参见文献[29, 54-56]，这里不多做描述了。在目前的大多数自适应系统中仅仅对均值进行自适应，而保持原始的协方差矩阵不变。因为很多的文献都提到，协方差矩阵的自适应不仅没有是原来的系统得到改善，甚至降低了系统的识别率[?]。文献[?]中对这种情况做了简单的分析，认为估计协方差矩阵时的未知参数比较多，而可获得的有限样本个数无法满足精确估计的要求。

3.2.3 先验知识 -- 初始模型与 τ

对于基于 MAP 的方法有一点最重要的制约，就是它要求对于先验分布 $p(\phi)$ 有一个精确的估计，而做到这一点又常常是很困难的。对先验分布 $p(\phi)$ 的估计具体对应到公式（3-15）中，主要是对初始模型的参数 μ_k 和模型先验分布的参数 τ_k 的估计。一般我们使用已训练好的包含了原始训练条件的特点的初始模型来估计先验分布 $p(\phi)$ 。

对于 μ_k 的估计，我们一般直接使用一个已训练好的初始模型的参数，或者是从多个已训练好的初始模型的参数中统计得到。有些文献[?, 65]则利用说话人聚类的方法提供更加精确的初始模型的估计。

对于 τ_k 的估计，也就是如何对先验知识加权，是 MAP 重估中十分重要的因素。它控制着自适应对先验信息的依赖程度，对自适应效果有很大的影响。那么如何得到它呢？一种常用的方法是根据说话人或环境相似度的衡量把初始的

训练数据聚类成初始的高斯先验，从而用这一系列模型来作为先验分布 $p(\phi)$ 的观察值，从中提取出先验分布。还有的方法是直接从训练数据中运用经验 Bayes 方法来提取先验分布的。当然在实际的说话人自适应应用中，采用固定的 τ_k 也取得了相当好的效果。下面具体给出两种求 τ_k 的方法。

一种方法[?]是从初始的一个非特定人系统（或多个的特定人系统的）参数中提取 τ_k。假设模型的均值是随机变量，而方差 σ^2 是已知和固定的。可以知道均值的先验分布也是高斯分布，设均值是 μ，方差是 ν^2。根据公式（3-14）得到下面公式：

$$\hat{\mu}_{MAP} = \frac{\sigma^2 \mu + n\nu^2 m}{\sigma^2 + n\nu^2} \tag{3-18}$$

其中 m 是状态对应的训练样本的均值，n 是样本的总数。而 σ^2、μ、ν^2 这些先验参数是通过非特定系统的混合参数（或一组特定人系统参数）统计得到的，具体公式如下：

$$\mu = \sum_{m=1}^{M} w_m \mu_m \tag{3-19}$$

$$\nu^2 = \sum_{m=1}^{M} w_m (\mu - \mu_m)^2 \tag{3-20}$$

$$\sigma^2 = \sum_{m=1}^{M} w_m \sigma_m^2 \tag{3-21}$$

其中 w_m 是第 m 个混合的高斯分布（或第 m 个模型高斯分布）的权重，$\sigma_m{}^2$ 和 ν_m 分别是第 m 个混合的高斯分布（或第 m 个模型高斯分布）的方差和均值。如果要改写公式（3-18）成公式（3-15）的话，做个 $\tau_k = \sigma^2 / \nu^2$ 的替换即可。

另一种方法[?]是利用自适应数据的均值 m_k 和初始模型的高斯均值 μ_k 的 Mahalanobis 距离的倒数来估算 τ_k。m_k 是通过 ML 估算出的自适应数据的均值，其包含了目标说话人的信息。如果有足够的数据，τ_k 可以表示一定的对于目标说话人的先验分布信息。因为如果自适应的均值和初始模型相差比较大，则他们分布的距离比较大，τ_k 比较小，于是自适应的均值在线性加权中起的作用大

一些。相反当自适应的均值和初始模型比较接近，则τ_k比较大，初始模型占的比重就大一些。在自适应数据不多时，可以对所有的高斯分布使用统一的τ_k，即τ，用来衡量目标说话人和初始模型的差异大小。

$$\tau = \frac{d\sum_k^{ALL} n_k}{\sum_k^{ALL} n_k (m_k - \mu_k)^{\mathrm{T}} \sum_k^{-1} (m_k - \mu_k)} \tag{3-22}$$

其中d是特征的维数，$\sum_k$是第k个高斯分布的协方差矩阵，$\sum_k^{ALL}$是对CDHMM的所有高斯分布求和。公式（3-22）表示τ是CDHMM中所有m_k和μ_k的Mahalanobis距离加权平均的倒数。本论文第五章实验中使用的方法就是这种方法。

3.2.4 向量域平滑（Vector Field Smoothing，VFS）

由于自适应数据的有限，HMM的均值有些没有对应的数据来自适应，有的方法对尚无训练数据的模型均值就不做转换。但引入向量域平滑（VFS）后就解决了这个问题。VFS是基于如下假设：从一个说话人的声学特征空间可以连续地转移到另一个说话人的特征空间。它通过对均值转移向量的插值和平滑来解决没有自适应数据或数据不均匀造成的自适应错误，如图3-2所示。其核心思想就是一种类间（inter-class）的平滑，这与MAP/ML方法解决的类内（intra-class）的训练不同。

（1） 转移向量的估计

先给出转移向量的定义。对于HMM的任一均值向量μ_k，可定义其转移向量如下：

$$v_k = \hat{\mu}_k - \mu_k \tag{3-23}$$

于是我们可以对有训练数据的高斯分布μ_p（$p \in K_1$，K_1是所有有训练数据的高斯分布的集合），计算出转移向量：

$$v_p = \hat{\mu}_p - \mu_p \tag{3-24}$$

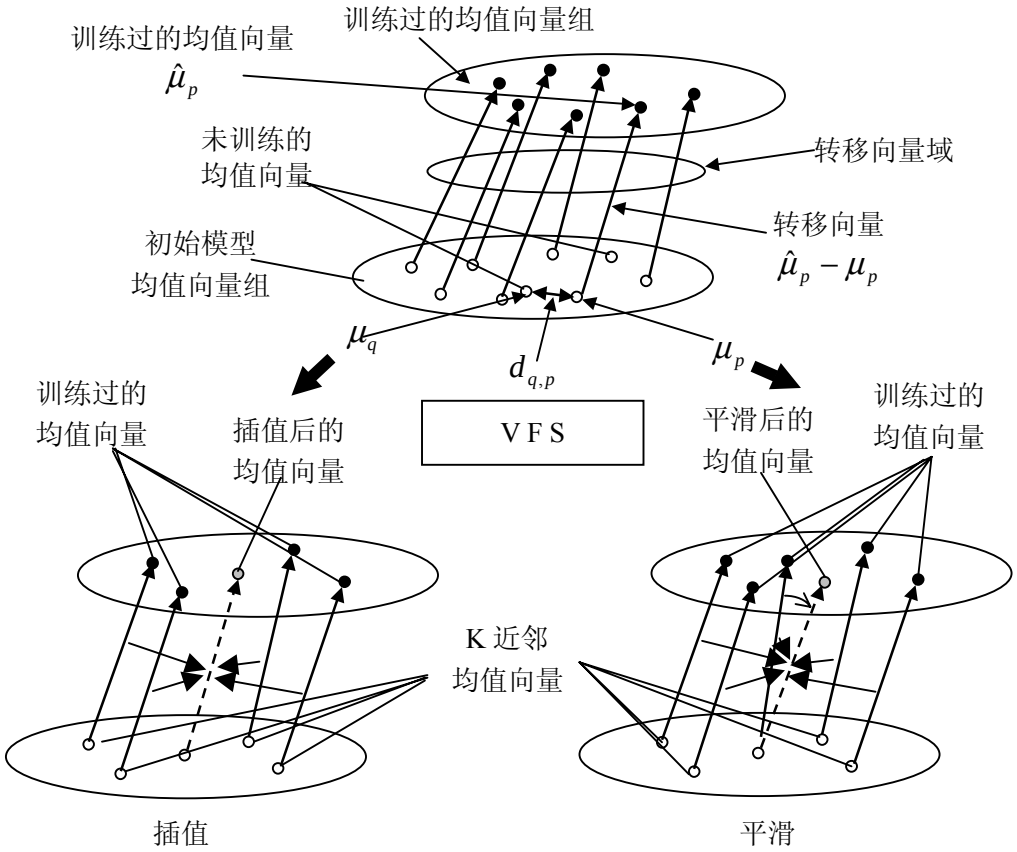


图 3-2　向量域平滑（VFS）方法的原理示意图

（2） 转移向量的插值

如果假设所有转移向量构成的向量域是平滑的，则可以用插值方法估算出未训练均值的转移向量。对于没有训练的高斯分布μ_q（$q \in K_2$，K_2是所有没有训练数据的高斯分布的集合），用下面的插值公式（3-25）通过训练过的相邻的转移向量v_k（$k \in N(q)$）计算μ_q的转移向量v_q：

$$v_q = \sum_{k \in N(q)} \lambda_{q,k} v_k \Big/ \sum_{k \in N(q)} \lambda_{q,k} \tag{3-25}$$

$$\lambda_{q,k} = \exp\left(\frac{-d_{q,k}^2}{s}\right) \tag{3-26}$$

$$d_{q,k}^2 = \sum_{l=1}^{L} \frac{(\mu_{q,l} - \mu_{k,l})^2}{\sigma_{q,l}^2} \tag{3-27}$$

其中 $N(q)$ 是 μ_q 的训练过的 k 近邻的集合。$\lambda_{q,k}$ 是基于 μ_q 和 μ_k 的马氏距离（Mahalanobis Distance）$d_{q,k}$ 的插值权重参数，体现了 μ_k 的转移向量对 μ_q 的影响程度，它随着 $d_{q,k}$ 的增大而减小。L 是特征的维数。s 是平滑的控制参数，控制着 $\lambda_{q,k}$ 随 $d_{q,k}$ 变化的速度，s 取值越大，转移向量之间的影响范围就越大，所以可以通过调整 s 的办法，控制 VFS 方法的平滑程度。如果 $s=0$ 时就相当没有平滑，当 $s=\infty$ 时就是线性插值。

这样得到转移向量 v_q 后就可以求出 $\hat{\mu}_q$ 了：

$$\hat{\mu}_q = \mu_q + v_q \tag{3-28}$$

（3） 转移向量的平滑

另外，因为对应于各个已训练均值的训练数据量不同，转移训练的置信程度也就各不相同，所以还可以把它们做平滑处理。对于所有训练过的转移向量 v_p（$p \in K_1$）用公式（3-29）进行平滑。平滑过程假设了所有的转移向量是受连续性约束的。换句话说，一个人的声学特征空间是可以连续转换到另一个人的。

$$v_p^S = \sum_{k \in N(p)} \lambda_{p,k} v_k \Big/ \sum_{k \in N(p)} \lambda_{p,k} \tag{3-29}$$

其中 $\lambda_{p,k}$ 与 $d_{p,k}$ 的计算方法和公式（3-26）和（3-27）一样。平滑公式（3-29）和插值公式（3-25）唯一不同的地方是 $N(q)$ 这里包含了 μ_p 自己，因为它训练过。而在插值中就不包括它自己。

然后可以求出平滑后的均值了：

$$\hat{\mu}_p^S = \mu_p + v_p^S \tag{3-30}$$

通过 VFS 技术的使用，可以从两个方面提高了 MAP 的性能：一方面解决了无自适应语料的模型调整问题，另一方面对训练语料不足的模型参数进行了补偿。

3.2.4 算法实现

本文分别实现了 MAP 和 MAP/VFS 的重估算法。具体实现算法如下：

MAP 自适应算法：

```
for 所有的混合成分，k
    for 所有观察序列，o
        for 所有帧，t
            计算状态发生概率 c_kt^o，公式（3-17）
        end
        记录样本数 n_k
    end
    计算出混合成分的样本均值 m_k
end

for 所有混合成分
    计算对应的权重 τ_k，公式（3-22）
end

for 所有混合成分
    自适应混合密度的均值 μ̂_k，公式（3-15）
end
```

MAP/VFS 自适应算法：

从初始模型中统计 $\lambda_{q,k}$，公式（3-26）和（3-27）

for 所有有观察序列的混合成分，p
  for 所有观察序列，o
    for 所有帧，t
      计算状态发生概率 c_{kt}^{o}，公式（3-17）
    end
    记录样本数 n_p
  end
  计算出混合成分的样本均值 m_p
end

for 所有有观察序列的混合成分，p
  计算对应的权重 τ_p，公式（3-22）
end

for 所有有观察序列的混合成分，p
  自适应有观察序列的混合密度的均值 $\hat{\mu}_p$，公式（3-15）
  计算均值转移向量 v_p，公式（3-24）
end

for 所有没有观察序列的混合成分，q
  插值计算均值转移向量 v_q，公式（3-25）
  计算出没有观察序列的混合成分的均值 $\hat{\mu}_q$，公式（3-28）
end

for 所有有观察序列的混合成分，p
  平滑计算均值转移向量 v_p^S，公式（3-29）
  计算出有观察序列的混合成分的均值 $\hat{\mu}_p^S$，公式（3-30）
end

3.3 最大似然线性回归（Maximum Likelihood Linear Regression，MLLR）

当 CDHMM 用在声学建模时，最重要的待自适应的参数就是输出高斯分布，如均值向量和协方差矩阵。我们可以用一系列线性转换函数对均值和协方差进行映射使得自适应数据的似然最大。这种方法就是最大似然线性回归(MLLR)方法，目前被广泛使用在说话人和环境自适应中。

这种方法是由 Leggetter[?, ?, 50, 66]提出的。它使用 ML 对自适应数据估算出一套转移参数$[A,b]$，再用转移参数把 SI 系统的参数 y 转换成自适应模型的参数 x：

$$x = Ay + b \tag{3-31}$$

下面我们分别介绍 MLLR 的原理，回归类的定义，转移矩阵的重估和具体实现。

3.3.1 简介

正如上面提到的，MLLR 是对于密度函数的均值向量进行自适应的技术，其他的参数在自适应模型中不作改变。尽管理想上所有参数应该都作改变，因为它们都是和说话人相关的，但由于在实际的应用中只有少量的自适应数据，所以没有足够的数据用来自适应所有的参数。由于转移概率和混合权重系数对最后的效果影响很小[?]，所以一般很少有对它们做转换的。而对于协方差矩阵，有人使用和均值转移矩阵相同的转移矩阵来进行线性转换。但 Leggetter[?]认为如果用均值的转移矩阵来转换协方差，还不如保持协方差不变，这样可以保持每个单独的分布。后来文献[?]又提出了其他的单独转换协方差矩阵的 MLLR 方法，取得了比较好的效果。但转换协方差矩阵使得待估参数变多，相应对自适应数据的要求就更大了，而且计算量也有所增加，所以一般系统还是只对均值进行自适应[3]。图 3-3 给出了在声学特征空间的转换均值向量的效果示意图。

[3] 在文献[?]中的实验里，对均值的自适应可以提高系统性能 12%，而增加对协方差矩阵的转移矩阵的估计只能得到额外的 2%的提高。

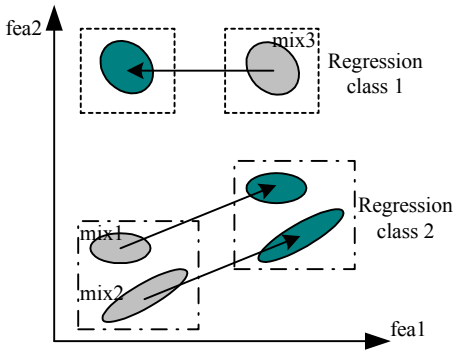


图 3-3　在声学特征空间转换均值向量的效果示意图：只转换 HMM 模型的均值向量相当于在声学特征空间中对混合成分的移动，不过没有改变他们的形状。图中表示的是有两维特征，三个混合成分，两个回归类的简单示意。

对于公式（3-31）具体到重估 CDHMM 模型的新均值 $\hat{\mu}_s$，可以表示成下面一系列公式。自适应混合成分 s 的均值向量是通过把初始模型的均值向量乘以转移矩阵计算出的：

$$\hat{\mu}_s = W_s \xi_s \tag{3-32}$$

其中 W_s 是混合成分 s 的转移矩阵，$\hat{\mu}_s$ 是自适应后的均值向量，ξ_s 是混合成分 s 的扩展均值向量，其定义如下：

$$\xi_s = [\omega, \mu_{s_1}, ..., \mu_{s_n}]^T = [\omega : \mu_s]^T \tag{3-33}$$

其中 μ_s 是原始的均值向量，而 ω 是偏移量，n 是特征维数。MLLR 的核心就是使用 ML 对自适应数据估算出转移矩阵 W_s。

3.3.2 回归类（regression classes）

一般情况下每一个混合成分对应一个转移矩阵，但这要求所有的类别都有自适应样本。类似 MAP 中 VFS，为了解决这个问题，Leggetter[?] 提出并引入了回归类的概念。一个回归类就是一个混合成分的集合，类中的所有混合成分使用同一个转移矩阵。这样做的好处是，一个回归类的转移矩阵可以使用回归类中所有的混合成分的自适应数据来估算。所以对于没有对应自适应数据的模型，也可以通过使用回归类的转移矩阵自适应。这个方法的问题在于如何选择回归

类的分法。

回归类是定义在混合成分的集合上，对于同一个状态的不同混合成分完全可以属于不同的回归类。这样对于一个给定的状态的分布自适应的灵活性大大增强。从声学特征空间可以看到，如果混合成分属于同一个回归类，他们移动的方向是相同的。而不同回归类的混合成分可能向不同的方向移动。

如何决定回归类的个数和如何分类变得十分的重要。理论上回归类的个数可以从所有混合成分共一个到每个混合各一个。但是在实际中，由于受到自适应数据量的上限，必须保证每个回归类有足够的自适应数据。在有些应用里，可以提前知道自适应数据的多少，这样回归类可以提前给出。而其他的应用则可以参用一种动态的分类方法。Leggetter[?]就给出了一种基于树的动态方法。树的根节点就是全局的回归类，每个叶子就是单个混合成分。混合成分的分类通过树的子树来完成。

在给混合成分分类时最关键的是把在自适应时可能进行相似转换的混合成分分在同一回归类。通常有两种方法来决定两个混合称分是否在一类：

声学特征：通过混合成分属于的声学单元的声学特征来分类
距离衡量：根据两个混合成分间的距离来分类，距离的选择由特征决定。

在我们第五章的实验中分别采用了这两种方法，都取得了一定的效果。

3.3.3 重估转移矩阵

下面我们给出如何重估转移矩阵。Leggetter[?]从单个高斯分布、观察值唯一、每个分布有自己的回归类的简单情况入手推出了各种情况的重估公式。重估分为几种情况：是否是多个回归类，是否是多混合成分，是否是多观察输入。文献[?, 50, 51]中给出了详细的讨论和推导。这里我们直接给出最一般的情况，即多回归类，多混合成分，多观察输入的情况。

正如节 3.3.1 中说的转移矩阵的重估是使用了最大似然（ML）估计的原理。在多回归类的情况下，目标就是得到回归类中所有元素共享的转移矩阵$\hat{W}_s$。

一个回归类是指 R 个混合成分$\{s_1k, s_2k, \ldots, s_Rk\}$的集合，其中$1 \le k \le M$，$M$是每个状态的混合个数。根据 ML 原理$\hat{W}_s$是使自适应模型参数$\hat{\lambda}$产生 P 个观察语句的概率最大的W_s：

$$\hat{W}_s = \arg\max_{W_s} P(O_P \mid \hat{\lambda}) \tag{3-34}$$

其中 O_P 是 P 个观察语句的集合，$O^{(1)}\ldots O^{(P)}$，每个语句的帧数不一定相等。观察序列 $O^{(p)}$有 T_p 帧（$O^{(p)} = o_1^{(p)}\ldots o_{T_p}^{(p)}$）。解决 ML 问题常用的方法是构造一个辅助函数：

$$Q(\lambda, \hat{\lambda}) = \sum_{\theta \in \Theta} \sum_{K \in \Omega_b} P(O_P, \theta, k \mid \lambda) \cdot \log(P(O_P, \theta, k \mid \hat{\lambda})) \tag{3-35}$$

其中 K 是一个给定的混合成分序列，Ω_b是所有可能的混合成分序列的集合，θ是一个给定的状态序列，Θ是所有可能的状态序列的集合。这样定义的辅助函数满足这样的特性，如果$\hat{\lambda}$使公式（3-35）最大，则：

$$Q(\lambda, \hat{\lambda}) \ge Q(\lambda, \lambda) \Rightarrow P(O_P \mid \hat{\lambda}) \ge P(O_P \mid \lambda) \tag{3-36}$$

也就是公式（3-34）中的概率也最大。所以可以通过求使辅助函数最大的$\hat{W}_s$来得到转移矩阵。其中概率$P(O_P, \theta, k \mid \hat{\lambda})$可以表示成模型参数的函数，于是可以求得辅助函数对于$\hat{W}_s$的偏导，使它为零，则可以得到下面等式：

$$\sum_{p=1}^{P}\sum_{t=1}^{T_P}\sum_{r=1}^{R} \gamma_{s_rk}^{(p)}(t) \Sigma_{s_rk}^{-1} o_t^{(p)} \xi_{s_rk}^{\mathrm{T}} = \sum_{p=1}^{P}\sum_{t=1}^{T_P}\sum_{r=1}^{R} \gamma_{s_rk}^{(p)}(t) \Sigma_{s_rk}^{-1} \hat{W}_s \xi_{s_rk} \xi_{s_rk}^{\mathrm{T}} \tag{3-37}$$

其中Σ_{s_rk}是第 s_rk 个混合成分的协方差矩阵，$\gamma_{s_rk}^{(p)}(t)$是引入的辅助变量，表示混合成分的发生概率，定义如下：

$$\gamma_{s_rk}^{(p)}(t) = \frac{1}{P(O^{(p)} \mid \lambda)} \sum_{\theta \in \Theta} \sum_{K \in \Omega_b} P(O^{(p)}, \theta_t = s_r, k_t = k \mid \lambda) \tag{3-38}$$

虽然这里对于协方差矩阵没有什么假设，不过本文后面讨论都是对于对角阵协方差矩阵的处理方法。如果想了解关于全协方差矩阵的处理方法，可以参考文献[?]。

为了计算方便，我们引入几个矩阵改写一下公式（3-38）。矩阵$V^{(r)}$和$D^{(r)}$定义如下：

$$V^{(r)} = \sum_{p=1}^{P}\sum_{t=1}^{T_P} \gamma_{s_rk}^{(p)}(t) \Sigma_{s_rk}^{-1} \tag{3-39}$$

$$D^{(r)} = \xi_{s_r k} \xi_{s_r k}^{\mathrm{T}} \tag{3-40}$$

于是公式（3-38）就改写成：

$$\sum_{p=1}^{P} \sum_{t=1}^{T_p} \sum_{r=1}^{R} \gamma_{s_r k}^{(p)}(t) \Sigma_{s_r k}^{-1} o_t^{(p)} \xi_{s_r k}^{\mathrm{T}} = \sum_{r=1}^{R} V^{(r)} \hat{W}_s D^{(r)} \tag{3-41}$$

公式（3-41）的右边用矩阵 Y 来表示，其元素由下式得到：

$$y_{ij} = \sum_{p=1}^{n} \sum_{q=1}^{n+1} w_{pq} \left[\sum_{r=1}^{R} v_{ip}^{(r)} d_{qj}^{(r)} \right] \tag{3-42}$$

$$y_{ij} = \sum_{q=1}^{n+1} w_{iq} \sum_{r=1}^{R} v_{ii}^{(r)} d_{jq}^{(r)} \tag{3-43}$$

其中 n 是特征的维数。注意后面的推导前提是矩阵$V^{(r)}$是对角阵，$D^{(r)}$是对称阵。矩阵$V^{(r)}$和$D^{(r)}$的乘积合成矩阵$G^{(i)}$，$1 \le i \le n$。

$$G^{(i)} = \left[g_{jq}^{(i)} \right] = \begin{bmatrix} \sum_{r=1}^{R} v_{ii}^{(r)} d_{11}^{(r)} & \sum_{r=1}^{R} v_{ii}^{(r)} d_{12}^{(r)} & \cdots & \sum_{r=1}^{R} v_{ii}^{(r)} d_{1n+1}^{(r)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{r=1}^{R} v_{ii}^{(r)} d_{n+11}^{(r)} & \sum_{r=1}^{R} v_{ii}^{(r)} d_{n+12}^{(r)} & \cdots & \sum_{r=1}^{R} v_{ii}^{(r)} d_{n+1n+1}^{(r)} \end{bmatrix} \tag{3-44}$$

最后定义一个 Z 矩阵，其元素 z_{ij} 是公式（3-41）的左边，所以有：

$$Z = \sum_{p=1}^{P} \sum_{t=1}^{T_p} \sum_{r=1}^{R} \gamma_{s_r k}^{(p)}(t) \Sigma_{s_r k}^{-1} o_t^{(p)} \xi_{s_r k}^{\mathrm{T}} \tag{3-45}$$

把 Z 和右边的 Y 相等起来，每个元素可以由下式给出：

$$z_{ij} = y_{ij} = \sum_{q=1}^{n+1} w_{iq} g_{jq}^{(i)} \tag{3-46}$$

这样转移矩阵重估$\hat{W}_s$可以按行来进行：

$$\hat{w}_i^{\mathrm{T}} = G^{(i)-1} z_i^T \tag{3-47}$$

其中 $\hat{w}_i$ 和 z_i 分别是矩阵$\hat{W}_s$和 Z 的第 i 行。

下面介绍对角的转移矩阵。上面描述的方法计算十分的复杂。不过我们可以通过把转移矩阵$\hat{W}_s$约束为一个对角阵来大大减少重估转移矩阵时的计算复

杂度。

$$\hat{W}_s = \begin{bmatrix} w_{1,1} & w_{1,2} & 0 & \cdots & 0 \\ w_{2,1} & 0 & w_{2,3} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ w_{n,1} & 0 & \cdots & 0 & w_{n,n+1} \end{bmatrix} \tag{3-48}$$

重写矩阵 $\hat{W}_s$ 中非零元素为一个向量 $\hat{w}_s$：

$$\hat{w}_s = \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{n,1} \\ w_{2,1} \\ \vdots \\ w_{n,n+1} \end{bmatrix} \tag{3-49}$$

通过定义和全矩阵同样的辅助函数 $Q(\lambda, \hat{\lambda})$，再经过类似的分析过程，可以得到和公式（3-37）相似的如下公式：

$$\sum_{p=1}^{P}\sum_{t=1}^{T_P}\sum_{r=1}^{R}\gamma_{s_r k}^{(p)}(t) D_{s_r k}^{\mathrm{T}} \Sigma_{s_r k}^{-1} o_t^{(p)} = \sum_{p=1}^{P}\sum_{t=1}^{T_P}\sum_{r=1}^{R}\gamma_{s_r k}^{(p)}(t) D_{s_r k}^{\mathrm{T}} \Sigma_{s_r k}^{-1} D_{s_r k} \hat{w}_s \tag{3-50}$$

其中矩阵 $D_{s_r k}$ 是这样定义的：

$$D_{s_r k} = \begin{bmatrix} \omega & 0 & \cdots & 0 & \mu_{s_r k_1} & 0 & \cdots & 0 \\ 0 & \omega & \ddots & \vdots & 0 & \mu_{s_r k_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \omega & 0 & \cdots & 0 & \mu_{s_r k_n} \end{bmatrix} \tag{3-51}$$

现在定义两个矩阵 $A(2n \times 2n)$和 $B(2n \times 1)$：

$$A = \sum_{p=1}^{P}\sum_{t=1}^{T_P}\sum_{r=1}^{R}\gamma_{s_r k}^{(p)}(t) D_{s_r k}^{\mathrm{T}} \Sigma_{s_r k}^{-1} D_{s_r k} \tag{3-52}$$

$$B = \sum_{p=1}^{P}\sum_{t=1}^{T_P}\sum_{r=1}^{R}\gamma_{s_r k}^{(p)}(t) D_{s_r k}^{\mathrm{T}} \Sigma_{s_r k}^{-1} o_t^{(p)} \tag{3-53}$$

这样转移向量 $\hat{w}_s$ 可以通过公式（3-50）得到：

$$B = A\hat{w}_s \tag{3-54}$$

$$\hat{w}_s = A^{-1}B \tag{3-55}$$

这就是常用的 MLLR 的自适应方法。

3.3.4 算法实现

本文分别实现了全转移矩阵和对角转移矩阵的重估算法。具体实现算法如下：

全转移矩阵的重估算法：

```
for 所有回归类
    for 所有的混合成分，s_r k
        for 所有观察序列，p
            for 所有帧，t
                计算状态发生概率 γ_{s_r k}^{(p)}(t)，公式（3-38）
            end
        end
        计算两个矩阵 V^{(r)} 和 D^{(r)}，公式（3-39）和公式（3-40）
    end

    for 每一维特征
        计算 G^{(i)}，公式（3-44）
        对 G^{(i)} 求逆
    end

    计算 Z，公式（3-45）
    计算 W_s，公式（3-47）

    for 所有混合成分
        自适应混合密度的均值，公式（3-32）
    end
end
```

对角转移矩阵的重估算法

```
for 所有回归类
    for 所有的混合成分，s_r k
        for 所有观察序列，p
            for 所有帧，t
                计算状态发生概率 γ_{s_r k}^{(p)}(t)，公式（3-38）
            end
        end
    end

    计算 A，公式（3-52）
    对 A 求逆
    计算 B，公式（3-53）
    计算 ŵ_s，公式（3-55）

    for 所有混合成分
        自适应混合密度的均值，公式（3-32）
    end
end
```

3.4 综述

本章详细介绍了我们实现的两种自适应方法：MAP 和 MLLR。这两个自适应的方法是目前说话人自适应领域使用比较广泛的方法，从第五章我们的实验中也可以看出它们有很好的自适应效果。MAP 方法通过把初始模型当作先验知识引入训练过程，致力于让模型参数对于自适应数据更精确。当自适应数据比较多时，效果很好，当数据很多时，理论上和 ML 等同。MLLR 方法是通过一个线性变换把初始的模型变换到新的说话人上，它利用 ML 原理重估出线性变换的转移矩阵。它的优点是适应速度比较快。对于自适应数据缺乏或不均匀的问题，MAP 和 MLLR 分别使用 VFS 技术和共享回归类技术来解决。本文在第五章也给出了相应的实验结果。

第四章 综合渐进自适应方法

本章描述了我们提出的一种改进并综合了 MAP 和 MLLR 的渐进自适应方法。它在自适应数据比较少的情况下也可以取得好的效果，而且这种综合方法对环境的自适应效果也很好。

4.1 引言

在给出和分析我们的综合渐进自适应方法之前，我们先来看两个相关的问题：环境自适应和 MAP/MLLR 的结合方法。因为我们希望新的渐进自适应方法不仅对说话人之间的差异有好的自适应能力，而且对环境有更强的适应能力。另外我们希望能够充分利用 MAP 和 MLLR 两种自适应方法的优点，弥补他们各自的缺点，达到更好的自适应效果。下面我们先分析一下环境自适应的特点。

4.1.1 环境自适应

同我们在第二章中提到的说话人差异类似，语音识别系统中也存在着环境的差异（Environment Variation）。我们生活的这个世界上到处都存在着各种不同的声源，我们无法避免语音识别系统工作在有噪音的环境中。当我们对计算机做语音输入时，可能有人在旁边交谈，或者有人在开门，或者空调刚好启动。当我们的语音识别系统工作在掌上电脑或者手机上时，这种环境噪声更会随着机主和机器的运动而发生变化。另外除了背景噪声，一些输入设备如麦克风产生的噪音或变化也要考虑。图 4-1 给出了一个麦克风之间差异的例子。由于环境差异的存在，环境自适应问题很自然就提出了。

目前对于环境噪音的自适应已经成为实现强健语音识别（Robust Speech Recognition）系统的一个急待解决的重要课题。虽然目前在对提取噪音不敏感

的新特征方面有很多详细的研究，但依赖特征的语音识别系统对于变化多端的现实世界还是不够鲁棒[67]。主要因为即使是训练条件和测试条件之间存在一个很小的差异，比如更换了一个麦克风或者有很小声的背景噪声，系统的识别性能都可能会大打折扣。所以目前一般使用的环境自适应的方法是：采集一定的测试环境下的数据，对系统参数进行自适应，从而达到对测试环境的更好匹配。这种自适应和说话人自适应很类似，于是很多系统也就使用了说话人自适应的技术来进行环境自适应[67, 77, 91, 97]。

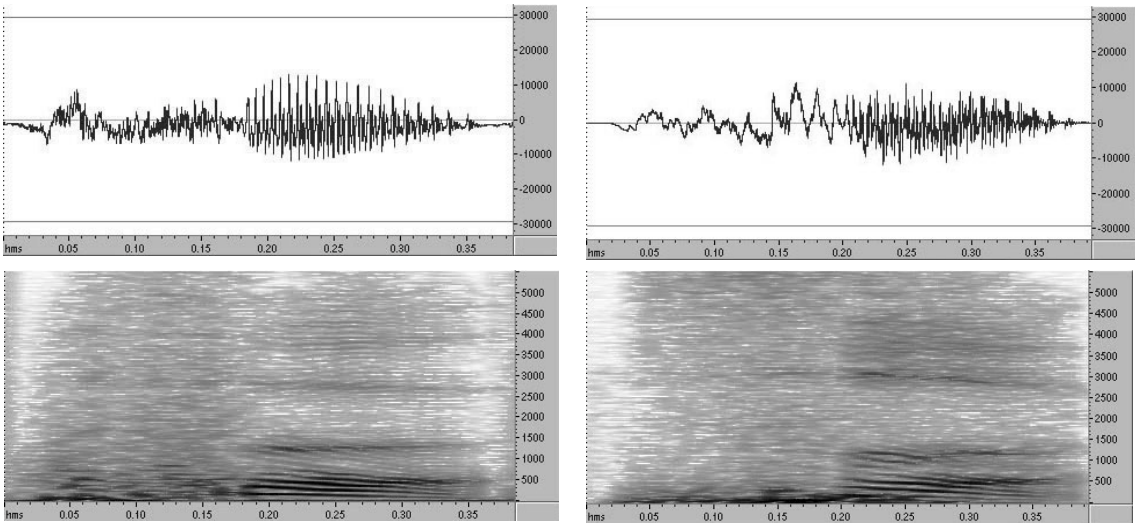


图 4-1　同一个说话人用不同的麦克风发数字“4”时的语音的时频波形图和语谱图。我们可以清楚看出环境的差异对语音的影响。

值得注意的是环境差异有一个特点，就是它对各个识别模型的影响相对比较统一。这种特性和说话人差异有些不同。所以在我们的综合渐进自适应方法中针对这个特点使用了简化的 MLLR（公用一个转移矩阵）来处理统一的环境差异，取得比较好的效果。

4.1.2 MAP 与 MLLR 的优缺点

在第三章中我们已经详细的描述了 MAP 和 MLLR 这两种常用的说话人自适应技术。这两种方法由于原理不同，所以也各自有自己的特点。

MAP 方法把初始模型提供的信息看作先验知识作为对自适应数据补充。

通过 Bayes 理论给出了结合先验知识和自适应数据的最优解。当自适应数据数量比较小时，这种结合更倾向于先验知识，从而避免了自适应数据估计的错误。当自适应数据数量增多时，这种结合就受先验知识牵制变小，从而防止了忽视自适应数据的后验知识。当自适应数据不断增加时，自适应效果将稳步提高。极端情况就是，自适应数据足够充足时，它将收敛在特定说话人系统。所以 MAP 有很好的渐进性，可以充分利用语音数据的细节信息。但它也有相应的缺点。第一，MAP 的自适应速度相对比较慢，需要的自适应数据比较多。当自适应数据比较少时，自适应效果不好。然而许多的语音识别应用，如语音指令控制（Command&Control）系统，都不可能提供足够的自适应数据。因为让用户在使用前经过一个复杂的训练过程是不现实的，这一点和听写机不一样。同时，由于自适应往往是在音素层次上进行，每个状态可利用的自适应数据比较有限，当数据量少时，也有可能造成与最初意愿相背的结果。第二，MAP 对初始模型的精确性要求比较高。因为 MAP 是通过把初始模型参数作为先验知识以某种形式“融入”到自适应样本中，所以先验知识的选择至为重要，在很大程度地影响了自适应效果。在初始的模型与新说话人的语音数据特性相差较远的情况下，识别系统性能的改善不大，较严重时甚至有可能使识别率降低。

MLLR 方法是通过一些线性转换来对初始模型进行自适应的。这种方法的优点是比较简单，而且自适应速度比较快。在自由参数比较少时，即使自适应数据量不足，MLLR 方法也可以获得较理想的效果。不过它在自适应数据量比较足的情况下，就不如 MAP 效果好。MLLR 最大的缺点在于不能充分利用自适应数据的信息。由于待估计的参数数量较少甚至有时仅有一个，因此即使新的说话人能够提供更多的语音样本，系统性能有可能过早地维持在一种饱和状态，无法随着个人语音信息的逐步增加而得到进一步的改善。

从上面的分析中可以看到，MAP 和 MLLR 方法在单独使用时存在各自的优势和缺点，并且是一种互补的关系。因此很自然的想法就是在自适应系统中综合使用两种方法，让它们互相发挥各自的长处，弥补不足。在下一章的实验中，我们可以看到这种结合 MAP 和 MLLR 的方法确实效果很好。进一步，在综合渐进自适应方法中，我们也分别利用了 MAP 和 MLLR 的特点，力图充分发挥它们各自的优点，使得渐进自适应更加鲁棒。

4.2 综合渐进自适应

所谓渐进（incremental）的自适应，就是在识别系统运行过程中逐渐调整参数，不断的使用新的数据来自适应。渐进自适应有时又被称为“动态自适应”或“在线自适应”。它的提出主要是因为许多实际的语音识别系统一次不能提供足够多的自适应数据，但可以不断的提供一定的自适应数据，用来渐进提高系统性能。这些渐进的自适应数据可以是用户分批专门提供的（有监督的），也可以是用户使用时的语音（无监督的）。如果是后者，即边使用边获得自适应数据，通常这种渐进自适应过程是可以不为用户所知道的。渐进自适应的好处是它可以不断的进行自适应，动态的提高系统性能，并且随着自适应数据的增加，效果趋近于 SD 系统。所以这种方法更适合实用系统。

MAP 方法具有良好的渐进性，当自适应数据逐渐增加时它的性能逐渐变好。在理论上，当训练数据趋近于无穷时，MAP 可以与用充分语料做最大似然方法训练所得的效果等价。所以，MAP 方法比较符合渐进自适应的目标要求，适于作为使系统由 SI 向 SD 渐进过渡的自适应手段。目前的渐进自适应系统也通常采用了 MAP 或类 MAP 的方法。

不过使用MAP的渐进自适应也存在我们前面一节分析提到的两个问题。一个是当自适应数据不足时，MAP自适应的效果很差。而在渐进自适应中，这个问题很可能发生，因为在每一次自适应时的自适应数据可能只有几个词或者短句子。另一个问题是MAP需要有比较精确的初始模型，这一点当自适应数据比较少时更显重要。我们从公式（3-15）中可以看到，如果自适应数据比较少，MAP的结果和初始模型的关系更大。于是怎样使得初始模型变的更精确变的很重要，它直接影响自适应效果。另外在环境相对嘈杂的情况下，自适应也希望有比较好的初始模型。

为了解决上面的问题，我们在基于MAP的渐进自适应中引入了一个简化的MLLR模块，用来处理环境差异和说话人生理差异，为MAP提供更好的初始模型。而且我们知道，MLLR是利用线性转移矩阵对CDHMM的参数进行转换的，它即使在自适应数据比较少时也可以比较快速的达到自适应效果，所以MLLR模块的引入也可以提高自适应速度。同时我们根据MAP和MLLR不同的特点，提出了一

种新的综合渐进策略。在后面一章的实验中，我们将可以看到这种新方法在强健语音识别中的良好性能。下面我们详细介绍这种新的综合渐进自适应方法。

4.2.1 整体框架

图 4-2 给出了我们的综合渐进自适应的整体框架。自适应系统由简化的 MLLR 自适应模块和渐进的 MAP 自适应模块组合而成。

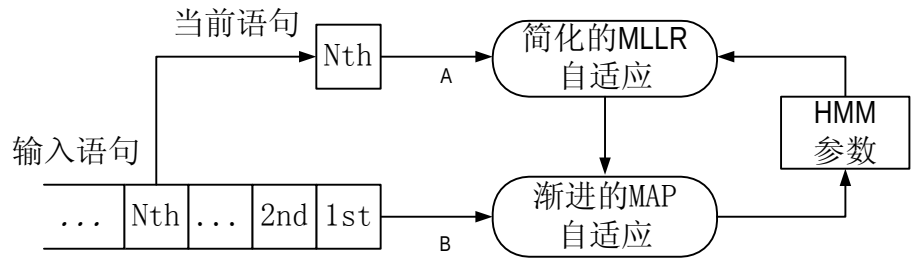


图 4-2 综合渐进自适应方法框架示意图

在下面的两节中我们分别介绍这两个模块和它们采用的综合渐进自适应策略。

4.2.2 MAP 模块与 MLLR 模块

MAP 和 MLLR 模块在综合渐进自适应中担任着不同的作用。我们在第二章讨论过说话人之间的差异问题，我们知道说话人之间的差异主要是由生理差异和说话习惯差异造成的。生理差异，它与说话人的声带形状、声道长度、尺寸等有关。这种差异与具体哪个发音无关，在这一点上它与环境差异很相似。而说话习惯的差异则是在音素单元层上的，由个人的说话风格、口音等造成。所以针对这些差异的特点和 MAP 与 MLLR 的各自优点，我们使用 MAP 模块来精细刻画基于音素层次的差异，而使用简化的 MLLR 模块来对付对于所有音素所共同存在的差异（生理差异和环境差异）。

简化的 MLLR 模块

在第三章中我们给出了 MLLR 方法的详细介绍与实现。在我们的综合渐进自适应方法中 MLLR 用来给 MAP 提供更精确的初始模型。这里它通过自适应数据对上一次自适应后的模型参数估算出转移矩阵 W。然后使用下面的公式进行第一步自适应。

$$\mu_{mllr,k}^{l} = W_{globe}^{l}\mu_{map,k}^{l-1} \qquad (4\text{-}1)$$

其中 $\mu_{map,k}^{l-1}$ 是上一次（l-1 次）自适应的结果，$\mu_{mllr,k}^{l}$ 是经过简化 MLLR 自适应的结果，也是 MAP 模块的输入。

这里对 MLLR 进行了两个简化。第一，由于自适应数据的缺少，为了减少待估算的参数个数，我们使用了对角阵的转移矩阵 W。这样做可以大大减轻了计算时的复杂性，提高了自适应的速度。第二，根据我们引入 MLLR 的目的是对付环境差异和说话人的生理差异这种高于音素层次的差异，我们对所有的模型使用了同一的转移矩阵 W_{globe}^{l}。这样所有模型的自适应数据共同参与重估转移矩阵 W_{globe}^{l}，也解决了自适应数据不足的问题。重估出来的 W_{globe}^{l} 用来对所有的 $\mu_{map,k}^{l-1}$ 进行自适应。

通过引入这个简化的 MLLR 模块，即提高了对环境自适应的能力，为 MAP 提供了更精确的初始模型，又加快了自适应的速度。在后面一章的实验中取得很好的效果。

渐进的 MAP 自适应模块

MAP 模块中，我们使用的就是第三章中讲的 MAP 方法。主要的重估公式是公式如下：

$$\mu_{map,k}^{l} = \frac{\tau_k^l \mu_{mllr,k}^{l} + n_k^l m_k^l}{\tau_k^l + n_k^l} \qquad (4\text{-}2)$$

其中 $\mu_{mllr,k}^{l}$ 是简化的 MLLR 模块的输出。MAP 模块就是以它作为初始模型，再利用自适应数据进行自适应。最后得到的 $\mu_{map,k}^{l}$ 实际是 MLLR 的结果 $\mu_{mllr,k}^{l}$ 和训练的自适应数据的均值 m_k^l 的一种线性加权。而在这个加权过程中 τ_k^l 是一个十

分重要的因素。它的得到方法我们在第三章中也已经介绍过了。可以使用从很多初始模型的参数或训练数据中计算出，也可以象我们实验中使用了初始模型均值 $\mu^{l}_{mllr,k}$ 和自适应样本均值 m^{l}_{k} 的 Mahalanobis 距离的倒数来估计。详细的计算公式请参考第三章中公式（3-15）到（3-22）。

MAP 模块的主要作用是较为精确的对说话人差异和环境差异进行自适应，特别侧重这些差异对每一个模型的不同的影响。同时 MAP 方法良好的渐进性也确保整个综合渐进自适应的效果随自适应数据增加而变得更好。

4.2.3 渐进的策略

渐进的策略，主要指在渐进自适应方法中如何使用自适应数据来对模型自适应。通常渐进的策略可以划分为两种基本的方法，如图 4-3 中的策略 A 和策略 B 所示[68] 。策略 A 是只使用当前一句（批）的自适应数据来重估最近已经自适应过的 CDHMM 模型参数。这种方法的优点是它的动态性能比较好，能够使系统不断适应新的变化。不过它的缺点也很明显，就是当目前的语句十分短小时，这种重估很不可靠，系统效果也会变差。另一种策略 B 是使用所有累计的语句来自适应 CDHMM 的参数。这种方法的好处是充分利用了所有的自适应数据，所以自适应性能比较好。但是这种方法的问题是必须在每一次自适应时计算所有的样本，这样计算量不断增加。

针对我们提出的综合渐进自适应方法，我们提出了一个新的综合渐进策略，见图 4-3 中的 C 策略。这种新策略是根据 MAP 模块和 MLLR 模块的不同用途和两种方法自身的特点对他们采用不同的渐进策略。

对于简化的 MLLR 自适应模块，我们只使用当前的语句来进行自适应。因为我们的 MLLR 方法中所有模型只使用了一个回归类，也就是说所有的模型的自适应数据都用来重估同一的转移矩阵。这样当前的语句的数据量应该是足够用来自适应的。这样做还有一个好处，就是适合于环境发生变化的情况。由于这个模块主要是对付环境差异的，而在实际的系统中，环境有可能发生变化，于是 MLLR 模块用当前的数据来自适应就可以动态调整系统的参数。

而对于 MAP 模块，它的主要目的是消除模型（音素）层的差异。为了能够对每个模型的细微特点进行自适应，同时考虑到 MAP 方法对自适应数据量的需要比较大，所以我们在 MAP 模块中使用所有累计的自适应数据。另外由于环境可能发生变化，所以 MAP 的自适应是在 MLLR 模块的结果上进行，这样保证了 MAP 的初始模型的精确性。当然为了解决计算所有样本的参数计算量太大的问题，我们采用了两种解决办法。一个是在计算时可以利用上一次 MAP 计算的一些中间值，从而减少总的计算量[?]；另一个是通过实验我们发现，综合方法在经过很少几次的渐进自适应就有很好的效果。所以不必计算所有累计的样本，而可以使用一个缓冲区，只计算最近的几次的累计样本。

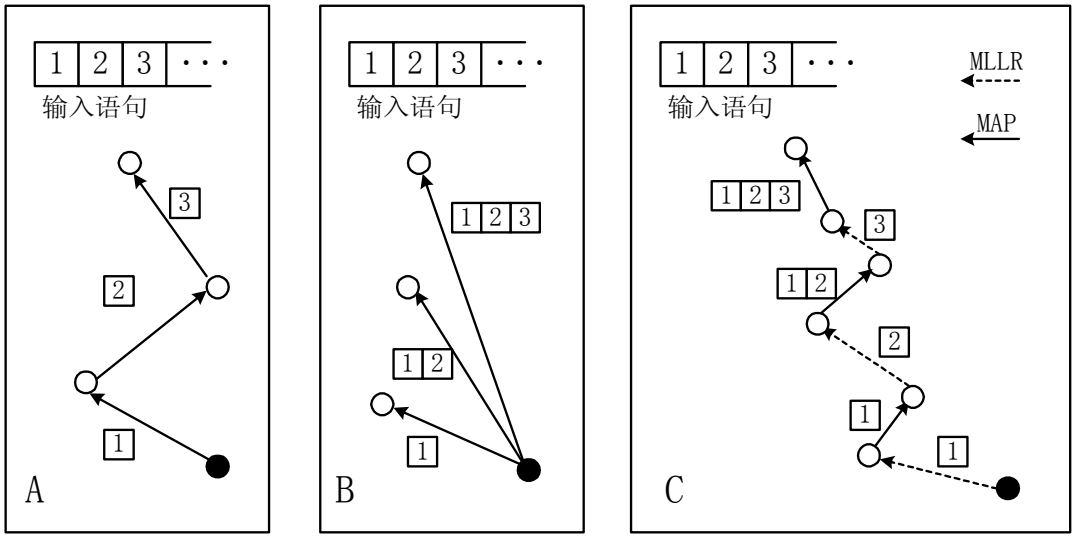


图 4-3 渐进自适应策略示意图： A 和 B 是常用的两种渐进策略，C 是我们使用的新的综合渐进策略。

4.3 综述

本章描述了我们提出的一种快速的适合于强健语音识别的综合渐进自适应方法。这种方法通过在渐进的 MAP 方法中引入了一个简化的 MLLR 模块，用来对付环境差异和说话人的生理差异。同时针对综合渐进自适应方法的特点我们配合使用了一种新的使用自适应数据的渐进策略。在后面的实验中，这

种综合方法即使在自适应数据比较少的情况下也可以取得好的效果。在无噪音和有噪音的环境中分别可以降低 23.03%和 29.69%的识别字错误率。所以我们可以说这种新的综合渐进自适应方法对于说话人差异和环境差异都有很好的效果，适合强健语音识别系统的要求。

第五章 实验与讨论

本章给出了基于 HMM 模型语音识别系统的自适应实验结果，并对各种方法的实验结果进行了比较和分析。下面首先介绍一下我们的自适应系统运行的实验环境。

5.1 实验环境

5.1.1 实验数据

我们的语音识别系统和自适应实验都是基于语音数据库 CIDS[69，70]（Chinese Isolate words, Digits and Syllables）进行的。CIDS 语音数据库是由我们实验室（清华大学智能技术与系统国家重点实验室）收集和整理。它的内容包含五个子数据库，共 102 个人的普通话语音，以及少量的英文数字语音。其详细组成情况见表 5-1。

表 5-1　语音数据库 CIDS 的具体内容组成

类　型		音 节	数 字	数字串	双字词	英文数字
说话人	男	62	62	62	62	14
	女	40	40	40	40	22
样本种类数		1322	10	70	1000	10
每人样本数		1	1	1	1	5
总样本数		134,844	1,020	7,140	102,000	1,800

表中的音节、数字、数字串和双字词都是汉语普通话语音。其中，数字串包含 140 个普通话数字串，每个串包含 5 个数字，每个数字{0~9}在所有数字串中的出现概率基本相同。音节由从几本词典中摘取出来的 1322 个带音调的普通话音节组成；如果不考虑音调，共有 411 个基本音节（base syllable）。双字词由从新闻报纸中摘取的 1000 个普通话孤立词组成，每个词的长度为 2 个音节，其中存在发音十分相似的词语。数据库的语音主要由 102 个人(40 女，62 男)提供，年龄分布在 17 到 35 岁之间，来自中国的 25 个省，其中有一些人带有轻微的地方口音。总的来说该数据库对汉语普通话音节进行了比较好的覆盖。在我们的开发和研究工作中经常使用数据库 CIDS，实验证明，它是一个令人满意的数据库。

除了语音数据库 CIDS 外，为了对环境自适应进行实验，我们又采集了一部分新的数据。数据内容也是汉语普通话的数字和数字串，每人 10 个孤立数字和 70 个数字串（同 CIDS 中）。环境分别选取：1）更换麦克风，2）录自掌上电脑（PalmPC），3）带背景噪音。为了区分方便，我们在后面的实验中，称这些数据的集合为测试集 B，而称选自 CIDS 中的干净语音数据的集合为测试集 A。

对于上述数据库中所有的语音数据，我们都采用表 5-2 列出的处理方法进行了预处理和特征分析。

表 5-2　语音信号分析

	语音信号分析
预加重（preemphasis）	$1-0.95z^{-1}$
采样频率（sampling rate）	11.025KHz
加窗（window functions）	哈明窗（Hamming window）
帧长（frame length）	23.4ms(256 个采样点)
特征（feature）	16 阶 LPC 倒谱（LPC-Cepstrum） +16 阶一阶差分 LPC 倒谱 +1 维差分对数能量，总共 33 维特征

5.1.2 实验系统框架

我们的实验系统都是基于 CDHMM 的识别系统。每个汉语音节（无音调的基本音节）或数字一个上下文无关的 HMM 模型。每个 HMM 模型有 3 个状态，每个状态的分布函数有 5 各高斯混合分量，高斯函数的协方差矩阵采用对角阵。HMM 的转移概率矩阵采用自左向右的无跳转的形式，见图 5-1。各个模型之间目前没有共享分布函数。

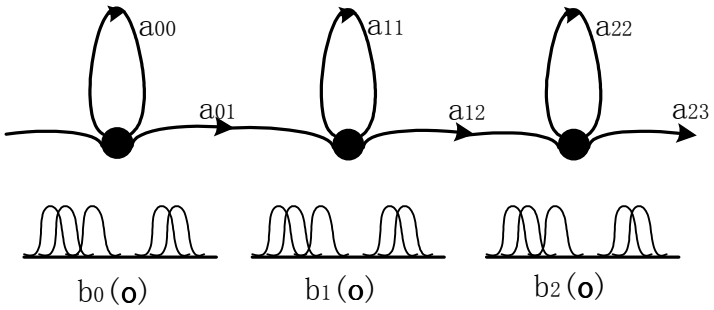


图 5-1　实验系统中 CDHMM 模型的示意图

自适应实验系统的框架如图 5-2 所示。实验中的非特定人系统都是使用 CIDS 数据库中的前 40 个男性语音训练而成的。自适应数据和测试数据则从训练集外的 20 个男性语音中选取。实验分别对连续数字识别系统、孤立音节识别系统和双字词识别系统进行了自适应实验。关于这些系统的组成、训练和识别方法请参考文献[71]，里面有较详细的介绍和讲解。

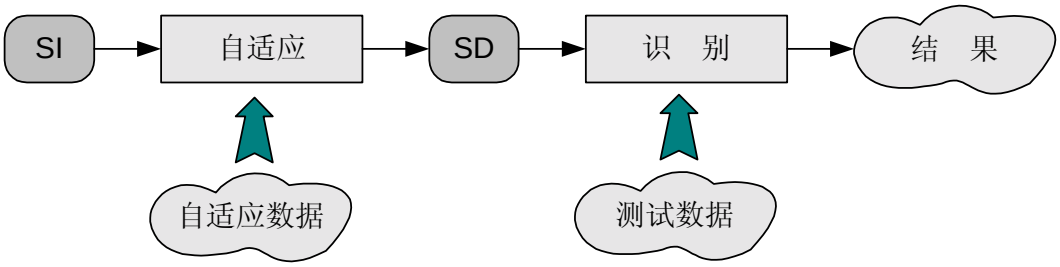


图 5-2　自适应实验系统的框架结构

5.2 实验与讨论

这一节里我们将分别给出基于 MAP 方法、MLLR 方法、综合渐进自适应方法的自适应实验结果，并对实验结果进行仔细的分析。

5.2.1 MAP 与 MLLR 方法的自适应实验

我们的自适应实验主要在连续数字语音识别系统上开展。我们的初始的非特定人系统是通过语音数据库 CIDS 的 40 个男性的语音训练而成，每人使用了 10 个孤立数字的语音和 40 串连续数字串。测试时对每个人的 70 个连续语音串进行识别（本文中数字识别的所有的测试数据都是如此测试），系统的识别效果如下表：

表 5-3　非特定人系统的性能

SI	*训练集（40 人）*	*测试集 A（12 人）*	*M1*	*M2*	*M3*
串错误率	38.07%	43.71%	55.71%	82.85%	40.00%
字错误率	11.58%	17.38%	21.42%	33.14%	12.00%

表中测试集 A 由 CIDS 数据库中训练集（40 人）外的 12 个人的语音组成，而 M1、M2 和 M3 是其中的三个人的语音。表格中的字错误率是采用如下公式计算:

$$E = \frac{R + D + I}{N} \qquad (3.1)$$

式中 E 是字错误率，R 是替换的个数，D 是遗漏的个数，I 是添加的个数，N 是总字符数。

从表 5-3 中可以清楚的看到，非特定人系统对于训练集，相对测试集有更好的效果[4]。没有训练的字错误率大约比训练过的高 50%。而对于个别的人，比如 M2，其识别效果就更难以接受了。这也是我们在第一、二章中提到引入自适应

[4]总体的识别错误率比较大，一个重要的原因是我们所录制的连续数字串不是随机的，它包含了很多容易识别错误的序列，诸如 11，22，55。这一类的组合出现次数太多，从直观上来看，这些组合也是很容易识别错误的。

技术的原因之一。所以，下面的自适应方法的实验就是针对测试集中的没有训练过的人的语音进行的。另外，在衡量连续语音识别的识别效果时，识别的字错误率是最主要的参数，于是我们也通过它的降低幅度来衡量自适应的效果。

5.2.1.1 MAP 与 MLLR 方法的自适应效果

首先我们通过对比实验来看看 MAP 和 MLLR 的自适应效果。这里的 MAP 自适应方法请参见第三章第 2 节，其中τ是使用 Mahalanobis 距离来估计的。而 MLLR 则使用的是对角转移矩阵，每个状态的混合分布共享同一个转移矩阵。自适应数据使用 10 个连续的数字串。实验的结果如表 5-4 所示。其中为了比较，SI 一栏给出初始的非特定人系统的性能；SD 一栏是使用该测试人的 40 个串训练的特定人系统的结果；而 ML 一栏是使用和自适应同等数据量（10 个串）的 ML 训练的结果。

表 5-4 MAP/MLLR 自适应性能

字错误率	**SI**	*ML*	*MAP*	*MLLR*	*SD*
测试集 A	17.38%	53.45%	13.00%	11.38%	8.55%
M1	21.42%	43.14%	16.85%	15.71%	9.71%
M2	33.14%	66.28%	22.85%	20.85%	14.85%
M3	12.00%	59.14%	9.42%	5.71%	10.28%

从结果中我们可以清楚的看到，通过 MAP 和 MLLR 自适应后的模型比初始的非特定人的模型字错误率有了很大的降低，分别降低了 25.20%和 34.52%（自适应效果的计算方法：(SA-SI)/SI）。而如果使用同样的数据在初始的系统上做 ML，错误率却明显提高，说明这么少的数据对于 ML 训练是不够的。当然这里自适应的效果还没有达到 SD 的水平，不过在下一个实验中可以看到，随着自适应数据的增多，自适应效果会向 SD 逼近的。

5.2.1.2 自适应数据量与自适应效果的关系

在这个实验里，我们不断的增加自适应数据，看看 MAP 和 MLLR 的自适应效果。实验中的方法同上个实验。实验结果见表 5-5 和表 5-6。

表 5-5 MAP 自适应性能随自适应数据量的变化

字错误率(%)	*SI*	*2*	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*	*20*
测试集 A	17.38	26.23	19.14	16.83	14.35	13.00	12.98	12.00	11.73	11.50	11.26
M1	21.42	19.42	17.42	16.85	16.28	16.85	16.28	16.00	15.14	15.14	14.85
M2	33.14	36.00	35.42	24.57	24.00	22.85	22.28	20.85	22.00	20.00	19.42
M3	12.00	11.42	10.85	10.00	8.00	9.42	9.14	6.96	5.42	5.42	4.85

表 5-6 MLLR 自适应性能随自适应数据量的变化

字错误率(%)	*SI*	*2*	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*	*20*
测试集 A	17.38	29.73	18.04	13.95	12.16	11.38	10.78	9.38	9.38	9.33	9.30
M1	21.42	33.71	22.85	18.28	15.71	15.71	14.00	12.85	13.42	13.14	13.14
M2	33.14	47.71	27.14	20.57	20.57	20.85	19.42	17.71	17.71	16.57	17.14
M3	12.00	22.00	8.28	6.85	5.71	5.71	5.42	5.14	5.42	5.42	5.42

从结果中可以看到，当自适应数据从 2 个增加到 20 个，自适应的效果无论是 MAP 还是 MLLR 都不断的提高。当自适应数据很少时，比如只有 2 个时，自适应的效果还不如不自适应的 SI 的效果，这是因为自适应数据太少的缘故。而当自适应数据增加到 14 个以后，效果改善的幅度也变得比较平缓，说明自适应随数据增加将逐渐趋于饱和。对比 MAP 和 MLLR 两种自适应方法，可以看到 MLLR 的自适应速度比较快，在自适应数据比较少的情况下效果要优于 MAP。实验数据的曲线图参见图 5-3。

5.2.1.3 MAP/MLLR 相结合的实验

在第四章中我们分析过 MAP 和 MLLR 的各自优缺点，他们有一定的互补性。所以很自然的可以同时结合使用这两种方法，这样既可以保证自适应的渐进性，又可以加快自适应的速度。这个实验就是 MAP 和 MLLR 相结合的实验。表 5-7 和表 5-8 分别给出了先做 MAP 再做 MLLR 和先做 MLLR 再做 MAP 两种方法的自适应性能。

表 5-7 MAP+MLLR 自适应的性能

字错误率(%)	*SI*	*2*	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*	*20*
测试集 A	17.38	31.57	18.11	13.09	11.40	10.07	9.73	8.92	8.73	8.47	8.57
M1	21.42	32.28	25.14	18.28	15.71	15.14	14.57	12.57	12.28	12.57	12.85
M2	33.14	43.42	29.42	22.00	19.71	19.71	16.57	16.57	15.42	15.14	13.42
M3	12.00	21.71	8.28	6.00	6.00	5.14	5.71	5.71	5.42	4.85	4.85

表 5-8 MLLR+MAP 自适应的性能

字错误率(%)	*SI*	*2*	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*	*20*
测试集 A	17.38	36.97	23.80	14.40	12.45	10.90	10.42	9.54	9.26	9.02	8.69
M1	21.42	32.57	23.14	18.57	18.00	17.14	16.28	14.00	13.42	14.00	13.71
M2	33.14	55.14	32.00	20.28	23.42	18.28	16.85	18.00	16.57	15.71	15.71
M3	12.00	25.71	7.42	7.71	6.28	5.71	5.71	5.42	5.14	4.84	4.57

从实验结果看，两种结合的方法都有好的自适应效果。从这两个表看，两者之间效果没有太大差别，所以先做哪种方法在这里并不重要。和表 5-5 和表 5-6 的结果比较，MAP 和 MLLR 结合的方法比单独使用 MAP 和 MLLR 的效果

要好，当然带来的是程序的复杂度相应的增加。我们可以把上面几个实验的结果汇集成图 5-3，这样上面的结论可以更加直观的看到。

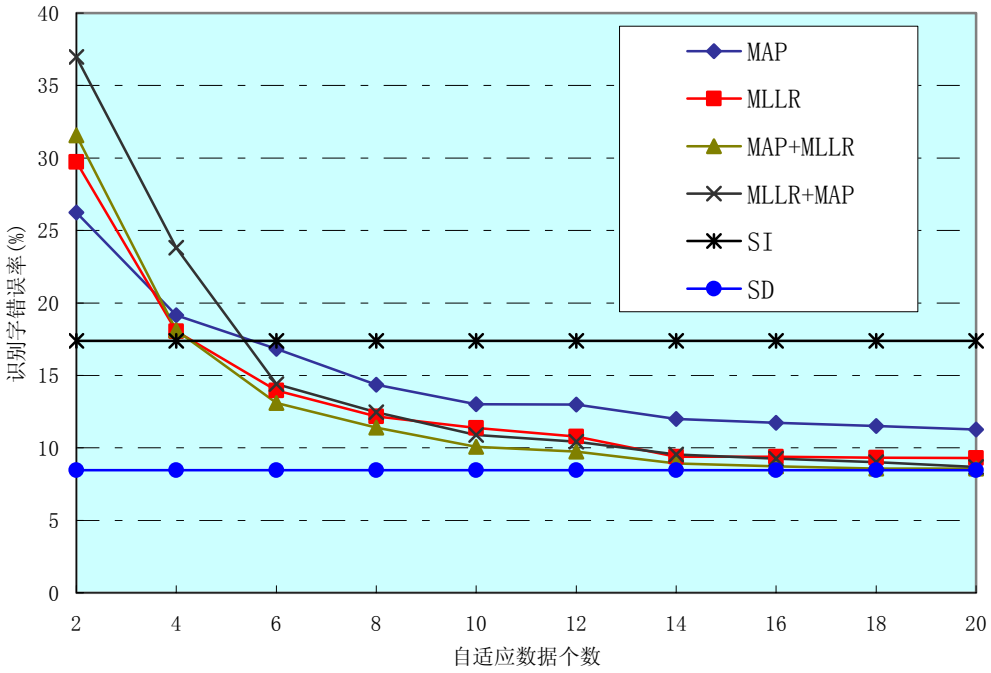


图 5-3　各种自适应方法效果比较

5.2.2 MAP/MLLR 对环境自适应的效果

环境自适应的实验我们是对测试集 B 进行的。自适应数据 10 个，实验结果见表 5-9。测试集 B 中的数据分为三类：1）使用不同的麦克风；2）使用 PalmPC 录制；3）在有噪音的环境下录制。表中详细给出了各种方法对各种数据的自适应效果。

表 5-9　MLLR 和 MAP 对环境自适应的性能（1）

字错误率	*SI*	*MAP*	*MLLR*	**MAP +MLLR**	*MLLR+MAP*
测试集 B（平均）	30.04%	24.76%	22.80%	24.19%	21.04%

换 MIC	25.14%	23.57%	22.85%	23.42%	22.00%
PalmPC	28.00%	20.28%	15.28%	17.57%	14.00%
噪音	33.71%	30.42%	30.28%	31.57%	27.14%

我们从实验结果看，几种的自适应方法对环境的自适应都有一定的效果。进一步由表 5-10 我们可以看到，比较而言 MLLR 和 MLLR+MAP 相对效果更加。主要原因是因为 MLLR 共享一定的回归类，对环境的共同影响（如更换麦克风造成的影响）刻画能力比较强，而 MAP 则更侧重于细致刻画每个模型的特点。由此我们在新的综合渐进自适应方法中使用了简化的 MLLR 模块来对环境差异进行自适应。

表 5-10 MLLR 和 MAP 对环境自适应的性能（2）

字错误率降低比率	**SI**	*MAP*	*MLLR*	*MAP+MLLR*	*MLLR+MAP*
测试集 B	Baseline	17.57%	23.86%	19.47%	29.96%

下面的表 5-11 是对于同一个说话人在不同环境下的识别效果和自适应效果，自适应数据仍然是 10 个。注意的是这个说话人是 SI 训练集中的一员，所以其干净语音的识别效果极好。而环境因素中的某一个环节稍有变动，如更换了录音的麦克风，其识别效果明显下降。好在我们可以通过自适应技术来进行一定的弥补。

表 5-11 对于同一个说话人的环境自适应的性能

字错误率	*SI*	*MAP*	*MLLR*
干净的语音	0.07%	0.06%	0.05%
换 MIC	18.00%	17.42%	16.85%
PalmPC	28.00%	20.28%	15.28%
噪音	28.85%	24.57%	28.00%

5.2.3 MAP/MLLR 对性别自适应

这个实验是这样进行的。我们使用男音的非特定人系统对女音进行自适应实验。男音的系统就是前面实验使用的系统，女音使用 CIDS 中的 15 个女音。自适应数据使用 10 个。自适应的效果由表 5-12 给出。我们可以看到 MAP 和 MLLR 都能大大降低错误率。这个实验只是为了从另一个方面说明 MAP 和 MLLR 的自适应效果，在实际系统中性别的差异一般不使用模型转换自适应技术来对付，而通常都是利用两个性别孤立的模型来解决的。

表 5-12 MLLR 和 MAP 对性别自适应的性能

字错误率	*SI*	*MAP*	*MLLR*
平均	69.62%	52.35%	35.40%
W1	66.57%	48.00%	33.14%
W2	58.00%	40.57%	36.85%
W3	78.85%	66.57%	40.85%

5.2.4 MAP/VFS 与 MLLR 共享回归类

这一部分的实验是为了检验向量域平滑（VFS）和 MLLR 中的共享类技术的效果。由于这两个技术是用来解决对于模型比较多，每个模型的自适应数据不足或没有的问题。所以下面的实验是基于音节和双字词识别的自适应实验，模型（音节）共有 411 个。这里我们使用了 CIDS 中的训练集外的一个说话人 M4[5]的语音作为测试数据。

5.2.4.1 MAP/VFS

首先我们对第三章 3.2.4 节中提到的向量域平滑（VFS）技术基于音节识别和双字词识别做了两个实验，结果见表 5-13 和表 5-14。

[5] 这个说话人可以算是非特定系统的 outlier，识别效果很差。

表 5-13 音节识别中 MAP/VFS 的效果（错误个数，共 1322 个音节）

自适应数据个数	**SI**	*MAP*	*MAP_VFS（100/30）*	*MAP_VFS（100/50）*	*MAP_VFS（50/30）*	*MAP_VFS（50/50）*
100	706	754	678	702	673	690
500	706	690	637	639	633	642

表 5-13 列出的是我们分别使用 100 个音节和 500 个音节作为自适应数据的对音节识别的实验。不使用 VFS 技术的 MAP 方法用 100 个自适应数据自适应后，错误个数反而变的比 SI 还多。这是因为很多模型没有相应的自适应数据对应，而只对有样本的模型自适应会影响其他模型的识别效果。相比用 500 个音节自适应的效果稍好一些。而使用了 VFS 技术后，自适应的效果都有了明显的好转。表 5-13 中第一行中括号中的数字（*a*/*b*）分别表示：*a* 是指公式（3-26）中的 *s* 的取值，*b* 是指公式（3-25）中的 *N(q)*的元素个数，即通过几近邻来插值和平滑。

表 5-14 双字词识别中 MAP/VFS 的效果（错误个数，共 1000 个词）

自适应数据类型	**SI**	*MAP*	*MAP_VFS（100/30）*	*MAP_VFS（50/30）*
孤立音节	461	478	435	435
双子词	461	479	436	431

表 5-14 列出的是我们分别使用孤立音节和双字词作为自适应数据的对双字词识别的实验结果。可以看到不使用 VFS 的 MAP 的自适应效果也都比 SI 差，而使用了 VFS 后，自适应的效果也都有了明显的好转，平均错误个数大约比不用 VFS 的下降了 10%。

5.2.4.2 MLLR 中的共享回归类

当使用 MLLR 方法自适应时，同样会遇到自适应数据缺乏和不均匀的情况，这一点我们可以从表 5-15 的自适应数据 100 个和表 5-16 的用双字词作自适应数据的结果中看到，引入 MLLR 后错误个数都比 SI 还高。这时我们就应该采

用我们在第三章 3.3.2 节中提到的共享回归类的技术。

共享回归类的技术就是在计算转移矩阵时，一些模型共用相同的转移矩阵。这里最关键的是如何共享，即怎样划分回归类。实验里我们使用了两种常用的方法。一个是根据声学特性人为分类，把 411 个无声调音节分为 34 个类（下面称为 C34）。另一种方法则根据对 HMM 模型参数的聚类得到，把所有音节归入 9 类（下面称为 C9）。这两种的方法的具体的分类方法请参考文献[72]，附录中给出了分类的结果。同时我们也采用了所有模型共用同一个转移矩阵，即所有音节属于同一个回归类（下面称为 C1）。

表 5-15 音节识别中 MLLR 共享回归类的效果（错误个数，共 1322 个音节）

自适应数据个数	*SI*	*MLLR*	*MLLR（C1）*	*MLLR（C9）*	*MLLR（C34）*
100	706	853	669	676	656
500	706	584	673	657	578

表 5-16 双字词识别中 MLLR 共享回归类的效果（错误个数，共 1000 个词）

自适应数据类型	**SI**	*MLLR*	*MLLR（C1）*	*MLLR（C9）*	*MLLR（C34）*
孤立音节	461	456	438	426	363
双字词	461	536	422	403	379

从表 5-15 和表 5-16 中，我们可以看到在使用共享回归类后，MLLR 的自适应效果都变好了，说明共享回归类的方法可以解决自适应不足和不匀的问题。而 C9 和 C34 的效果都比较好，说明两种分类方法都是可取的。比较细微的差别 C34 更好一些，这是因为我们的 C9 聚类使用了最简单的聚类方法，还有待提高细化。从两张表中也可以发现即使使用统一的转移矩阵（即 C1）也可以达到一定的效果，于是在后面的综合渐进自适应方法中我们就使用了这个方案。

5.2.5 综合渐进自适应方法

为了检验我们的综合渐进自适应方法，我们分别对无噪声的干净语音（测试集 A 中的 10 个人的语音）和在有噪声的环境下的语音（测试集 B 中的 6 个人的语音）进行了综合方法与 MAP 渐进自适应方法的比较实验。实验中，我们以每 4 句语音组为一组，进行了连续的 3 组渐进自适应实验。实验结果如表 5-17 和图 5-4 和图 5-5，其中的识别字错误率是在测试集上的平均值。

表 5-17 中给出了在经过两组数据的渐进自适应后的识别字错误率。我们可以看到，即使在噪音的环境下，我们的综合方法也明显好于 MAP 的效果。在 8 个自适应数据时，对于测试集 A 和 B 分别降低了 SI 的字错误率 23.03%和 29.69%，平均比渐进 MAP 方法约好 10%。

表 5-17　综合渐进自适应方法效果的比较：SI 系统、MAP 及综合渐进自适应在经过 8 个自适应数据后的识别字错误率（%）

字错误率（%）	*SI*	*MAP*	*综合方法*
测试集 A (10 人平均)	17.41	13.85	13.40
测试集 B (6 人平均)	30.04	24.71	21.12

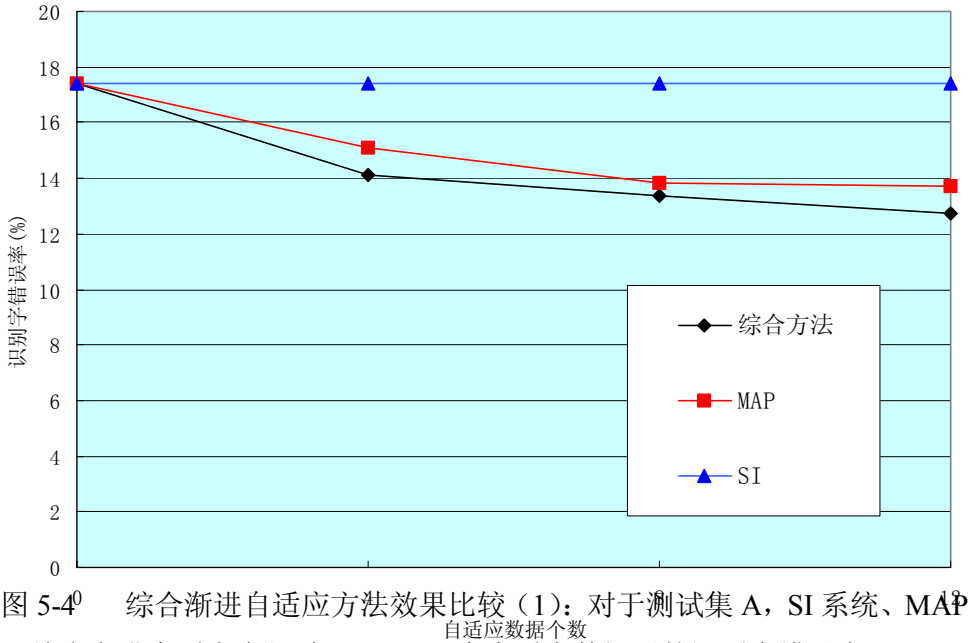


图 5-4 综合渐进自适应方法效果比较（1）：对于测试集 A，SI 系统、MAP 及综合渐进自适应在经过 4、8、12 个自适应数据后的识别字错误率（%）

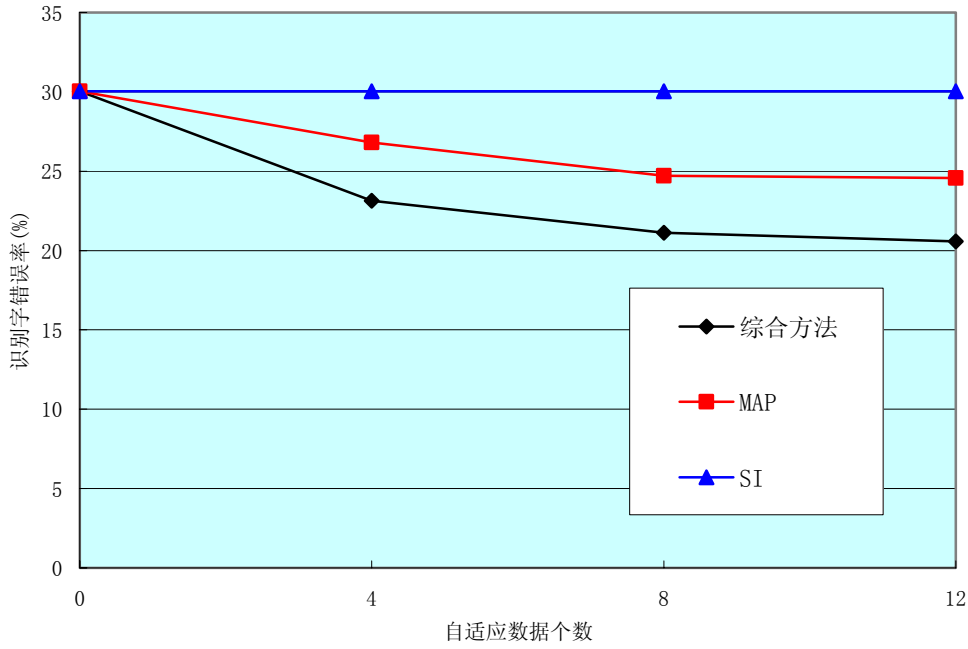


图 5-5 综合渐进自适应方法效果比较（2）：对于测试集 B，SI 系统、MAP 及综合渐进自适应在经过 4、8、12 个自适应数据后的识别字错误率（%）

图 5-5 和图 5-6 给出的是所有的实验结果，从图我们可以看到随着自适应数据的增加，自适应效果更好。不过在 8 句以后这种改善就不是很明显了，所以我们认为在综合方法中 MAP 部分保证 8 到 10 句的缓存就可以了。同时我们的

综合方法不论是在自适应数据比较少的时候，还是处在有噪声的环境下都比原有的 MAP 渐进方法要好。这说明引入 MLLR 方法在渐进 MAP 方法中对处理说话人差异和环境差异都取得了很好的效果。也证明这种综合渐进自适应方法的可行性和鲁棒性。

5.3 综述

通过本章的各种实验，我们可以得出这样的结论。MAP 和 MLLR 两种方法对于说话人自适应和环境自适应都有较好的效果，而我们提出的综合渐进自适应方法比渐进的 MAP 方法有更好的效果。

在我们的说话人自适应实验中，MAP 和 MLLR 分别可以降低识别字错误率 25.20%和 34.53%。比较而言两种方法各自有自己的特点，MLLR 自适应速度比较快，而 MAP 有很好的渐进性。实验还证明如果把这两种方法结合起来使用自适应效果会更好。同时在环境自适应的实验中，这两种方法分别降低识别字错误率 17.57%和 23.86%，显示了对环境的自适应的良好效果。最后，我们还通过基于音节模型的音节与双字词识别实验，验证了 VFS 和共享回归类方法对自适应数据不足和不匀问题解决的有效性。

进一步，我们提出了新的综合渐进自适应方法，在无噪音和有噪音的渐进自适应情况下分别可以降低识别字错误率 23.03%和 29.69%，比渐进 MAP 方法要好大约 10%的效果。实验证明这种新的方法十分适合强健语音识别系统的要求。

第六章 总 结

本文研究了语音识别领域中一个非常重要的实用性技术—语音自适应问题。通过对说话人和环境引起的声学差异的讨论，我们分析和实现了两种常用的自适应方法：MAP 和 MLLR。同时提出了一种对环境鲁棒的综合渐进自适应方法。本文具体研究了如下几个方面：

第一，分析了说话人差异和环境差异对于语音识别系统的影响，研究和分析了各种常见的自适应方法，讨论了它们的优缺点。

第二，研究和实现了基于最大后验概率（MAP）方法的自适应方法。在数字串识别系统中识别字错误率比 SI 系统降低了 25.20%。同时实现了基于 MAP 方法的向量域平滑（VFS）技术，解决了自适应数据不足和不匀的问题。

第三，研究和实现了基于最大似然线性回归（MLLR）方法的自适应方法。在数字串识别系统中识别字错误率比 SI 系统降低了 35.42%。同时实现了基于 MLLR 方法的回归类共享技术，解决了自适应数据不足和不匀的问题。

第四，研究了基于上述两种方法的说话人自适应技术对环境自适应的效果。在噪音环境下 MAP 和 MLLR 分别使得系统识别字错误率比 SI 系统降低了 17.57%和 23.86%。而两种方法的结合方法效果更佳。

第五，提出了一种新的快速的综合渐进自适应方法，并配以新的渐进自适应策略。通过引入一个简化的 MLLR 模块处理环境差异和说话人生理差异，大大提高了传统的 MAP 渐进方法的性能，适合强健语音识别系统的要求。这种综合方法即使在自适应数据比较少的情况下也可以取得好的效果。在无噪音和有噪音的环境中分别可以降低 23.03%和 29.69%的识别字错误率。

参考文献

[1] Rabiner L, Juang B. Fundamentals of Speech Recognition. Englewood Cliff, New Jersey: Prentice-Hall, 1993

[2] Olson H, Belar H. Phonetic Typewriter. J. Acoust. Soc. Am, 1956, 28(6):1072~1081

[3] 方特(瑞典), 高奋(瑞典). 言语科学与言语技术. 北京：商务印书馆, 1994

[4] Fant G.. Acoustic Theory of Speech Production. Hague: Mouton, 1970

[5] Flanagan J. Speech Analysis, Synthesis and Perception. New York: Springer-Verlag, 1972

[6] Itakura F. Minimum Prediction Residual Applied to Speech Recognition. IEEE Trans. Acoustics, Speech, Signal Proc., 1975, ASSP-23(1): 67~72

[7] Sakoe H, Chiba S. Dynamic Programming Algorithm for Spoken Word Recognition. IEEE Trans. Acoustics, Speech, Signal Proc., 1978, ASSP-26(1): 43~49

[8] Baker J. The DRAGON System – an overview. IEEE Trans. Acoustics, Speech, Signal Proc., 1975, ASSP-1: 24~29

[9] Jelinek F. Continuous Speech recognition by Statistical Methods. Proc. IEEE, 1976, 64(4): 532~556

[10] Lesser V, Fennell R, Erman L, et al. The Hearsay-II Speech Understanding System. IEEE Trans. Acoustics, Speech, Signal Proc., 1975, ASSP-23(1): 11~24

[11] Jelinek F, Bahl L, Mercer R. Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech. IEEE Trans. Information Theory, 1975, IT-21: 250~256

[12] Rabiner L, Levinson S, Rosenberg A, et al. Speaker Independent Recognition of Isolate Words using Clustering Techniques. IEEE Trans. Acoustics, Speech, Signal Proc., 1979, ASSP-27(8): 336~349

[13] Linde Y., Buzo A., Gray R.M.. An Algorithm for Vector Quantization. IEEE Trans. On COM, Jan. 1980, 28(1)

[14] Rabiner L, A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. Proc. IEEE, 1989, 77(2): 257~286

[15] Jelinek F. Statistical Methods for Speech Recognition. London: MIT Press, 1998

[16] Lee K, Hon H, Reddy D. An Overview of the SPHINX Speech Recognition System. IEEE Trans. Acoustics, Speech, Signal Proc., 1990, 38: 600~610

[17] Lee K, Automatic Speech Recognition – the Development of the SPHINX System. Boston: Kluwer Academic Publishers, 1989

[18] Joe Tebelskis, “Speech Using Neural Networks”, CMU-CS-95-142, May 1995.

[19] M. Cohen, H. Franco, N. Morgan, et al. Combining Neural Networks and Hidden Markov Models for Continuous Speech Recognition, Proceedings of the DARPA Speech and Natural Language Workshop, Harriman, NY, 1992.

[20] H. Bourlard, N. Morgan. Continuous Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, 1994

[21] 俞铁城, 周健来, 宋岩涛. 基于神经网络/隐马尔可夫模型的混合语音识别方法的研究现状. 第五届全国人机语音通讯学术会议论文集, 哈尔滨, 1998 年 7 月. 18~21

[22] Miastkowski S. Can we talk? Voice-recognition Packages. PC World. 1999, 17(1): 127~136

[23] Huang X, Acero A, Alleva F, et al. Microsoft Windows Highly Intelligent Speech Recognizer: Whisper. In: IEEE, eds. Proc. ICASSP-95. New York: IEEE, 1995, I: 93~96

[24] Hwang M, Rosenfeld R, Thayer E, et al, Improving Speech Recognition Performance via Phone-dependent VQ codebooks and Adaptive Language Models in SPHINX-II. In: IEEE, eds. Proc. ICASSP-94. Adelaide: IEEE, 1994, I: 549~552

[25] 朱凌云(责任编辑). 语音识别. 个人电脑. 2000, 6(2): 81~91

[26] Chris J Leggetter. Improved Acoustic Modeling for HMMs Using Linear Transformations: [PhD Thesis]. University of Cambridge, February 1995. 80~137

[27] 杨行峻, 迟惠生等. 语音信号数字处理. 北京: 电子工业出版社. 1995. 334～335.

[28] 戴礼荣. 人机语声对话特点及系统设计. NCMMSC-96, 1996. 22~26

[29] X.D.Huang, A.Acero, H.Hon, et al. Spoken Language Processing. New Jersey: Prentice Hall PTR, 2000, 401~403, 429~437

[30] Ron Cole, Lynette Hirschman, et al, The Chanllenge of Spoken Language Systems: Research Directions fro the Nineties. IEEE Trans. on Speech and Audio Processing, 1995, 3(1): 1~7

[31] 张帆. 语音识别话者自适应谱变换分块变换: [硕士学位论文]. 北京: 清华大学电子工程系, 1997

[32] F. Macro, M.M. Anna. Fast Speaker Adaptation: Some Experiments on Different Techniques for Codebook Adaptation and HMM Parameters Estimation. In: IEEE, eds. Proc. ICASSP. May 1991. 849~852

[33] B.S. Atal. Automatic Recognition of Speakers from Their Voices. Proc. IEEE, 1976, 64(4): 460~475

[34] F. Fallside, W.A. Woods. Computer Speech Processing. Prentice-Hall, London, 1985

[35] F. Nolan. The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge, 1983

[36] 牛小川, 徐波. 说话人自适应策略与方法的研究与实验. 第五届全国人机语音通讯会议论文集, 哈尔滨. 1998. 181~186

[37] Zheng Rong, Wang Zuoying. Speaker Adaptation: An Overview. Chinese Journal of Electronics, 1998, 7(2): 121~127

[38] 王霞. 声学模型及其评价方法的研究: [硕士学位论文]. 北京: 清华大学计算机科学与技术系, 1999

[39] Qiguang Lin, Chiwei Che. Normalizing the Vocal Tract Length for Speaker Independent Speech Recognition. IEEE Signal Processing Letters, 1995, 2(11): 201~203

[40] 陈景东, 徐波, 黄泰翼. 基于 Mellin 变换的语音新特征与说话人自适应技术的比较. 第五届全国人机语音通讯会议论文集, 哈尔滨. 1998. 86~91

[41] Chen Hingdong, Xu Bo, Huang Taiyi. A New Speech Feature Insensitive to the Variation of Different Speakers. Chinese Journal of Electronics, 1999, 8(1): 67~72

[42] Chen Hingdong, Xu Bo, Huang Taiyi. A Novel Robust Speech Feature Based on the Mellin Transform and Speaker Normalization. Proc. ISCSLP98. Singapore: 1998. 191~195

[43] A.Imamura. Speaker-Adaptive HMM-Based Speech Recognition with A Stochastic Speaker Classifier. In Proc. IEEE Int. Conf. Acoustic, Speech, Signal Proc., 1991, 841~844

[44] L. Mathan, L. Miclet. Speaker Hierarchical Clustering for Improving Speaker-Independent HMM Word Recognition. In: IEEE, eds. Proc. ICASSP90. 1990, 149~152

[45] T. Kosaka, S. Sagayama. Tree-Structured Speaker Clustering for Fast Speaker Adaptation. In: IEEE, eds. Proc. ICASSP94, 1994, 1: 245~248

[46] Chin-Hui Lee. Learning from Surprises-Statistics and Speech/Speaker Recognition. Speech Lab 20th Anniversary Celebration, Tsinghua University, Beijing, 1999

[47] Y. L. Chow, et al. BYBLOS: The BBN Continuous Speech Recognition System. In: IEEE, eds. Proc. ICASSP87. 1987, 89~92

[48] G. Rigoll. Speaker Adaptation for Large Vocabulary Speech Recognition Systems using Speaker Markov Models. In: IEEE, eds. Proc. ICASSP89. 1989, 5~8

[49] C.J. Leggetter, P.C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Model. Computer Speech and Language, 1995, 9:171~185

[50] C.J. Leggetter, P.C. Woodland. Speaker Adaptation of HMMs Using Linear Regression. Technical Report CUED/F-INFENG/TR181, Cambridge Univ., Jun. 1994

[51] Heidi Christensen. Speaker Adaptation of Hidden Markov models using Maximum Likelihood Linear Regression: [MSc Thesis]. Aalborg University, Jun. 1996

[52] M.J.F. Gales, P.C. Woodland. Mean and Variance Adaptation within the MLLR Framework. Computer Speech and Language, 1996, 10:249~264

[53] M.J.F. Gales. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Computer Speech and Language, 1998, 12: 75~98

[54] Jean-Luc Gauvain, Chin-Hui Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Trans. on Speech and Audio Proc., 1994, 2(2): 291~298

[55] C-H. Lee, C-H. Lin, B-H. Juang. A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models. IEEE Trans. on Signal Proc., 1991, 39(4): 806~814

[56] Seyed Mohammad Ahadi-Sarkani, Bayesian and Predictive Techniques for Speaker Adaptation: [PhD Thesis]. Cambridge Univ., 1996

[57] 李虎生, 杨明杰, 刘润生. 汉语数码语音识别自适应算法. 电路与系统学报, 1999, 4(2): 1~6

[58] Qiang Huo. Adaptive learning and Compensation of Hidden Markov Model for Robust Speech Recognition. Proc. ISCSLP98. Singapore: 1998. 31~43

[59] Qiang Huo, Chin-Hui Lee. On-Line Adaptive Learning of Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate. IEEE Trans. on Speech and Audio Proc., 1997, 5(2): 161~171

[60] Qiang Huo, Chorkin Chan, Chin-Hui Lee. On-Line Adaptation of SCHMM Parameters Based on the Segmental Quasi-Bayes Learning for Speech Recognition. IEEE Trans. on Speech and Audio Proc., 1996, 4(2): 141~144

[61] Qiang Huo, Chin-Hui Lee. On-Line Adaptive Learning of Correlated Continuous Density Hidden Markov Models for Speech Recognition. IEEE Trans. on Speech and Audio Proc., 1998, 6(4): 386~397

[62] Jun-ichi Takahashi, Shigeki Sagayama. Vector-field-smoothed Bayesian Learning for Fast and Incremental Speaker/Telephone-channel Adaptation. Computer Speech and Language, 1997, 11: 127~146

[63] Masahiro Tonomura, Tetsuo Kosaka, Shoichi Matsunaga. Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Estimation. Computer Speech and Language, 1996, 10: 117~132

[64] P. F. Brown, C-H. Lee, J. C. Hooper. Bayesian Adaptation in Speech Recognition. In: IEEE, eds. Proc. ICASSP83. 1983, 761~764

[65] Mukund Padmanabhan, Lalit R. Bahl, David Nahamoo, et al. Picheny. Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems. IEEE Trans. on Speech and Audio Proc., 1998, 6(1): 71~77

[66] C. J. Leggetter, P.C. Woodland. Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression. Proc. ICSLP94. Yokohama: 1994

[67] Ramalingam Hariharan, Olli Viikki. On Combining Vocal Tract Length Normalisation and Speaker Adaptation for Noise Robust Speech Recognition. Proceedings of Eurospeech ’99, Budapest, 1999

[68] Michal Schüβler, Florian Gallwitz, Stefan Harbeck. A Fast Algorithm for Unsupervised Incremental Speaker Adaptation. In: IEEE, eds. Proc. ICASSP97. Munich: IEEE, 1997, 1019~1023

[69] 张怡颖. 语音识别和说话人识别的研究: [博士学位论文]. 北京: 清华大学计算机科学与技术系, 1999

[70] 贾宾. 语音识别的声学建模及其应用研究: [博士学位论文]. 北京: 清华大学自动化系, 2000

[71] 王昱. 连续语音识别系统的研究与实现: [学士学位论文]. 北京: 清华大学计算机科学与技术系, 1998

[72] 郭锐. 一种关键词确认方法的研究: [学士学位论文]. 北京: 清华大学计算机科学与技术系, 1999

[73] 郭庆. 声学模型中帧间相关性和自适应问题的研究: [博士学位论文]. 北京: 清华大学计算机科学与技术系, 1999

[74] Guoqiang Li, Limin Du, Yanjun Xu, et al. Maximum Likelihood Smoothes and Predictions for Fast Speaker Adaptation. Proc. ISCSLP98. Singapore: 1998. 212~215

[75] Dimitry Rtisher, David Nahamoo, Michael Picheny. Speaker Adaptation via VQ Prototype Modification. IEEE Trans. on Speech and Audio Proc., 1994, 2(1): 94~96

[76] Erwin W. Drenth, Bernhard Ruber. Context-dependent Probability Adaptation in Speech Understanding. Computer Speech and Language, 1997, 11: 225~252

[77] Francoise Beaufays, Mitch Weintraub. Model Transformation for Robust Speaker Recognition from Telephone Data. In: IEEE, eds. Proc. ICASSP-97. April 21-24 1997

[78] Eric Thelen, Xavier Aubert, Peter Beyerlein. Speaker Adaptation in the PHILIPS System for Large Vocabulary Continuous Speech Recognition. In: IEEE, eds. Proc. ICASSP97. Munich: IEEE, 1997, 1035~1038

[79] S.M. Ahadi, P.C. Woodland. Combined Bayesian and Predictive Techniques for Rapid Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language, 1997, 11: 187~206

[80] Tomoko Matsui, Sadaoki Furui. N-Best-based Unsupervised Speaker Adaptation for Speech Recognition. Computer Speech and Language, 1998, 12: 41~50

[81] Tetsuo Kosaka, Shoichi Matsunaga, Shigeki Sagayama. Speaker-independent Speech Recognition Based on Tree-structured Speaker Clustering. Computer Speech and Language, 1996, 55~74

[82] Qiang Huo, Chorkin Chan, Chin-Hui Lee. Bayes Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition. IEEE Trans. on Speech and Audio Proc., 1995, 3(5): 334~345

[83] Hui Jiang. Robust Speech Recognition Based on a Bayesian Prediction Approach. IEEE Trans. on Speech and Audio Proc., 1999, 7(4): 426~440

[84] Ying Hao, Ditang Fang. Speech Recognition Using Speaker Adaptation by System Parameter Transformation. IEEE Trans. on Speech and Audio Proc., 1994, 2(1): 63~68

[85] Vassilios V. Digalakis, Dimitry Rtischev, Leonardo G. Neumeyer. Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures. IEEE Trans. on Speech and Audio Proc., 1995, 3(5): 357~365

[86] Vassilis D. Diakoloukas, V. V. Digalakis. Maximum-Likelihood Stochastic-Transformation Adaptation of Hidden Markov Models. IEEE Trans. on Speech and Audio Proc., 1999, 7(2): 177~187

[87] Vassilios V. Digalakis. Online Adaptation of Hidden Markov Models Using Incremental Estimation Algorithm. IEEE Trans. on Speech and Audio Proc., 1999, 7(3): 253~261

[88] Yasunaga MIYAZAWA, Tun-ichi TAKAMI, Shigeki SAGAYAMA, et al. Unsupervised Speaker Adaptation Using All-Phoneme Ergodic Hidden Markov Network. IEEE Trans. INF. & SYST., 1995, E78-D(8): 1044~1049

[89] Ronald A.Cole, Joseph Mariani, Hans Uszkoreit, et al. Survey of the State of the Art in Human Language Technology. November 21, 1995

[90] S.Goronzy, R.Kompe. A MAP-like Weighting Scheme for MLLR Speaker Adaptation. Proceedings of Eurospeech ’99, Budapest, 1999

[91] Y. Gong, John J, Godfrey. Transforming HMMs for Speaker-independent Hands-free Speech Recognition in the Car. In: IEEE, eds. Proc. ICASSP99. 1999

[92] Patrick Nguyen, Christian Wellekens, Jean-Claude Junqua. Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environments. Proceedings of Eurospeech ’99, Budapest, 1999

[93] Alexander Fisher, Volker Stahl. Database and Online Adaptation for Improved Speech Recognition in Car environments. In: IEEE, eds. Proc. ICASSP99. 1999

[94] P. Price, W. M. Fisher, J. Bernstein, et al. The DARPA Resource Management Database for Continuous Speech Recognition. In: IEEE, eds. Proc. ICASSP88. 1988, 1: 651~654

附录 MLLR 共享回归类

在第三章中我们提到了关于 MLLR 共享回归类的技术。这里我们给出我们在第五章实验中所使用的两种共享回归类的划分方法。正如第三章我们说的一般可以使用声学特性或者距离衡量来划分回归类。

声学特征，通过各个模型在声学上的特征分类，一般这种方法使用人为分类。在我们的是实验中是通过韵母对 411 个音节分类的。汉语拼音共有 37 个韵母，其中个别韵母在 411 个拼音中并没有出现，因而可以略去不用；同时某些韵母（如 i）与声母联合发音时有较大差别，也应该将这样的拼音予以分开；此外还有一些近似音（如 in 与 ing）合并的情况。所以最终我们选取了 34 个类别，具体分类如表附-1。

表附-2　按声学特性分类的回归类，总共 34 个类

类别	对应音节
0	a,ba,pa,ma,fa,da,ta,na,la,ga,ka,ha,zha,cha,sha,za,ca,sa
1	ai,bai,pai,mai,dai,tai,nai,lai,gai,kai,hai, zhai,chai,shai,zai,cai,sai
2	ao,bao,pao,mao,dao,tao,nao,lao,gao,kao, hao,zhao,chao,shao,rao,zao,cao,sao
3	An,ban,pan,man,fan,dan,tan,nan,lan,gan,kan han,zhan,chan,shan,ran,zan,can,san,
4	ang,bang,pang,mang,fang,dang,tang,nang,lang, gang,kang,hang,zhang,chang,shang,rang,zang,cang,sang
5	o,bo,po,mo,fo,lo,yo
6	ou,pou,mou,fou,dou,tou,nou,lou,gou,kou, hou,zhou,chou,shou,rou,zou,cou,sou
7	e,er,me,de,te,ne,le,ge,ke,he,zhe,che,she,re,ze,ce,se
8	ei,bei,pei,mei,fei,dei,tei,nei,lei,gei,

	kei,hei,zhei,shei,zei
9	en,ben,pen,men,fen,den,nen,gen,ken,hen, zhen,chen,shen,ren,zen,cen,sen
10	zi,ci,si,zhi,chi,shi,ri
11	yi,bi,pi,mi,di,ti,ni,li,ji,qi,xi
12	ya,dia,lia,jia,qia,xia
13	ye,bie,pie,mie,die,tie,nie,lie,jie,qie,xie
14	yao,biao,piao,miao,diao,tiao,niao,liao,jiao,qiao,xiao
15	you,miu,diu,niu,liu,jiu,qiu,xiu
16	yan,bian,pian,mian,dian,tian,nian,lian,jian,qian,xian
17	yin,bin,pin,min,nin,lin,jin,qin,xin,ying bing,ping,ming,ding,ting,ning,ling,jing,qing,xing
18	yang,niang,liang,jiang,qiang,xiang
19	wu,bu,pu,mu,fu,du,tu,nu,lu,gu,ku,hu,zhu, chu,shu,ru,zu,cu,su
20	wa,gua,kua,hua,zhua,chua,shua,rua
21	dong,tong,nong,long,gong,kong,hong,zhong, chong,rong,zong,cong,song
22	wai,guai,kuai,huai,zhuai,chuai,shuai
23	wei,dui,tui,gui,kui,hui,zhui,chui,shui,rui,zui,cui,sui
24	wan,duan,tuan,nuan,luan,guan,kuan,huan,zhuan, chuan,shuan,ruan,zuan,cuan,suan
25	wen,weng,dun,tun,nun,lun,gun,kun,hun, zhun,chun,shun,run,zun,cun,sun
26	wang,guang,kuang,huang,zhuang,chuang,shuang
27	beng,peng,meng,feng,deng,teng,neng,leng, geng,keng,heng,zheng,cheng,sheng,reng,zeng,ceng,seng
28	yu,nv,lv,ju,qu,xu
29	yue,nve,lve,jue,que,xue
30	yuan,juan,quan,xuan
31	yun,jun,qun,xun
32	yong,jiong,qiong,xiong
33	wo,duo,tuo,guo,kuo,huo,nuo,luo,zhuo,chuo, shuo,ruo,zuo,cuo,suo

距离衡量，通过对隐含马尔可夫模型的参数进行基于距离的聚类实现分类。这种方法基于 411 个拼音如果存在近似发音，那么在特征空间上就表现为参数特征向量的距离接近的这一设想。我们经过实验通过基于距离的迭代聚类把 411 个音节聚成了 9 个类别，具体如表附-2 所示。对分类的结果作一个粗略分析，就可以看出元音，即韵母在聚类中起到关键作用，有相同韵母的音节一般集中分布在 1~2 类当中，在其它类别中仅有零星存在，如[ai]，仅有[shai]分在第 6 类，其余带有韵母[ai]的音节全在第 0 类，而含有[ei]的音节，大致平均地分布在第 4 类和第 7 类。这一方面说明韵母相同的音节的特征向量在特征空间上距离接近，易于聚为一类；另一方面，在韵母起决定作用的基础上，部分声母对于分类也会产生重大影响。尤其是[x]、[s]、[sh]这三个轻擦音声母，以它们开头的音节集中分布在第 6 类和第 7 类，表明在这类音节中声母对参数特征向量所起的作用更为突出，对它们的特征描述主要表现在对声母的描述。

表附-2　按距离衡量分类的回归类，总共 9 个类

类别	代表韵母	对应拼音
0	ai,an, uai,uan, en	ai,an,bai,ban,ben,cai,can,chai,chan,chuai,chuan, cuan,dai,dan,den,duan,en,er,fan,gai,gan,guai,guan, hai,han,hen,huai,huan,kai,kan,ken,kuai,kuan,lai, lan,luan,mai,man,nai,nan,nuan,pai,pan,pen,ran,ruan,sai,san,tai,tan,tuan,wai,wan,zai,zan,zhai,zhan, zhuai,zhuan,zuan,
1	e,eng	beng,ce,ceng,che,cheng,de,deng,e,feng,ge,geng,he, heng,ke,keng,le,leng,meng,ne,neng,peng,re,te,teng, ze,zeng,zhe,zheng,zi,
2	ao,iao	ao,bao,biao,cao,chao,dao,diao,gao,hao,jiao,kao,lao, liao,mao,miao,nao,niao,pao,piao,qiao,rao,sao,tao, tiao,xiao,yao,zao,zhao,
3	a,ang, ia,iang, ua,uang	a,ang,ba,bang,ca,cang,cha,chang,chua,chuang,da,dang,dia,fa,fang,ga,gang,gua,guang,ha,hang,hua,huang, jia,jiang,ka,kang,kua,kuang,la,lang,lia,liang,ma, mang,na,nang,niang,pa,pang,qia,qiang,rang,rua, ta,tang,wa,wang,ya,yang,za,zang,zha,zhang,zhua, zhuang,
4	ei,ong, ui,un,ou iu,en	bei,dei,diu,dong,dui,dun,ei,gei,lei,liu,long,lun, mei,men,miu,nei,nen,niu,nong,nou,nun,ren,reng,rong, rou,rui,run,wei,wen,yo,yong,you,zei,

5	o,u,ong, ou,uo	bo,bu,chong,chou,chu,chuo,cong,cou,cu,cuo,dou,du, duo,fo,fou,fu,gong,gou,gu,guo,hong,hou,hu,huo, kong,kou,ku,kuo,lo,lou,lu,luo,me,mo,mou,mu,nu,nuo, o,ou,po,pou,pu,ru,ruo,tong,tou,tu,tuo,weng,wo,wu, zhong,zhou,zhu,zhuo,zong,zou,zu,zuo,
6	iu,iong **s,sh,x	jiong,jiu,qiong,qiu,sa,sang,se,seng,sha,shai,shan, shang,shao,she,sheng,shou,shu,shua,shuai,shuan,su, shuang,shuo,song,sou,suan,suo,xia,xiang,xiong,xiu,
7	i,ui,ei,un,en **s,sh,x	cen,chen,chi,chui,chun,ci,cui,cun,fei,fen,gen,gui, gun,hei,hui,hun,jin,jing,ju,juan,jue,jun,kei,kui, kun,pei,qi,qian,qie,qin,qing,qu,quan,que,qun,ri, sen,shei,shen,shi,shui,shun,si,sui,sun,tei,tui,tun, xi,xian,xie,xin,xing,xu,xuan,xue,xun,zen,zhei,zhen, zhi,zhui,zhun,zui,zun,
8	i,ian,ie in,ing v,ve	bi,bian,bie,bin,bing,di,dian,die,ding,ji,jian,jie, li,lian,lie,lin,ling,lv,lve,mi,mian,mie,min,ming, ni,nian,nie,nin,ning,nv,nve,pi,pian,pie,pin,ping, ti,tian,tie,ting,yan,ye,yi,yin,ying,yu,yuan,yue,yun

关于聚类的具体方法和实现可以参考文献[72]。

图表索引

图索引

表索引