基于小数据量的方言背景普通话语音识别声学建模研究

(申请清华大学工学博士学位论文)

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研　究　生 ： 刘 林 泉

指 导 教 师 ： 吴 文 虎 教　授

副指导教师 ： 郑　　方 研究员

二〇〇七年四月

Research on A Small Data Set Based Acoustic Modeling for Dialectal Chinese Speech Recognition

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

by

Liu Linquan

(Computer Science and Technology)

Dissertation Supervisor: Professor Wu Wenhu

Associate Supervisor: Professor Zheng Fang

April, 2007

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：(1) 已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；(2) 为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；(3) 根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

(保密的论文在解密后遵守此规定)

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘　要

本文研究方言背景普通话语音识别。文中以一个标准普通话识别器（以下称为基线系统）为基础，研究利用少量方言背景普通话训练数据（小于 1 小时），进行声学建模，目标是实现一个高效的语音识别器，既能显著提高方言背景普通话的识别率，又能不降低标准普通话的识别率。本文的成果如下：

1. 提出了改进的上下文相关权重策略，解决了发音字典中音节对应于不合理的声韵母组合现象。运用改进的上下文相关权重策略，新的发音字典与基线系统的发音字典相比，对方言背景普通话的相对词错误率下降了 12.4%；同时，对标准普通话的识别率没有降低[①]。

2. 提出了状态相关的基于基元的模型归并 SDPBMM 策略。SDPBMM 能很好的解决基线系统声学模型对方言背景普通话空间覆盖度不够的问题，显著的提高了方言背景普通话的识别率；同时，又保证了标准普通话的识别率不降低。与基线系统声学模型相比，对于方言背景普通话的相对词错误率降低了 50.4%，对于标准普通话的相对词错误率降低了 32.3%[①]。实验证明，其性能优于自适应、重训练和模型插值方法。

3. 提出了基于距离度量的识别网络生成策略。此策略应用于 1)生成方言背景普通话发音字典；2) 在 SDPBMM 中，作为状态归并的准则之一。解决了基于小数据量的发音建模过程中的数据稀疏问题。

4. 提出了抑制高斯混合扩张的选择策略，得到了可选择的 SDPBMM。文中采用标准普通话和方言背景普通话模型状态之间的声学距离为准则，进行有选择的归并，将需要参与归并的状态和不需要参与归并的（冗余的）状态分开。实验表明，在模型规模降低的同时，不降低性能，有效地缓解了识别率的提高和模型规模扩张之间的矛盾。

关键词：方言背景普通话；连续语音识别；声学建模；发音字典；发音建模；小数据量

[①] 文中使用了标准普通话和闽南普通话为测试数据，每个集合由 1,200 个短语组成。

Abstract

In this dissertation, research on acoustic modeling for dialectal Chinese speech recognition based on a small data set is focused on. In combination with a standard Chinese recognizer, two objectives are set for dialectal Chinese acoustic modeling. One is how to make full use of a small dialectal Chinese data set (less than one hour); the other aims to improve the performance for dialectal Chinese speech recognition significantly while no degradation for standard Chinese speech recognition occurs. Some contributions of this dissertation are listed as follows:

A weighting method, ***improved context-dependent weighting*** (ICDW) method, is proposed on the basis of classical *context-dependent weighting* (CDW) method. The ICDW method can deal well with the illegal combinations of Initial and Final for a certain syllable in the lexicon. Compared experimentally with the lexicon from the standard Chinese recognizer, a relative SER reduction of 12.4% was achieved by the ICDW method on dialectal Chinese, meanwhile, no degradation was caused by the ICDW method on standard Chinese.

A novel, simple but effective acoustic modeling method, ***state-dependent phoneme-based model merging*** (SDPBMM), is proposed. The SDPBMM can increase dramatically the coverage of standard Chinese acoustic model; as a result, some significant improvement for dialectal Chinese speech recognition is achieved while no degradation for standard Chinese is introduced. Experimentally, compared with the standard Chinese acoutic model, on the one hand, a relative WER reduction of 50.4% was achieved by the SDPBMM on dialectal Chinese; on the other hand, a relative WER reduction of 32.3% was also achieved by SDPBMM on standard Chinese. In addition, the SDPBMM outperformed some other acoustic modeling methods on the basis of adaptation, retraining or interpolation.

Some research on pronunciation modeling for dialectal Chinese based on a small development data set is carried out. To deal with data sparseness in the process of pronunciation modeling, ***a recognition network generation method based on a certain distance measure*** is proposed. The method is adopted 1) in the generation of dialectal Chinese lexicon; 2) in the SDPBMM acting as one of the two merging criteria.

One side effect of SDPBMM is Gaussian mixture expansion problem. To solve the problem, ***selective-SDPBMM*** is proposed which can differentiate the states that need merging from those that need no merging in accordance with some certain distance measures. The selective-SDPBMM can decrease the scale of Gaussian mixtures while it brings no performance deterioration for both dialectal and standard Chinese speech recognition.

Key words: dialectal Chinese; continuous speech recognition; acoustic modeling; pronunciation modeling; a small development data set

目 录

主要符号对照表

AM	声学模型（Acoustic Model）
AUP	累积一元概率（Accumulated Uni-gram Probability）
CER	字错误率（Character Error Rate）
CDW	上下文相关权重策略（Context-Dependent Weigting）
CM	置信度（Confidence Measure）
CMN	倒谱均值归一化（Cepstrum Mean Normalization）
GIF	广义声韵母（Generalized Initial/Final）
GMEP	高斯混合扩张问题（Gaussian Mixture Expansion Problem）
GMM	高斯混合模型（Gaussian Mixture Model）
HMM	隐马尔可夫模型（Hidden Markov Model）
ICDW	改进的上下文相关权重策略（Improved Context-Dependent Weighting）
IF	汉语声母或韵母（Initial/Final）
LM	语言模型（Language Model）
LVCSR	大词表连续语音识别（Large Vocabulary Continuous Speech Recognition）
MAP	最大后验概率（Maximum *a Posteriori*）
MFCC	Mel 倒谱系数（Mel Frequency Cepstrum Coefficient）
MLLR	最大似然线形回归（Maximum Likelihood Linear Regression）
Mono-XIF	上下文无关（或一元）扩展声韵母模型
WER	词错误率（Word Error Rate）
SDPBMM	状态相关基于基元的模型合并（State-Dependent Phoneme-Based Model Merging）
SER	音节错误率（Syllable Error Rate）
Tri-XIF	上下文相关（或三元）扩展声韵母模型
XIF	扩展声母或韵母（eXtended Initial/Final）

第 1 章 绪 论

1.1 研究的背景和意义

本文针对方言背景普通话语音识别，研究基于少量方言背景普通话数据的声学建模方法。所谓方言背景普通话，文（陈亚川，1991）给出的定义为“民族共同语（现代汉语）在各方言地区的运用中带有一定程度的地方特色的普通话”。方言背景普通话具有以下鲜明的特点（姚佑椿，1989；齐沪扬，1999）：

1. 受母方言的影响很大。母方言的知识对于学习使用普通话，在语音和语言上必然产生一定程度的干扰作用。这种干扰作用直接表现为用母方言的语言规律去替代普通话的语言规律。例如，普通话的声韵调时常会用母方言的近似音或者变异音来替代。

2. 方言背景普通话是学习普通话的人在普通话输入的基础上形成的一种与标准普通话有不同程度的差异的语言系统。它是方言区的广大人民学习普通话的过程逐步偏离母语方言，向普通话靠拢的一种中介语言现象。

3. 方言背景普通话的多样性体现在两个方面。一是方言的多样性，不同地域的方言，就会产生不同的方言背景普通话；二是同一个方言区，其方言背景普通话也是一个从方言到普通话的过渡集合，各种不同程度差异的普通话都可能出现。同一个人，在说书面语的时候接近普通话，在口语中却又接近方言。

4. 方言背景普通话是学习普通话过程中产生的“偏误”。这种“偏误”是有规律可循的，不同于偶然发生的语言错误。

5. 方言背景普通话有其不易改变的顽固性。顽固程度大体是语音最甚，词汇次之，语法再次。这种顽固性在不同场合，不同情绪下表现程度不同。

方言背景普通话是一种客观存在的语言现象，必然会长期存在，随着社会的发展、交流与融和，还会不断发展完善。开展方言背景普通话语音识别的研究，对于完善和推进语音识别的研究和应用具有重要的现实意义。

1. 标准普通话识别器对于方言背景普通话的识别率差强人意。一般情况下，即使很鲁棒的标准普通话识别器被用来识别方言背景普通话时，也有 20%~50% 的识别率降低。例如，在 NIST 1997 广播新闻评测任务（NIST 1997 Broadcast News evaluation task）（NIST, 1997）中，当使用标准普通话识别器测试标准普通话时，

其字错误率 CER 为 21.38%。但是，当用同一个识别器去识别上海普通话时，对朗读式和自然发音式的上海普通话测试集，其字错误率分别上升为 61.89%和 72.17%。

2. 探索标准普通话和方言背景普通话的语音识别的相同点和不同点。由方言背景普通话的定义，我们知道，方言背景普通话是普通话在方言区的一种变体，两者之间必然存在很多的相同点，否则就会影响相互之间的交流；另外，也存在很多的不同点，否则就不能称之为方言背景普通话了。在语音识别领域，研究它们之间的异同点，是不断完善普通话识别器，提高方言背景普通话的识别率的必要途径。

3. 方言背景普通话的语音识别是目前研究的热点和难点问题。现阶段，即使在语言学领域，对于方言背景普通话的研究依然存在很多争论，许多问题尚无定论，而其实际应用技术－语音识别更是处于起步和探索阶段。2004 年，在约翰·霍普金斯大学举行的研讨会（Workshop，2004）中首次提出了“方言背景普通话语音识别框架”的概念，研究如何从一个标准普通话识别器出发建立一个方言背景普通话识别器。此框架中提出了从声学模型、发音字典、语言模型和解码器四个层次对方言背景普通话进行研究。考虑到标准普通话和方言背景普通话在语音层面差异最大，目前大多数机构都把研究重点集中在声学模型或者发音字典。例如，清华大学计算机系（李净，2005）、清华大学电子系（Zhang, 2006）、中科院自动化所（刘明宽，2002）、中科院声学所（Pan，2006）、香港科技大学（Fung，2005）、香港中文大学（Lee，2002）、社科院语言所（李爱军，2003）、微软亚洲研究院（Huang，2004）、云南大学（普园媛，2005）等单位在对普通话语音识别开展研究的同时，对上海普通话、广东普通话、闽南普通话、云南普通话等方言背景普通话在语音层面也开展了研究工作。

4. 推进语音识别技术向实际应用的需要。由于方言口音是我国人民日常交流中普遍存在的现象，如何增强标准普通话识别器的鲁棒性，提高在方言地区的适用性，是推进语音识别技术逐步走向实际应用的关键技术。

本文的工作是对“方言背景普通话语音识别框架”（Workshop，2004；李净，2005）（如图 1.1）中研究任务的延续和发展。在以往的工作中，通过声学模型、发音字典、语言模型以及解码器对方言背景普通话语音识别进行了开创性的研究。在图 1.1 中，中间的虚线方框表示以往研究中曾经采用的方法。在此框架指导下，本文对方言背景普通话语音识别中的声学建模部分进行更加深入

的研究，这是由标准普通话和方言背景普通话之间最大的差异体现在语音层面的特点决定的，是整个框架中的核心和难点。另外，本文开展的方言背景普通话语音识别的应用背景之一就是基于嵌入式设备的语音识别系统。在此应用环境中，通常情况下，并不采用针对特定方言背景普通话的语言模型，因而声学模型的性能在很大程度上决定了整体的性能。建立鲁棒的方言背景普通话声学模型是提高汉语语音识别系统在方言地区的适用性的基础。本文在声学建模研究过程中一方面对以往的方法进行改进，另一方面也提出一些新方法，借此来建立更鲁棒的方言背景普通话语音识别的声学模型。

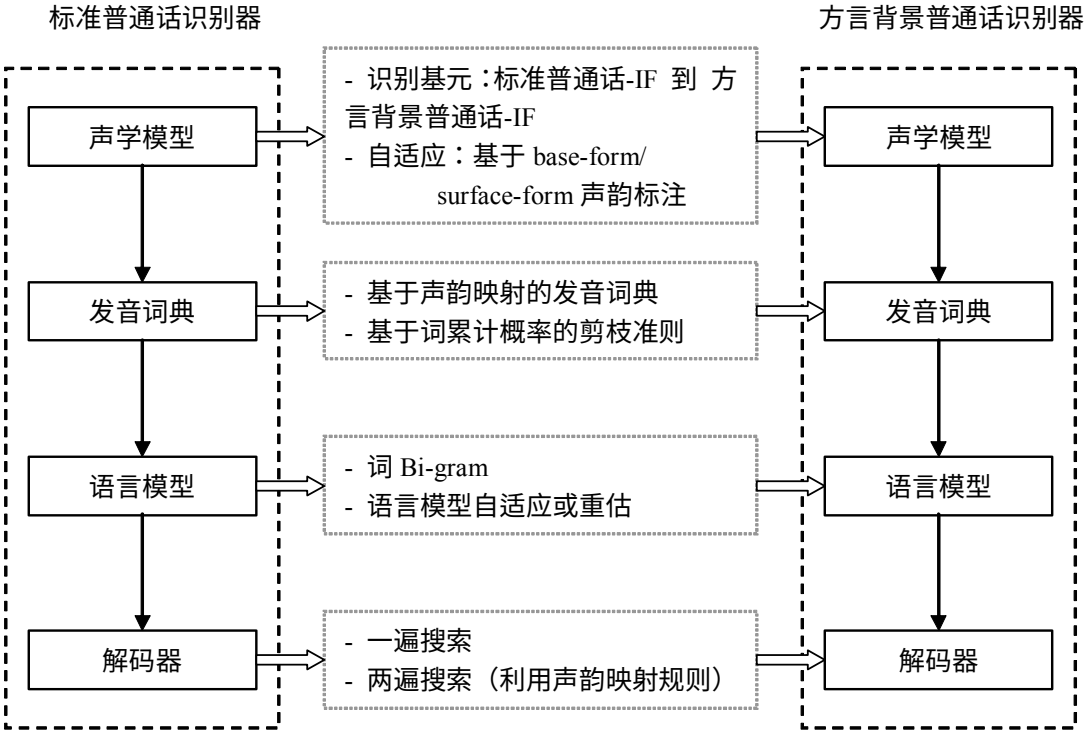


图1.1 方言背景普通话语音识别框架

1.2 带口音的语音识别研究现状

为了更好的跟踪、借鉴、吸收、对比他人的研究成果，本文在方言背景普通话在声学层面关注的目标综合了方言背景普通话语音识别、带口音（accented）的语音识别和非母语（non-native）语音识别三个领域，因为这三个领域中，说话人语音都与标准语音存在一定的口音差异。在这三个领域中，主要的研究方向如图 1.2 所示：

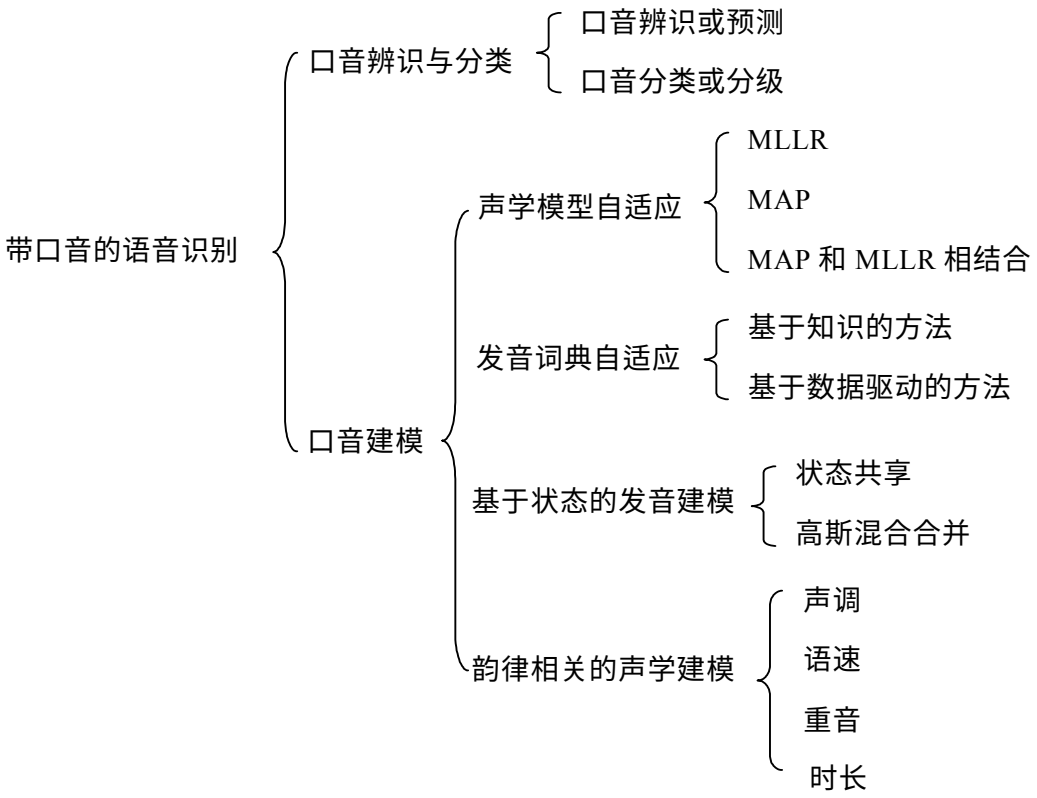


图1.2 方言背景普通话语音识别在声学层面主要研究方向

如图 1.2 所示，近年来，在语音层面，对带有一定口音的语音识别研究主要集中在两个方面：1）口音辨识（dialect/accent identification）与分类（classification）；2）口音建模。其中前者可以作为后者的前端加以应用，后者在很大程度上依赖于前者的识别结果，后者的性能对于整个系统的性能起着关键的作用。

1.2.1 口音辨识与分类

口音辨识与分类是近年来的研究热点之一。口音辨识与分类的结果可以作为特定口音或方言背景语音识别器的前端，依据其辨识或分类结果选择特定的模型或发音词典进行识别，以提高识别性能（Lim，2005；Huang，2005；Tobias，2005；Ma，2006；Raux，2004；Yuan，2005）。这些方法包括，借助声学模型、语言模型和基于潜在语义分析（latent semantic analysis，LSA）的高阶统计信息来进行汉语方言辨识（Lim，2005）；利用韵律结构（如 F0 曲线，语速，句子时长等）、基于词和子词（sub-word，如音素）的建模或分类方法，来进行英语

的口音分类（Huang，2005；Ma，2006），在（Angkititrakul，2006）中，就基于音素的建模在口音辨识中的应用进行了深入的研究，利用音素轨迹作为附加的口音信息，采用最大似然估计进行口音辨识，对标准英语、以汉语为母语的英语、以泰语为母语的英语、以法语为母语的英语、以土耳其语为母语的英语进行了分类，获得了 78.73%的正确率；采用高斯混合模型（GMM）进行汉语普通话口音分类（Chen, 2001, 赵征鹏, 2005）；在(Chen, 2001)中，利用 GMM 进行建模，对带上海、北京、广东、台湾口音的普通话进行辨识，并明确指出文中提出的方法只需要 3~5 个句子就能获得很好的正确率，男女说话人的分类错误率分别为 15.5%和 11.7%；利用基于母语知识的分类和基于声学模型的聚类方法对说话人进行划分（Tobias，2005）；利用韵律信息进行口音预测（Yuan，2005）；根据发音习惯对说话人进行分类，从而选择特定的发音词典（Ruax, 2004），选择不同的自适应方法，甚至不同的语言模型。另外，在同一种口音内部，对说话人的口音等级进行划分，如采用 GMM 方法进行口音等级建模，或通过观测特定的音素或声韵母（如吴方言中的/z/，/c/，/s/）在句子中出现的频率来进行口音等级划分，然后使用特定的声学模型和发音词典进行识别（Zheng, 2005；Sproat, 2004）。在（Zheng, 2005；Sproat, 2004）就对上海普通话数据，依据口音轻重的不同划分为 6 类，对每一类采用不同的声学模型与之相匹配。总的来说，口音辨识与分类对于改善人机交互，提供语音识别器前端是很有意义的，其结果作为带口音语音识别系统的前端，另一方面，语音识别的性能在很大程度上依赖于口音辨识的结果，如果口音辨识的结果不理想，则最终的识别性能也很难提高。

1.2.2 声学模型自适应

对于口音建模研究，声学模型自适应是最常用而有效的方法之一。如图 1.2 所示，声学模型自适应最基本的方法有最大似然线性回归 MLLR（Leggetter，1995）和最大后验概率 MAP（Gauvain，1994）方法，许多新的自适应方法都是从二者中派生出来的（He，2007；Wu，2004; Cheng, 2006; 丰洪才, 2005; 穆向禹, 2003）。MLLR 是一种基于变换的方法，对数据量依赖较小，常用于数据量较少的情况或进行快速自适应。而 MAP 结合先验知识，在数据量相对较多的情形下能够取得较好的性能。在文（He, 2007）中研究了如何利用少量数据进行

MAP 自适应的方法。最近的一些研究成果往往将两者结合使用，将基于 MLLR 的自适应的结果作为 MAP 的先验知识，而后再进行基于 MAP 的自适应。例如在（Sproat，2004）中，使用 6.3 小时的自适应数据，基于 MLLR 和 MAP 相结合的方法就比单独使用 MAP 和 MLLR 词错误率分别多降低了 1.7%和 4.1%。在文（丰洪才, 2005）中通过在渐进的 MAP 方法中引入一个简化的 MLLR 模块，提出一种适合于鲁棒语音识别的快速综合渐进自适应语音识别方法，并在实验基础上得出结论，MLLR 和 MAP 的结合比单独的 MAP 平均性能提高了 10.0%。

1.2.3 发音字典自适应

发音词典自适应常采用发音建模（pronunciation modeling，PM）相关技术，主要研究由说话方式、语速、口音、情绪等带来的影响，在带口音的语音识别中被广泛应用（Huang，2004；Goronzy， 2004；Wester，2003；Livescu，2000；潘复平，2005）。文（Lussier, 2003）对发音建模进行了全面的回顾和阐述，对比了基于专家知识和数据驱动两种发音建模方式的优缺点。对发音建模中，发音变化的获得，发音变化的概率估计，采用何种剪枝策略来降低发音字典的混淆度等难点问题进行了深入的探讨；文（Huang，2004）从理论上证明了口音是说话人之间最主要的差异，提出了发音字典自适应（pronunciation dictionary adaptation，PDA）的方法，在上海普通话测试集上，音节错误率由 23.18%降至 19.96%。在文（刘明宽，2002）中提出了以汉语音节为单位的混淆矩阵作为发音变化的生成策略，进而构造适合于上海普通话特点的发音字典。通过应用发音字典，使普通话识别器对于上海普通话的相对音节错误率下降了 20.0%。在（宋战江, 2001）中着重研究了发音变化概率估计策略，提出了基于声韵母的上下文相关的权重策略（CDW），更加精确地刻画了自然发音方式（spontaneous）中的变化概率。在（Li，2006）则对发音字典自适应中的剪枝策略进行了研究，提出了累积一元概率（AUP）的剪枝策略，相对于常用的累积概率的剪枝策略，音节错误率下降了 1.2%。

1.2.4 基于状态的发音建模

发音字典自适应的方法主要是将发音建模的成果在字典中体现。而实际上，

许多发音变化并没有泾渭分明的区别。例如字典自适应较好地解决了发音中的音素变化（phone change）现象，却不能很好的解决声音变化（sound change）问题。为此有些研究者就将发音变化在 HMM 的状态一级，甚至高斯混合一级进行建模来解决声音变化的问题（Liu，2006；Hain，2005；Kam，2003；Saraclar，2000；Liu，2004；Fung，2005），并取得了很好的效果，在很大程度上，对于带口音的语音识别效果优于发音字典自适应技术，当然两者往往互为补充，结合使用。在文（Saraclar, 2000）中，作者将发音建模同声学建模相结合，在状态一级通过高斯混合共享来体现发音变化，在自然发音方式的语音识别中将字错误率降低了 1.7%。作为对文（Saraclar, 2000）的完善和改进，文（Liu, 2004）提出了局部变化音素模型（partial change phone model，PCPM）来有针对性的解决声音变化的问题。为了进一步提高系统的鲁棒性，降低模型参数规模，文中采用决策树作为状态高斯混合共享的策略。通过对自然发音式的语音测试结果表明，音节错误率下降了 2.39%。因而，对于带口音的语音识别的声学建模的研究中，如何在声学建模过程中更好和发音建模相结合是目前的一个研究方向，也是能显著提高识别效果的方法之一。本文的方向之一也是在这方面展开研究，着重解决方言背景普通话中的声音变化带来的影响。

1.2.5 韵律相关的声学建模和发音建模

最新的一些研究文献开始关注韵律信息在声学模型中的应用以及其对发音变化的影响。通常情况下，声调、重音、停顿、语速、词的时长等韵律信息都可以用来进行建模（Vergyri，2003；Milone，2003；Stephenson，2004；Dong, 2005；Chen, 2006；王作英，2003）。但在汉语语音识别中，目前发表的论文，绝大都数用到的韵律信息就是声调（tone）（Lee，2002；Zhou，2004；Cao, 2004；钟金宏, 2001；Pan, 2006），建模过程中，声调特征一般由基频 F0 和能量来刻画。在现阶段，在建模过程中，韵律特征始终作为一种附加或者额外信息，其应用场合更多的是被限定在声调识别、孤立词识别或者短语识别，在大词表连续语音识别中还没有被广泛应用。其原因在于，一方面由于稳定韵律特征比较难以提取，另一方面如何把韵律信息同传统的声学建模相结合也是一个急需解决的问题（Peng，2004；Tang，2005），因而韵律信息在带口音的语音识别中的应用尚未全面展开。但声调作为汉语区别于西方语音的一个显著特征，在未来的

汉语语音识别中应该会发挥更重要的作用。

1.3 研究的依据、目标和难点

1.3.1 语音识别原理

连续语音识别的目标就是，对于给定的输出观察矢量 $\mathbf{X}=X_1X_2\cdots X_n$，通过最大后验概率准则得到最优词序列 $\hat{\mathbf{W}}=W_1W_2\cdots W_m$，即公式(1-1)和(1-2)，

$$\hat{\mathbf{W}}=\arg\max_{\mathbf{W}} P(\mathbf{W}\mid\mathbf{X}) \tag{1-1}$$

利用贝叶斯准则，上式可以表示为

$$\hat{\mathbf{W}}=\arg\max_{\mathbf{W}}\frac{P(\mathbf{X}\mid\mathbf{W})P(\mathbf{W})}{P(\mathbf{X})}=\arg\max_{\mathbf{W}} P(\mathbf{X}\mid\mathbf{W})P(\mathbf{W}) \tag{1-2}$$

在式（1-2）中，$P(\mathbf{X}|\mathbf{W})$表示声学模型匹配得分，$P(\mathbf{W})$表示语言模型得分，$P(\mathbf{X})$对结果没有影响，因此可以在实际计算时忽略掉。语言模型 $P(\mathbf{W})$按如下方式定义，见公式（1-3）和（1-4）（Huang，2001；李净，2005），

$$P(\mathbf{W})=P(W_1W_2\cdots W_m)=\prod_{i=1}^{m}P(W_i\mid H_i) \tag{1-3}$$

其中 W_i表示当前词，而 H_i表示其历史。例如常用的 tri-gram 统计语言模型仅考虑历史中的最近两个词，即

$$P(\mathbf{W})=P(W_1)P(W_2\mid W_1)\prod_{i=3}^{m}P(W_i\mid W_{i-1},W_{i-2}) \tag{1-4}$$

由于本文关注的焦点是声学模型的研究，所以对于语言模型并没有太多涉及，这里为了公式的完整性，也将其列出。

本文将关注的焦点转向声学模型 $P(\mathbf{X}|\mathbf{W})$，$P(\mathbf{X}|\mathbf{W})$表示的是声学得分，它与声学模型、发音词典和解码器有着密切关系。本文在公式（1-2）的基础上，引入音节序列 $\mathbf{Y}$，从而将 $P(\mathbf{X}|\mathbf{W})$扩展为

$$P(\mathbf{X}\mid\mathbf{W})=\sum_{\mathbf{Y}}P(\mathbf{X}\mid\mathbf{Y},\mathbf{W})P(\mathbf{Y}\mid\mathbf{W}) \tag{1-5}$$

其中，序列 $\mathbf{Y}=Y_1Y_2\cdots Y_n$，表示词序列 **W** 对应的音节序列，而 $P(\mathbf{Y}|\mathbf{W})$表示的是词序列 **W** 产生音节序列 **Y** 的概率。在汉语中，由于多音词的存在，同一个词序列可能对应多个音节序列，引入序列 **Y** 可以表示这种情况。同时，本文引入序列 **Y** 也是为了表示音节相关的发音变化。

将音节序列 **Y** 进一步转换为声韵母序列 **B**，$\mathbf{B}=B_1B_2\cdots B_n$。可以得到式（1-6），

$$P(\mathbf{X}\mid\mathbf{W})=\sum_{\mathbf{Y}}\sum_{\mathbf{B}}P(\mathbf{X}\mid\mathbf{B},\mathbf{Y},\mathbf{W})P(\mathbf{B}\mid\mathbf{Y},\mathbf{W})P(\mathbf{Y}\mid\mathbf{W}) \tag{1-6}$$

式（1-6）中，可将序列 **B** 被视为标准发音序列。为了表示多发音现象，本文在式（1-6）的基础上引入序列 **S**，$\mathbf{S}=S_1S_2\cdots S_n$，用以表示可能的实际发音序列，则 $P(\mathbf{X}|\mathbf{B},\mathbf{Y},\mathbf{W})$可表示为

$$P(\mathbf{X}\mid\mathbf{B},\mathbf{Y},\mathbf{W})=\sum_{\mathbf{S}}P(\mathbf{X}\mid\mathbf{S},\mathbf{B},\mathbf{Y},\mathbf{W})P(\mathbf{S}\mid\mathbf{B},\mathbf{Y},\mathbf{W}) \tag{1-7}$$

从而将式（1-6）进一步表示为

$$\begin{aligned}P(\mathbf{X}\mid\mathbf{W})=\sum_{\mathbf{Y}}\sum_{\mathbf{B}}\sum_{\mathbf{S}}&P(\mathbf{X}\mid\mathbf{S},\mathbf{B},\mathbf{Y},\mathbf{W})P(\mathbf{S}\mid\mathbf{B},\mathbf{Y},\mathbf{W})\\&\cdot P(\mathbf{B}\mid\mathbf{Y},\mathbf{W})P(\mathbf{Y}\mid\mathbf{W})\end{aligned} \tag{1-8}$$

由于词序列 **W** 到音节序列 **Y** 和标准声韵母序列 **B** 的对应关系相对确定，那么，$P(\mathbf{B}|\mathbf{Y},\mathbf{W})$和 $P(\mathbf{Y}|\mathbf{W})$就相对稳定，对结果影响较小，此处将其简写为 $P_{\mathbf{W},\mathbf{Y}}(\mathbf{B})$，在实际应用中可以简化处理或者忽略（Huang, 2004）。同时，再将上式中的模型打分条件进行简化，即假设模型仅依赖于实际（方言背景普通话）声韵母发音序列 **S**。从而得到

$$P(\mathbf{X}\mid\mathbf{W})=\sum_{\mathbf{Y}}\sum_{\mathbf{B}}\sum_{\mathbf{S}}P(\mathbf{X}\mid\mathbf{S})P(\mathbf{S}\mid\mathbf{B},\mathbf{Y},\mathbf{W})P_{\mathbf{W},\mathbf{Y}}(\mathbf{B}) \tag{1-9}$$

式（1-9）中，$P(\mathbf{X}|\mathbf{S})$表示的是声学模型匹配分数，$P(\mathbf{S}|\mathbf{B},\mathbf{Y},\mathbf{W})$的物理意义就是发音变化的概率，它们是语音识别声学建模过程中的两个核心问题，因而本文在方言背景普通话语音识别声学模型的研究工作也是围绕这两个方面展开的。

1.3.2 研究的目标及难点

在本节中，将兼顾研究的前瞻性和系统的实用性的原则，确定本文的研究目标和侧重点。

1. 如何充分利用小数据量的方言背景普通话开展研究。这是由方言背景普通话的多样性决定的，中国有九大方言区，次方言区有四十多个，再细分多达上千种（候精一，2002）。因而如何充分利用少量的方言背景普通话数据建立方言背景普通话识别器，在经济上，在可行性上有着巨大的意义。同以往的方法的相比，使用方言背景普通话的数据量要更小，在（Li，2006；Zheng，2005；Sproat，2005；Fung，2005）虽然也是所谓小数据量，但其数据量一般都在 4~6 小时，本文所指的小数据量就是少于 1 小时的语音数据，如何在这样的小数据量下获得大数据量下的系统性能是本文要解决的难点。

本文之所以采用 1 小时的方言背景普通话数据集，是在实验对比的基础上作出的。在语音识别中，系统的性能与训练数据多少有着很密切的关系，对于方言背景普通话语音识别而言，最有效的方式就是采集足够多的方言背景普通话数据进行重新训练建模。在以往的研究中，我们对不同规模的上海普通话数据集（Li，2003）通过重新建模进行了对比，测试数据是上海普通话，其结果如图 1.3 所示：

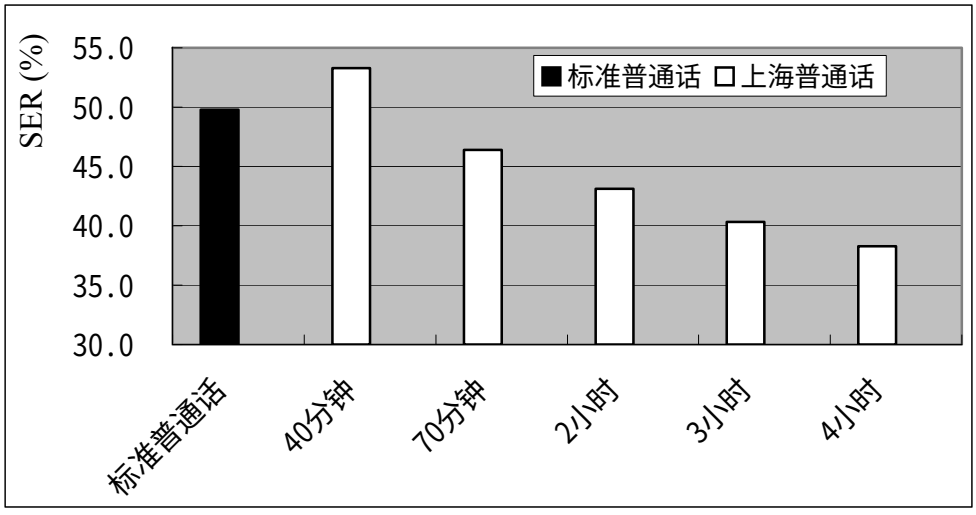


图1.3 不同规模的方言背景普通话数据集重新建模

在图 1.3 中，最左侧黑色立柱代表标准普通话识别器，在文（李净，2005）中就采用了此识别器作为基线系统，其对上海普通话识别的音节错误率是 49.8%。右侧的白色立柱是利用不同规模的上海普通话进行重新建模后的识别结果。可以看到，当上海普通话数据超过 1 小时，重新建模方法对于上海普通话的识别率就超过了标准普通话识别器。这就是本文采用约 1 小时方言背景普通话作为小数据量数据集的直接原因。

2. 如何在提高方言背景普通话识别率的同时保证标准普通话的识别率的不降低。同以往的研究工作不同，本文中，不再一味强调方言背景普通话识别率的提高，而是强调“ 两者兼顾 ”，即提高方言背景普通话识别率的同时保持标准普通话识别率不降低，甚至还要有所提高。这是基于方言背景普通话是普通话在方言地区应用时的一种变体来考虑的，既然它是一种接近于普通话的语言现象，就应该保持标准普通话识别器的性能，因而在研究的目标中就明确规定，要首先保证普通话的识别率，这是实际应用的根本，因为方言口音是一个动态变化，其与普通话的接近程度受到说话人主客观的影响，如何保持对标准普通话的高识别率的同时，对不同轻重口音的方言背景普通话的识别率也有显著提高，是研究的核心和难点。以往的方法，例如自适应、重训练等只能保证单向提高，就是对于要识别的目的语音效果较好，对于原始语音识别往往不尽人意。

3. 本文关注的说话人是在某一方言区长期（15 年以上）生活，以该方言为日常交流语言的人。例如，本文中所采用闽南普通话数据是说话人在书面语的，也就是朗读式发音。将更多的精力关注于口音对于识别率的影响，而淡化其它语言现象的干扰；同时这也是由现阶段研究的实用性所决定的。对应于（Li，2006；Zheng，2005；Sproat，2004）的口音分类就是 2B, 2C, 3A, 3B 级别，就是中度口音和重度口音类别中口音较轻的群体。

1.4 研究的总体思路和创新点

1.4.1 研究的思路及框架

在语音学中，日常的发音从音素的层次来看，标准的识别基元（通常称为基准形式，即 base-form）与实际观测到的识别基元（通常称为表象形式，即 surface-form）之间存在着如下两类变化（Decker，1999；Chen，2000）：

1．声音变化，即一个音素产生略微的发音变化，例如发生了鼻音化、圆唇化、央化、清化、浊化等。此时虽然从听觉上或语谱图上可以觉察或观察到一些略微的畸变（相对于标准普通话发音），但是仍可判断出所发的音属于期望的音素。这往往是由口音、较快的发音速度及前后音素间的协同发音造成的；

2. 音素变化，即一个音素完全变为另一个音素，或者发生音素的插入或删除现象。此时从听觉上或语谱图上，相对于标准普通话发音，都可以觉察或观察到明显的区别。这种情况往往是由较快的发音速度、不同的口音背景或非标准读音造成的；

通过语音学知识可知，由于受母方言的影响，方言背景普通话和标准普通话之间在语音学领域主要存在这两类差别：声音变化和音素变化。因而本文的声学建模研究也是从这两方面作为切入点的，这与公式（1-9）的形式化描述是一致的。

针对方言背景普通话中的音素变化，本文采用了方言背景普通话发音字典的策略，其目标就是生成反映方言背景普通话发音特点的字典。在方言背景普通话发音字典的生成过程中，提出了改进的上下文相关的权重策略（ICDW），它是对常用的上下文相关权重策略的一种改进（宋战江，2001；Zheng，2002），解决了方言背景普通话发音字典中可能出现的声韵母不合理组合的现象。该方法首先在较大的方言背景普通话数据集上进行了验证，证明此方法的有效性。在此基础上，将改进的上下文相关的权重策略应用到基于小数据量的方言背景普通话发音字典的生成中，因为基于小数量的研究是本文的支撑点之一。在基于小数据量的方言背景普通话发音字典生成过程中，由于数据量的急剧减少会出现数据稀疏问题，容易造成发音变化概率估计出现较大偏差，为解决数据稀疏问题，本文又提出了基于距离度量的识别网络生成策略。关于此部分的研究思路如图 1.4 所示：

在图 1.4 中，方言背景普通话发音字典 $L1$ 是在较大数据量基础上实现的，它不是本文研究的重点，它只是基于小数据的方言背景普通话发音字典 $L2$ 的参考值，也就是说，方言背景普通话发音字典要在小数据量的条件下，达到大数据量条件下相近的性能。在图 1.4 左侧部分中，在方言背景普通话发音字典生成中，提出了改进的上下文相关权重策略，在基于小数据量的方言背景普通话发音字典生成中，提出了基于距离度量的识别网络生成策略，并与改进的上下文权重策略进行了结合。

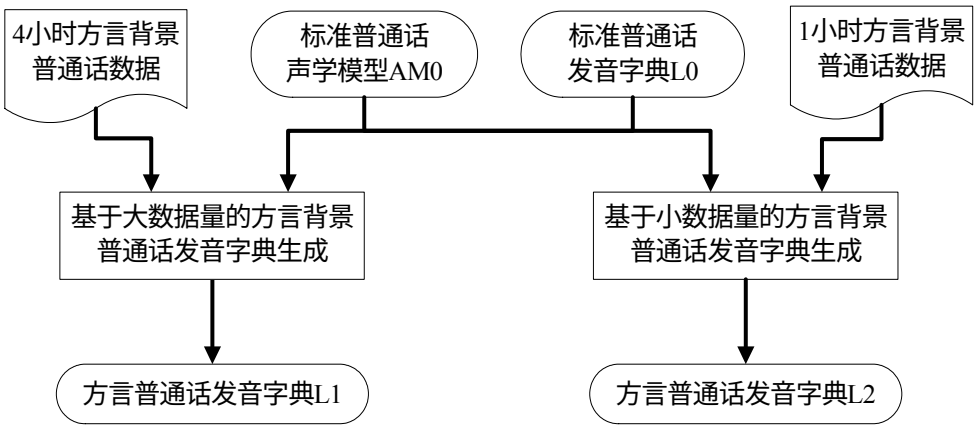


图1.4 方言背景普通话发音字典研究思路

在本文中，针对声音变化现象，提出了一种简单而有效的声学建模方法，状态相关基于基元的模型归并 SDPBMM，在此方法中，来自于方言背景普通话的上下文无关的声韵母 HMM 和来自于标准普通话的上下文相关的声韵母 HMM 在模型状态级，依据的一定的准则（状态级的发音建模和模型插值），将其所包含的高斯混合进行归并。此部分的研究思路流程如图 1.5 所示，图中灰色部分，**状态相关基于基元的模型归并**和针对 SDPBMM 中存在的**高斯混合扩张问题（GMEP）的处理**是作者在此部分的创新工作。此方法的好处在于：1）解决方言背景普通话小数据量的问题，因为训练鲁棒的上下文无关的声韵母 HMM 需要的数据量较小；2）扩大了标准普通话声韵母 HMM 在声学空间上的覆盖度，使进一步提高标准普通话的识别率成为可能；3）另一方面，借助于标准普通话上下文相关的声韵母 HMM 的模型精度，通过与标准普通话声韵母 HMM 的结合，也显著提高了方言背景普通话的识别率。

由于在 SDPBMM 的策略中存在高斯混合扩张的问题，为解决此问题，在基于状态距离度量的基础上，本文提出了可选择的 SDPBMM，依据声学空间上的相似度来对状态进行鉴别，从而决定是否参与归并。这样一来，既可以降低声学模型规模，减少模型的冗余度，又可以保证声学模型的性能不下降。

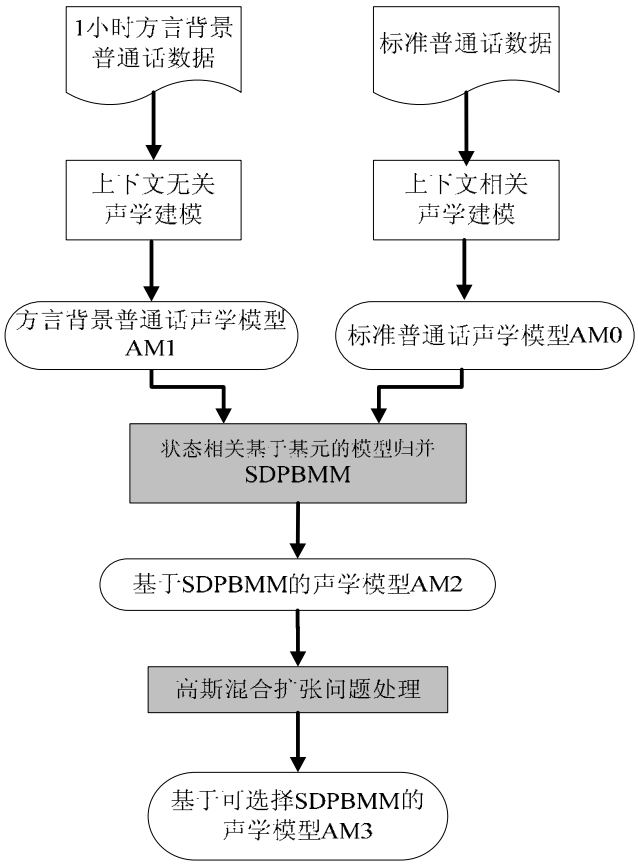


图1.5 针对方言背景普通话中声音变化的声学建模

1.4.2 论文创新点概述

围绕小数据量和提高方言背景普通话识别率的同时不降低标准普通话识别率的目标，本文提出了若干新思路并通过实验证明了其有效性，同时也为方言背景普通话语音识别的后续研究积累了一定的经验。其创新点可以概括为如下四点：

1. 针对方言背景普通话中的音素变化现象，对发音字典自适应进行了深入的研究。在上下文相关权重策略的基础上，提出了改进的上下文相关权重策略，解决了发音字典中音节对应不合理声韵母组合的现象。

2. 针对方言背景普通话中的声音变化现象，提出了状态相关基于基元的模型归并策略，它能很好的解决标准普通话模型在方言背景普通话空间覆盖度不够的问题，显著的提高了方言背景普通话的识别率；同时，它又保证了标准普

通话的识别率的不降低。其性能优于自适应、重训练和模型插值方法。

3. 解决了小数据量情况下，鲁棒的发音建模问题。难点是如何解决数据稀疏问题，为此本文提出了基于距离度量的识别网络生成策略来实现基于小数据量的发音建模。此方法应用于 1) 方言背景普通话发音字典; 2) SDPBMM 中，作为状态归并的准则之一。

4. 为解决 SDPBMM 方法中的高斯混合扩张问题，提出了可选择的 SDPBMM。在降低声学模型规模的同时，保证了识别率不降低，有效地缓解了识别率的提高和模型规模扩张之间的矛盾。

1.5 论文结构安排

论文的内容和结构是围绕方言背景普通话声学建模中音素变化和声音变化两个基本问题的解决而展开的。

第 2 章中，首先对标准普通话声学建模过程中的一些难点问题进行简要的介绍；其次对研究中采用的标准普通话和闽南普通话数据进行了简要介绍，并在一定原则的基础上，对数据集进行了划分。在此基础上对标准普通话识别器的性能进行了验证。

第 3 章中，主要工作是方言背景普通话发音字典的生成。首先，在基于较大数据量基础上，提出了改进的上下文相关权重策略；其次，以此为参考上限，在小数据量基础上提出了基于距离度量的识别网络生成策略，并在方言背景普通话发音字典生成中，与改进的上下文权重策略进行了结合。

第 4 章中，状态相关基于基元的模型归并 SDPBMM 及高斯混合扩张处理。首先，对 SDPBMM 进行的深入的研究和对比；其次对 SDPBMM 中存在的高斯混合扩张问题进行了处理，提出了可选择 SDPBMM 的方法。

第 5 章中，对各种方法之间的联系和效果等进行归纳和总结，

第 6 章中，对本文的工作进行了总结，并对未来的工作进行了展望。

在附录 A 中，为了进一步证明 SDPBMM 的有效性，增强说服力，相关策略在上海普通话数据集上进行了验证和分析。

第 2 章 标准普通话识别器

2.1 标准普通话声学建模

通常情况下，一个大词表连续语音识别系统（LVCSR）由声学模型、语言模型、发音字典、解码器等部分构成。本文只对基于汉语普通话的大词表连续语音识别声学建模中的一些难点问题进行简要阐述，详细信息可参见（Huang，2001；Young，2002；李净，2005）。

2.1.1 识别基元

语音识别中，识别基元的选择可以是基于语音学知识的（例如，音节、音素、声韵母等）（杨行峻，1995），也可以是基于数据驱动方式的，例如，状态级基元 senone（Hwang, 1993）。同时，基元的选择也依赖于具体的应用，对于小型系统，如语音命令与控制系统等，一般可以采用较长的语音段作为基元，如词或者音节；而对于大词表连续语音识别，一般选用较短的语音段作为基元，且为了描述连续语音中的协同发音现象，往往还要考虑上下文相关性，如常用的三元音素基元（triphone）。汉语的发音是基于音节结构的，直接将音节作为识别基元是一个显见的选择（郑方，1999），但由于其个数较多（汉语音节的全集为 418 个无调音节，或大约 1200 多个有调音节），需要有很大规模的数据库才能避免训练数据的稀疏；音素通常可以认为是最小的发音单位，在西方语言的语音识别中被广泛采用。现代汉语中，标准普通话中有 35 个音素，数目较少，便于建立上下文相关模型。但音素并没有完全反映出汉语语音的特点，而且相对于声韵母，音素显得不够稳定，这就给标注带来了困难，也会影响模型的稳定性。标准普通话中有 59 个无调声韵母（IF），如表 2.1 所示。声韵结构是汉语音节特有的结构，使用声韵母作为识别基元具有以下优点：1）声韵更能反映汉语的特点，汉语中的汉字是单音节结构的，而音节又具有独特的声韵结构；2）有许多语音学知识的研究成果是基于声韵母的，它们可以用来指导声学模型训练；3）基元数目和语音段长度比较适当。与音节比，便于建立上下文相关模型；与音素比，稳定性好。

表2.1 标准声韵母基元列表

声母基元（21）	韵母基元（38）
b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s	a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, ii, iii, ia, ian, iang, iao, ie, in, ing, iong, iou,u, ua, uai, uan, uang, uei, uen, ueng, uo,v, van, ve, vn

在实际应用中，很多研究者都在表 2.1 所示的标准声韵母集合的基础上进行了一定的改进，例如（Li，2001）就提出了扩展声韵母 XIF 集合的概念，增加了{_a，_o，_e，_y，_w，_v}，6 个零声母。这样一来，每个汉语音节就真正的由一个声母（或零声母）和一个韵母构成，避免了单韵母音节的出现，其目的就是使建模过程更加一致，在进行上下文相关建模时能够降低混淆度（每个声母后面只能接韵母，韵母后面只能接声母），这样做的好处之一就是缓解了数据稀疏问题。考虑到汉语是有声调的，文（Cao, 2004）采用了带调的韵母作为识别基元。文（宋战江，2001）考虑到自然发音方式中发音变化更为灵活的特点，针对标准普通话声韵集覆盖度不够的问题，提出了广义声韵母 GIF 的概念。文（Liu，2003）通过置信度的方法自动生成声韵母集合。文（Chen，2002）利用卡方检验（chi-square testing）统计方法来降低基元之间的混淆，得到了更稳定的基元集合。在方言背景普通话或者带口音的汉语语音识别中，也存在标准普通话声韵母对方言背景普通话的发音特点刻画不够精确的问题，文（李净，2005）通过利用专家知识，针对上海普通话的发音特点，对标准声韵母集合进行了必要的扩展。

2.1.2 上下文相关建模中的状态共享

在连续语音中，协同发音是十分普遍的现象，因此，建立上下文相关模型来刻画协同发音，提高识别率是非常必要的。为解决上下文相关建模时的数据稀疏问题，同时也为了解决有些上下文相关基元模型在训练数据中没有出现的问题，基于决策树的状态共享是最常用的策略（Young，1994；Reichl，2000； Hwang，1996；Liu，2004a）。

基于决策树的状态共享策略已经广泛地应用于改善大词表连续语音识别系统的声学模型性能。决策树是一个二叉树，每个非叶子结点都绑定

着一个“是/否”问题，所有允许进入根结点的 HMM 状态要回答结点上绑定的问题，根据回答的结果选择进入左枝还是右枝。最后，每个进入根结点的 HMM 状态都会根据对一系列结点问题的回答进入设定的一个叶子结点。进入同一个叶子结点的 HMM 状态会被认为是相似的，叶子结点中所包含的参数将依据一定的规则由进入该结点的多个状态所共享。其基本原理如图 2.1 所示：

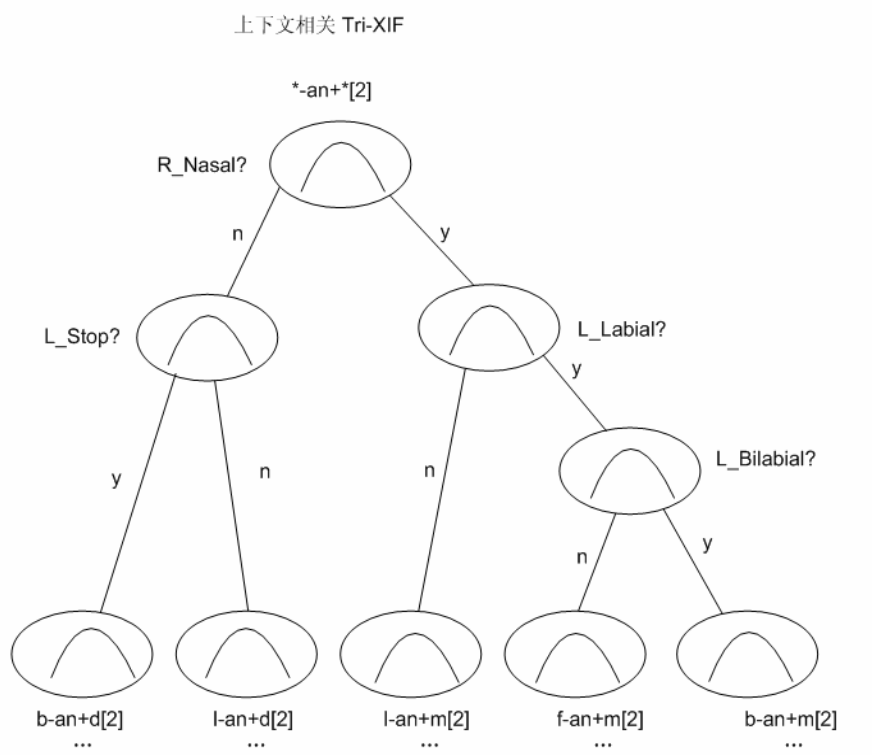


图2.1 基于决策树的状态共享的结构图

图 2.1 中，*-*an*+*[2]，代表所有以韵母 *an* 为中心基元的上下文相关三元扩展声韵母 Tri-XIF 模型的第二个状态，这些状态在初始阶段被放在决策树的根部，认为初始时所有的状态都属于同一个类。其中，符号“-”左面代表中心基元的左相关基元，对于图中的韵母 *an*，其左相关基元就是一个声母；同样，符号“+”右面代表中心基元的右相关基元。每个非叶结点绑定一个问题（例如，*R_Nasal* 表示当前中心基元的右相关基元是鼻音吗？），这些状态通过回答一系列的问题，最终将被划归为若干类，即图中叶子结点。处于叶子结点中的状态将共享相同的模型参数（对于 HMM

就是高斯混合）。

基于决策树的状态共享策略应用于上下文相关声韵母建模时通常有以下几个主要问题：问题集的设计，状态共享策略，结点分裂和停止分裂准则及剪枝准则。

1. 问题集的设计。问题集设计是否合理直接影响着系统的识别性能，因为决策树是根据问题的结果来决定状态之间是否共享相同的模型参数，通常这些问题都是来自于语音学的先验知识。文（Li，2001）根据语音学知识设计了标准普通话的声韵母建模的问题集，根据发音部位、发音方法的不同、其中的声母类有 22 个，韵母类有 39 个。按照声韵母作为问题类，就是认为属于同一类的上下文的声韵母对于中心基元（声韵母）在发音上具有相似的影响（Young, 1994）。在（Chen，2002）也根据类似的方法设计了 90 个问题，分为通用类、辅音类和元音类。在（Beulen，1998）中，给定音素集合情况下，提出了一种问题集自动生成策略，解决了基于语音知识的问题集灵活性不足的缺点。

2. 状态共享策略。状态共享策略指的是哪些基元的哪些状态可以被共享到一起，而哪些不允许。一般地，只有中心基元相同而上下文不同的基元才会进行共享，且主要考虑两种策略，一种是状态相关的，一种是状态无关的。如果进行状态相关的共享策略，则只有对应的状态才可能共享到一起。反之，如果进行状态无关的共享策略，则不同状态也可能进行共享，如，第一个状态和第三个状态可能共享到一起。考虑到语音信号的时序性特点，一般采用的是状态相关的共享策略，例如图 2.1 就是以韵母 *an* 为中心基元的上下文相关声韵母 HMM 的第二个状态之间进行共享。

3. 结点分裂和停止分裂准则。在构建决策树时，首先要将所有中心基元相同的 HMM 的同一状态放入一个状态共享池中，对应图 2.1 中就是所有以 *an* 为中心基元的 tri-XIF HMM 中的第二个状态，然后根据一定的分裂准则进行逐级分裂，当满足停止分裂准则时，分裂过程停止。其中常用的方法有：1）最大似然准则（maximum likelihood criterion, ML），即选择结点分裂后似然分增加最大的问题作为本结点绑定的问题（Young，1994）。决策树的停止分裂一般采用似然分数阈值和训练样本数目两个条件进行控制; 2）贝叶斯信息准则（Bayesian information criterion, BIC）（Rissanen，1984），在分裂过程中引入了一个惩罚因子，用来在似然分

数和模型复杂度之间进行平衡，通过调整这个因子来调整决策树的规模。3）聚类有效性准则（cluster validity criterion）（Chien, 2005），此方法采用 Hubert's Γ statistic 作为结点分裂的准则，采用 T2-statistic 作为停止分裂准则。本文出于快速和计算简便的考虑，采用了最大似然准则。

2.1.3 高斯混合分裂

在大词表连续语音识别系统中，为提高识别率，一般情况下，每个基元模型的状态中都包含多个高斯混合。在基于决策树的状态共享过程为了计算简便和考虑到训练数据量的限制，一般都采用单高斯混合。在完成状态共享之后，在上下文相关的声学建模中，要进行高斯混合分裂。分裂发生在参数重估（re-estimation）之前，其分裂公式如（2-1）（Simonin，1996）：

$$\begin{cases} \mu_k^{'} = \mu_k + \varepsilon \cdot (\Sigma_k) \\ \mu_k^{''} = \mu_k - \varepsilon \cdot (\Sigma_k) \\ \Sigma_k^{'} = \Sigma_k^{''} = \Sigma_k \\ w_k^{'} = w_k^{''} = w_k / 2 \end{cases} \tag{2-1}$$

其中， μ_k，Σ_k，w_k 分别是待分裂的高斯混合的均值、方差和权重，分裂后的参数依次为$\mu_k^{'}$，$\mu_k^{''}$，$\Sigma_k^{'}$，$\Sigma_k^{''}$，$w_k^{'}$，$w_k^{''}$。ε是均值扰动因子，(Σ_k)表示Σ_k的对角阵。在实际应用中，往往选择权重最大的高斯混合进行分裂（Young, 1994），而当高斯混合的权重低于某一阈值时，分裂将不再进行。一个状态中含有的最佳高斯混合的数目往往和训练数据量有关系，如果训练数据不足，反复用之迭代，则很容易造成过训练（overtraining），反而造成识别率下降。在大词表连续语音识别中，通常根据训练数据量和模型复杂度来综合决定状态中包含的高斯混合数目的多少。

2.2 数据库及数据划分

本文中采用了两个数据库，一个是标准普通话数据库（DB_STD），另一个是闽南普通话数据库（DB_MIN），根据研究方案初始的设计思想，就是用鲁棒的标准普通话识别器结合少量的方言背景普通话数据来实现一个方言背景普通

话识别器。因而在数据库设计中，标准普通话数据库要远远大于闽南普通话数据库。DB_STD 的候选语料来源于人民日报语料库，该语料库共有 800,000 个句子，每个句子包含 15～20 个汉字。DB_STD 需要从中选择 26,400 个句子，在设计时应用鼓励低频单元 ELF 算法（Li，2003）来达到右相关声韵母基元均衡。DB_STD 的主要信息如表 2.2 所示：

表2.2 标准普通话数据库DB_STD的主要信息

条 目	内 容
说话人	132 人，男女均衡，10~50 岁，普通话发音标准
语音内容	200 句子/人，10 数字/人，26 字母/人，另有 32 人，每人还包含 200 个短语
采样率	16,000Hz, 16bit, 麦克风
标注	汉字，音节和声韵母标注

闽南普通话数据库 DB_MIN 设计思想与 DB_STD 是相同的，其语料同样来自于人民日报，只是增加了 700 个体现闽南普通话发音特点的短语。其主要信息如表 2.3 所示：

表2.3 闽南普通话数据库DB_MIN的主要信息

条 目	内 容
说话人	36 人，男女均衡，20~30 岁,70%中度口音，30%重度口音，厦门出生并居住 15 年以上
语音内容	200 句子/人，10 数字/人，26 字母/人，200 短语/人
采样率	16,000Hz, 16bit, 麦克风
标注	汉字，音节和声韵母标注（其中声韵母标注是由标准普通话识别器通过强制对齐得到）

由表 2.2 和 2.3 可以看出，标准普通话数据库 DB_STD 和闽南普通话数据库 DB_MIN 具有很多相似之处，其语料来源一致，录音环境完全相同，不存在信道差异，其差异几乎被限定在方言口音上，即方言背景普通话和标准普通话的差异集中于声学层面，这样一来，就排除了信道差异的影响，排除了语言、语法上的差异，将研究的关注点完全集中在声学层面上，这对开展方言背景普通话语音识别声学建模的研究是十分理想的实验环境。本文将两个数据库划分为 4 个互不相交的数据集，所有的数据集都是男女均衡的，其基本信息如表 2.4 所示：

表2.4 数据集的划分

名 称	用 途	内 容
TRAIN_STD	标准普通话训练集	120 人，200 句子/人，共 24,000 个句子，约 25 小时
TEST_STD	标准普通话测试集	12 人，100 短语/人，共 1,200 句，包含 1,126 个不同的短语，约 70 分钟
DEV_MIN	闽南普通话开发集	20 人，50 句子/人，共 1,000 句，约 1 小时
TEST_MIN	闽南普通话测试集	16 人，75 个短语/人，共 1,200 句，约 70 分钟; 其中包含 1,129 个不同的短语，60%中度口音，40%重度口音

对于数据集的划分基于以下原则：

1. 利用大量的标准普通话数据训练的鲁棒的标准普通话声学模型，即 TRAIN_STD 数据集；

2. 利用少量的方言背景普通话数据作为开发集，即 DEV_MIN，约 1 小时；

3. 两个测试集，TEST_STD 和 TEST_MIN，具有相同的数据量及相同的复杂度的词表。

2.3 标准普通话识别器性能

在建立标准普通话识别器的过程中，鲁棒的声学模型对系统的性能有着至

关重要的影响。本文采用了文（Li，2001）中的扩展声韵母 XIF 集作为识别基元集合，来进行上下文相关的声学建模。以数据集，TRAIN_STD，作为训练集，提取了 39 维的 MFCC 特征（Davis，1980），包括 13 维倒谱特征，以及一阶差分Δ和二阶差分ΔΔ。在计算特征时，去掉了约 100Hz 以下的低频部分，并使用了 1 秒窗宽进行倒谱均值归一化（CMN）（Viikki，1998）；对于每一个声韵母 HMM 采用了图 2.2 中 a）所示的拓扑结构；考虑到长、短静音的现象，对于静音 HMM 采用了图 2.2 中 b）所示的拓扑结构。由 2.2 图可以知道，每一个声韵母 HMM 由三个物理状态组成。具体的建模过程在文（李净，2005）有着详尽的描述，在此不再赘述。

a) b)

图2.2 声韵母和静音HMM拓扑结构图

本文关注的焦点是声学建模，为了排除语义、语法等高层信息的影响，又考虑到具体的应用环境，在研究过程中没有使用语言模型。由于测试集都是由命令短语组成，本文统一以词错误率 WER 作为评价标准。建模过程中利用了 HTK3.2（Young，2002）作为建模工具。标准普通话识别器基本信息如表 2.5 所示：

表2.5 标准普通话识别器

名 称	内 容
上下文无关 mono-XIF 数目	65
上下文相关 tri-XIF 数目	10,629
物理状态数目	3,803
高斯混合数目	53,382，每个状态包含 14 个高斯混合
发音字典	402 个无调音节，每个音节对应一种发音

在标准普通话声学建模过程中，针对在节 2.1.3 提到的高斯混合分裂，本文通过实验对比的方法来确定 HMM 中的每个状态应包含的最佳高斯混合数目，对比实验中采用 TEST_STD 作为测试集，其结果如图 2.3 所示。其中，横轴代表每个状态包含的高斯混合数目，纵轴代表词错误率。

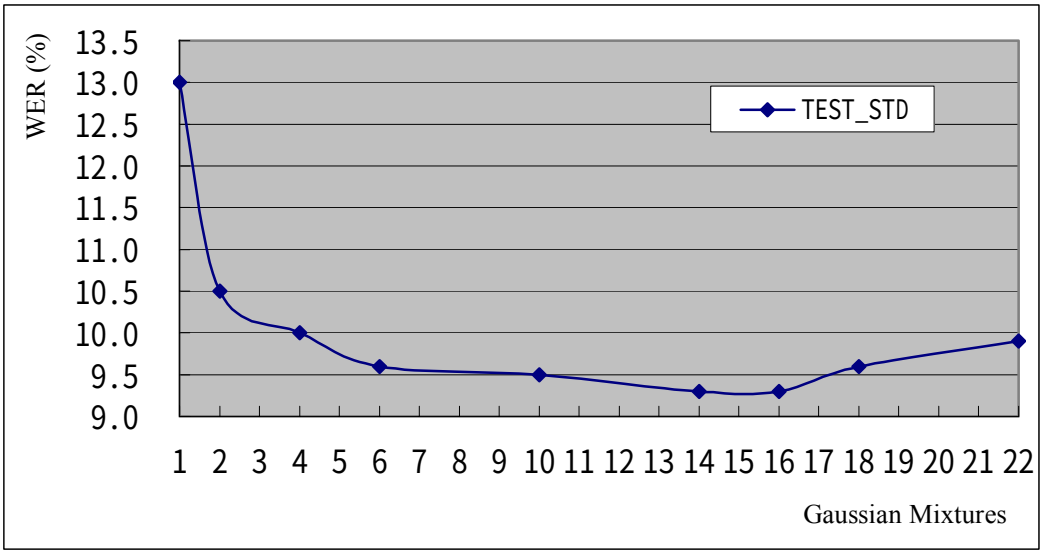


图2.3 HMM状态包含的高斯混合数目对识别性能的影响

从图 2.3 中，可以看出，随着每个状态包含的高斯混合数目的增加，模型精度提高，识别率也相应的提高；同时也可以看到当高斯混合数目增大到一定数值的时候，识别率不再有明显提高甚至还有可能降低，如图中所示，当高斯混合数目超过 14 的时候，识别率不再明显提高，当高斯混合数目超过 16 时，识别率有所下降。这与文（Gouws，2004）实验结果是一致的，此文中也证实当高斯混合的数目超过 14 之后，识别率不再有明显提高。这种现象很大程度是因为高斯混合分裂时，需要更多的训练数据来进行模型优化，而实际训练数据不足造成的。根据对比结果，在本文的标准普通话声学模型中，每个状态包含 14 个高斯混合。另外，可以得出结论，单纯增加状态中包含的高斯混合数目是不能有效提高识别率的，这个结论对后面的研究很有借鉴意义，在第 4 章这个结论将与 SDPBMM 的结果进行对比。

在标准普通话声学建模完成之后，本文对标准普通话识别器在标准普通话和闽南普通话上分别进行了测试，其结果如表 2.6 所示：

表2.6 标准普通话识别器性能

识别器	WER	
	TEST_STD	TEST_MIN
标准普通话	9.3%	23.4%

由表 2.6 可以看出，对于标准普通话和闽南普通话而言，它们之间在错误率上存在 14.1%的差距，可以认为，这个差距完全是由方言口音引起的，本文所做的工作就是在不降低标准普通话识别率的前提下，尽量缩小这个差距。此外，这里的标准普通话识别器的结果将作为本文的基线系统（baseline）。

2.4 闽南普通话上下文无关声学建模

对于闽南普通话，采用 DEV_MIN 作为开发训练集来进行上下文无关 mono-XIF HMM 声学建模，这样做的原因有：1）数据量的限制，上下文无关建模比上下文相关建模需要的数据少得多；2）根据以往的经验，在小于 1 小时的数据量情况下，因为数据稀疏问题，上下文相关建模不会显著提高识别性能（参见本文第 1 章图 1.3）。表 2.7 给出了闽南普通话上下文无关声学模型的测试结果，在此声学模型中，每个声韵母 HMM 的状态采用了与标准普通话声韵母 HMM 状态完全相同的拓扑结构，即图 2.2 所示，另外，每个状态含有 8 个高斯混合，实验中与标准普通话声学模型一样，采用了标准普通话发音字典。

表2.7 闽南普通话mono-XIF声学模型性能

声学模型	WER	
	TEST_STD	TEST_MIN
闽南普通话 mono-XIF	26.7%	22.6%

可以看出，基于少量方言背景普通话数据进行声学建模，即便对方言背景普通话识别结果也很不理想。此外，在第 3 章和第 4 章中，都将用到这个声学模型来生成发音字典和完成状态相关基于基元的模型归并 SDPBMM。

第 3 章　方言背景普通话发音字典

3.1　引言

与方言背景的影响相比，发音变化的概念更为宽泛。它们之间有许多共通之处，但又不尽相同。发音变化一般是由口音、语速、说话习惯、上下文不同等造成的，常表现为声学层面的发音变异，如音素或声韵母的替代、插入、删除错误等，即所谓的音素变化。对于此类“错误”，一般采用字典自适应的方法进行处理，即在发音字典中加入反映说话人发音特点的实际发音，因为在音素变化中，一个音素（或声韵母）往往被另外一个发音相近的音素（声韵母）所替换。在方言背景普通话语音识别中，由于说话人受到同一种母方言的影响，其发音特点除了带有群体性特点之外，往往还带有一定的规律性和倾向性，文（Goronzy，2001；Huang，2004）认为这种规律性通过方言背景普通话发音字典来体现是最合适的。

本章的前半部分将对方言背景普通话发音字典进行深入的研究和阐述，着重解决如何得到准确的发音变化权重问题，提出了改进的上下文相关的权重策略，以及采用更合理的剪枝策略来降低发音字典的混淆度；方言背景普通话发音字典的生成是在较多数据的基础上，通过数据驱动的方式实现的。在本章的后半部分，研究的重点是如何采用少量的数据来建立方言背景普通话发音字典，目标是将前半部分基于大数据量的研究成果应用到基于小数据量的发音字典生成中，同时还要保证识别率不降低。前半部分的结果将作为后半部分（即基于小数据量的方言背景普通话发音字典）的参考值，也就是识别率的上限。

在本章的节 3.2 中，首先对字典自适应中的三个重点问题进行总结分析；在此基础上，在节 3.3 中提出了改进的上下文相关的权重策略；节 3.4 中对不同方法得到的方言背景普通话发音字典进行了对比；在节 3.5，针对基于小数据量的方言背景普通话发音字典生成过程出现的数据稀疏问题，提出了基于距离度量的识别网络生成策略，并定义了对距离度量准则的评价策略；节 3.6 对基于小数据量的发音字典进行了测试；节 3.7 给出了本章的小结。

3.2 发音字典自适应中的重点问题

广义上，对于发音字典的构造有两种方法：基于专家知识（knowledge-based）（Hoste，2004）和基于数据驱动（data-driven）（Wester，2003；Strik，1999；Amdal，2003）。在基于专家知识的发音字典自适应方法中，字典中的每一个实际发音项由语音学家给出，或者遵循一定的发音变化规则，通过规则的限制，决定一个标准普通话的音节/声韵母在方言背景普通话中的实际发音。这些规则往往同语音的上下文相联系。例如：$A \rightarrow B / C\_D$，表示满足左相关 C 和右相关 D 的情况下，A 发音变化为 B。基于专家知识的字典自适应方法具有以下优点：1）通用性，针对特定的方言背景普通话发音特点总结带有规律性的规则，从而很容易推断出实际发音；2）不需要方言背景普通话语音数据。这是其相对于数据驱动方法的一大优点。在实际应用中，尤其是在没有方言背景普通话数据情况下，通过对规则的应用来生成特定的方言背景普通话发音字典，从而达到提高方言背景普通话识别率的目的，这是一种高效而又经济的方法。也存在一些缺陷：1）过于依赖专家的先验知识。这些知识或者规则是语音专家经过长期观察得到的。对于一些小的方言区，受限于专家人数和研究时间，要得到客观准确的语音学知识可能很困难；2）专家知识可能与待识别语音不匹配，毕竟专家知识是通过大量观察得到的，带有很强的通用性，而实际情况又是千差万别的，难免会出现与实际的识别语音不一致的情况；3）利用专家知识很难得到准确的发音变化概率，对于这一点，在基于专家知识的方法中一般采用等概率的方式来处理。在基于数据驱动的方法中，发音字典所需要的信息来源于训练数据，通过一定的策略来获取实际的发音变化规则。基于数据驱动的方法有以下优点：1）不需要太多的人工干预，节省资源；2）能与待识别语音数据很好的匹配；3）具有很好的可推广性，很容易得到不同的标准普通话与方言背景普通话之间的发音变化规律。4）能较精确的刻画发音变化概率。当然它也存在一些不足：1）需要较多的数据，特别是考虑上下文语境的影响，需要刻画发音变化的规律越精细，需要的数据量也越大。2）由于引入大量的发音变化项，导致发音字典的混淆度增加，反而使识别正确率下降。3）对训练数据有很强的依赖性，不同的训练数据得到的发音字典有所不同。

由于本文的研究目标是方言背景普通话语音识别中的通用方法，并不涉及具体方言背景普通话的专家知识，所以采用基于数据驱动的方式来得到方言背景普通话的发音字典。在基于数据驱动的发音字典自适应过程中需要着重解决

以下三个基本问题（Tsai，2007；Lussier，2003）：

3.2.1 如何获得发音变化

发音字典自适应中，发音变化的获得一般是利用标注，也就是通过基准形式标注和表象形式标注的对齐来实现。表象形式的标注的来源，一是由语音工作者进行人工标注；二是利用现成的识别器进行识别。在基于数据驱动的方式中，采用第二种方法，即采用基于音素或者声韵母的识别器（phone recognizer）对方言背景普通话数据进行无任何限制的基于音素（unconstrained phone loop）的解码。而后利用动态规划（Huang，2001）的算法将对基准形式和表象形式的标注进行对齐，得到从基准到表象的发音变化。获得这些发音变化常用的方法有，混淆矩阵（宋战江，2001；Huang，2004；刘明宽，2002；潘复平，2005）；决策树（Humphriesy，1996；Hoste，2004）；基于最大似然估计（Holter，1999）；基于置信度（Williams，1998）。其中混淆矩阵应用最为广泛，本文也采用了这种方法。在汉语语音识别中，在获得描述变化的时候一般是以音节或者声韵母为单位。考虑到数据量的限制，采用以声韵母为单位的发音变化规则更有助于减小发音字典的混淆度，另外需要的数据量也小很多，在很大程度上能避免数据稀疏问题。本文中将利用标准普通话识别器对方言背景普通话数据集进行基于声韵母的解码，即所谓的音素识别（phoneme recognition）（Lussier，2003），换言之，声韵母集合中任何一个声韵母都可能是识别结果。本文选择得分最高的声韵母作为识别结果，即得到 1-best 的表象形式的标注。此后通过动态规划将基准形式标注与表象形式标注进行对齐，以此为基础得到基于声韵母的混淆矩阵。

3.2.2 如何得到准确的发音变化概率

假设一段语音数据的基准形式的发音是 $B=b_1b_2\ldots b_N$，其对应的一种表象形式的发音为 $S=s_1s_2\ldots s_N$，若应用基元独立性假设，则发音变化概率的计算公式为：

$$P\left(S|B\right)\approx\prod_{j=1}^{N}P\left(s_j|b_j\right) \tag{3-1}$$

如果考虑到上下文语境的影响，则公式（3-1）中的基元变化概率 $P\left(s|b\right)$ 可以表示为：

$$P\left(s|b\right)=\sum_{c}P\left(s|b,c\right)P\left(c|b\right) \tag{3-2}$$

其中，c 代表当前考察的基准形式基元 b 的上下文相关的基元，既可以是左相关，也可以是右相关。例如，在汉语中，如果当前基元是韵母，那么其在基准形式的左、右相关就是声母（或者零声母）[②]，这就是所谓的上下文相关策略。如果不考虑上下文的影响，那么就是所谓的上下文无关策略。显然上下文相关策略对于发音变化概率的描述要精确，但同时也需要更多的数据。特别的，对于方言背景普通话语音识别中，以汉语音节为单位的发音字典，公式（3-1）可表示为：

$$P\left(Syl_{dc}|Syl_{sc}\right)\approx P\left(I_{dc}|I_{sc}\right)\cdot P\left(F_{dc}|F_{sc}\right) \tag{3-3}$$

公式（3-3）中，Syl 代表汉语音节，sc 和 dc 分别表示标准普通话和方言背景普通话，I 和 F 分别代表声母和韵母（下同）。对于声韵母之间的发音变化，如果采用上下文相关策略，则表示为：

$$P\left(IF_{dc}|IF_{sc}\right)=\sum_{C}P\left(IF_{dc}|IF_{sc},C_{sc}\right)P\left(C_{sc}|IF_{sc}\right) \tag{3-4}$$

其中，IF 表示一个声母或者韵母，C 表示基准形式标注的上下文相关声韵母。一般在实际中，采用左相关，这是假设一个声韵母受最紧邻的前一个声韵母的影响要更大，尤其在一个音节内部的发音变化，这种现象更加明显。当然如果同时考虑左右相关，那么需要的数据量更大，很有可能出现数据稀疏问题，那样一来，概率估计就会出现偏差，就失去采用上下文相关策略的意义了，所以必须在准确性和数据量之间进行折衷。文（宋战江，2001；Zheng，2001）都采

[②] 这里为了简化计算，没有考虑插入、删除的情况。

用了左相关的策略。

3.2.3 采用何种剪枝策略

在基于数据驱动的发音字典自适应过程中会产生大量可能的发音变化。发音变化趋势不够集中的主要原因，是由于在基于音素识别的解码过程中的识别器错误，以及方言背景普通话与标准普通话发音“偏差”造成的。如何选择最具代表性的发音变化，就是所谓的剪枝策略，因为过多的发音变化会造成混淆，降低识别率，另外也会增大搜索空间，增加时间开销（Strik，1999）。常用的剪枝策略有以下几种：

1. 基于概率值的剪枝策略（Riley，1991）。将每一个基元的发音变化，根据概率从大到小排序，若其概率值大于一个预先设定的阈值，则保留，否则丢弃；

2. 基于固定数目的剪枝策略（Lussier，1999）。依据概率值将每一个基元发音变化按降序排列，前 n 个保留，大于 n 的丢弃；

3. 累积概率策略（宋战江，2001）。每一个基元的发音变化根据概率从大到小排序后，按大到小的顺序依次累加，当其和超过某一个阈值的时候就停止，其和在阈值范围内的发音变化则保留，其余的丢弃；

4. 基于 uigram 的剪枝策略（李净，2005）。根据音节或者声韵母在训练数据中出现的频率决定是否拥有多个发音变化。换句话说，常用词在最终的发音字典中更有可能包含多个发音变化。在文（李净，2005）中通过应用此方法，同方法 1 相比，音节错误率多降低了 0.8%；

5. 相对最大概率策略(relative-to-maximum)（Stolcke, 2000）。取每一个基元的发音变化概率的最大值 P_{max} 作为参考值，其它发音变化的概率值大于 $\alpha \cdot P_{\max}\left(0<\alpha \leq 1\right)$ 时则保留，否则丢弃。这个策略就是保留概率值相对于最大概率值的一定区间范围内的发音变化。

3.3 改进的上下文相关权重策略

在上下文相关的权重策略中，在公式（3-3）中，应用了独立性假设，虽然在公式（3-4）中考虑到了基准形式（标准普通话）的声韵母之间的关联关系，

但没有考虑到表象形式（方言背景普通话）声韵母之间的关联关系，因而会存在一个问题，无法区分不合理的声韵母组合。在汉语中，虽然一个音节由一个声母(或者零声母)和一个韵母组成，但是某些声母和韵母的组合是不合理的。例如，如果采用公式（3-3），以无调音节 *chang* 为例，在标准普通话的发音中，它是由声母 *ch* 和韵母 *ang* 组成，表 3.1 表示 *ch* 和 *ang* 在方言背景普通话的发音变化及其概率：

表3.1 *ch*和*ang*的发音变化

标准发音	概率	实际发音	标准发音	概率	实际发音
ch	0.7	ch	ang	0.6	ang
	0.3	c		0.4	iang

那么根据公式（3-3），标准普通话音节 *chang* 就有以下四种可能的声韵母组合方式，以概率值降序排列如表 3.2。

表3.2 音节*chang*可能的发音变化组合

标准音节	概 率	发音变化组合	备 注
chang	0.42	ch ang	
	0.28	ch iang	**不合理**
	0.18	c ang	
	0.12	c iang	**不合理**

从表 3.2 中可以清楚看到，如果利用基于概率值的剪枝策略，（如节 3.2.3 中的剪枝策略 1-5），排名第二的 *ch iang* 发音变化组合就有可能包含在最终的发音字典中，而实际上，这种发音变化组合方式是不合理的，在实际发音中也是不存在，为了解决该类问题，本文提出了改进的上下文相关的权重策略。

诚然，如果直接考虑到方言背景普通话声韵母之间的关联关系，基于音节之间的发音变化可表示为公式（3-5）：

$$P\left(Syl_{dc}\left|Syl_{sc}\right.\right)=P\left(I_{dc},F_{dc}\left|I_{sc},F_{sc}\right.\right)=P\left(I_{dc}\left|I_{sc},F_{sc}\right.\right)P\left(F_{dc}\left|I_{dc},I_{sc},F_{sc}\right.\right) \quad (3\text{-}5)$$

在公式(3-5)中,方言背景普通话声母的发音变化概率,$P\left(I_{dc}\left|I_{sc},F_{sc}\right.\right)$,若只考虑标准普通话的左相关的情况,则取决于标准普通话的声母 I_{sc} 和韵母 F_{sc};对于公式(3-5)中的方言背景普通话的韵母部分的发音变化概率,$P\left(F_{dc}\left|I_{dc},I_{sc},F_{sc}\right.\right)$,则不仅取决于标准普通话的韵母 F_{sc} 和标准普通话的左相关声母 I_{sc},还取决于方言背景普通话的左相关的声母 I_{dc}。理论上,虽然公式(3-5)准确的刻画了基于音节的方言背景普通话的发音变化概率,但实际上,由于需要计算的参数规模的增加,相应发音建模需要的方言背景普通话数据量也将成倍的增加,数据稀疏的问题愈加严重,在文(Riley,1991),就采用了这种概率计算公式,实验证明没有取得预期的效果;另一方面,这也与本文所强调的基于小数据量的主旨思想背道而驰。基于以上考虑,本文在公式(3-4)的基础上引入了方言背景普通话的声韵母转移因子 α,定义了新的权重策略,称为**改进的上下文权重策略 ICDW**[③],表示为公式(3-6)。

$$W_{dc|sc}=\alpha\cdot P\left(I_{dc}\left|I_{sc}\right.\right)\cdot P\left(F_{dc}\left|F_{sc}\right.\right)\quad\left(0\le\alpha\le1\right) \quad (3\text{-}6)$$

公式(3-6)中,$P\left(I_{dc}\left|I_{sc}\right.\right)$和 $P\left(F_{dc}\left|F_{sc}\right.\right)$的计算依然遵循公式(3-4),即上下文相关策略;另外,α 定义为:

$$\alpha=P\left(F_{dc}\left|I_{dc}\right.\right) \quad (3\text{-}7)$$

公式(3-7)中的转移因子α可以通过方言背景普通话标注得到;另外,利用方言背景普通话中的声韵母的转移概率α作为调节因子,考察在方言背景普通话中声韵母的关联关系,对于可能出现的概率很低的声韵母组合或者不合理的组合进行了限制,很大程度上,解决了公式(3-3)所带来的不合理声韵组合的问题。例如,在实际发音中,$P\left(ang\left|ch\right.\right)\gg P\left(iang\left|ch\right.\right)$,因而新的排名如表 3.3 所示:

[③]这里不再是一个严格意义上的“概率”,而是一种“权重”,因而称为权重策略。

表3.3 应用改进的上下文相关权重策略之后的结果[④]

标准音节	概 率	发音变化组合	排 名
chang	0.75	ch ang	1
	0.22	c ang	2
	0.02	ch iang	3
	0.01	c iang	4

相应的，为与改进的上下文相关权重策略相结合，本文采用了基于相对最大概率的剪枝策略，这也是由于 $\sum W_{dc|sc}$ 很有可能不等于 1，如果以某一确定阈值进行剪枝，就有可能将权重值很小的发音变化（可能是不合理的声韵母组合）保留在发音字典之内（例如，累积概率策略；基于固定数目的剪枝策略），为了增强灵活性，所以采用了相对最大概率的策略，这样保证了权重处于相对于最大值的一定范围内的音节才能够保留下来。在得到了最终的发音字典之后，对每一个音节的可能的多个发音变化进行了归一化处理，使其满足概率之和为 1 的要求。

3.4 基于大数据量的方言背景普通话发音字典的测试

在本节中，将利用节 3.3 所提出的改进的上下文相关权重策略来建立闽南普通话发音字典。实验是在较大数据量的闽南普通话基础上实现的，其目的有二：一是将本文提出的改进的上下文相关权重策略与以往的方法进行对比，验证其有效性；二是为本文的核心思路之一，基于小数据量的方言背景普通话发音字典提供一个识别率参考上限，以验证基于小数据量的发音字典的有效性。

这里采用了一个过渡性的闽南普通话数据集，DEV_MIN_4H，其基本信息如表 3.4 所示，它是第 2 章中数据集 DEV_MIN 的一个超集。

[④] 表中“概率”是权重经过归一化处理后的结果。

表3.4 过渡性数据集DEV_MIN_4H[⑤]

名 称	用 途	内 容
DEV_MIN_4H	闽南普通话开发集	20 人，200 句子/人，共 4,000 句，约 4 小时

基于数据集 DEV_MIN_4H 的闽南普通话发音字典的生成流程如图 3.1 所示。在图 3.1 中，左侧的说明部分表示在闽南普通话发音字典生成过程中所采用的具体方法。

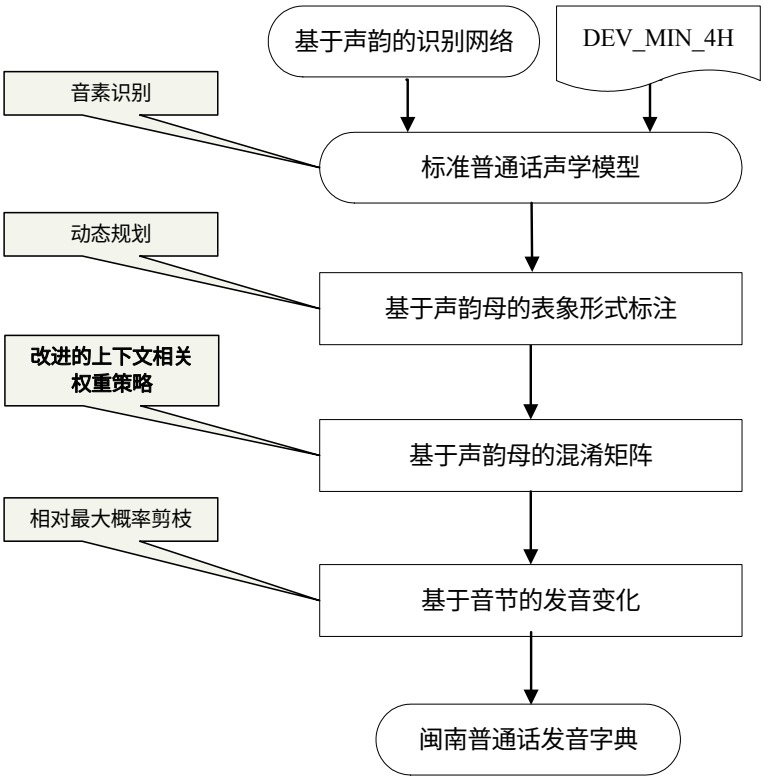


图3.1 基于大数据量的方言背景普通话发音字典生成

按照图 3.1 所示流程生成的闽南普通话发音字典，与以往方法主要的不同就是采用了改进的上下文相关的权重策略。图 3.2 列出了改进的上下文相关策略和等概率、上下文无关策略、上下文相关策略的对比结果，这里采用的是标准普通话声学模型，使用的测试集是闽南普通话测试集，

⑤ 此数据集包含了闽南普通话数据库训练集的几乎所有数据。

TEST_MIN。

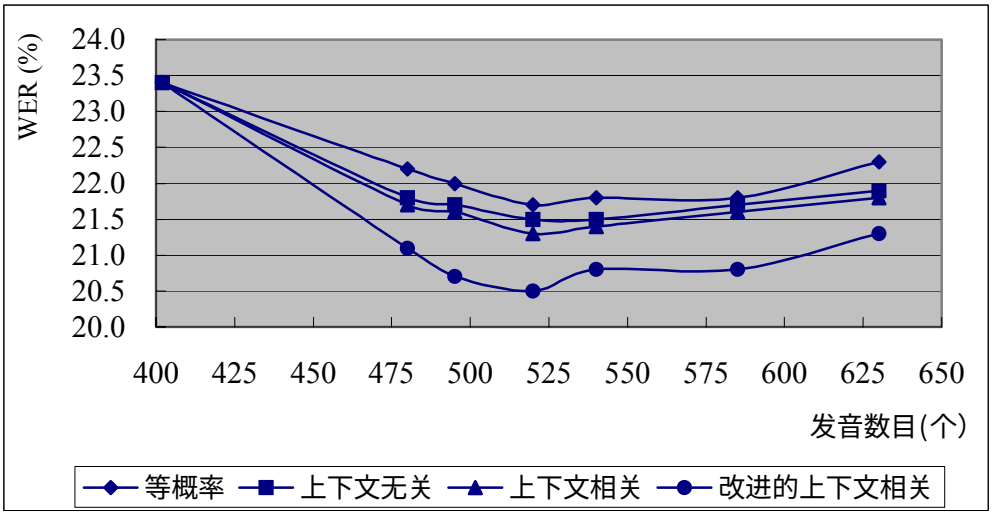


图3.2 不同的权重策略通过TEST_MIN测试结果对比

图 3.2 中，横轴代表发音字典中包含的发音项数目，其中最左上端的重合点代表标准普通话发音字典（每个音节对应一种发音）。从图中可以看到：

1. 准确的发音变化概率有助于提高识别率。随着发音变化概率精度的提高，错误率也在降低。就识别效果而言，考虑到上下文相关策略在性能上优于上下文无关策略的实际情况，说明精细准确的概率能更准确的刻画标准普通话与方言背景普通话之间的发音变化趋势。

2. 改进的上下文权重策略对于错误率的降低最为显著，这说明正确的发音变化组合对于降低发音字典的混淆度，提高识别率，有着重要的作用。

3. 当发音字典包含 520 个左右发音项的时候，对方言背景普通话的识别错误率达到最低，也就是说，当方言背景普通话发音字典的规模是标准普通话发音字典的 1.3 倍的时候，识别性能最好，这与文（李净，2005）中的结论是一致的。

此外，本文的另外一个目标就是在提高方言背景普通话识别率的同时不降低标准普通话的识别率，为此对图 3.2 中性能最佳的发音字典，即基于改进的上下文相关权重策略和包含 520 个发音项，分别在 TEST_STD 和 TEST_MIN 进行了测试，结果如表 3.5 所示：

表3.5 方言背景普通话发音字典在TEST_STD和TEST_MIN测试结果

发音字典	WER	
	TEST_STD	TEST_MIN
闽南普通话（520）	9.3%	20.5%

综合图 3.2 和表 3.5 的测试结果可以知道，相对于标准普通话发音字典，利用方言背景普通话发音字典，方言背景普通话的词错误率由 23.4%降低为 20.5%，相对错误率降低了 12.4%，与上下文相关的策略相比，相对错误率也多降低了 3.8%。同样是此方言背景普通话发音字典，在对标准普通话进行识别时，没有造成性能的下降，说明没有引入新的混淆，达到了本文的目的。可以得出这样的结论，方言背景普通话发音字典对于提高方言背景普通话的识别率是一种行之有效的方法，尤其对于方言背景普通话中的音素变化进行了有效的刻画。

以上的方言背景普通话发音字典的生成是在较大的数据集，（DEV_MIN_4H）上实现的，如果同样依据图（3-1）的流程，采用用 1 小时的数据集（DEV_MIN）重新生成一个闽南普通话发音字典，同样用 TEST_MIN 进行测试，其词错误率为 21.9%，这比基于 DEV_MIN_4H 数据集生成的发音字典，词错误率上升了 1.4%。造成这种现象的原因就是因为数据稀疏，发音变化趋势分散，概率估计不精确所致。如何在小数据的情况下得到鲁棒的方言背景普通话发音字典将是本章后半部分的研究重点。

3.5 基于距离度量的识别网络生成策略

本文的主要研究方向之一就是如何利用少量的方言背景普通话数据生成发音字典，将方言背景普通话的音素变化通过发音字典体现出来，进而提高系统的识别率。如何得到正确的发音变化，将是解决数据稀疏的关键。通过前一节的研究，可以知道：

1. 降低音素识别过程中的错误率，由于对方言背景普通话进行无任何限制的基于声韵母的解码，在其识别结果与基准形式标注之间肯定会产生不一致。这些不一致来源于两个方面，一是方言背景普通话的发音特点，

由于受到方言的影响其发音本来就同标准普通话的发音存在差别；二是来自于解码器本省的错误，由于解码器本身不够鲁棒，另外由于训练数据和识别数据的差异（信道差异，说话人差异等），都会造成识别错误。例如在（Gruhn，2004）中就指出，45%左右的不一致，就是由识别器引入的；另外用标准普通话的识别器来对方言背景普通话来进行解码，这种错误也是难以避免的。本文在发音字典生成过程中的一个目标，就是凸显第一种不一致的比例，降低第二种不一致的比例，这样一来，就能更容易通过数据驱动的方式，找到方言背景普通话的发音变化趋势。

2. 从节 3.4 的结果知道，当方言背景普通话发音字典的规模是标准普通话发音字典的 1.3 倍的时候，识别率最高，也就是说，每一个音节平均对应 1.3 个发音，这样一来，除了单发音项之外，绝大多数音节只有两个发音项，准确得到这两个发音变化是建立方言背景普通话发音字典的最关键因素，其次才是精确的概率。

3. 为得到准确的发音变化，重要的一步就是在解码过程中，降低识别错误率，凸显体现方言背景普通话发音特点的结果。其手段之一，就是在解码过程中引入识别网络的限制，上一节中使用的音素识别虽然针对大数据量是有效的方法，但往往引入很多与发音特点无关的错误，如果能将一个标准普通话声韵母在语音学上相近的方言背景普通话发音作为识别候选加入到识别网络中，就能解决发音变化趋势分散的问题，同时所需要的数据量也可大大降低，从而达到解决数据稀疏问题的目的。举例说明上面阐述的基本思想，如表 3.6 所示：

在表 3.6 中，如果利用音素识别, 原则上，声母 *ch* 有可能被识别成右边的任何一个声韵母。当然，如果数据量的大的时候，某种发音变化的趋势就非常明显，比如由于方言口音的影响，大多数 *ch* 都被识别成 *ch* 或者 *c*，但在数据量小的情况下，这种发音变化受到“噪音”（识别错误）的干扰，其趋势就不那么明显，这对于最终生成的发音字典的准确性就会有影响。现在要做的就是如何将变化趋势限定在一定范围内，如表 3.6 中的黑体标示的声母子集。为此本文提出了基于距离度量的识别网络生成策略，即通过计算模型之间的声学距离来确定它们之间的相似度，将相似度最高的声韵母子集作为可能的识别结果。距离越小，说明标准普通话和方言背景普通话之间相似度越高，发生发音变化的可能性越大；反之，则认

为两者之间存在较大的差异，发生发音变化的可能性越小。这样一来，最有可能的发音变化将在这些子集中产生，也就是对应表 3.6 中的黑体部分。

表3.6 标准声母*ch*在方言背景普通话中可能的发音变化

基准发音	表象发音
ch	**ch** **c** **sh** **s** … a o e

表 3.7 列出了常用的距离度量准则，本文在对这些准则进行对比的基础上提出了距离度量准则的评价策略，作为采用何种距离度量准则的依据。

表3.7 常用的距离度量准则

名 称	公 式
欧式(Euclidean)距离 （MacQueen，1967）	$dis(A,B)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{N} w_{Bj} d_{A,B}(i,j)\right]$， 其中，$d(i,j)=\left(\mu_i-\mu_j\right)^T\left(\mu_i-\mu_j\right)$
马氏(Mahalanobis)距离 （王仁华，1992）	$dis(A,B)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{N} w_{Bj} d_{A,B}(i,j)\right]$， 其中，$d(i,j)=\left(\mu_i-\mu_j\right)^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}\left(\mu_i-\mu_j\right)$

表3.7（续）常用的距离度量准则

名　称	公　式
Bhattachyaryya 距离（Kosaka，1996）	$dis(A,B)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{N} w_{Bj} d_{A,B}(i,j)\right]$，其中，$d(i,j)=\frac{1}{8}(\mu_i-\mu_j)^T\left(\frac{\Sigma_i+\Sigma_i}{2}\right)^{-1}(\mu_i-\mu_j)+\frac{1}{2}\ln\frac{\left\|(\Sigma_i+\Sigma_j)/2\right\|}{\left\|\Sigma_i\right\|^{1/2}\left\|\Sigma_j\right\|^{1/2}}$
K-L 距离（Zhou，2004）	$dis(A,B)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{N} w_{Bj} d_{A,B}(i,j)\right]$，其中，$d(i,j)=\frac{1}{2}\left(\mu_i-\mu_j\right)^T\left(\Sigma_i^{-1}+\Sigma_j^{-1}\right)\left(\mu_i-\mu_j\right)+\frac{1}{2}tr\left[\Sigma_i^{-1}\Sigma_j+\Sigma_i\Sigma_j^{-1}-2I_d\right]$
类散度距离（彭煊，2005）	$dis'\left(A,B\right)=\frac{1}{2}\left(\frac{dis(A,B)}{dis(A,A)}+\frac{dis(B,A)}{dis(B,B)}\right)$，其中，$dis(X,Y)$可以是前面 4 种距离度量准则中的任一种

表 3.7 中，M 和 N 分别表示一个 HMM 中状态 A 和 B 所包含高斯混合数目，w_{Ai} 和 w_{Bj} 分别表示状态 A 的第 i 个高斯混合和状态 B 的第 j 个高斯混合的权重。μ和Σ分别代表高斯分布的均值和方差。$d(i,j)$表示任意两个高斯混合之间的距离，对于 $d(i,j)$常用的方法有 K-L 距离、欧式距离、马氏距离以及 Bhattachyaryya 距离等。在类散度距离度量准则中，$dis(A,B)$可以是以上四种的距离中的任意一个。此表给出的是任意两个 HMM 状态之间的距离，如果计算任意两个具有相同拓扑结构 HMM 之间的距离，其公式定义如下：

$$dis\left(M^{(sc)},M^{(dc)}\right)=\frac{1}{K}\sum_{i=1}^{K}dis\left(M_i^{(sc)},M_i^{(dc)}\right) \tag{3-8}$$

本文利用公式（3-8）来计算标准普通话和方言背景普通话之间模型距离。其中 $M^{(sc)}$ 表示标准普通话 HMM，$M^{(dc)}$ 表示方言背景普通话 HMM，K 表示 HMM 中包含的状态数目。本文应用以上几种距离度量方法分别计算一个方言背景普通话神韵母 HMM 和一个标准普通话声韵母 HMM 之间的声学距离，例如，方言背景普通话 *ch* 和标准普通话 *ch* 之间的距离，从语音学角度而言，虽然存在一定的发音差异，但根本上还是同一个声母，理论上其声学距离应该最小；另一方面，标准普通话的 *ch* 在方言背景普通话中很多情况下被发音成 *c*，那么标准普通话 *ch* 和方言背景普通话 *c* 之间的声学距离应该也很小。为衡量不同的距离度量准则对于标准普通话和方言背景普通话之间模型距离的可信程度，本文提出了距离度量准则的评价策略，将其量化为以下公式：

$$Score=\sum_{x\in C}\left(N-rank(x)\right) \tag{3-9}$$

公式（3-9）中，x 表示一个声韵母，C 表示标准普通话声韵母集合，N 表示每个标准普通话声韵母所对应的方言背景普通话中从语音学角度相近的声韵母的数目，也是说，距离最近的前 N 个，$rank(x)$ 表示 x 在这 N 个候选项中的排名，其排名由距离大小决定，取值区间为[0~N-1]，如果 x 不在这 N 个中，则规定 $rank(x)=N$，那么 $Score$=0。这样一来，$Score$ 的值越大，说明与标准普通话发音相同的方言背景普通话声韵母排名靠前的概率越大，这样做的目的就是首先保证发音一致的发音变化不被排斥在候选集合之外。本文以此来衡量某种准则对于模型距离度量的精细程度。

提出公式（3-9）其依据在于：由节 3.4 的结果知道，当方言背景普通话发音字典的规模为标准普通话发音字典的 1.3 倍时，系统识别性能最好，分析得到的发音字典可以发现，很多的音节只有一个发音，即标准普通话发音，其它的音节一般拥有两个发音变化，其中一个是与标准普通话发音相同的。这与方言背景普通话是普通话变体的事实是一致的。基于这样的观察结果，本文认为首先保证与标准普通话发音相同的发音变化出现在发

音字典中对于提高方言背景普通话的识别率，同时又不降低标准普通话的识别率是至关重要的。因而采用了公式（3-9），其提出的背景就是以此为依据的。

通过应用公式（3-9），在确定某种距离度量准则之后，基于小数据量的方言背景普通话发音字典生成方法的流程如图 3.3 所示：

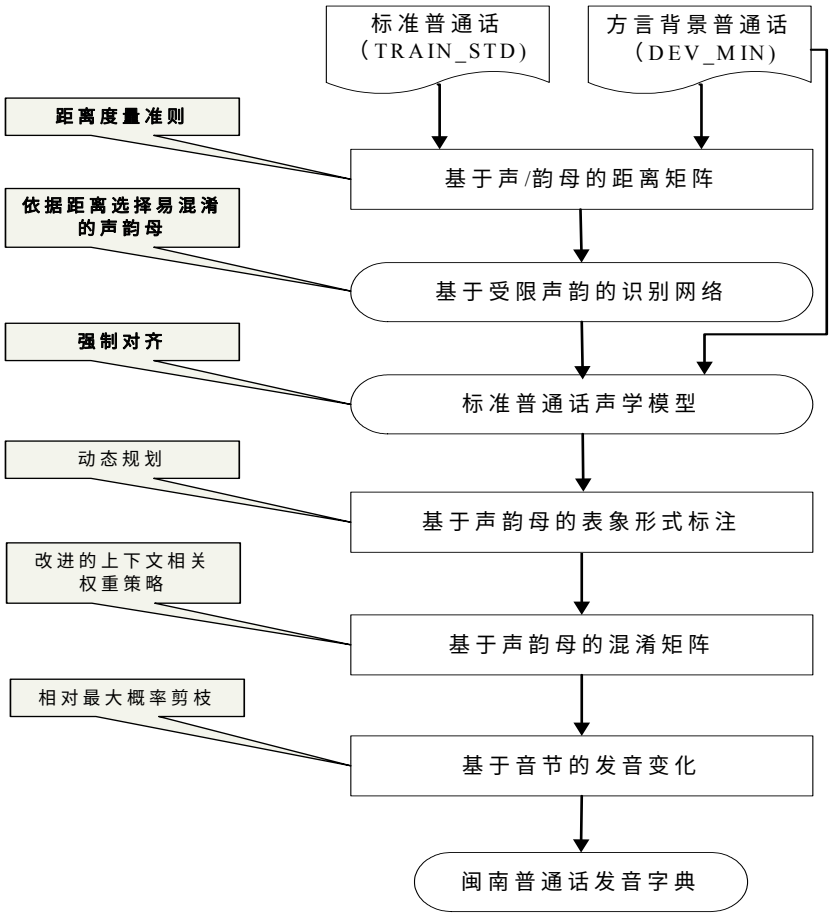


图3.3 基于小数据量的方言背景普通话发音字典生成方法

在图 3.3 中，与图 3.1 中基于大数据量的方言背景普通话发音字典生成方法相比，不同的实现步骤有（对应于图 3.3 左侧的加黑斜体部分）：

1. 首先，利用标准普通话数据和少量方言背景普通话数据训练基于声韵母的 HMM，这里采用的是上下文无关的声韵母 mono-XIF 模型，一是受到方言背景普通话数据量的限制，二也是为了计算简便。为了计算精确，这里的 mono-XIF 模型应该包含多个高斯混合。

2. 应用一定的距离度量准则，得到声母和韵母两个距离矩阵。这里假定声母和韵母之间是不会相互混淆的。理论上，绝大多数情况下，距离矩阵中对角线的值应该最小（发音相同），当然也有一些标准普通话声韵母与方言背景普通话的发音相近的声韵母距离值最小。

3. 距离值按照升序排列，将距离最小的前 N 个作为发音相近的声韵母集合，并与标准普通话发音字典相结合构建受限的识别网络。例如，根据距离度量，与标准普通话 *ch* 发音相近的闽南普通话发音就可能有 *ch*, *c*, *sh*, *s* 等。

4. 在受限的识别网络的基础上，利用标准普通话上下文相关模型进行强制对齐（forced alignment），得到所谓的方言背景普通话标注。

此后的步骤与图 3.1 是相同的，在发音字典生成过程中，应用了节 3.3 中的改进的上下文相关权重策略，最终得到方言背景普通话的发音字典。

3.6 基于小数据量的方言背景普通话发音字典的测试

在此实验中，采用标准普通话训练集 TRAIN_STD 进行上下文无关的声韵母建模（这是上下文相关声韵母建模的中间结果），采用少量闽南普通话数据 DEV_MIN 进行上下文无关的声韵母声学建模（见节 2.4）。每个 mono-XIF 模型包含了 8 个高斯混合。在此基础上对表 3.7 中的距离度量准则进行了对比。当取距离最小的前 5 个作为最容易混淆的声韵母子集时，应用公式（3-9）的结果如表 3.8 所示：

表 3.8 中，类散度距离准则中分别采用了 4 种不同的方法（括号中所示）来计算两个状态之间的距离。结果表明：

1. 距离度量准则越精确，得分越高。在欧式距离中，因为不考虑方差的影响，只考虑均值的差异，很容易产生混淆，分辨率不高；马氏距离中，虽然考虑了方差的影响，但对于均值相同的情况，马氏距离就无法度量方差的差异。对于 Bhattachyaryya 距离准则要更精确，其前半部分与马氏距离相同，后半部分还重点考察协方差矩阵的差异；在更大程度上强调两个模型的相似度，因而得分最高。

2. 类散度距离中，不仅计算两个状态之间的距离，还计算状态本身的距离，进一步强调两个模型的相似度，并将模型之间的差异进行了归一化

处理，具体分析可参见（彭煊，2005）。

表3.8 不同的距离度量准则的得分

距离度量准则	Score	距离度量准则	Score
欧式距离	282	马氏距离	294
Bhattachyaryya	299	K-L 距离	292
类散度距离（欧式距离）	304	类散度距离（马氏距离）	318
类散度距离（Bhattachyaryya）	324	类散度距离（K-L 距离）	316

在表 3.8 中，可以看出，尤以类散度距离与 Bhattachyaryya 距离的结合得分最高。综合以上的结果，本文中采用了类散度（Bhattachyaryya 距离），作为基于小数据量的发音字典生成策略的距离度量准则。在此基础上，生成了声母和韵母两个距离矩阵。公式（3-9）中取 N=3，即对每一个标准声韵母选择 3 个距离最接近的声韵母作为候选建立识别网络。根据图 3.3 中的流程，得到基于声韵母的发音变化如表 3.9 所示：

表3.9 标准普通话与闽南普通话基于声韵母的发音变化

标准发音	实际发音	概率(%)	标准发音	实际发音	概率(%)	标准发音	实际发音	概率(%)
a	a	82.54	ai	ai	81.32	an	an	75.01
ao	ao	77.37	ang	ang	67.94	ang	iang	10.80
b	b	83.36	c	c	61.78	c	ch	10.52
ch	ch	45.92	ch	c	45.92	d	d	82.76
e	e	71.49	er	e	46.27	er	er	24.11
f	f	83.21	g	g	84.48	h	h	73.48

表3.9 （续） 标准普通话与闽南普通话基于声韵母的发音变化

标准发音	实际发音	概率(%)	标准发音	实际发音	概率(%)	标准发音	实际发音	概率(%)
i	i	82.90	ia	ia	79.43	ian	ian	71.11
iang	iang	86.69	iao	iao	81.03	ie	ie	69.61
ii	ii	71.67	iii	ii	21.14	iii	iii	26.83
in	in	64.49	in	ing	16.22	ing	ing	74.77
ing	in	15.02	iong	iong	72.96	iou	iou	79.93
j	j	86.67	k	k	81.04	l	l	59.55
m	m	80.44	n	n	72.79	n	l	21.20
o	o	64.28	o	u	15.12	ong	eng	12.85
ong	ong	71.16	ou	ou	65.84	p	p	76.64
q	q	76.47	r	r	44.55	r	l	10.71
s	s	64.84	sh	x	18.48	sh	sh	20.39
sh	s	20.39	t	t	74.45	u	u	77.76
ua	ua	72.20	uai	uan	11.00	uan	uan	73.16
uang	uang	70.55	uei	uei	80.41	uen	uen	62.68
ueng	eng	50.00	ueng	ueng	25.00	uo	uo	79.74
v	v	73.87	van	van	56.23	van	vn	11.15
ve	ve	73.53	vn	vn	67.11	x	x	84.72
z	z	70.76	zh	z	25.07	zh	zh	41.32

在文（Li，2005）中也给出了标准普通话和闽南普通话之间声韵母发音变化情况，此结果是在 50 个说话人，共 2,200 句基础上进行人工标注，通过语音学者的观察得到的（如表 3.10 所示）。与表 3.9 相比可以发现，本文所提出的方法，即便只取前 3 个候选项，也全部覆盖了表 3.10 所给出的发音变化情况。这说明，基于类散度距离度量准则在距离度量准则评价策略中得分最高，最大限度的覆盖了一个标准普通话发音在方言背景普通

话中的近似发音。其发音变化与语音专家给出的基本是一致的。例如，在一些介绍闽南普通话的书籍中，时常会举例，“粉红凤凰飞”被说成“哄红哄黄灰”。但是无论是文（Li，2005）中，还是本文的结论，发现在闽南普通话声母 *f* 和 *h* 的混淆程度并不十分明显。而在表 3.9 中，声母 *sh* 和 *x* 容易混淆，这是由于闽南普通话中“是（*shi*）”很容易发成“系（*xi*）”，而在表 3.10 并没有给出这种现象。由此可见基于数据驱动的方式更能反映数据本身的发音特点，与测试数据更匹配。

表3.10 语音专家给出的标准普通话与闽南普通话基于声韵母的发音变化[6]

标准发音	实际发音	标准发音	实际发音	标准发音	实际发音	标准发音	实际发音
a	a	ai	ai	an	an	ang	ang
ao	ao	b	b	c	c	ch	c
ch	ch	d	d	e	e	ei	ei
en	en	eng	eng	er	er	f	f
g	g	h	h	i	i	ia	ia
ian	ian	iang	iang	iao	iao	ie	ie
ii	ii	iii	iii	in	in	in	ing
ing	in	ing	ing	iong	iong	iou	iou
j	j	k	k	l	l	l	n
m	m	n	l	n	n	o	o
ong	ong	ou	ou	p	p	q	q
r	l	r	n	r	r	s	s
sh	s	sh	sh	t	t	u	u
ua	ua	uai	uai	uan	uan	uang	uang
uei	uei	uen	uen	ueng	ueng	uo	uo
v	v	van	van	van	ian	ve	ve

[6]文（Li，2005）未给出发音变化概率

表 3.10（续）语音专家给出的标准普通话与闽南普通话基于声韵母的发音变化

标准发音	实际发音	标准发音	实际发音	标准发音	实际发音	标准发音	实际发音
vn	vn	x	x	z	z	zh	z
zh	zh						

利用基于距离度量的识别网络生成策略，解决了数据稀疏问题，根据图 3.3 生成了方言背景普通话发音字典，在此过程中，还应用了节 3.3 中的改进的上下文相关权重策略。另外，此发音字典同样包含 520 个发音（节 3.4 结论），通过与标准普通话声学模型相结合，在 TEST_STD 和 TEST_MIN 上的测试结果如表 3.11 所示:

表3.11 基于小数据量的方言背景普通话发音字典测试结果

发音字典	WER	
	TEST_STD	TEST_MIN
闽南普通话（520）	9.3%	20.9%

与表 3.5 的结果相比，可以看出：

1. 基于小数据量生成的方言背景普通话发音字典，其性能接近基于大数据量的生成方法。说明基于距离度量的识别网络生成策略很好的解决了数据稀疏问题。

2．对于闽南普通话测试集，基于小数据量的发音字典已经接近于上限参考值(词错误率 20.5%)，但仍有 0.4%的差距，这主要还是由于数据量不足，权重估计不够精确造成的；另一方面，由于数据量的限制，对于训练样本较少的声韵母的模型就会不精确，造成距离度量也不精确，这样一来某些发音变化就不能出现在受限的识别网络中。

3. 方言背景普通话发音字典中多发音的引入没有造成标准普通话识别率的下降，这说明适当的增加发音变化，不会增加字典的混淆度。在一定策略控制下，发音字典不但可以提高方言背景普通话的识别率，同时还能够保证标准普通话识别率不降低。

3.7 本章小结

1. 首先在大数据量的基础上，对方言背景普通话的发音字典进行了研究，并将结果作为基于小数据量的发音字典的参考上限。

2. 为了解决音节对应不合理声韵母组合问题，本文提出了改进的上下文相关权重策略，应用此策略，在方言背景普通话测试集上获得了最低的错误率。词错误率由 23.4%降低到 20.5%，相对错误率降低了 12.4%。

3. 为生成基于小数据量的发音字典，本文采用距离度量准则来构建受限的识别网络，大幅降低了由标准普通话识别器本身所引入的错误，使方言背景普通话的发音变化分布更加集中。

4. 对比了几种距离度量策略，提出了一个距离度量准则评价策略，应用此评价策略来确定具体的距离距离度量准则，在此基础上生成了受限的识别网络。

5. 基于小数量的发音字典获得了与基于大数据量的发音字典相近的性能，证明了基于小数据量的发音字典生成策略的有效性。

第 4 章　状态相关基于基元的模型归并

4.1　引言

为了解决标准普通话与方言背景普通话之间的声音变化问题，本章将对一种简单新颖而有效的声学建模方法，状态相关基于基元的模型归并 SDPBMM，进行研究。在 SDPBMM 中，利用少量方言背景普通话数据，在 HMM 状态一级，同时考虑到方言背景普通话和标准普通话的发音特点，对两者之间的声音变化进行建模。实验结果表明，SDPBMM 能显著的提高方言背景普通话的识别率，同时对于标准普通话的识别率也有明显提高。

在 SDPBMM 中，在进行状态归并的同时，造成 HMM 状态所包含的高斯混合模型数目增加，本文称之为高斯混合扩张问题，为了降低模型规模，同时又不降低识别率，本文提出了可选择的 SDPBMM，即以状态之间的声学距离为准则，进行有选择的归并，将需要归并的状态和不需要归并的（冗余的）状态分开。

本章的第 1 部分首先对常用的带口音语音识别声学建模方法进行了介绍，在对比现有方法优缺点的基础上，提出了本文的建模思路。第 2 部分阐述了 SDPBMM 的基本思想、原理、及两个归并准则，即基于小数据量的发音建模和插值因子的获得，并对 SDPBMM 进行全面、系统的测试验证。第 3 部分针对 SDPBMM 中的高斯混合扩张问题进行处理，得到了可选择的 SDPBMM 模型，并进行了验证。

4.2　常用带口音语音识别声学建模方法及本文思路

本节将同时关注于带口音的语音识别、方言背景普通话语音识别和非母语语音识别中声学语音建模问题，因为这些语音在声学层面有许多相似之处，都是与标准发音有一定的差异。在对一些最新的建模方法进行介绍之后，在总结前人研究工作基础上提出了适合于本文研究目的的方言背景普通话语音识别声学建模方法。

4.2.1 自适应方法

对于基于少量数据的声学建模中，自适应方法通常是最常用且有效的方法之一。结合一定的方言背景普通话数据，通过自适应，可以将一个标准普通话声韵母模型转换为一个方言背景普通话声韵母模型。一般最基本的方法有最大后验概率 MAP（Gauvain，1994）方法和最大似然线性回归 MLLR（Leggetter，1995）方法。

由以往的研究成果知道，对于 MAP 存在以下的优缺点（王昱，2000；何磊, 2001；Huang，2001）：

1. 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的补充，通过贝叶斯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。当自适应数据不断增加时，可以充分利用语音数据的细节信息，自适应效果将稳步提高；

2. 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。同时，由于自适应往往是在音素或者声韵层次上进行，每个状态可利用的自适应数据比较有限，当数据量少时，其效果往往不理想。另外，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，所以初始模型的选择至关重要，在很大程度上影响自适应效果。

同样，对于 MLLR 也存在以下的优缺点：

1. 需要进行自适应的参数少，速度快。一般情况下，只对概率密度函数的均值向量进行自适应，其它的参数在自适应模型中不作改变，即使自适应数据量不足，MLLR 方法也可以获得较理想的效果；

2. MLLR 最大的缺点在于不能充分利用自适应数据的信息。在自适应数据量比较充足的情况下，系统可能过早地维持在一种饱和状态，这样一来，其效果就不如 MAP 好。

由于 MAP 和 MLLR 都存在一定的优缺点，最新的研究中往往将两者结合来实现基于小数据的模型自适应。首先利用 MLLR 进行自适应得到一个较为精确的初始模型，再利用 MAP 进行自适应。例如在（Sproat，2004）中，使用 6.3 小时的自适应数据，基于 MLLR 和 MAP 相结合的方法就比单独使用 MAP 和

MLLR 字错误率分别降低了 1.7%和 4.1%。在文（Oh，2007）也采用了两者相结合的方式，3 小时的带韩语口音的英语数据对基于标准英语的声学模型进行自适应，比单纯的 MAP 和 MLLR，在词错误率多降低了 0.4%和 1.0%。本文将对以上三种自适应方法在少量方言背景普通话数据上进行对比，采用性能最好的方法作为本文的自适应方法。

4.2.2 重训练

重训练对于提高方言背景普通话识别率是一种最直接的方法。其中的策略之一就是完全利用方言背景普通话进行重新建模。我们以往的工作中证明，当采用 70 分钟的上海普通话重新建模时，其对上海普通话的识别率超过了 30 个小时的标准普通话声学模型（参见图 1.3），文（Wang, 2003）也给出了类似的结果，利用 52 分钟的德语口音的英语数据训练得到的模型性能，优于 34 小时标准英语训练得到的模型。另外一种策略是将标准语音数据和带口音的语音数据进行混合用来训练新的声学模型。文（Wang，2003）将 52 分钟德语口音的英语数据与 34 小时标准英语数据进行混合后重训练，对于德语口音的词错误率由 49.3%（基于标准英语的声学模型）降为 42.7%。另外，也可以在基于标准语音的声学模型的基础上，加入一定量的带口音的语音数据，而后再进行迭代训练。在文（Tomokiyo，2001）中，在标准英语训练得到的声学模型中，加入 3 小时的日语口音的英语数据，再进行两次额外的迭代，可以将词错误率由 63.0%降为 48.0%[⑦]。

4.2.3 状态级的发音建模

对于实际发音中的声音变化，不具类似于音素变化中 $A \rightarrow B$ 显著特征，因而无法通过发音字典来解决的，一些研究者就提出了在状态一级进行发音建模来解决声音变化的问题，称之为隐式（implicit）发音变化建模（Hain，2005）。文（Saraclar，2000）就提出了状态级的发音模型（state-level

[⑦]由于测试环境不同，这些识别结果不能直接进行对比。

pronunciation model, SLPM），其核心思想就是将实际发音中容易发生混淆的音素（例如 *ae→eh*）的状态中的高斯混合进行共享，例如 *hh-ae+d* 与 *hh-eh+d* 进行状态共享，*ae-d+sil* 与 *eh-d+sil* 进行状态共享，这些状态共享都是基于上下文相关 triphone 模型的。实验结果表明，在自然发音式的语音识别中采用上述思想，字错误率降低了 1.7%。作为对 SLPM 思想的继承和发展，文（Liu，2004；Fung，2005；Liu，2003）提出了局部变化音素模型 PCPM，在状态共享过程中降低了混淆度，同时也降低了模型的规模。在 PCPM 中，针对易混淆的声韵母，例如 *n→l* 和 *l→n*，考虑到声音变化是实际发音部分偏离标准发音的事实，提出了增加 *n_l* 和 *l_n* 形式的新基元，通过音素解码得到新基元的标注，而后对新基元利用决策树进行状态共享，在此基础上只对新基元应用类似于 SLPM 的状态间的高斯混合共享。文（Oh，2007）中，在上下文相关的声学建模时，将某个音素的状态与其发音变化的状态混合在一起进行基于决策树的状态共享，而后再进行迭代。例如，如果有类似 $b\rightarrow b'$的发音变化，那么上下文相关音素模型，*-*b*+*[2]及其发音变化*-b'+*[2]将采用决策树进行状态共享。此文中采用了 3 小时的韩语口音的英语数据，得到的 6 个发音变化参与了母语与非母语之间的状态共享，在基于韩语口音的英语测试中，将相对词错误率降低了 53.43%。在声学模型的状态一级体现发音变化是目前声学建模中研究的一个新方向，结果表明对于带口音的语音识别的效果大大优于发音字典自适应的方法。

4.2.4 模型插值

利用插值（interpolation）对标准声学模型和带口音的声学模型的参数进行结合，通过调整插值因子实现对带口音语音的最佳识别效果。利用标准模型训练充分和带口音模型发音变化显著的特点，达到平滑模型参数并进而提高识别性能的目的。文（Wang，2003）就利用下面公式来提高对非母语语音的识别率。

$$P'(O)=W_{native}P_{native}(O)+W_{non-native}P_{non-native}(O) \qquad (4\text{-}1)$$

其中，W_{native}是标准语音模型的插值因子，$W_{non-native}$带口音模型的插值因子，且 $W_{native}+W_{non-native}=1$。$P_{native}(O)$和 $P_{non-native}(O)$分别表示标准语音模型和带口

音语音模型。文中实验表明，当 $W_{native}=0.56$时，对于非母语语音识别效果最好，词错误率由 49.3%（母语语音训练得到的模型）降至 36.0%。另外，在文（Tomokiyo，2001）和（Livescu，1999）都曾经采用过类似的方法，在文（Tomokiyo，2001）中实验对比的基础上，得到 $W_{native}=0.75$时，插值模型性能最好，可见，插值的因子的确定，依赖于具体数据，通常根据实际的测试数据和识别任务来确定，不同的识别任务其值也不同。

4.2.5 本文的声学建模思路

通过以上的分析，结合本文的两个研究目标，即基于小数据量以及提高方言背景普通话识别率的同时不降低标准普通话识别率，提出本文的研究思路。

1. 首先，自适应方法虽然可以充分利用少量的自适应数据来有效提高方言背景普通话的识别率，但由于自适应方法是将一个标准普通话模型转变为一个方言背景普通话模型，就是使最终的模型最大程度上接近于自适应（方言背景普通话）数据，其识别率的提高往往是“单向”的，即对标准普通话的识别效果就不够理想。

2. 重训练的方法虽然简单，但须重新建模，费时费力，往往这种方法对于方言背景普通话识别率的提高并不显著，毕竟相比于标准普通话，方言背景普通话数据量太小了。另外，它对识别率的提高也没有做到“两者兼顾”，对于标准普通话的识别率会有所下降。

3. 虽然基于状态的发音建模对于描述方言背景普通话中的声音变化是非常好的方法，但目前的方法都是基于大量数据的，而且经常需要重新迭代，其建模方法过于复杂，实际应用起来繁琐，不利于推广。

4. 模型插值简单实用，通过调整插值因子来调整标准普通话和方言背景普通话在插值模型中的权重，达到对方言背景普通话与标准普通话识别效果的平衡，其缺点是没有考虑标准普通话和方言背景普通话之间的发音变化，往往对于方言背景普通话的提高不够显著。

综合以上分析，本文的思路是将基于状态的发音建模和模型插值方法加以综合，使用小的数据量，实现对标准普通话和方言背景普通话识别率的“双向”提高。为此提出了状态相关基于基元的模型归并 SDPBMM。

4.3 状态相关基于基元的模型归并

在声学建模过程，状态相关基于基元的模型归并将在模型状态一级同时对标准普通话和方言背景普通话的发音进行建模，以期实现对标准普通话和方言背景普通话识别率，“双向”提高的目的。

4.3.1 SDPBMM的基本思想

在基于声韵母的汉语语音识别中，将来自于标准普通话的上下文相关模型 tri-XIF 与来自于方言背景普通话的上下文无关模型 mono-XIF（与 tri-XIF 的中心基元相同）在状态相同的前提下，依据一定的准则进行归并，归并后的状态包含来自于标准普通话和方言背景普通话的高斯混合，这就是所谓的**状态相关基于基元的模型归并 SDPBMM**。在 SDPBMM 中，所谓状态相关是指来自于不同 HMM 模型的相同状态（参见节 2.1.2）；在基元的选择上也体现了相关性，涉及到一个标准普通话基元及其在方言背景普通话中的发音相同的基元和发音变化的基元。为了减小模型规模，提高系统鲁棒性，SDPBMM 是在声学建模过程中基于决策树的状态共享完成之后实现的。具体的示意如图 4.1 和 4.2 所示：

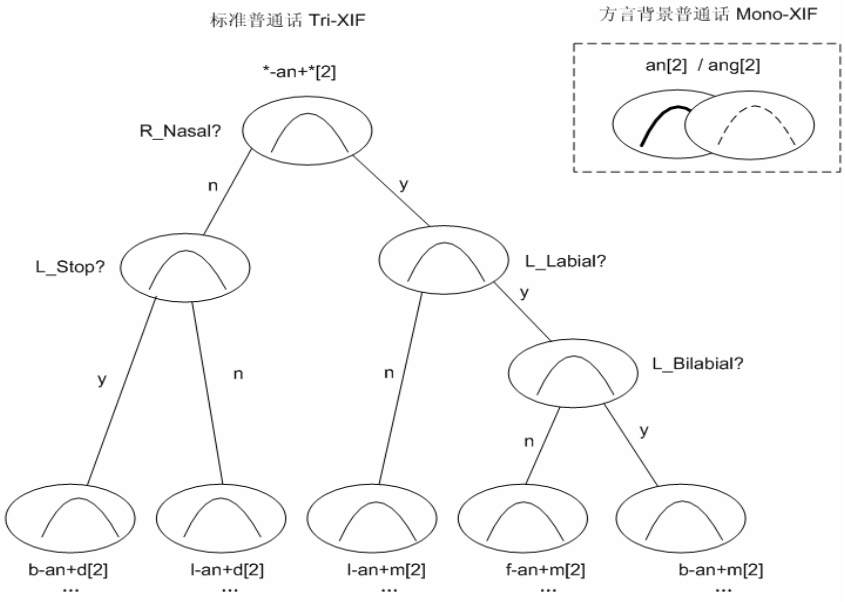


图4.1 SDPBMM之前的拓扑图

在图 4.1 中，所有中心基元为韵母 *an* 的上下文相关模型 tri-XIF 的第二个状态（*-*an*+*[2]），利用决策树的方式进行模型状态共享（参见节 2.1.2），图中的叶子结点表示要进行共享的物理状态，也就是多个在语音学上相近的 Tri-XIF 要共享这一个物理状态。在图 4.1 中，右上角表示的是方言背景普通话上下文无关模型 mono-XIF，*an* 和 *ang* 的第二个状态，其中方言背景普通话 *an* 是与标准普通话 tri-XIF，*-*an*+*，的中心基元是相同的，方言背景普通话 *ang* 是标准普通话 *an* 对应的发音变化（也可能没有）。在 SDPBMM 中，本文采用了状态级的发音建模和插值因子作为状态归并的准则。其归并过程如图 4.2 所示：

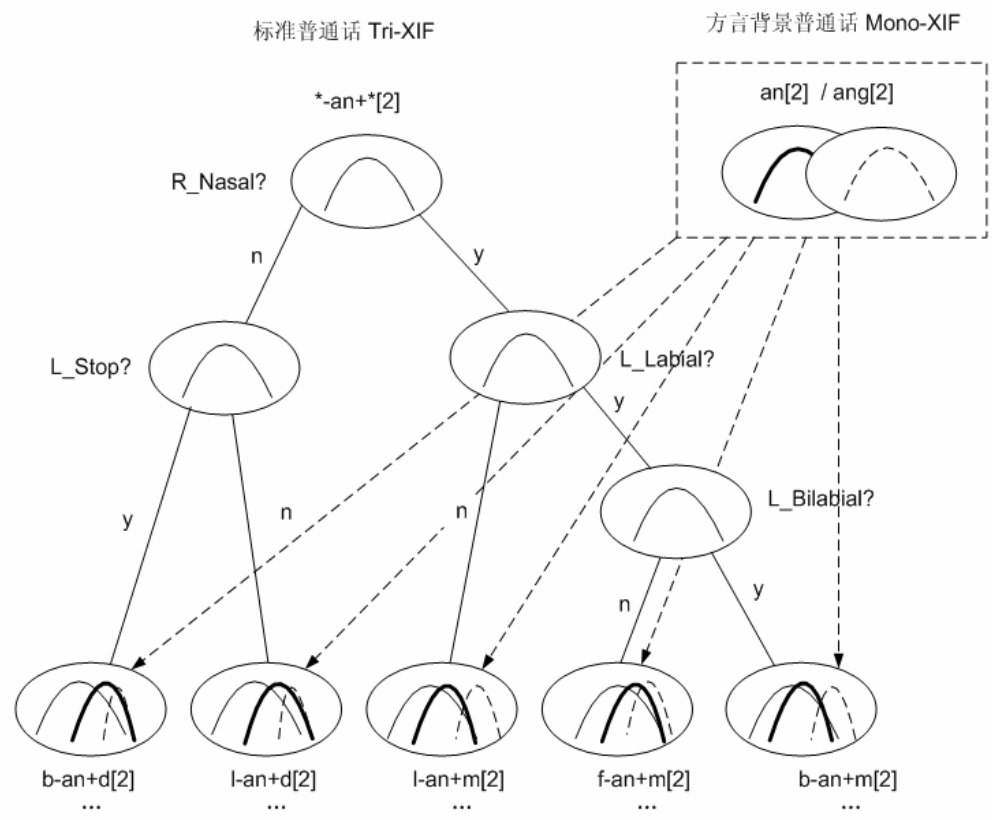


图4.2 SDPBMM之后的拓扑图

在图 4.2 中，可以看到将方言背景普通话 mono-XIF, *an* 和 *ang*，与标准普通话 tri-XIF，*-*an*+*，在同一状态上（第二个状态）进行了归并。归并后的状态由多个高斯混合组成，这些高斯混合分别来自标准普通话 *-*an*+*[2]，方言背景普通话 *an*[2]和标准普通话 *an* 在方言背景普通话的发

音变化 *ang*[2]，如图中浅黑色实线、粗黑色实线和浅黑色虚线所示。

可以注意到，在 SDPBMM 中，并没有改动标准普通话声学建模过程中决策树的拓扑结构，保持了原有的拓扑结构，这对标准普通话 tri-XIF 的状态共享没有造成不利影响。另外，对于标准普通话声韵母在方言背景普通话中的发音变化是在状态一级进行刻画的，图 4.2 中的 *ang* 是标准普通话 *an* 在方言背景普通话中的发音变化，当然在可能很多标准普通话声韵母没有对应的方言背景普通话的发音变化，这样一来，就退化为只与方言背景普通话中发音相同的声韵母状态之间的归并。

4.3.2 SDPBMM的形式化描述

在 HMM 中一个状态的输出概率密度函数（p.d.f）是通过多个高斯混合分布来描述的，若用 x 表示输入特征向量，s_i 表示第 i 个状态，则 p.d.f，$p(x|s_i)$，可以表示为公式（4-2）：

$$p(x|s_i)=\sum_{k=1}^{K} w_{ik} N\left(x;\mu_{ik};\Sigma_{ik}\right) \tag{4-2}$$

其中，K 表示状态中包含的高斯混合的数目，w_{ik} 表示第 i 个状态的第 k 个高斯混合的权重。为了简便起见，后面将用 $N_{ik}(\cdot)$ 代表 $N\left(x;\mu_{ik};\Sigma_{ik}\right)$。

在 SDPBMM 中，归并后第 i 个状态的 p.d.f，$p'\left(x|s_i\right)$可表示为公式（4-3）：

$$p'\left(x|s_i\right)=\lambda p\left(x|s_i^{(sc)}\right)+\sum_{m=1}^{M}(1-\lambda)\, p\left(x|s_i^{(sc)},s_{im}^{(dc)}\right) p\left(s_{im}^{(dc)}|s_i^{(sc)}\right) \tag{4-3}$$

公式（4-3）中，$s_i^{(sc)}$和 $s_i^{(dc)}$分别表示来自于标准普通话和方言背景普通话模型的第 i 个状态，λ表示插值因子，M 表示参与归并的方言背景普通话状态的数目，也就是状态 $s_i^{(sc)}$ 对应的方言背景普通话在状态一级发音变化的数目。其中 $p\left(s_{im}^{(dc)}|s_i^{(sc)}\right)$表示状态级的第 m 个发音变化的概率。对公式（4-3）应用独立性准则，这里假定标准普通话和方言背景普通话模型状态相互独立，得到公式（4-4）：

$$p'\left(x\left|s_i\right.\right)=\lambda p\left(x\left|s_i^{(sc)}\right.\right)+\sum_{m=1}^{M}(1-\lambda)\,p\left(x\left|s_{im}^{(dc)}\right.\right)p\left(s_{im}^{(dc)}\left|s_i^{(sc)}\right.\right) \tag{4-4}$$

将公式（4-2）代入公式（4-4）中，展开后得到公式（4-5）：

$$\begin{aligned}p'\left(x\left|s_i\right.\right)=&\sum_{k=1}^{K}\lambda w_{ik}^{(sc)}N_{ik}^{(sc)}(\cdot)+\\&\sum_{m=1}^{M}\sum_{n=1}^{N}(1-\lambda)\cdot P\left(s_{im}^{(dc)}\left|s_i^{(sc)}\right.\right)\cdot w_{imn}^{(dc)}N_{imn}^{(dc)}(\cdot)\end{aligned} \tag{4-5}$$

其中，K 和 N 表示来自于标准普通话和方言背景普通话，同是第 i 个状态中分别包含的高斯混合数目；对于标准普通话，新的高斯混合权重为 $w_{ik}^{\prime(sc)}=\lambda w_{ik}^{(sc)}$，显然它不仅取决于初始的权重，还受到插值因子的制约；对于方言背景普通话，新的高斯混合权重为 $w_{imn}^{\prime(dc)}=(1-\lambda)\cdot P(s_{im}^{(dc)}\left|s_i^{(sc)}\right.)\cdot w_{imn}^{(dc)}$，它取决于三个参数，即初始权重，插值因子和发音变化概率。由于是概率，则 $\sum_{m=1}^{M}P\left(s_{im}^{(dc)}\left|s_i^{(sc)}\right.\right)=1$ 要满足。

从公式（4-5 ）可以看出，若不考虑发音变化，则 $P\left(s_{im}^{(dc)}\left|s_i^{(sc)}\right.\right)=1$，那么就退化为模型插值（见公式（4-1））。在 SDPBMM 中既不改变决策树的拓扑结构，也不需要修改解码器和发音字典，更重要的是它不需要重新迭代，简单而快捷，需要的数据量也很少。另外，在一定程度上，SDPBMM 可以看作是对标准普通话声学模型适当的扩展，它增加了标准声韵母模型的空间覆盖度，使其能够更鲁棒的识别与标准发音有一定“偏差”的方言背景普通话发音，同时，方言背景普通话模型也可以借助标准普通话上下文相关模型的精度，来提高方言背景普通话的识别率。

4.3.3 SDPBMM中归并准则的确定

在 SDPBMM 中，作为归并两个准则：一是插值因子λ；二是发音变化概率 $P(s_{im}^{(dc)}\left|s_i^{(sc)}\right.)$ 如何确定，是首先要解决的问题。对于状态级的发音变化 $P(s_{im}^{(dc)}\left|s_i^{(sc)}\right.)$，由于受到数据量的限制，通常情况下，采用基于声韵母的发音变化来近似；另外，文（Saraclar，2000）也证明了在声学建模过程中无论是采用基于音素之间的发音变化，还是采用基于状态之间的发音变化，

对于最终的识别率几乎无影响。这样一来，要解决的问题就是基于小数据量的发音建模，核心难点仍然是数据稀疏问题。这个问题在节 3.5 中通过基于距离度量的识别网络生成策略已经很好地加以解决，不过，这里生成的是标准普通话和方言背景普通话声韵母之间的发音变化。对于插值因子λ，一般情况下，通过实验方式来确定（Wang, 2003; Tomokiyo，2001；Witt，1999），本文也是如此。

本文中，将采用“两步走”的策略来确定这两个准则。首先先确定标准普通话和方言背景普通话之间的发音变化，$P(s_{im}^{(dc)}\left|s_i^{(sc)}\right.)$，而后再调整插值因子λ，使基于 SDPBMM 的声学模型性能达到最优。也就是说，这两个准则的确定，都是通过数据驱动的方式实现的。

4.4 有关SDPBMM的测试与分析

在与 SDPBMM 有关的测试验证中，依然采用由标准普通话数据集，TRAIN_STD，训练得到的上下文相关的声韵母 HMM（即节 2.3 中标准普通话声学模型），作为图 4.1 和 4.2 中的“标准普通话 Tri-XIF”；相应地，采用闽南普通话数据集，DEV_MIN，训练得到上下文无关的声韵母 HMM（参见节 2.4）作为图 4.1 和 4.2 中的“方言背景普通话 mono-XIF”。在 SDPBMM 中，核心问题就是两个归并准则的确定。

4.4.1 确定归并准则

首先要确定标准普通话和闽南普通话之间的发音变化，$P(s_{im}^{(dc)}\left|s_i^{(sc)}\right.)$，采用如图 3.3 所示的，基于距离度量的识别网络生成策略来解决数据稀疏问题，依然采用上下文权重策略来估计概率，这里只需得到声韵母之间的发音变化，而不是节 3.5 中最终得到的基于汉语音节的发音变化。相应地，采用了基于概率值和确定数目相结合的剪枝方式来控制发音变化的数量，以减少混淆。表 4.1 中列出了包含一个以上发音变化的声韵母的列表，对于只有一个发音变化的声韵母，则 $P(s_{im}^{(dc)}\left|s_i^{(sc)}\right.)=1$。

表4.1 SDPBMM归并准则之发音变化的确定

标准普通话	概 率	闽南普通话	标准普通话	概 率	闽南普通话
c	0.787	c	ch	0.372	c
	0.213	ch		0.628	ch
h	0.857	h	ii	0.858	ii
	0.143	f		0.142	iii
iii	0.360	ii	in	0.754	in
	0.640	iii		0.246	ing
ing	0.201	in	iong	0.837	iong
	0.799	ing		0.163	iou
n	0.149	l	ong	0.236	eng
	0.851	n		0.764	ong
p	0.831	p	r	0.162	l
	0.169	t		0.152	n
				0.686	r
s	0.820	s	sh	0.371	s
	0.180	sh		0.629	sh
ve	0.152	ie	z	0.786	z
	0.848	ve		0.214	zh
zh	0.358	z			
	0.642	zh			

其次，采用表 4.1 所示的结果，在确定 $P(s_{im}^{(dc)} \mid s_i^{(sc)})$ 之后，对于插值因子 λ，本文采用在实验比对的基础上来确定。对于不同的λ，其在 TEST_STD 和 TEST_MIN 的测试结果如图 4.3 所示。

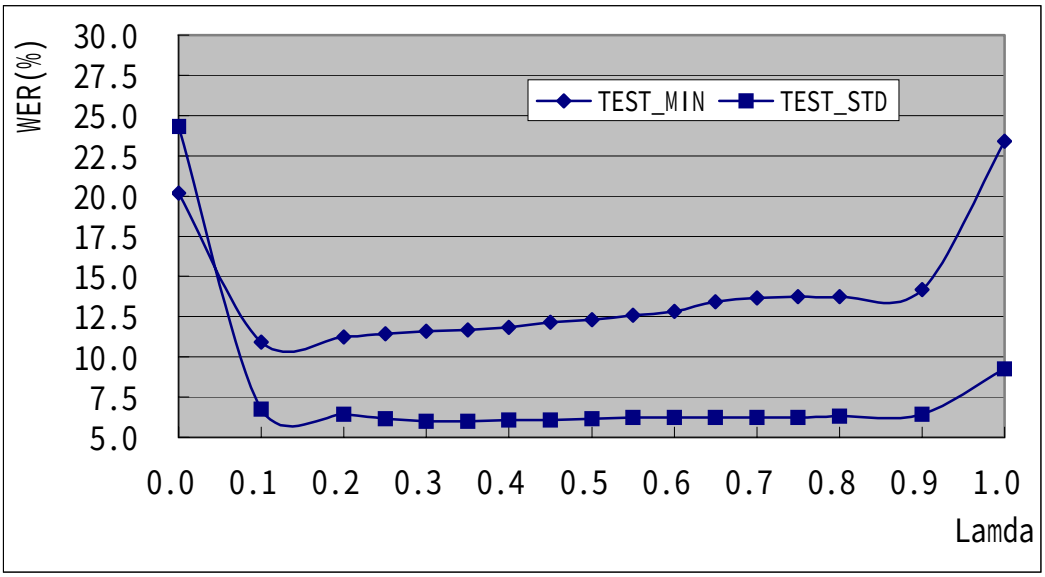


图4.3 确定归并准则λ的最优值

图 4.3 中，横轴代表不同的λ取值，从图中可以看出：

1. 当λ=0 时，就相当于闽南普通话的上下文相关的 HMM（此处借助的是标准普通话上下文相关的状态共享结构），其对闽南普通话的识别效果好于标准普通话，但都不理想。当λ=1，SDPBMM 退化为标准普通话上下文相关声学模型，识别结果与第 2 章中表 2.6 的结果是相同的。

2. 对于闽南普通话的识别，当λ位于区间 0.1~0.9 时，词错误率随着λ的增大而增大，这说明对于闽南普通话的识别过程中，标准普通话模型所占权重越小，闽南普通话模型所占权重越大，识别率越高，这与实际经验是一致的。

3. 对于标准普通话的识别，当λ位于区间 0.1~0.9 时，词错误率最低点为 0.30 和 0.35。但总体变化趋势比较平稳，不如闽南普通话识别过程中，对λ值的变化敏感。这说明，SDPBMM 中通过状态归并来共享模型参数，的确提高了标准普通话模型的空间覆盖度，对于标准普通话的识别率有所提高，对于λ值也不甚敏感。

4. 通过调整插值因子λ，可以保证基于 SDPBMM 的声学模型对标准普通话和方言背景普通话都有较好的识别正确率。本文中取λ=0.3 作为 SDPBMM 中插值因子的默认值。

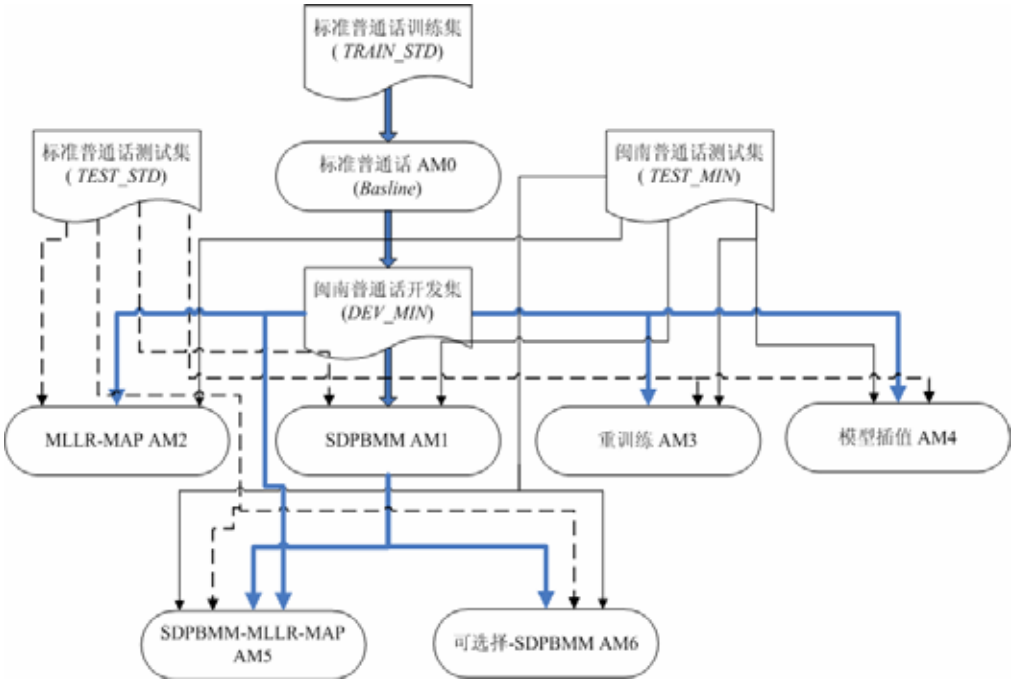


图4.4 有关SDPBMM的实验设计框架

为了能更好地与其它声学建模方法相对比，本文的实验设计如图 4.4 所示，后面的实验都是依照此图开展的。

在图 4.4 中，基于 SDPBMM 的声学模型对应于 *AM*1。测试中，都采用了标准普通话发音字典。不同方法之间对比结果的分析，统一放在节 4.4.5 中进行。对于 SDPBMM 的测试结果如表 4.2 所示：

表4.2 SDPBMM对于TEST_STD和TEST_MIN测试结果

声学模型	**WER**	
	TEST_STD	**TEST_MIN**
SDPBMM (AM1)	6.0%	11.6%

需要说明是，在 SDPBMM 中，由公式（4-5）可以知道，在进行状态归并中，势必造成每一个通过决策树共享的状态所包含的高斯混合数目的增加，但是，由节 2.3 中的结论知道，单纯提高模型状态中包含的高斯混合数目是不能有效提高识别率的（见图 2.3）。

4.4.2 SDPBMM与自适应

本文首先对基于 MAP、MLLR、MLLR 与 MAP 结合，三种自适应方法进行了对比，从中选择性能最好的作为本文的自适应方法。在自适应中，同样采用了 DEV_MIN，作为自适应数据。在三种自适应方法中，都只对均值向量进行更新。在 MLLR 中，使用了 65 个回归类，每一个声韵母对应一个回归类。图 4.5 列出了三种自适应方法对于 TEST_MIN 测试结果。

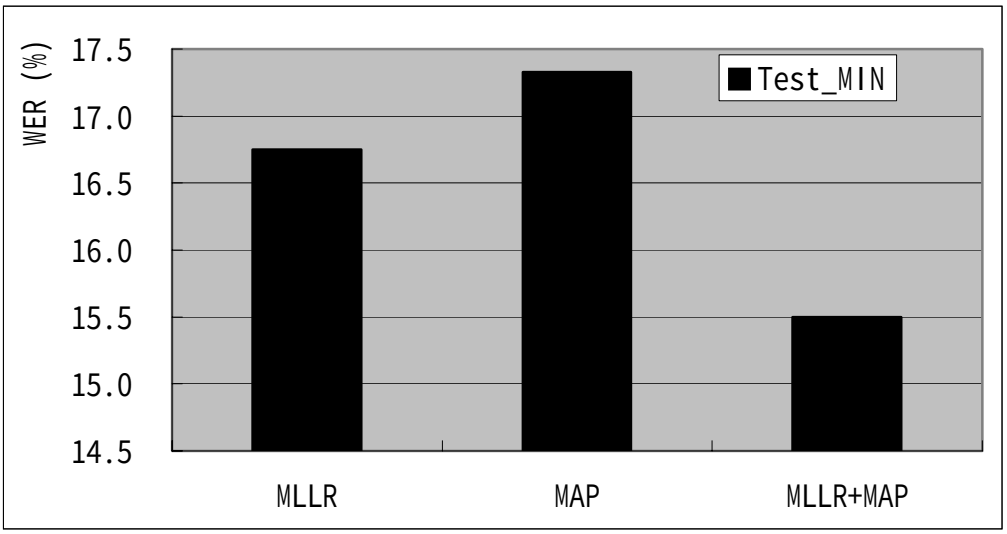


图4.5 MLLR、MAP和MLLR+MAP性能比较

由图 4.5 可以看到，在少量自适应数据情况下，MLLR 性能确实优于 MAP。其中性能最好的是 MLLR 与 MAP 相结合的方式，即 *MLLR+MAP*，这与首先进行 MLLR 自适应，得到 MAP 自适应所需要的精确初始模型有关系。相对于 MLLR，MLLR+MAP 可以将词错误率多降低 1.3%。通过实验，本文将把 MLLR 和 MAP 相结合的方式作为默认的自适应方法，对应于图 4.4 中的 *AM*2。

基于 MLLR+MAP 的自适应方法对于 TEST_STD 和 TEST_MIN 的测试结果如表 4.3 所示：

表4.3 自适应方法对于TEST_STD和TEST_MIN测试结果

声学模型	WER	
	TEST_STD	TEST_MIN
MLLR+MAP (AM2)	10.9%	15.5%

4.4.3 SDPBMM与重训练

为了同重训练方法进行对比，本文采用了文（Wang, 2002）中数据混合方式的重训练(pooled retraining)。在此实验中将 TRAIN_STD 和 DEV_MIN 中的数据进行了混合，采用了与标准普通话声学建模相同的方法和同等的模型规模，对应于图 4.4 中的 *AM*3。表 4.4 列出了基于数据混合方式重训练的结果。

表4.4 重训练对于TEST_STD和TEST_MIN测试结果

声学模型	WER	
	TEST_STD	TEST_MIN
重训练 (AM3)	10.4%	19.2%

4.4.4 SDPBMM与模型插值

在模型插值实验中，将标准普通话上下文相关声学模型 AM0 与闽南普通话上下文无关的模型进行了模型插值。与 SDPBMM 相同，这里也取 λ=0.3。插值模型对应于图 4.4 的 *AM*4。表 4.5 列出了插值模型的测试结果。

表4.5 模型插值对于TEST_STD和TEST_MIN测试结果

声学模型	WER	
	TEST_STD	TEST_MIN
模型插值 (AM4)	7.8%	12.9%

4.4.5 与SDPBMM对比实验结果分析

为了更好的说明问题，图 4.6 列出了标准普通话声学模型，自适应模型，重训练、插值模型和 SDPBMM 对比结果。

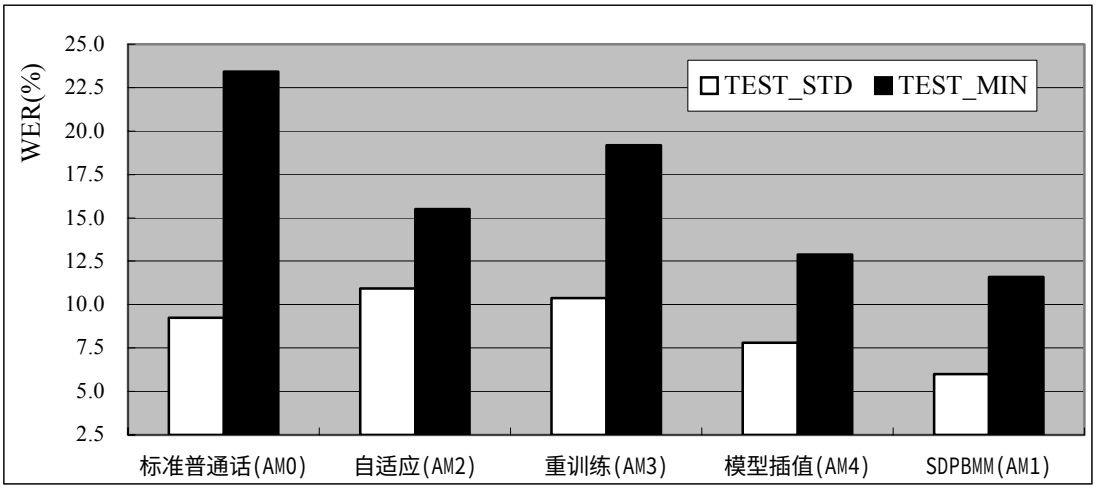


图4.6 多种声学建模方法性能对比

从图 4.6 可以看出:

1. 相对于标准普通话声学模型,其它几种建模方法都能降低闽南普通话的识别率，其中效果最显著的是 SDPBMM，词错误率由 23.4%降为 11.6%，相对错误率降低了 50.4%。其它的建模方法在小数据量条件下，性能不如 SDPBMM。

2. 相对于标准普通话声学模型，在对标准普通话的识别中，自适应和重训练方法都不能保证识别正确率不降低，这与实际经验是一致的。模型插值和 SDPBMM 不但没有造成标准普通话识别率的降低，还提高了识别率，可见若要保持标准普通话识别率不降低，适当扩大声学模型的规模是必要的；另外，因为在测试过程中，采用了受限的词表，混淆度比不受任何限制的词表要低，这也是造成标准普通话识别率显著提高的部分原因。同样，SDPBMM 的效果也是最显著的，词错误率由 9.3%下降为 6.0%，相对词错误率降低了 32.3%。

3. 在 SDPBMM 中，在状态一级进行建模，方言背景普通话模型借助于标准普通话模型的精度提高对方言背景普通话的识别率；同样，标准普

通话模型借助于方言背景普通话模型，涵盖更多的发音变化，使标准普通话模型有了更大的声学空间覆盖度，因而也提高了对标准普通话的识别率。

4.4.6 SDPBMM与自适应方法相结合

实际中，存在一些有效建模方法会与已有的一些常用的方法相互抵触，其性能有可能相互抵消。为了验证此点，本文又进行了 SDPBMM 与自适应方法相结合的实验。就是在 SDPBMM 模型的基础上，利用 DEV_MIN 数据集，再进行基于 MLLR 与 MAP 相结合的自适应，对应于图 4.4 中的 *AM*5。其测试结果如表 4.6：

表4.6 SDPBMM与自适应结合方法对TEST_STD和TEST_MIN测试结果

声学模型	**WER**	
	TEST_STD	**TEST_MIN**
SDPBMM_MLLR+MAP (AM5)	6.8%	11.0%

与基于 SDPBMM 声学模型 AM1 相比，AM5 将闽南普通话的词错误率进一步降低，从 11.8%降为 11.0%。但对于标准普通话的词错误率由 6.0%上升为 6.8%。这与自适应方法的特性是一致的，即识别率的提高是“ 单向 ”的，但由于 SDPBMM 增大了对于标准普通话的空间覆盖度，自适应之后的声学模型也没有造成标准普通话识别率的显著下降。

4.5 可选择的SDPBMM

由公式（4-3）知道，在 SDPBMM 中，归并状态的共享参数不仅包含了来自于标准普通话的高斯混合，还包含了来自于方言背景普通话的高斯混合，这就造成了模型规模的扩大，本文称这种现象为**高斯混合扩张问题**。为了抑制 SDPBMM 中存在的高斯混合扩张，降低模型规模，本文提出了可选择的 SDPBMM 方法，其核心思路就是在状态归并参数共享的过程中，依据一定的策略将需要参与归并的状态和不需要参与归并的状态（冗余的

状态）区分开来，提高 SDPBMM 中状态归并的针对性，从而达到降低模型规模的目的。基本思想如图 4.7 所示：

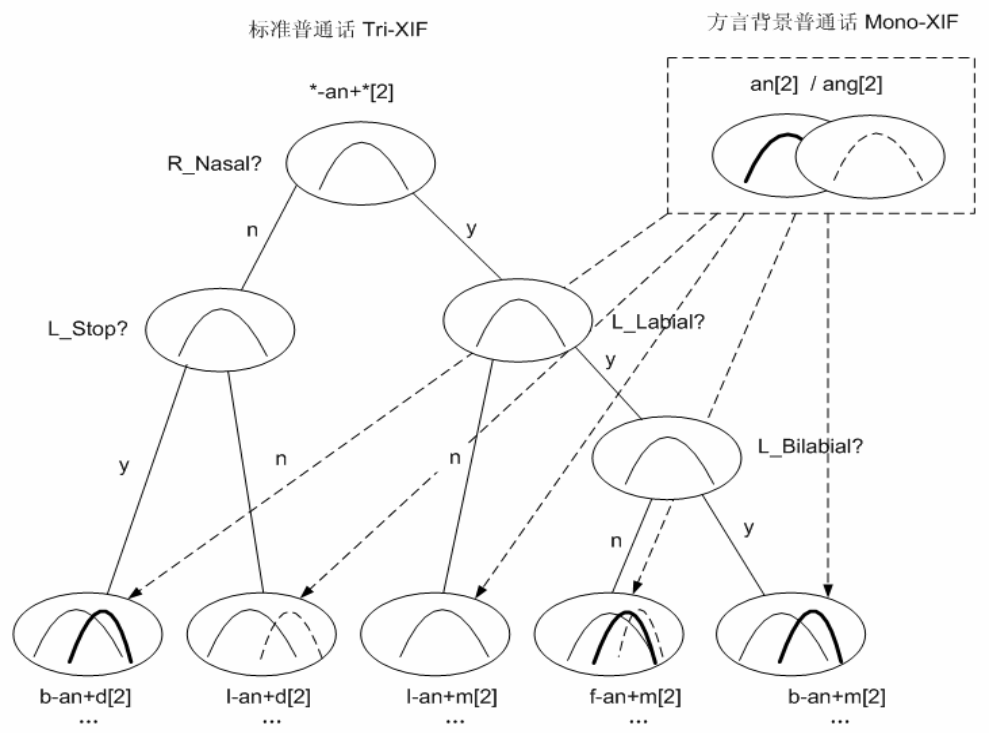


图4.7 可选择的SDPBMM

从图 4.7 可以看出，在进行状态归并的时候，不再像图 4.2 那样将标准普通话的 tri-XIF 中的共享状态与方言背景普通话的 mono-XIF 状态进行完全归并，而是进行了一定的选择，例如，图中状态 *b-an+d*[2]就只与 *an*[2]进行了归并，*f-an+m*[2]就与 *an*[2]和 *ang*[2]都进行了归并，而 *l-an+m*[2]就没有同任何一个状态归并。

本文采用了距离度量的策略来实现可选择 SDPBMM 建模，考虑到 SDPBMM 是为了增加标准普通话模型的空间覆盖度，同时又不引入新的混淆，因此对不同的声韵母采用了不同的距离度量策略。对于标准普通话 tri-XIF 以及与其中心基元相同的方言背景普通话 mono-XIF 的组合，即图 4.7 中的*-*an*+*[2]与 *an*[2]，称之为**类别 1**；对于标准普通话 tri-XIF 及其中心基元对应的方言背景普通话发音变化的组合，即图 4.7 中的*-*an*+*[2]与 *ang*[2]，称之为**类别 2**。

对于类别 1，在状态归并中希望突出两者之间的差异，这样一来，归并后的状态空间覆盖度就大。非对称的马氏距离采用方言背景普通话模型的协方差矩阵来计算距离，强调标准普通话和方言背景普通话状态之间的差异性（Liu，2005；Tasi，2003），因而本文采用非对称的马氏距离。其定义如公式（4-6）和（4-7）。公式（4-6）中，A 和 B 分别代表来自于标准普通话和方言背景普通话的模型状态，它们包含 M 和 N 个高斯混合。公式（4-7）表示非对称的马氏距离，用来计算任意两个高斯混合的距离。

$$dis(A,B)=\sum_{i=1}^{M}w_{Ai}\left[\sum_{j=1}^{N}w_{Bj}d_{A,B}(i,j)\right] \tag{4-6}$$

$$d(i,j)=\left(\mu_i-\mu_j\right)^T\Sigma_j^{-1}\left(\mu_i-\mu_j\right) \tag{4-7}$$

对于类别 1，其归并策略如表达式（4-8），其中α为阈值，当标准普通话和方言背景普通话的状态距离大于阈值α时，则说明要参与状态归并（用 1 表示），否则不参与状态归并（用 0 表示）。

$$\begin{cases}1, & dis\left(s^{(sc)},s^{(dc)}\right)\geq\alpha\\ 0, & dis\left(s^{(sc)},s^{(dc)}\right)<\alpha\end{cases} \tag{4-8}$$

对于类别 2，在状态归并中希望缩小两者之间的差异，以期望不要因为差异过大而引入混淆（声学空间出现交叠），比如标准普通话识别中的 *an* 和 *ang* 的混淆，因此采用了类散度距离（Bhattachyaryya 距离）（公式定义见表 3.7）作为状态选择准则，因为此距离强调两者之间的相似性，不凸显其差异。那么对于类别 2，其归并策略如表达式（4-9）。其中β也是一个阈值，此表达式与（4-8）正好相反，当标准普通话和方言背景普通话状态距离在阈值β范围内的时候，则两者状态归并，否则无归并发生。

$$\begin{cases}1, & dis\left(s^{(sc)},s^{(dc)}\right)\leq\beta\\ 0, & dis\left(s^{(sc)},s^{(dc)}\right)>\beta\end{cases} \tag{4-9}$$

4.5.1 可选择SDPBMM性能测试

在实验中，对于阈值的确定，依然采用了类似于相对最大概率的策略，对于类别 1，其阈值 $\alpha = 0.3 \cdot d_{\max}$（其中 $d_{\max}$ 是类别 1 中的距离最大值），就是说两个状态距离大于α时，将它们进行归并；对于类别 2，其阈值 $\beta = 0.7 \cdot d'_{\max}$（其中 $d'_{\max}$ 是类别 2 中的距离最大值），就是说两个状态距离小于β时，将它们进行归并。通过调整阈值，实验发现将 SDPBMM 声学模型的规模降低 30%的时候识别效果最好。由于 SDPBMM 的基本性能和特点都在前面进行了阐述，本节只对 SDPBMM 和可选择的 SDPBMM 进行了对比，可选择 SDPBMM 模型对应于图 4.4 的 *AM*6。可选择的 SDPBMM 在标准普通话和闽南普通话上的测试结果如表 4.7 所示：

表4.7 可选择SDPBMM的测试结果

声学模型	**WER**	
	TEST_STD	**TEST_MIN**
可选择 SDPBMM (AM6)	6.4%	11.5%

由表 4.7 知道，相比于 SDPBMM，可选择的 SDPBMM 的规模降低了 30%，而性能没有显著变化，对于方言背景普通话，词错误率几乎无变化，这说明在归并过程有些状态的归并是存在一定的冗余的，降低模型规模并不影响识别率。对于标准普通话，词错误率略有上升，从 6.0%到 6.4%，原因还是来自于模型规模的缩小，可能的发音变化覆盖度减小。

总之，可选择的 SDPBMM 在适度的降低了模型规模的同时保持了 SDPBMM 特性。在后面的实验中将以可选择的 SDPBMM 作为默认的 SDPBMM 模型。

4.6 小结

为了解决标准普通话发音和方言背景普通话发音之间的声音变化问题，SDPBMM 利用 1 小时的方言背景普通话数据进行声学建模，实验证明 SDPBMM 具有以下特点：

1. SDPBMM 是一种简单而有效的声学建模方法，其建模过程中保持了标准普通话模型的基本结构，不需要重训练和再迭代，也不需要大量的数据。

2. SDPBMM 显著的降低了方言背景普通话的错误率，相比于标准普通话声学模型，其相对词错误率降低了 50.8%，性能优于自适应、重训练和模型插值方法。

3. SDPBMM 也提高了标准普通话的识别率，克服了自适应，重训练只能单向提高识别正确率的不足。实验中，相比于标准普通话声学模型，相对词错误率降低了 31.2%。

4. SDPBMM 还能很好的同自适应方法想结合，可以进一步提高方言背景普通话的识别率。在实际中，SDPBMM 可以作为自适应的前端处理方法而应用于语音识别系统中。

5. SDPBMM 本身也固有的存在高斯混合扩张问题，本文对属于不同类别的状态采用了不同的距离度量策略来控制状态之间的归并，从而实现有选择的 SDPBMM 方法，实验证明，在模型规模降低 30%的情况下，性能没有受到影响。

6. SDPBMM，不针对某种特定的方言背景普通话，不借助方言相关的专家知识，因而很容易推广到某种方言背景普通话的声学建模中，甚至应用于其它基于子词（如音素）的声学建模中。

第 5 章　声学建模方法综合

5.1　引言

在方言背景普通话的语音识别研究中，针对方言背景普通话在声学层面，与标准普通话相比，出现的音素变化和声音变化。本文分别提出了基于小数据量的方言背景普通话发音字典和一种新颖的建模方法 SDPBMM 来对这两类变化进行分别建模。在本节将对这两个方向的建模方法进行综合、对比及分析，并将与自适应方法进行结合，最后回顾各种方法组合对于标准普通话和方言背景普通话性能的影响。

5.2　方言背景普通话发音字典与SDPBMM结合

这里将采用节 3.6 中，基于少量闽南普通话得到的发音字典（包含 520 个发音）以及节 4.5 中的可选择的 SDPBMM 声学模型进行结合，并测试。本文把这两者的结合称之为方言背景普通话识别器。表 5.1 列出方言背景普通话识别器在数据集 TEST_STD 和 TEST_MIN 上的测试结果。

表5.1 发音字典与SDPBMM结合的测试结果

识别器	**WER**	
	TEST_STD	**TEST_MIN**
闽南普通话发音字典 +SDPBMM	6.7%	11.3%

一方面，对比表 5.1 和表 4.7 的结果可以知道，采用闽南普通话发音字典之后，进一步提高了对于方言背景普通话的识别率，词错误率由 11.5%降低到 11.3%；另一方面，对于标准普通话的识别率有所下降，词错误率由 6.4%上升为 6.7%。这说明，方言背景普通话发音字典很好的反映了方言背景普通话中的音素变化，但对于标准普通话的识别，引入了一定的混淆，因为有一些声音变化已经在 SDPBMM 中解决了。

另一方面，对比表 5.1 和表 3.11 的结果可以知道，在采用同一个闽南普通话发音字典的情况下，对于闽南普通话，词错误率由 20.9%降低为 11.3%；对于标准普通话，词错误率由 9.3%降低为 6.7%。由此可见，对于系统性能的提高主要来自于有效的建模方法，源自于对方言背景普通话中声音变化的很好解决，这与文（Hain，2005；Kam，2003；Saraclar，2000；Liu，2004；Fung，2005）认为的状态级的发音建模比字典级的发音建模能更显著的提高识别率的结论是一致的。

5.3 方言背景普通话识别器与自适应的结合

通过前面的实验可以知道，声学模型自适应方法能进一步提高系统对于方言背景普通话的识别率，但会降低标准普通话的识别率。总体上，同标准普通话识别器相比，方言背景普通话识别器不但没有造成标准普通话识别率的降低，还显著地提高了标准普通话的识别率。在对标准普通话识别率较为满意的前提下，若想进一步缩小标准普通话和方言背景普通话识别率的差距，可再进行声学模型自适应。本文对节 5.2 方言背景普通话识别器中的声学模型 SDPBMM 又进行了基于 MLLR 与 MAP 相结合的自适应，其在 TEST_STD 和 TEST_MIN 上的识别结果如表 5.2 所示：

表5.2 发音字典与SDPBMM结合的测试结果

识别器	**WER**	
	TEST_STD	**TEST_MIN**
闽南普通话发音字典 +SDPBMM+MLLR+MAP	7.3%	10.8%

表 5.2 中的结果对于方言背景普通话的识别率有进一步提高，对标准普通话识别率有一定下降，这与实际期望的是相一致的。

总之，在本文的基线系统（标准普通话识别器）中，标准普通话和方言背景普通话两者识别率之间存在 14.1%的差距,经过多种方法综合之后，两者之间的差距变为 3.5%;更为重要的是，这是在标准普通话与方言背景普通话识别率都有所提高的情况下获得的。

5.4 各方法有效性概览

为了对本文所采用方法的性能有一个全面而清晰的了解，图 5.1 列出了各种方法组合对标准普通话和方言背景普通话的识别结果。

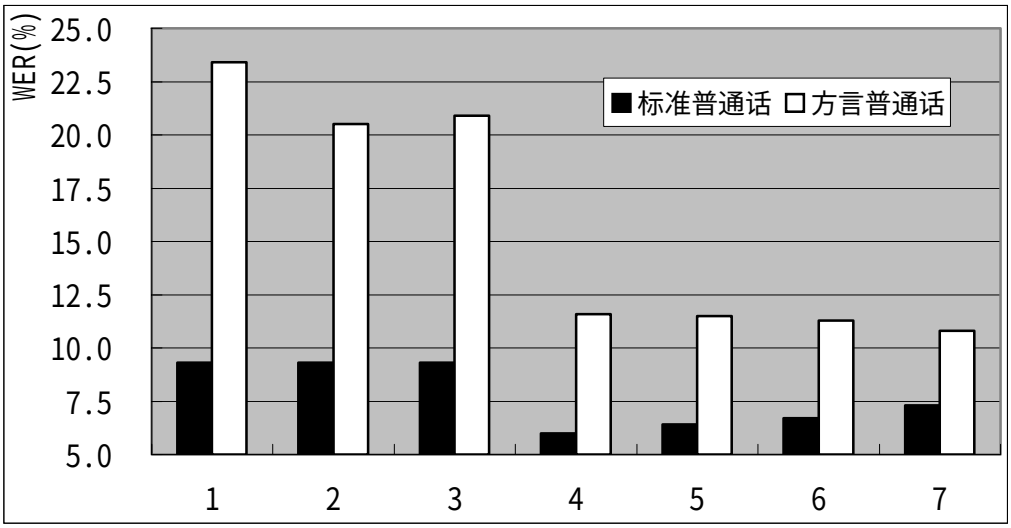


图5.1 本文所采用的建模方法有效性概览

在图 5.1 中，横轴代表了不同组合的测试项目，具体实验环境的说明如表 5.3 所示：

表5.3 对图5.1不同实验环境的说明

实验	实验环境	备 注
1	标准普通话声学模型+ 标准普通话发音字典	标准普通话识别器（基线系统）
2	标准普通话声学模型+ 闽南普通话发音字典 1	基于 4 小时闽南普通话数据，采用改进的上下文相关权重策略
3	标准普通话声学模型+ 闽南普通话发音字典 2	基于 1 小时闽南普通话数据，采用改进的上下文相关权重策略 （默认的闽南普通话发音字典）
4	SDPBMM+ 标准普通话发音字典	基于 1 小时闽南普通话数据

表5.3（续） 对图5.1不同实验环境的说明

实验	实验环境	备 注
5	可选择 SDPBMM+ 标准普通话发音字典	基于 1 小时闽南普通话数据
6	可选择 SDPBMM+ 闽南普通话发音字典 2	采用实验 3 中的闽南普通话发音字典
7	可选择 SDPBMM+MLLR+MAP 闽南普通话发音字典 2	采用实验 3 中的闽南普通话发音字典

图 5.1 清晰地的显示了缩小标准普通话和方言背景普通话识别率差距的过程。两者识别率差距的缩小意味着系统对于方言背景语音不再敏感（识别率不再大幅降低）。系统的鲁棒性显著提高，这对于一个实用的语音识别系统是至关重要的。

5.5 讨论

由图 5.1 可以清楚的看到，对于方言背景普通话语音识别而言，SDPBMM 的效果显著地的优于发音字典。采用相同的数据集 DEV_MIN，SDPBMM 可以将相对词错误率降低 50.8% ;而发音字典可以将相对词错误率降低 10.7%。这说明：

1）发音字典不能解决声音变化问题。发音字典只能解决确定性的发音变化，例如 *zh→z*，*sh→s*，*ch→c* 等，而位于两者之间的声音变化，往往无能为力，发音字典充其量利用概率来刻画音素变化的可能性，不能从模型本身提高对发音的辨识度。而 SDPBMM 拥有更大的声学空间覆盖度，对于不同程度的相对于标准普通话发音的“偏离”都能较好的解决。

2）声音变化出现的频率大大高于音素变化。方言普通话越标准，声音变化的现象越普遍。在书面语形式（如朗读方式）的方言背景普通话中，由于说话人在录音时注意到是在讲普通话，其结果是发音更接近于标准普通话，而远离方言发音习惯，这也是 SDPBMM 优于发音字典的一个重要原因。而在自由交谈的方式中，方言的影响力将增强，其中音素变化的概率也将提高。

第 6 章 总结与展望

6.1 总结

本文针对方言背景普通话语音识别的声学建模展开研究，围绕标准普通话和方言背景普通话之间在声学层面的两类变化，即音素变化和声音变化，提出了若干新方法、新策略，并通过实验证明了其有效性，同时也为后续的方言背景普通话语音识别的研究积累了一定的经验。

概括来说，本文的工作重点与贡献主要体现在如下几个方面：

6.1.1 充分利用小数据量进行声学建模

本文始终把基于小数据量的声学建模作为一个关注点，在建模过程中探索如何充分利用少量方言背景普通话数据，解决小数据量建模过程中面临的数据稀疏问题。作为一种通用方言背景普通话语音识别建模方法，方法的经济性是一个必须面对的问题。采用少量方言背景普通话数据，快速而有效与鲁棒的标准普通话识别器相结合，达到显著提高方言背景普通话识别率的同时不降低标准普通话识别率的目标，这是减少经济成本，利用已有成果，产品实用化的具体要求。

6.1.2 提出改进的上下文相关权重策略

针对方言背景普通话发音字典生成过程中字典中的音节对应不合理声韵母组合的现象，提出了改进的上下文相关权重策略来降低发音字典的混淆度，提高方言背景普通话的识别率。与上下文相关的权重策略相比，相对词错误率多降低了 3.8%；与标准普通话发音字典相比，相对词错误率下降了 12.4%。同时，对于标准普通话语音的识别率没有降低。

6.1.3 提出基于距离度量的识别网络生成策略

为了解决基于小数据量的方言背景普通话发音字典生成中面临的数据稀疏问题，本文提出了基于距离度量的识别网络生成策略。在方言背景普通话发音字典生成中，采用以声学距离为依据生成受限的识别网络，从而达到限制发音变化趋势分散的目的，很大程度上解决了数据稀疏问题。同时，对不同的距离度量方法进行了对比，提出了距离度量准则评价策略。在此基础上生成的方言背景普通话发音字典，在性能上接近基于大数据量的发音字典。

6.1.4 提出状态相关基于基元的模型归并策略

为了解决方言背景普通话中的声音变化问题，尤其在少量方言背景普通话数据条件下，本文提出了状态相关基于基元的模型归并策略 SDPBMM，在标准普通话模型状态中引入了方言背景普通话模型状态参数。实验证明，SDPBMM 是一种简单而有效的建模方法，在小数据量条件下，系能优于自适应、重训练和模型插值方法。与标准普通话声学模型相比，对于方言背景普通话的相对词错误率降低了 50.4%，对于标准普通话的相对词错误率降低了 32.3%，达到了显著提高方言背景普通话识别率的同时不降低标准普通话识别率的目标。

6.1.5 提出可选择的SDPBMM

为了解决 SDPBMM 中的高斯混合模型扩张问题，本文提出了可选择的 SDPBMM 策略，其核心思路，就是利用一定的准则将参与归并中的状态进行鉴别，通过剔除冗余状态达到降低模型规模的目的。文中采用了距离度量作为鉴别准则，对不同的状态采用不同的距离度量准则。实验表明，在模型规模降低 30%的情况下，模型性能不受影响。

6.2 展望

对于方言背景普通话的语音识别，虽然经过本人和前人连续几年的研究，取得了一些成果（宋战江, 2001; 李净，2005；Sproat，2004），但还存在许多不

足，很多关键问题还有待解决，这方面的研究工作还将继续下去。下一阶段的研究工作将在以下几个方向展开：

6.2.1 鲁棒的标准普通话声学模型

鲁棒的标准普通话识别器是进行汉语语音识别的基础，只有建立高性能的标准普通话识别器，才能将标准普通话语音识别中的相关技术应用于方言背景普通话语音识别，才能更好的同方言背景普通话语音识别技术相融和。例如本文中，如果没有鲁棒的标准普通话识别器为基础，对于方言背景普通话的提高也是有限度的，还是离实际应用有差距的。毕竟普通话是我国的通用语言，方言背景普通话是普通话的地方变体，实用的方言背景普通话识别器是以标准普通话识别器为基础的。在语音识别系统中，声学模型的占有至关重要的地位，是整个识别系统的基础，直接影响着系统的整体性能。如何采用更加有效的建模方法是这方面研究的主要课题。区分性训练（discrimative training）虽然具有更好的可分性，不仅对本类的数据具有很好的描述性，对于不同类别之间也有很好的描述性，但由于存在计算量大，收敛速度慢的不足，一直只应用于中小规模的语音识别系统（Schlüter，2001）。目前随着研究的深入和发展，区分性训练正逐步应用于大词表连续语音识别中（Woodland，2002；Doumpiotis，2006）。目前这部分工作还处在起步阶段，还只是局限在英语语音识别中，对汉语语音识别还未全面展开（Wu，2005）。我们将在这方面展开研究工作，建立更加鲁棒的标准普通话声学模型。

6.2.2 韵律信息在方言背景普通话语音识别中的应用

声调是汉语的典型特征，标准普通话与方言背景普通话的差异很大程度上反映在声调的差异上，目前声调在语音识别中的应用还局限在标准普通话（Zhou，2004）或者方言（例如粤语）语音识别（Lee，2002）中，在方言背景普通话语音识别中的应用还很少，与标准普通话语音识别同时应用还未发现有这方面的文献。另外，文（Li，2003）中，作者在对比标准普通话和上海普通话的发音特点基础上，提出这两者之间的差异更多的体现在重音（stress）上面。目前，如何提取反映重音变化的特征以及如何

同传统的建模方法相结合，都是未来要研究课题。另外很多语音学学者在重读（王蓓，2002）、共振峰（于珏，2004；陈娟文，2004）、声韵母（周萍，2006）等方面开展了标准普通话和方言背景普通话的对比研究，如何将这些研究成果应用于语音识别中也是未来面临的挑战。

6.2.3 识别基元的自动扩展和评价

由于标准普通话与方言在声韵母集合上的差异，例如上海话就有 34 个声母，54 个韵母组成，那么用标准普通话声韵母集合很显然不能完全涵盖方言背景普通话的发音特点，对方言背景普通话声韵母集合进行适当扩展对于提高识别率有很好的帮助。以往的方法更多的借助于语音学者的先验知识（李净，2005；宋战江，2001）来进行识别基元的扩展，是一种基于专家知识的方式。另外，也有基于距离度量（Liu，2006a；Tsai，2003）和基于置信度的识别基元自动生成策略（Liu，2003），但这些方法都对数据都有很大的依赖性，对于已有的语音知识应用不够。总体来说，生成方法还比较简单、粗糙。此外，如何评价扩展识别基元集合的混淆度，以及与标准普通话识别基元集合的关系都是要研究的问题。

6.2.4 基于高斯混合的建模方法

本文中提出了建模方法 SDPBMM 是基于模型状态级，在状态级对来自于标准普通话和方言背景普通话的模型状态的高斯混合，依据一定的准则进行归并。在此过程中，并没有对参与归并的高斯混合进行鉴别，这是造成模型规模扩大的主要原因。在未来，将对参与归并的高斯混合进行鉴别，在一定准则的限定下，从多个高斯混合集合中选择最能提高识别率的子集。这样一来，就可以真正的在提高识别率的同时，不扩大模型规模。此问题的难点就是鉴别准则的确定，先期实验表明，单纯的距离度量策略不能满足上述目标。

6.2.5 继续开展方言背景普通话语言模型的研究

作为“方言背景普通话语音识别框架”的重要组成部分，我们将继续开展方言普通话语言模型的研究。在以往的工作中，通过收集一定数据量的方言词汇来对标准普通话语言模型进行自适应。例如，在文（李净，2005）就在 15K 的标准普通话语言模型中加入了 200 个吴方言常用词汇来训练 uni-gram 和 bi-gram 模型。由于方言普通话与标准普通话在词汇上的差异小于在语音上的差异，而训练语言模型比训练声学模型需要更多的数据。另外，我国采用统一的文字，现成的文本大都是标准普通话的，方言普通话的文本几乎没有。如果重新训练新的方言背景普通话语言模型就存在着“费力不讨好”的矛盾，即收集大量的文本语料，但识别效果并不明显。因而，如何利用方言知识，利用语义规则是解决数据稀疏问题是一条现实而可行之路。在标准普通话和方言背景普通话在词汇的差异主要表现在以下几个方面：1）普通话的复音，发为单音。例如，“衣服”对应闽南背景普通话的“衫”；2）普通话的单音，发为复音。例如，“床”对应闽南普通话的“眠床”；3）词素排列次序不一样；例如，“前头”和“头前”等。4）利用方言词汇替代标准普通话词汇，例如标准普通话的“知道”，在上海普通话中对应于“晓得”，这就是一些对应规则。利用方言知识和少量数据进行语言模型的建模是未来的工作方向之一。

6.2.6 开展基于自然发音的方言背景普通话语音识别

本文的研究主要集中在朗读式发音的方言背景普通话语音识别上，相对于自然发音的方式，其发音变化，协同发音等现象要少很多。如何将已有的研究成果应用于自然发音方式的方言背景普通话语音识别，同时提出适合于自然发音的建模方法也是下一步的工作任务。2004 年，由美国自然科学基金（NSF）资助的 Workshop 在约翰·霍普金斯大学的语言和语音处理中心举行，其中一个主要研究课题就是“吴方言背景普通话研究”，对方言背景普通话语音识别展开了一系列研究，取得了一些成果（Sproat，2004）。对于自然发音方式的上海普通话的字符错误率达到了 43.7%，可见还有很大提升空间。目前的研究发现，一个鲁棒的标准普通话识别器对

于方言背景普通话识别率的提高起着关键作用，因为有许多发音变化并不是由方言口音引起的，这些应由标准普通话识别器来解决。

参考文献

陈娟文. 2004. 上海普通话和普通话韵律特征对比研究[硕士学位论文]. 浙江大学.

陈亚川. 1991. “地方普通话”的性质特征及其它. 世界汉语教学，1:12-17.

丰洪才，卢正鼎. 2005. 基于 MAP 和 MLLR 的综合渐进自适应方法研究. 计算机工程，31(5):4-7.

何磊. 2001. 语音识别中的说话人鲁棒性和自适应技术研究 [博士学位论文]. 清华大学计算机科学与技术系.

侯精一. 2002. 现代汉语方言概论，上海教育出版社.

李爱军, 王霞, 殷治纲. 2003. 汉语普通话和地方普通话的对比研究，第六届全国现代语音学学术会议.

李净，郑方，张继勇等. 2004. 汉语连续语音识别中上下文相关的声韵母建模. 清华大学学报，44(1): 61-64.

李净. 2005. 吴方言背景普通话语音识别研究[博士学位论文]. 清华大学计算机科学与技术系.

刘明宽，徐波，黄泰翼等. 2002. 音节混淆字典以及在汉语口音自适应中的应用研究. 声学学报, 27(1):53-58.

穆向禹, 贾磊, 张树武,徐波. 2003. 基于目标驱动的多层 MLLR 自适应算法. 中文信息学报，17(6):39-46.

潘复平, 赵庆卫, 颜永红. 2005. 一种用于方言口音语音识别的字典自适应技术. 计算机工程与应用，23:4-9.

彭煊，王炳锡. 2005. 基于高斯混合模型差别度量的说话人聚类. 计算机工程与应用, 5:99-102.

普园媛, 杨鉴, 尉洪等. 2005. 云南民族口音汉语普通话语音识别研究, 计算机工程与应用, 11:45-47.

齐沪扬. 1999. 就“方言背景普通话”答客问. 修辞学习，4:26-27.

宋战江. 2001. 汉语自然语音识别中发音建模的研究[博士学位论文]. 清华大学计算机科学与技术系.

王蓓，吕士楠，杨玉芳. 2002. 汉语语句中重读音节音高变化模式研究. 声学学报, 27(3):234-240.

王仁华，何林顺，黎建宁. 1992. 等方差加权倒谱失真测度及其在说话人识别中的应用. 电子学报, 20(8):49-55.

王昱. 2000. 语音识别自适应技术的研究与实现[硕士学位论文].清华大学计算机科学与技术系.

王作英，李健. 2003. 汉语连续语音识别的语速自适应算法. 声学学报，28(3):229-234.

杨行峻，迟惠生等. 1995. 语音信号数字处理. 电子工业出版社.

姚佑椿. 1989. 开展对“地方”普通话的研究. 语文研究与应用，3:18-21.

于珏. 2004. 上海普通话与标准普通话元音系统的声学对比研究[硕士学位论文], 浙江大学.

赵征鹏，杨鉴. 2005. 基于高斯混合模型的非母语说话人口音识别. 计算机工程，31(6):148-150.

郑方，牟晓隆，徐明星等. 1999. 汉语语音听写机技术的研究与实现. 软件学报，10(4): 436~444.

钟金宏. 2001. 基于音节的汉语连续语音声调识别方法研究[博士学位论文]. 合肥工业大学。

周萍. 2006. 上海地区普通话水平测试中韵母发音偏误的比较研究[硕士学位论文]. 华东师范大学;

Amdal I, Lussier E F. 2003. Pronunciation variation modeling in automatic speech recognition. Telektronikk, 2:70-82.

Bates R A，Ostendorf M，Wright R. 2007. Symbolic phonetic features for modeling of pronunciation variation. Speech Communication, 49:83-97.

Beulen K，Ney H. 1998. Automatic question generation for decision tree based state tying. ICASSP’98.

Cao Y, Zhang S W, Huang T Y *et al*. 2004. Tone modeling for continuous mandarin speech recognition. International Journal of Speech Technology, 7:115-128.

Chen, K, Johnson M H, Cohen A *et al*. 2006. Prosody dependent speech recognition on radio news corpus of American English. IEEE Transactions on Audio, Speech and Language Processing, 14(1): 232-245.

Chen T, Huang C, Chang E, *et al*. 2001. Automatic accent identification using Gaussian mixture model. IEEE workshop on ASRU’2001.

Chen Y J, Wu C H, Chiu Y H, *et al.* 2002. Generation of robust phonetic set and decision tree for Mandarin using chi-square testing. Speech Communication, 38: 349-364.

Chen X X, Li A J, et al. 2000. An application of SAMPA-C for standard Chinese. International Conference on Spoken Language Processing (ICSLP'2000), 4:652-655.

Cheng S S, Xu Y Y, Wang H M et al. 2006. Automatic construction of regression class tree for MLLR via model-based hierarchical clustering. ISCSLP'2006, Sigapore.

Chien J T. 2005. Decision tree state tying using cluster validity criteria. IEEE Tranactions on Speech and Audio Processing, 13(2):182-193.

Davis S B, Mermelstein P. 1980. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustic, Speech and Signal Processing, 28(4):357-366.

Decker A M, Lamel L. 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication, 29: 83-98.

Dong H H，Tao J H，Xu B. 2005. Chinese prosodic phrasing with a constraint-based approach. Interspeech'05, Lisbon.

Doumpiotis V, Byrne W. 2006. Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition. Speech Communication, 48:142–160.

Fung P, Liu Y. 2005. Effects and modeling of phonetic and acoustic confusions in accented speech. Journal of Acoustic Society of America. 118(5): 3279-3293.

Gauvain, J L, Lee, C H. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, 2(2): 291-299.

Goronzy S, Marina S, Wolfgang W. 2001. Is non-native pronunciation modelling necessary? Eurospeech'2001, Aalborg, Denmark.

Goronzy S, Rapp S, Kompe R. 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Communication 42:109-123.

Gouws E, Wolvaardt K, Kleynhans N *et al*. 2004. Appropriate baseline values for HMM-based speech recognition.
http://www.csir.co.za/websource/ptl0002/docs/HLT/gouwse04HMMbaselines.pdf

Gruhn R, Markov K, Nakamura S. 2004. A statistical lexicon for non-native speech recognition. ICSLP'04.

Hain T. 2005. Implicit modelling of pronunciation variation in automatic speech recognition. Speech Communication, 46:171-188.

He X D, Zhao Y X. 2007. Prior knowledge guided maximum expected likelihood based model selection and adaptation for nonnative speech recognition. Computer Speech and Language, 21:247-265.

Holter T, Svendsen T. 1999. Maximum likelihood modelling of pronunciation variation. Speech Communication，29:177-191.

Hoste V, Daelemans W, Gillis S. 2004. Using rule-induction techniques to model pronunciation variation in Dutch. Computer Speech and Language, 18:1-23.

Huang C, Chen T, Chang E. 2004. Accent issue in large vocabulary continuous speech recognition", International Journal of Speech Technology, 7:141-153.

Huang R, Hansen J H L. 2005. Dialect/accent classification via boosted word modeling. ICASSP'2005, 1:585-588.

Huang X D, Acero A, Hon H W. 2001. Spoken language processing: A guide to theory, algorithm and system development. Prentice Hall.

Humphriesy J J, Woodland P C, Pearcez D. 1996. Using accent-specific pronunciation modeling for robust speech recognition. ICSLP'96.

Hwang M Y, Huang X D, Alleva F A. 1993. Predicting unseen triphones with senones. ICASSP'93, 311~314.

Hwang M Y, Huang X D, Alleva, F A. 1996. Predicting unseen triphones with Senones, IEEE Transactions on Speech and Audio Processing, 4(6):412-419.

Kam P, Lee T, Soong F K. 2003. Modeling Cantonese pronunciation variation by acoustic model refinement. EuroSpeech, 2003, Geneva.

Kessens J M, Cucchiarini C, Strik H. 2003. A data-driven method for modeling pronunciation variation. Speech Communication, 40:517-534.

Kosaka T, Matsunaga S, Sagayama S. 1996. Speaker-independent speech recognition based on tree-structured speaker clustering, Computer Speech and Language, 10:55-74.

Lee T，Lau W，Wong Y W *et al*. 2002. Using tone information in Cantonese continuous speech recognition. ACM Transactions on Asian Language Information Processing, 1(1):83-102.

Leggetter C J, Woodland P C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer Speech and Language, 9:171-185.

Li A J，Wang X. 2003. A Contrastive Investigation of Standard Mandarin and Accented Mandarin. Eurospeech'03, Geneva.

Li A J, Fang Q, Xu R Y, *et al*. 2005. A contrastive study between Minnan-accented Chinese and standard Chinese, O-COCOSDA.

Li J, Zheng F, Zhang J Y *et al*. 2001. The definition and extension of the question set for decision tree based state tying in Chinese speech recognition. International Conference on Chinese Computing, 106-110, Singapore.

Li J, Zheng T F, Xiong Z Y, *et al*. 2003. Construction of large-scale Shanghai Putonghua speech corpus for Chinese speech recognition. Oriental-COCOSDA, 62-69, Singapore.

Li J, Zheng T F, Byrne B *et al*. 2006, A dialectal Chinese speech recognition framework", Journal of Computer Science and Technology, 21(1): 106-115.

Livescu K. 1999. Analysis and modeling of non-native speech for automatic speech recognition. Master thesis, Massachusetts Institute of Technology.

Livescu K, Glass J. 2000. Lexical modeling of non-native speech for automatic speech recognition. ICASSP'00, Istanbul.

Lim B P, Li H-Z, Ma B. 2005. Using local and global phonotactic features in Chinese dialect identification. ICASSP'2005, 1:577-580.

Liu C J, Yan Y H. 2004a. Robust state clustering using phonetic decision trees, Speech Communication, 42:391-408.

Liu L Q, Zheng T F, Wu W H. 2006a. Automatic Initial/Final generation for dialectal Chinese speech recognition. ICSLP'06, Pittsburg.

Liu L Q, Zheng T F, Wu W H. 2006b. State-dependent phoneme-based model merging for dialectal Chinese speech recognition. ISCSLP '2006, Singapore.

Liu Y, Fung P. 2003. Automatic phone set extension with confidence measure for spontaneous speech, Eurospeech'2003, Geneva.

Liu Y，Fung P. 2004. Pronunciation Modeling for Spontaneous Mandarin Speech Recognition. International Journal of Speech Technology, 7:155-172.

Liu Y, Fung P. 2005. Acoustic and phonetic confusions in accented speech recognition. Interspeech'05, 3033-3036.

Lussier E F, Williams G. 1999. Not just what, but also when: Guided automatic pronunciation modeling for broadcast news. DARPA Broadcast News Workshop, Herndon, Virginia.

Lussier E F. 2003. A tutorial on pronunciation modeling for large vocabulary speech recognition. Lecture Notes in Computer Science, 2705, 38-77.

Ma B, Zhu D L, Tong R. 2006. Chinese dialect indentificaiton using tone features baesd on pitch flux. ICASSP'2006, 1029-1032.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium, 1:281-297.

Milone D H, Rubio A J. 2003. Prosodic and Accentual Information for Automatic Speech Recognition, IEEE Transactions on Speech and Audio Processing. 11(4): 321-333.

NIST. 1997. The 1997 Hub-4NE evaluation plan for recognition of Broadcast News, in Spanish and Mandarin.

http://www.nist.gov/speech/tests/bnr/hub4ne_97/current_plan.htm.

Oh Y R, Yoon J S, Kim H K. 2007. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. Speech Communication, 49: 59-70.

Pan F P, Zhao Q W, Yan Y H. 2006. Improvements in tone pronunciation scoring for strongly accented mandarin speech. ISCSLP'06, Sigapore.

Peng G, Wong W S Y. 2004. An innovative prosody modeling method for Chinese speech recognition. International Journal of Speech Technology, 7: 129-140.

Raux A. 2004. Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition. Interspeech'2004, 613-616.

Reichl W, Chou W. 2000. Robust decision tree state tying for continuous speech recognition. IEEE Transactions of Speech and Audio Processing, 8(5): 555-566.

Riley M. 1991. A statistical model for generating pronunciation networks. IEEE ICASSP'91, 737–740.

Rissanen J. 1984. Universal coding, information, prediction, and estimation. IEEE Transactions of Information Theory, 30:629-636.

Saraclar M, Nock H, Khudanpur S. 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language, 14:137–160.

Schlüter R, Macherey W, Müller B. *et al*. 2001. Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication, 34(3): 287-310.

Simonin J，Bodin S, Jouvet D *et al*. 1996. Parameter tying for flexible speech recognition. ICSLP'96.

Sproat R, Zheng F, Gu L, *et al*. 2004. Dialectal Chinese speech recognition: Final Report. CLSP Summer Workshop. 1-82.

Stephenson T A, Doss M M, Bourlard H. 2004. Speech recognition with auxiliary information. IEEE Transactions on Speech and Audio Processing, 12(3):189-202.

Stolcke A, Bratt H. Butzberger J. *et al.* 2000. The SRI March 2000 Hub-5 conversational speech transcription system. NIST Speech Transcription Workshop, College Park, Maryland.

Strik H, Cucchiarini C. 1999. Modeling pronunciation variation for ASR: A survey of the literature, Speech Communication，29: 225-246.

Tang M. 2005. Large Vocabulary Continuous Speech Recognition Using Linguistic Features and Constraints [Doctoral thesis]. MIT.

Tobias C, Rainer G, Satoshi N. 2005. Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models. Interspeech'2005, 1509-1512.

Tomokiyo L M. 2001. Recognizing non-native speech: characterizing and adapting to non-native usage in LVCSR. Ph.D. Thesis, Carnegie Mellon University, USA.

Tsai M Y，Lee L S. 2003. Pronunciation variation analysis based on acoustic and phonemic distance measures with application examples on Mandarian Chinese. ASRU '03, 117-121.

Tsai M Y, Chou F C, Lee L S. 2007. Pronunciation modeling with reduced confusion for Mandarin Chinese using a three-stage framework. IEEE Transactions on Audio, Speech and Language Processing, 15(2): 661-675.

Vergyri D, Stolcke A, Gadde V *et al*. 2003. Prosodic knowledge sources for automatic speech recognition. IEEE ICASSP'03, Hong Kong.

Viikki O, Laurila K. 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, 25:133-147.

Wang Z R, Schultz T, Waibel A. 2003. Comparison of acoustic model adaptation techniques on non-native speech. ICASSP, 540-543.

Wester M. 2003. Pronunciation modeling for ASR–knowledge-based and data-rerived methods", Computer Speech and Language, 17:69-85.

Williams G. Renals S. 1998. Confidence measures for evaluating pronunciation models. Proc. ESCA Workshop Model on Pronunciation Variation Automatic Speech Recognition, 151–155.

Witt S, Young S. 1999. Offline acoustic modeling of non-native accents. Eurospeech'99.

Woodland P C，Povey D. 2002. Large scale discriminative training of hidden Markov models for speech recognition. Computer Speech and Language, 16: 25-47.

Workshop. 2004. http://www.clsp.jhu.edu/ws2004/groups/ws04casr/.

Wu X T, Yan Y H. 2004. Speaker adaptation using constrained transformation. IEEE Trans on Speech and Audio Processing, 12(2):168-174.

Wu Y J, Kawai H, Ni J F *et al*. 2005. Discriminative training and explicit duration modeling for HMM-based automatic segmentation. Speech Communication, 47: 397-410.

Xu R, Wunsch D. 2005. Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3):645-678.

Young S, Odell J J, Woodland P C. 1994. Tree-based state tying for high accuracy acoustic modeling. Proc of ARPA Human Language Tech Workshop, 307-312.

Young S, Evermann G, Hain T, *et al*. 2002. The HTK Book (for HTK Version 3.2.1). Cambridge University.

Yuan J H, Brenier J M, Jurafsky D. 2005. Pitch accent prediction: Effects of genre and speaker. Interspeech'2005, 1409-1412.

Zhang C, Wang J, Xiao X *et al*. 2006. Pronunciation variation modeling for Mandarin with accent, ICSLP'06.

Zheng F, Song Z J, Fung P *et al*. 2001. Modeling pronunciation variation using context-dependent weighting and B/S refined acoustic modelling. In: EuroSpeech' 01, 1:57-60.

Zheng F, Song Z J, Fung P, *et al*. 2002. Mandarin pronunciation modeling based on CASS corpus. Journal of Computer Science & Technology, 17(3): 249-263.

Zheng Y L, Sproat R, Gu L, *et al*. 2005. Accent detection and speech recognition for Shanghai-accented Mandarin. Interspeech'2005, 217-220.

Zhou J L, Tian Y, Shi Y *et al*. 2004. Tone articulation modeling for mandarin spontaneous speech recognition. IEEE ICASSP'04, Montreal.

Zhou K S H, Chellappa R. 2004. Kullback-Leibler distance between two Gaussian densities in reproducing kernel Hilbert space. ISIT'2004, Chicago.

附录 A　基于少量上海普通话的 SDPBMM

A.1　引言

从正文的结果可以知道，基于小数据量的 SDPBMM 能很好的解决标准普通话和方言背景普通话之间的声音变化问题，能显著的提高方言背景普通话识别率，同时不降低标准普通话的识别率，在此附录中，本文将进一步对 SDPBMM 进行的验证。这里将采用与正文中不同的标准普通话识别器，结合更少的上海普通话数据（40 分钟）来实现 SDPBMM。在此基础上，将对 SDPBMM 进行系统的测试。另外将和以往的结果（李净，2005）对比，以进一步证明方法的有效性。

A.2　标准普通话与上海普通话

本节中采用的标准普通话数据来源于约翰·霍普金斯大学 CLSP 研究中心提供的 MBN（Mandarin Broadcast News）（HUB-4，1997）数据库，其中包含 30 个小时高质量宽带语音数据和详细的声韵标注信息。上海普通话来自于“吴方言背景普通话数据库 WDC”（Li，2003；Xiong，2003），此数据库包括 100 个说话人，50 个男声，50 个女声，都出生于上海，且在上海居住、生活，其父母来自上海及其它吴方言区。在 2004 年夏天，在约翰·霍普金斯大学举行的研讨会（Workshop，2004），其中一个主要研究课题，“吴方言背景普通话研究”，就采用了这两个数据库。本文的数据库划分如表 A.1 所示：

与正文的数据划分方式相同，本附录中也采用了四个互不相交的数据集，即标准普通话训练集（TRAIN_MBN）、标准普通话测试集（TEST_MBN）、上海普通话开发集（DEV_SH）和上海普通话测试集（TEST_SH）。分别来自于数据库 MBN 和 WDC。表 A.1 中，数据集 TRAIN_MBN、TEST_MBN 和 TEST_SH 与文（Sproat，2004；李净，2005）同功能的数据集完全相同，即采用了与以往研究工作中相同的标准普通话识别器。不同的是，本文采用了更小的上海普通话开发数据集，DEV_SH，而文（Sproat，2004）采用了 6.3 个小时的开发集，文（李净，2005）采用了 1.4 小时的开发集。

表A.1　标准普通话与上海普通话数据集的划分

名　称	用　途	内　容
TRAIN_MBN	标准普通话训练集（MBN）	约 30 小时，共 34,493 句子
TEST_MBN	标准普通话测试集（MBN）	约 1.5 小时，共 1,260 句子
DEV_SH	上海普通话开发集（WDC）	10 人，男女均衡，共 510 句，约 40 分钟，中度和重度口音
TEST_SH	上海普通话测试集（WDC）	20 人，男女均衡，共 995 句，约 1 小时，中度和重度口音

在基于 TRAIN_MBN 的标准普通话声学建模过程中，声学特征为 39 维的 MFCC，包括 13 维倒谱特征，以及一阶差分和二阶差分, 并且利用倒谱均值归一化方法(CMN) 来对特征进行归一化。使用 HTK3.2 进行上下文相关基于声韵母的 HMM 建模，每个 HMM 采用从左到右，无回跳的拓扑结构，另外，每个 HMM 包含 3 个有效状态。采用决策树进行模型状态共享。初始的标准普通话的识别基元集合由 65 个声韵母组成(包含 6 个零声母)，即 XIF 集合。这样建立起来的模型将作为标准普通话的声学模型。其包含 3,230 个物理状态，每个状态包括 14 个高斯混合，本附录将此标准普通话普通话模型作为基线系统 AM0。识别字典包含 406 个无调音节。实验中暂时没有使用语言模型，其目的就是在没有语言知识介入情况下，考察声学建模方法的效果，因而本节以音节错误率 SER 作为识别性能衡量标准。此标准普通话识别器对于 TEST_MBN 和 TEST_SH 的测试结果如表 A.2 所示，将此结果作为本节的基线系统的性能。

从表 A.2 可以看到，同样采用标准普通话识别器，标准普通话和上海普通话之间在音节错误率上存在 19.3%的差异。由于数据库 MBN 和 WDC 之间的信道差异很小，因而可以认为这个差距主要来自于方言口音，文（Huang，2004）理论上证明口音是造成说话人差异的最主要因素。

表A.2 标准普通话识别器对于TEST_MBN和TEST_SH的识别结果

声学模型	SER	
	TEST_MBN	TEST_SH
标准普通话	30.5%	49.8%

同样，利用上海普通话数据集 DEV_SH 训练了上下文无关的声韵母模型 mono-XIF，采用了标准普通话模型相同的特征参数以及状态拓扑结构，每个 mono-XIF 状态包含 6 个高斯混合。其对 TEST_MBN 和 TEST_SH 的识别结果如表 A.3 所示：

表A.3 上海普通话上下文无关声学模型对于TEST_MBN和TEST_SH的识别结果

声学模型	SER	
	TEST_MBN	TEST_SH
上海普通话	78.4%	54.1%

可见基于这么少量的方言背景普通话数据建立的声学模型，对于标准普通话和方言背景普通话识别率都极不理想。

A.3　SDPBMM的实现

在 SDPBMM 中，首先是两个归并准则，即基于状态的发音变化和插值因子的确定。对于基于状态的发音变化同样用基于声韵母的发音变化来近似，利用基于距离度量的识别网络生成策略来解决数据稀疏问题。表 A.3 列出了包含两个以上发音变化的声韵母列表，与语音专家给出的发音变化（Li，2003）基本是一致的。

在获得发音变化的基础上，通过实验对比的方式确定插值因子λ，本节中λ=0.72。与正文中的基于闽南普通话的 SDPBMM 中λ=0.30 相比，可见插值因子的确定与具体的数据有关系，也与具体的识别任务有关系。

表A.3 标准普通话和上海普通话声韵母之间的发音变化

标准普通话	概率	上海普通话	标准普通话	概率	上海普通话
c	0.736	c	ch	0.546	c
	0.264	ch		0.454	ch
en	0.572	en	eng	0.564	eng
	0.428	eng		0.436	en
ia	0.731	ia	ie	0.797	ie
	0.269	iang		0.203	ve
ii	0.764	ii	iii	0.571	ii
	0.236	iii		0.429	iii
in	0.432	in	ing	0.333	in
	0.568	ing		0.667	ing
s	0.729	s	sh	0.477	sh
	0.271	sh		0.523	s
uan	0.753	uan	zh	0.624	z
	0.247	an		0.376	zh
ve	0.750	ve	r	0.716	r
	0.250	ie		0.284	l

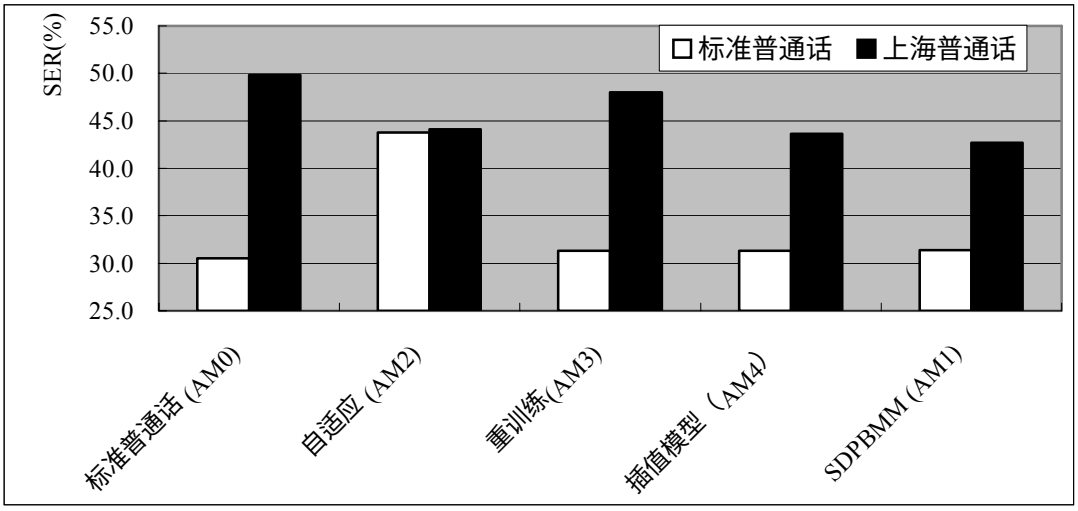
A.4　有关SDPBMM的测试与结果

针对 SDPBMM，本节设计了一系列的测试，首先是与重训练、自适应、模型差值的对比；其次是与自适应和语言模型的结合；最后是多种建模方法的综合以及与文（李净，2005）中的结果对比。

A.4.1 SDPBMM与其它建模方法对比

利用同样的上海普通话开发数据集 DEV_SH，本文对重训练（AM3）、

自适应（AM2）、模型差值（AM4）和 SDPBMM（AM1）进行对比。根据正文节 4.4 关于自适应方法的对比实验，考虑到上海普通话数据量的限制，MLLR 的效果优于 MAP，这里的自适应方法采用 MLLR；重训练方法是基于数据混合方式的，即 TRAIN_MBN 和 DEV_SH 的混合；模型插值中，插值因子λ=0.72。表 A.1 列出了在 TEST_MBN 和 TEST_SH 上的识别结果，其中最左侧代表标准普通话声学模型 AM0。



图A.1 声学建模方法对比

从图 A.1 可以得出与正文节 4.4 中基于闽南普通话的 SDPBMM 得出相似的结论。

1. 相比于标准普通话声学模型 AM0，其它几种建模方法都能提高上海普通话的识别率，其中重训练只能小幅的提高上海普通话的识别率，效果不太显著，效果最显著的仍然是 SDPBMM，音节错误率由 49.8%降为 42.7%，相对音节错误率降低了 14.3%。相比于自适应方法和模型插值方法，音节错误率分别多降低了 1.4%和 0.9%（绝对值）。

2. 相比于标准普通话声学模型，在对标准普通话的识别中，自适应方法造成标准普通话识别率的显著下降，这与实际经验是一致的。重训练方法方法对于标准普通话的识别率没有明显下降，这主要是因为上海普通话的数据量很小，没有对标准普通话识别造成副作用。模型插值和 SDPBMM 也没有造成标准普通话识别率的降低。

A.4.2 SDPBMM与自适应和语言模型的结合

首先，SDPBMM 与自适应方法进行了结合，采用了 MLLR 自适应方法，这里分别对两种不同的建模组合进行了对比，即 SDPBMM 的基础上进行 MLLR 自适应（SDPBMM+MLLR，AM5）和 MLLR 自适应之后再进行 SDPBMM（MLLR+SDPBMM，AM6）。由于采用了自适应策略，其对于标准普通话的识别率势必下降，所以本小节只在上海普通话上进行了测试，结果列于表 A.4。

表A.4 SDPBMM与自适应方法相结合的测试

测试集	**SER**	
	SDPBMM+MLLR（AM5）	**MLLR+SDPBMM（AM6）**
TEST_SH	40.8%	41.8%

从表 A.4 可以知道，SDPBMM 与 MLLR 可以很好的结合，这因为 SDPBMM 主要解决是语音学差异中的声音变化问题，而自适应不仅可以减小语音学差异还可以减少信道差异，因此它们的结合能进一步提高识别率。实验结果表明，无论哪种建模顺序，都可以有效的提高方言背景普通话的识别率。相比之下，SDPBMM+MLLR 的效果更显著，因此 SDPBMM 更适合作为自适应方法的前端处理方法。

其次，SDPBMM 与语言模型进行结合，这里采用了（李净，2005）第 5 章中的语言模型，采用了标准普通话发音字典，以字符错误率 CER 作为识别率评价标准，表 A.4 列出了测试结果。

表A.4 SDPBMM与语言模型相结合的测试

识别器	**CER**	
	TEST_MBN	**TEST_SH**
标准普通话（AM0）+LM	39.4%	61.8%
SDPBMM（AM1）+ LM	38.7%（**-0.7%**）	53.1%（**-8.7%**）

可以看出，SDPBMM 能很好地同语言模型相结合，得到了与前面实验类似的结果，对于上海普通话，相对字错误率降低了 14.1%；对于标准普通话，相对字错误率降低了 1.8%。

A.4.3 综合性能

在本实验中，为了同以往的结果进行对比，本文在声学建模过程中将 SDPBMM、基于 MLLR 的自适应和语音模型相结合，构造一个完整的上海普通话识别器，作为同文（李净，2005）中方法对比的平台。文（李净，2005）中声学建模中主要采用了基于专家知识的扩展基元和基于 MLLR 的自适应（参见图 1.1）。这里统一采用了文（李净，2005）中提出的基于累积一元概率 AUP 的上海普通话发音字典；同一个数据集 DEV_SH 作为开发集。对于 TEST_SH 的识别结果如表 A.5 所示：

表A.5 综合性能对比

测试集	**CER**	
	SDPBMM+MLLR+LM（本文）	**扩展基元+MLLR+LM（李净，2005）**
TEST_SH	49.8%	52.4%

可以看到，在同样的数据量下，本文的方法要优于文（李净，2005）中的方法，相对字错误率多降低了 5.0%。在下一步的研究中，对于方言背景普通话语音识别，在不降低标准普通话识别率的前提下，也将引入扩展的识别基元。

A.5 总结

通过在基于不同的标准普通话识别器，针对不同的方言背景普通话（闽南普通话、上海普通话）的测试表明，SDPBMM 是一种简单而有效的声学建模方法，同以往方法相比，在采用少量的方言背景普通话数据的情况下，对方言背景普通话取得了更加显著的识别效果，相对音节错误率降低了 14.3%，相对字错误率降低了 14.1%。更重要的是，SDPBMM 没有

造成标准普通话识别率的下降。

A.6　参考文献

李净. 2005. 吴方言背景普通话语音识别研究[博士学位论文]. 清华大学计算机科学与技术系.

Li A J，Wang X. 2003. A Contrastive Investigation of Standard Mandarin and Accented Mandarin. Eurospeech'03, Geneva.

Li J, Zheng T F, Xiong Z Y, *et al*. 2003. Construction of large-scale Shanghai Putonghua speech corpus for Chinese speech recognition. Oriental-COCOSDA, 62-69, Singapore.

HUB-4. 1997. http://www.nist.gov/speech/publications/darpa97/pdf/stern1.pdf.

Sproat R, Zheng F, Gu L, *et al*. 2004. Dialectal Chinese speech recognition: Final Report. CLSP Summer Workshop. 1-82.

Xiong Z Y, Zheng F, Li J, *et al*. 2003. An automatic prompting texts selecting algorithm for di-IFs balanced speech corpus. NCMMSC7, 252-256, Xiamen.

致　谢

衷心感谢我的导师吴文虎教授和郑方研究员五年来对我的悉心指导和关怀。吴文虎教授渊博的专业知识、循循善诱的教导方式和平易近人的待人原则，使我受益匪浅；郑方教授敏捷的思维、严谨求实的科学作风、一丝不苟的学术态度，更是我在科研工作中的榜样和镜子。总之，两位导师无论在学业上还是在生活上都给予了我莫大的关心和帮助。在此，谨向两位恩师致以最诚挚的谢意！

感谢语音技术中心的其它老师，包括方棣棠教授、李树青教授、徐明星副教授、邬晓钧老师，以及实验室的李净、吴根清、熊振宇、孙辉、邓菁、刘建、扈浩、潘玉春等同学，他们与我进行了许多有益的讨论，同时给予我很多工作上的支持，在此一并向他们表示感谢。

最后，衷心感谢我的父母，他们无私的爱和默默的关怀，一直伴随着我的奋斗历程。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1974 年 5 月出生于河北省阜平县。

1992 年 9 月考入解放军（南京）通信工程学院计算机系，1996 年 7 月毕业并获得工学学士学位。

1996 年 9 月进入解放军（南京）通信工程学院计算机系攻读硕士，1999 年 4 月毕业并获得工学硕士学位。

1999 年 5 月至 2002 年 8 月，解放军（武汉）通信指挥学院任教。

2002 年 9 月考入清华大学计算机系攻读博士至今。

发表的学术论文

[1] **Linquan Liu**, Thomas Fang Zheng, Wenhu Wu. State-Dependent Phoneme-Based Model Merging for Dialectal Chinese Speech Recognition, *ISCSLP*, Singapore, 2006. （Also collected by *Lecture Notes in Artificial Intelligence*, 4274, pp. 282-293, 2006. ）（**SCI 检索，**Accession Number: 9242075）

[2] **Linquan, Liu**, Thomas Fang Zheng, Wenhu Wu. English Alphabet Recognition Based on Chinese Acoustic Modeling, *ISCSLP*, Singapore, 2006.

[3] **Linquan, Liu**, Thomas Fang Zheng, Wenhu Wu. Automatic Initial/Final Generation for Dialectal Chinese Speech Recognition, *ICSLP*, Pittsburgh, 2006. （**EI** 检索）

被录用的学术论文

[1] **刘林泉，**郑方，吴文虎. 基于小数据量的方言背景普通话语音识别声学建模研究.（已被**清华大学学报**录用，**EI** 检索源刊）

[2] Linquan Liu, Thomas Fang Zheng, Makoto Akabane, Ruxin Chen,Wenhu Wu. Using a Small Development Data Set to Build a Robust Dialectal Chinese Speech Recognizer, *Interspeech***,** Antwerp**, 2007.** （**EI** 检索）

其他学术论文

[1] Yuchun Pan, Mingxing Xu, **Linquan Liu**, Peifa Jia. Emotion-detecting Based Model Selection for Emotional Speech Recognition, *Computational Engineering in Systems Applications* (*CESA2006*), Beijing, 2006. （**EI** 检索）

在读期间完成的其它研发工作

1. 基于汉语标准普通话语音识别的声学模型研究项目—设计、组织完成建立大型标准普通话语料库和建立鲁棒的声学模型，此项目由韩国 Wonkwang 大学资助。
2. 基于 SONY PlayStation2®的标准普通话语音识别—主要完成人，主要解决基于 SONY PlayStation2®的语音识别中如何建立简小而鲁棒的声学模型。此项目由日本索尼计算机娱乐公司资助。
3. 基于 SONY PlayStation2®的带方言背景普通话的语音识别—主要完成人，目标是在 SONY PlayStation2®的游戏平台中，显著提高方言背景普通话的识别率，同时保证标准普通话识别率不降低。此项目由日本索尼计算机娱乐公司资助。