吴方言背景普通话语音识别研究

李　净

Research on Wu Dialectal Chinese Speech Recognition

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

by

Li Jing

(Computer Science and Technology)

Dissertation Supervisor: Professor Wu Wenhu

Associate Supervisor: Professor Zheng Fang

November, 2005

摘　要

本文研究方言背景普通话语音识别。方言背景是汉语连续语音识别应用中必须面对的问题。本文以吴方言为例，研究如何利用方言知识，以较低成本由标准普通话识别器构建方言背景普通话识别器的方法，并且该方法可以方便地从一种方言背景扩展到其它方言背景。本文贡献如下：

（1）**提出一种可扩展的方言背景普通话语音识别框架。**一方面，汉语方言众多，对每种方言背景普通话构建识别器成本太高；另一方面，方言背景普通话受标准普通话和方言背景的双重影响，还具有从方言背景出发向标准普通话过渡的动态特性，故而为每种方言背景普通话单独构建识别器难以收到令人满意的效果。本文提出一种框架，利用少量数据和方言知识，从声学模型、发音词典、语言模型、解码器四个层面，将标准普通话识别器变换为方言背景普通话识别器。此框架整体性强、成本低、易扩展。

（2）**提出抑制高频单元（RHF）和鼓励低频单元（ELF）两种均衡语料选择算法并加以对比。**算法用于解决语料库设计中的单元覆盖率和均衡度问题：RHF 算法可以有效地保证各个单元的均衡度；ELF 算法则显著地提高了低频单元的覆盖率。二者相比，ELF 更适用于训练用语料库设计。

（3）**提出基于扩展声韵母的上下文相关声学建模方法和模型优化策略。**与标准声韵母相比，扩展声韵母规范了声学建模中音节的声韵结构，使上下文相关的声韵母三元组数目降低 75%，音节错误率降低 12.84%。在吴方言背景普通话识别中，对自然发音和朗读式语音，其字错误率分别降低 4.11%和 9.36%。

（4）**提出基于 surface-form/base-form 的声学自适应和基于声韵映射规则的多发音词典生成方法，用以描述方言背景对声学层的影响。**声韵母映射规则是从专家知识和少量数据中获取的。实验表明，使用约 1 小时语音数据和相关知识，即可将自然发音和朗读式语音字错误率降低 16.32%和 36.58%。

（5）**提出了基于累积一元概率(AUP)的多发音剪枝准则。**引入多发音词典对方言背景建模是必要的，但过多的发音入口（Entry）会增加混淆和解码开销。基于 AUP 准则，仅对高频词进行多发音扩展，可在几乎不降低识别率的前提下，将词表长度降低一半以上，使其长度仅为单发音词典的 1.2 倍。

关键词：方言背景普通话；连续语音识别；发音变化建模；语料库设计

Abstract

This work primarily concentrates on dialectal Chinese speech recognition, which is almost unavoidable for LVCSR(Large Vocbulary Continuous Speech Recognition). Taking the Wu-dialect as the target language, we attempt to establish a WDC speech recognizer from an available PTH speech recognizer, based on the Initial-Final structure of the Chinese language, in combination with dialect-specific linguistic knowledge. Moreover, it's shown that the proposed framework can be transferred with little effort to other languages as well as other dialects. The contributions of this work are as follows:

1. A highly scalable framework for dialectal Chinese speech recognition. In China there exist many dialects, such that it's impossible to build a dialect-specific recognizer for each dialect due to the huge cost and the time consumption. On the other hand, with the popularity of Putonghua, there are many words imported from Putonghua into the various dialects. The dialectal Chinese is affected by both of the Putonghua and the dialect. Therefore, it's practical to build dialectal Chinese recognizers on the basis of readily available Putonghua recognizers. Motivated by such an assumption, we come up with a general framework for dialectal Chinese recognition. Within the framework, it's shown how a Putonghua speech recognizer can be transformed into a dialectal Chinese recognizer via acoustic modeling, pronunciation lexicon modeling, language modeling and a decoder. The proposed framework is of simple architecture, low cost, high scalability and easy deployment.

2. Two automatic speech corpus selection algorithms called RHF (Restricting High-Frequency units) and ELF (Encouraging Low-Frequency units). Two algorithms are proposed and compared. The ELF is much better than the RHF method for designing a corpus for training purposes, for it is guaranteed that most low-frequency units will appear a certain number of times.

3. XIF(eXtented Initial/Finals) based context-dependent modeling and optimization method. In contrast to the standard IFs, the XIF take the combination of Initial and Final into account; what's more, it regularizes these combinations so that each Chinese syllable is composed of an Initial and a Final. In doing so, under the context-dependent scenario, the number of triphones is reduced significantly, from approximately 120k to 30k, while at the same time the improvement in performance is achieved effectively. In Putonghua recognition, the adoption of XIFs could lead to 12.84% in SER reduction. Accordingly, in Wu-dialectal Chinese recognition, 4.11% and 9.36% WER were obtained for the spontaneous and the read-style speech respectively.

4. Base-form/surface-form based AM adaptation for WDC recognition and a multi-pronunciation lexicon generating algorithm based on IF-Mapping Rules. The IF-Mapping rules are collected from expert knowledge and a small speech corpus. Combined with an IF-Mapping rules based multi-pronunciation lexicon, acoustic adaptation can primely solve the pronunciation variation on the acoustic level. Totally, 16.32% and 36.58% CER(Character Error Rate) reduction can be obtained.

5. Multi-Pronunciation Expansion (MPE) based on Accumulated Uni-gram. More pronunciations help model pronunciation variations, but also lead to more confusion; there must be some restrictions over the added variability. To implement this, we take the accumulated uni-gram probability (AUP) as our criterion. That is to say, the criterion can determine the multi-pronunciation according to its Uni-gram probability. However, for the ones lower than the predefined threshold, no pronunciation modeling is applied. In terms of the experiments, the scale of the lexicon is halved without apparent degradation in performance. On average, each word consists of 1.2 pronunciation entries.

Key words: dialectal Chinese speech; continuous speech recognition; pronunciation variation modeling; speech corpus design

目 录

主要符号对照表

ANN	人工神经元网络 (Artificial Neural Networks)
ASR	自动语音识别 (Automatic Speech Recognition)
AUP	累积一元概率 (Accumulated Uni-gram Probability)
CDCPM	中心距离连续概率模型 (Center-Distance Continuous Probability Model)
CER	字错误率 (Character Error Rate)
CMN	倒谱均值归一化 (Cepstrum Mean Normalization)
ELF	鼓励低频单元 (Encouraging Low-Frequency units)
FA	强制对准 (Forced Alignment)
GMM	高斯混合模型 (Gaussian Mixture Model)
HMM	隐马尔可夫模型 (Hidden Markov Model)
HZ	汉字 (Hanzi)
IF	汉语声母或韵母 (Initial/Final)
IPA	国际音标符号集合 (International Phonetic Alphabets)
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 倒谱参数 (Mel Frequency Cepstrum Coefficient)
MLLR	最大似然线形回归 (Maximum Likelihood Linear Regression)
NLP	自然语言处理 (Natural Language Processing)
PM	发音建模 (Pronunciation Modeling)
PTH	标准普通话，简称为普通话 (PUTONGHUA)
PTH-IF	标准普通话声母或韵母
PY	拼音 (Pinyin)
RHF	抑制高频单元 (Restricting High-Frequency units)
SAMPA	国际上通行的可机读语音键盘符号系统 (Speech Assessment

	Method Phonetic Alphabets)
SAMPA-C	在 SAMPA 基础上制定的针对汉语普通话的语音标注符号系统
SER	音节错误率（Syllable Error Rate）
WDC	吴方言背景普通话，指生活在吴方言区，以吴方言为母语的人们所说的普通话 (Wu-dialectal Chinese)
WDC-IF	吴方言背景普通话声母或韵母
WDC-XIF	吴方言背景普通话扩展声韵母
Wu-IF	吴方言声韵母
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

本文研究吴方言背景普通话语音识别研究方法。其中“吴方言”，或称为“吴语”，是语言学中的名词，指的是吴语区的方言，而吴语区主要包括现在的上海、苏州、杭州、宁波等地。“吴方言背景普通话”（WDC，Wu-dialectal Chinese）指的是生活在吴方言区，并以吴方言为母语的人们所说的普通话，即，受吴方言背景影响而带有吴方言口音的普通话。具体地说，本文研究的是，非特定人、大词表、吴方言背景普通话连续语音识别，属于通常所说的大词表连续语音识别（LVCSR, Large Vocabulary Continuous Speech Recognition）。

本文在实验中所使用的方言背景语音数据是上海普通话，但本文题目仍采用“吴方言”，有如下两个原因。首先，吴语区中不同的方言虽然存在一定的差异，但也有很多共同之处，尤其是在说普通话时，存在很多共同的规律。因此，本文中的许多工作，例如基元选择，常用吴方言词表选取等，都是基于整个吴语区的。并且，本文研究的是一种可扩展的方言背景普通话识别框架，虽然某些具体操作与特定的方言相关，但方法本身可以方便地应用于其他方言，对于同一方言区（如吴语区）内的方言更是如此。其次，我们知道，上海是由一个小渔村发展而来的，约 90%的人口是从周边地区（主要是吴语区）迁移过来的[1][2]。因此，上海方言可以认为是吴方言交汇的结果，更有语言学专家称上海话为“吴方言中的普通话”，这也表明上海话与吴方言的密切关系。基于以上原因，本文题目中采用“吴方言”的词语。

近二十年来，随着计算机技术的飞速发展和隐马尔可夫模型（HMM，Hidden Markov Model）[3]的广泛应用，语音识别技术得到了极大的推动。自动语音识别（ASR，Automatic Speech Recognition）的研究重点也从特定人、小词表、孤立词语音识别向非特定人、大词表、连续语音识别转移。一些著名的研究机构和公司，如，英国剑桥大学[4]，IBM 公司[5]，CMU 大学[6]，微软公司[7]，贝尔实验室[8]等，都建立了自己的语音识别系统，并努力将其实用化。很多文章称，在经过适当的自适应后，其大词表连续语音识别系统对特定说话人的识别率可以达到 95%以上[9][10][11]。语音识别技术的进步也推动了相关的应用研究，近年来兴

起的对话系统（航空订票系统等）、语音检索、语音翻译等[12][13][14]，都是以语音识别技术为基础的。

语音识别技术的进步是显见而可喜的。但是，我们也看到，一个在实验室环境下工作得很好的语音识别系统，在实际应用时，其性能就变得差强人意，而其中一个非常重要的原因就是口音问题。以英语为例，多数人所说的英语并不象播音员那样标准、清晰，而是受其母语的影响很大，带有各种各样的口音。比如，带西班牙口音的英语、日式英语等。即便说话人的母语同为英语，但由于所处地域的差异，人们在发音、用词等方面也存在着非常大的区别，如美式英语、英式英语、澳式英语等，其差异是非常大的。口音问题给语音识别带来了很大的困难，使得系统性能急剧下降，甚至完全不可用。

汉语语音识别中同样存在着口音问题，即方言背景普通话语音识别，并且比其它语言更为复杂。标准普通话（简称为普通话，或缩写为 PTH）是我国的官方语言，随着普通话在全国范围内的大力推广，越来越多的人学会了说普通话，并在日常生活中使用它，普通话在人们日常生活中的地位也越来越重要。因此，汉语语音识别常常以标准普通话，或带有轻微口音的普通话为研究对象。而国内语音识别的主要研究机构，如，清华大学语音技术中心，清华大学电子系语音实验室，中科院自动化所，中科院声学所等，都开发有自己的普通话连续语音识别系统。但实际上，很多人所说的普通话并不标准，往往受到其方言背景的影响而带有较重的口音，有时甚至就是普通话和方言的混合体，这就给语音识别带来了很大的困难。

中国的方言是非常丰富的，一般可以分为九大方言区，除了北方官话（北方地区），还有，吴（江苏南部，浙江，上海等），粤（广东，香港，南宁，广西等），闽（福建，广东汕头，海南海口，台湾台北），客家（广东梅县，台湾新竹），湘（湖南），赣（江西），徽（安徽）和晋（山西）。这九大方言区又可以进一步划分为约四十个小的方言区。各地方言的差异是很大的，从用词、发音，到句法、语法等，都会有所不同，有时甚至超过不同语言（如英语和法语）之间的差异。虽然不同的方言和标准普通话使用同样的汉字进行书写，而且，大部分方言区也在讲普通话，但人们所说的普通话还是受到了方言的严重影响。这种差异给自动语音识别带来了很大的困难。因此，当我们用标准普通话识别器来识别带口音的普通话时，识别器的性能便会急剧下降。

这里给出一个具体的例子。在 NIST 1997 广播新闻评测任务（NIST 1997

Broadcast News evaluation task[15]）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，Character Error Rate）为 21.38%。但是，当我们用此标准普通话识别器去识别吴方言背景普通话时，其字错误率则急剧上升。对朗读式和自然发音式吴方言背景普通话测试集，其字错误率分别变为 61.89%和 72.17%。对于汉语语音识别性能的评测，字错误率 CER 已成为广泛应用的方式，因此，本文也主要给出字错误率结果。由以上结果可以看出，方言背景对语音识别的影响是非常严重的。

基于以上原因，口音问题已经成为近几年来语音识别领域的研究热点之一。而在汉语中，普通话受方言背景影响较之其他语言更为严重，因此对语音识别的影响也更大。但这方面的研究相对比较欠缺，它又是从研究到实用过程中不得不面对的问题，十分值得进行深入的研究。2004 年，由美国自然科学基金（NSF）资助的 Workshop 在 JHU（The Jhons Hopkins University）的 CLSP 研究中心（The Center for Language and Speech Processing）举行，其中一个主要研究课题就是“吴方言背景普通话研究”（Dialectal Chinese Speech Recognition）[16][17]，而本文中的一部分工作就是在这次 Workshop 上完成的。

1.2　前人的工作

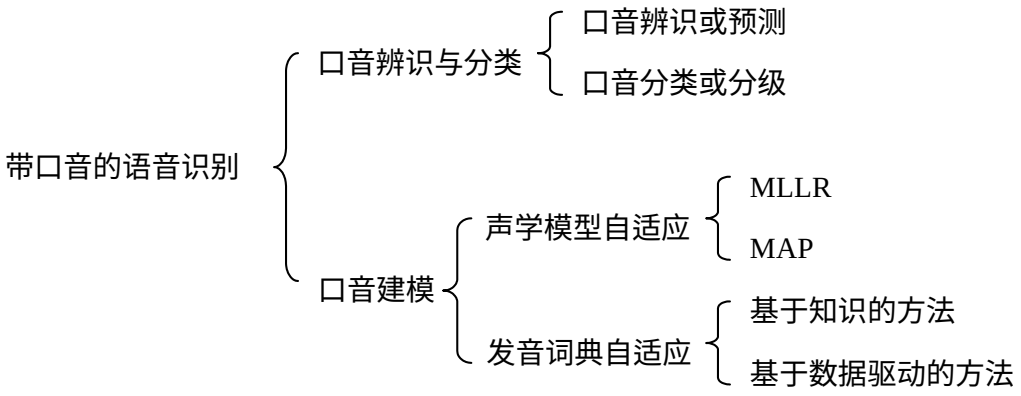


图1.1　带口音的语音识别研究的两个方面

如图 1.1 所示，近年来，对带口音的语音识别的研究主要集中在两个方面：（1）口音辨识（Dialect/Accent Identification）与分类；（2）口音建模。其中前

者可以作为后者的前端加以应用。另外，还有最新的一些研究关注于韵律规则对发音变化的影响，同样有助于口音建模研究[18][19]，但这一方面的工作在带口音的语音识别中尚未全面展开。

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与方言语音识别器的前端，利用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[20][21][22][23][24][25][26][27]。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（LSA，Latent Semantic Analysis）的高阶统计信息来进行汉语方言辨识[20]；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[21]；采用高斯混合模型（GMM）进行汉语普通话口音分类[22]；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[23]；利用韵律信息进行口音预测[24][25]；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[26]；以及对说话人的口音等级（受母语或方言背景影响程度）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，然后使用特定的模型和发音词典进行识别[17][27]，等。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，其结果也可以作为带口音语音识别系统的前端。口音辨识和分类应用于语音识别所面临的一个问题是，如何解决正确率与实用性之间的矛盾。分类时使用的测试数据多，则正确率高，但实时性差，不易实用；使用的数据少，则实时性强，但正确率低，对实际应用同样存在着问题。现有的方法中，基于特定音素或声韵母频率的方法是一个动态的方法，随着观测数据的增加，可以给出不同的置信度，是一个有潜力的方法[17][27]。

对于口音建模研究，现有的方法可以归结为以下两个主要方面：声学模型自适应（Acoustic Adaptation）和发音词典自适应（Lexicon Adaptation）[26][28][29][30][31][32][33][34]。如图 1.1 所示，声学模型自适应常采用最大似然线性回归（MLLR，Maximum Likelihood Linear Regression）[35][36]和最大后验概率（MAP，Maximum a Posteriori）[37][38]方法。这两种自适应方法是当前最为有效的自适应方法，许多新的自适应方法都是从二者中派生出来的。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 则适用于数据量相对较多的情形。发音词典自适应常采用发音变化建模

（PVM，Pronunciation Variation Modeling）相关技术，主要研究由说话方式、语速、口音等带来的影响，其研究成果可以借鉴来解决方言背景问题。发音变化建模也常被称为发音建模（PM，Pronunciation Modeling）。

斯坦福大学的Jurafsky是“ Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics ”（简称为 SLP）一书的作者，他对发音变化建模的重点和难点进行了阐述[39]。Strik 对发音变化建模相关工作进行了详细的总结，给出了一篇很好的综述，介绍了近年来发音变化建模研究的各种方法，并加以分类和总结[40]。许多基于知识和基于数据驱动的建模方法都得到了广泛的应用[41][42]。对于发音变化的描述，还可以在不同级别进行，如词、音节、音素等[43][44][45]，甚至音素内部，如状态或高斯混合（Gassian Mixture）一级[46]，并在口音建模中得到应用[47]。

表 1.1 给出了若干带口音的英语语音识别系统的结果，这些系统是针对自然式发音的（Spontaneous speech），其识别难度要高于朗读式发音，因此识别率也相对较低。表中列出了口音种类、建模所用的数据量，以及词错误率（WER，Word Error Rate）变化等[28][29][30]。

表1.1　带口音的自然发音式英语识别系统词错误率

文献	口音	带口音的数据	词错误率（口音一致）	词错误率（口音不一致）
[28]	西班牙	20 小时 (训练)	39.2%	68.5%
[29]	日　本	3 小时 (MLLR)	52.5%	63.1%
[30]	德　国	52 分钟 (MLLR)	43.5%	49.3%

表中第二列为口音种类；第三列指的是用于进行训练或自适应的“ 带口音的数据 ”的总量（总时间）。使用这些数据可以重新训练或自适应得到“ 口音一致的模型 ”，从而进行“ 口音一致 ”的测试。最后两列给出了识别器的识别结果——词错误率。表中，“ 口音一致 ”表示用带口音的识别器来测试带口音的英语，“ 口音不一致 ”则表示用标准的识别器来测试带口音的英语。以西班牙口音（表中第一行）为例，当口音一致时，词错误率为 39.2%，但当口音不一致时，词错误率便上升为 68.5%。由此可以看出，口音问题仍旧是自动语音识别的重点和难点之一，尤其是对于自然发音的语音。

在国内，研究语音识别多是标准普通话或纯方言的语音识别，而对方言背

景带来的问题一直没有展开广泛地研究。直到近几年，研究者们才开始关注这个方面的问题，并加强了语音数据库的建设和相关的研究。2003 年，清华大学语音技术中心与得意公司合作，为 JHU 大学设计并采集了吴方言背景普通话语音库（WDC），社科院语言所进行了手工标注。2004 年，社科院语言所在国家 863 高技术项目支持下完成了四大地方（上海、广州、重庆和厦门）普通话语音语料库（RASC863）的采集[48]，此库共有 800 个说话人的语音。微软亚洲研究院采集了四种口音（北京、上海、广东、台湾）的普通话语音库，共有 1,440 个说话人。在研究方法上，微软亚洲研究院主要进行声学自适应和发音词典自适应（PDA），在朗读式语音上取得了很好的结果，可以使错误率降低 28.4%。需要说明的是，这些结果在一定程度上是说话人相关的，因为训练集中包含了测试集中说话人的其它语音[31][32]。中科院自动化所尝试了基于音节发音变化词典的方法进行特定方言的语音识别研究[49][50]。

考察现有的研究方法和结果，我们可以得出如下结论：

（1） 口音问题是语音识别中的一个重要问题，且难度很大

从前面给出结果可以看出，当口音不一致时，系统的误识率是非常高的。因此口音问题可以说是语音识别中一个非常重要的问题，不得不去解决。同时口音问题也是个很复杂的问题，解决的难度很大，对于汉语更是如此。口音问题具有多样性、多级性、不稳定性等特点，给研究带来了很大的困难。

第一、多样性：不同说话人的母语或方言背景差异大，种类多，从而很难找到统一的方法进行处理，一般都要单独或分类处理。

第二、多级性：说话人口音轻重、对语法等的掌握程度存在着巨大差异。如果我们将方言定义为 0，将标准普通话定义 1，用来表示普通话的标准程度。则人们所说的普通话可能是 0 和 1 之间的任何值。在我国的普通话水平考试中，也将普通话水平分为三级六等，从高到低依次为 1A，1B，2A，2B，3A 和 3B。因此，相对于标准普通话和方言的识别研究，口音问题则要复杂得多。

第三、不稳定性：由于说话人对发音、词汇、语法等的掌握程度差异很大，其普通话也会表现得很不稳定。例如，在很多方言中，非卷舌音/z/和卷舌音/zh/是不区分的，都发为/z/，在说普通话时，多数说话人多会将/zh/发为/z/，但也会出现一种情况，即说话人不是很确定该如何发音，从而将/z/发为/zh/，属于一种矫枉过正的现象，这使得发音规律在一定程度上变得不稳定。

（2） 汉语中对方言背景普通话研究还比较欠缺。

在汉语语音识别中，多是将标准普通话和特定的方言（如粤语）视为两类进行研究。对于标准普通话，在建立语音库时，常常选择发音标准、清晰的说话人进行录制，说话人口音很轻或基本没有口音，如我们常用的 863 数据库；对于方言，常常将其视为一门新的语言，专门为其定义基元、发音词典，按照与标准普通话类似的方式建模和识别。而针对方言背景普通话的研究则非常少。实际上，人们所说的普通话受方言背景影响很大，从而严重影响了识别器的性能，这是语音识别实用化过程中必须解决的重要问题之一。近年来，国际上对于口音问题的研究日益加强，而国内相关研究却仍止步不前。

（3） 现有的许多研究方法缺乏整体性和通用性。

现有的方法大多把口音作为一个孤立的问题来研究，而对方言或母语这个大背景考虑不够深入，这也使得研究缺乏整体性。同时，现有的研究往往基于大量的数据，因而成本很高。虽然数据越多就越容易进行模型的训练或自适应，但汉语中方言众多，差异又大，要对每种方言都收集足够多的语音数据并进行标注，其时间和资金成本是很高的，这也使得方法本身难以推广。另一方面，有些方法又过于依赖口音的特性，很难应用到其它口音上，这也使得方法的通用性不够好。

1.3　研究的重点、难点和方法

本文选定方言背景普通话识别为研究方向，这是一个具有很强的研究价值和实际意义的课题。对于方言背景普通话识别器的构建，可以收集足够多的数据，定义特定的基元和词典，按照标准普通话的方式进行训练或进行自适应，从而得到特定的方言背景普通话识别器。毫无疑问，这种方法直接而有效，对于特定的方言背景普通话，可以取得较好的结果。但是，这类方法往往成本较高、通用性差、难以扩展。

不同于以往的方法，本文重点研究：如何使用较低成本构建一个方言背景普通话识别器，并且方法本身可以方便地扩展到其它方言。它强调两点内容：一是低成本，一是易于扩展。一般地，识别器的构建成本主要体现在数据库的采集和标注方面。而我国方言种类众多，如果对每种方言都要采集大量数据并进行标注的话，识别器的构建成本会变得非常高。因此，如何设计和选择语音库，如何从中提取有用信息等都是本文的研究重点之一。同时，各种方言虽然

不同，但它们对口音等的影响也具有一定的规律。本文希望研究的方法能够体现这种规律性，使得方法不仅对特定的方言有效，同时还能够方便地扩展到其它方言中。针对低成本、易于扩展的目标，本文着重考虑以下几个方面：

（1）采用变换的方式构建方言背景普通话识别器

如果有足够多的高质量语音数据，我们就可以按照标准普通话的构建方式，为每种方言重新建立一个特定的解码器。而实际上，为每种方言都采集大量语音数据，这种做法的成本太高了；另一方面，虽然各种方言背景对普通话的影响各不相同，但它们也存在着许多共性。比如，书写所用的文字与标准普通话相同，说话时都以标准普通话为目标，尽量把话说标准，等。可以认为，方言背景普通话是标准普通话受特定方言背景影响发生偏移或者变换，因此，我们可以采用变换的方式得到特定的方言背景普通话识别器。这样既可以降低成本，也可以使标准普通话识别器及其研究成果发挥作用。

（2）使用少量语音数据

数据库的设计、采集、标注等，都要耗费很多的人力物力。如果每种方言都需要采集大量语音数据的话，方法本身便不易推广。而且我们发现，对于吴方言背景普通话识别器，当带口音的数据超过 6 个小时以后，自适应方法与重新训练相比已不再占有明显优势。也就是说，这种情况下就不必采用变换的方式了，而直接训练模型即可。

（3）使用声韵母映射规则

本文认为，方言对标准普通话发音的影响主要表现在声韵层，而且这种规律对大多数方言都是适用的。这一点也可以从手工标注的差异中得到验证。因此，本文采用声韵映射规则（IF-Mapping Rules）来指导声学自适应、多发音词典的构建等。规则可以由专家知识来提供，也可以从数据中统计得到。虽然规则本身是与特定的方言相关的，但方法本身很容易推广到其它方言中。

（4）常用方言词汇

汉语方言中，常常存在一些方言特有的词汇，而且还会经常用于方言背景普通话中。如吴方言中的“晓得”，就是“知道”的意思。虽然每种方言会有自己特定的常用方言词汇，但收集方言词汇的方法是与特定的方言无关的，可以应用于所有方言。

这里给出本文研究的理论基础和形式化描述。连续语音识别的目标就是，对于给定的输出观察矢量 $\mathbf{X} = X_1 X_2 \cdots X_n$，通过最大后验概率准则得到最优词序

列 $\hat{\mathbf{W}} = W_1 W_2 \cdots W_m$，即公式(1-1)和(1-2)[51]，

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

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，$P(\mathbf{X}|\mathbf{W})$表示的是词序列 **W** 产生观察矢量 **X** 的概率，即，声学模型匹配得分；$P(\mathbf{W})$表示的是词序列 **W** 的先验概率，即，语言模型得分。而 $P(\mathbf{X})$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）[52]，

$$P(\mathbf{W}) = P(W_1 W_2 \cdots W_m) = \prod_{i=1}^{m} P(W_i \mid H_i) \tag{1-3}$$

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

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

方言背景普通话中仍旧保留了许多特定的方言词汇，这些词汇在原来的词表中没有被收录进来，更没有概率表示。另外，方言背景普通话在语序等方面与标准普通话存在着差异，也会在一定程度上影响 Bi-gram 概率。因此，有必要考察常用方言词汇，并对语言模型 $P(\mathbf{W})$进行重估或自适应。

$P(\mathbf{X}|\mathbf{W})$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 **Y**，从而将 $P(\mathbf{X}|\mathbf{W})$化为

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

其中，序列 $\mathbf{Y} = Y_1 Y_2 \cdots Y_{ny}$，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（或音素）序列 **B**，$\mathbf{B}=B_1B_2\cdots B_{nb}$。可以得到式（1-6），

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

如果将式（1-6）中的条件进行简化，即，假设序列 **B** 仅依赖于序列 **W**，且对应的声学模型仅依赖于序列 **B**，则可以得到更为简单的形式。而现有大多数语音识别系统中的声学打分都可以用类似的形式来表示。如式（1-7），

$$P(\mathbf{X}|\mathbf{W})=\sum_{\mathbf{B}}P(\mathbf{X}|\mathbf{B})P(\mathbf{B}|\mathbf{W}) \tag{1-7}$$

式（1-6）中，可将序列 **B** 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**，$\mathbf{S}=S_1S_2\cdots S_{ns}$，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，

$$P(\mathbf{X}|\mathbf{B},\mathbf{Y},\mathbf{W})=\sum_{\mathbf{S}}P(\mathbf{X}|\mathbf{S},\mathbf{B},\mathbf{Y},\mathbf{W})P(\mathbf{S}|\mathbf{B},\mathbf{Y},\mathbf{W}) \tag{1-8}$$

从而式（1-6）化为

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

将 $P(\mathbf{B}|\mathbf{Y},\mathbf{W})P(\mathbf{Y}|\mathbf{W})$简写为 $P_{\mathbf{W},\mathbf{Y}}(\mathbf{B})$，同时再将上式中的模型打分条件进行简化，即，假设模型仅依赖于序列 S，得到

$$P(\mathbf{X}|\mathbf{W})=\sum_{\mathbf{Y},\mathbf{B},\mathbf{S}}P(\mathbf{X}|\mathbf{S})P(\mathbf{S}|\mathbf{B},\mathbf{Y},\mathbf{W})P_{\mathbf{W},\mathbf{Y}}(\mathbf{B}) \tag{1-10}$$

式（1-10）中，$P(\mathbf{X}|\mathbf{S})$表示的是实际发音序列 **S** 产生观察矢量 **X** 的概率，即实际发音序列 **S** 的声学匹配得分。$P(\mathbf{S}|\mathbf{B},\mathbf{Y},\mathbf{W})$表示由序列 **W**，**Y**，**B** 产生实际发音序列 **S** 的概率，此项与发音变化规则和多发音词典相关。而本文在声学层的工作主要集中在这两部分，即，（1）声学模型训练与自适应；（2）多发音词典。

（1）对于声学模型，本文认为，方言背景普通话受到了标准普通话和方言背景的双重影响，仅用标准普通话基元来描述是不够的，因此引入了部分方言相关的基元。以吴方言为例，序列 **S** 可以采用两种形式，一是

仅使用标准普通话声韵母（PTH-IF）来表示，记为 $\mathbf{S}^{\mathbf{PTH\text{-}IF}}$（简写为 $\mathbf{S}^{\mathbf{PTH}}$）；一是使用吴方言背景普通话声韵母（WDC-IF，Wu-Dialectal Chinese Initial/Finals）来表示，记为 $\mathbf{S}^{\mathbf{WDC\text{-}IF}}$（简写为 $\mathbf{S}^{\mathbf{WDC}}$）。则有

$$P(\mathbf{X}\mid\mathbf{W})=\begin{cases}\sum\limits_{\mathbf{Y},\mathbf{B},\mathbf{S}^{\mathbf{PTH}}} P(\mathbf{X}\mid\mathbf{S}^{\mathbf{PTH}})P(\mathbf{S}^{\mathbf{PTH}}\mid\mathbf{B},\mathbf{Y},\mathbf{W})P_{\mathbf{W},\mathbf{Y}}(\mathbf{B}) & (1\text{-}11)\\ \sum\limits_{\mathbf{Y},\mathbf{B},\mathbf{S}^{\mathbf{WDC}}} P(\mathbf{X}\mid\mathbf{S}^{\mathbf{WDC}})P(\mathbf{S}^{\mathbf{WDC}}\mid\mathbf{B},\mathbf{Y},\mathbf{W})P_{\mathbf{W},\mathbf{Y}}(\mathbf{B}) & (1\text{-}12)\end{cases}$$

依据上式，本文分别采用两种方式进行模型自适应，以得到方言背景普通话声学模型。$\mathbf{S}^{\mathbf{PTH}}$ 对应的模型采用标准发音来监督自适应，而 $\mathbf{S}^{\mathbf{WDC}}$ 对应的模型采用实际发音来监督自适应（详见第 3 章）。

（2）对于多发音词典，本文仅考察邻近的上下文对声韵映射的影响，而不考虑更远的上下文，仅考虑替代（Substitution）错误，而不考虑插入删除（Insertion/Deletion）错误（替代错误是方言背景普通话识别中的主要错误）。因此在 $P(\mathbf{S}|\mathbf{B},\mathbf{Y},\mathbf{W})$中，实际发音 **S** 和标准发音序列 **B** 是等长的，即 $ns=nb$。假设 **B** 序列中，当前声韵母为 B_i，相邻的前一个声韵母为 $B_{i\text{-}1}$，而 B_i 所在的词和音节分别为 W_{Bi}，Y_{Bi}。则产生特定的实际发音序列 **S** 的概率为

$$P(\mathbf{S}\mid\mathbf{B},\mathbf{Y},\mathbf{W})\approx\prod_{i=1}^{ns}P(S_i\mid B_i,B_{i-1},Y_{B_i},W_{B_i}) \qquad (1\text{-}13)$$

再引入不同的上下文相关性假设：$C1$，$C2$ 和 $C3$，从而将 $P(S_i\mid B_i,B_{i-1},Y_{B_i},W_{B_i})$ 化为不同的形式。其中，

$C1$：S_i 仅依赖于 B_i 和 W_{Bi}，而与音节 Y_{Bi} 无关；

$C2$：S_i 的依赖于 B_i，左边相邻的上下文 $B_{i\text{-}1}$ 和 W_{Bi}，而与音节 Y_{Bi} 无关；

$C3$：S_i 依赖于 B_i 及其所在音节 Y_{Bi} 和词 W_{Bi}。

暂不考虑词 W_{Bi} 带来的影响（W_{Bi} 的作用不同于其它上下文，后文将给出具体说明），则 $C1$，$C2$ 和 $C3$ 分别表示：上下文无关声韵母映射规则，上下文左相关声韵母映射规则，以及音节相关声韵母映射规则。基于上述假设，式（1-13）可被表示为

$$P(\mathbf{S}\mid\mathbf{B},\mathbf{Y},\mathbf{W})\approx\begin{cases}\prod_{i=1}^{ns}P(S_i\mid B_i,W_{B_i}) & (C1) \quad (1\text{-}14)\\ \prod_{i=1}^{ns}P(S_i\mid B_i,B_{i-1},W_{B_i}) & (C2) \quad (1\text{-}15)\\ \prod_{i=1}^{ns}P(S_i\mid B_i,Y_{B_i},W_{B_i}) & (C3) \quad (1\text{-}16)\end{cases}$$

此处还可以进行其它形式的上下文相关性假设，从而得到不同的表示方式。另外，根据式（1-11）和（1-12），还可以选用不同的方式来表示序列 **S**，从而得到不同的映射规则和多发音词典，如 PTH-IF 基元集合内部映射，或 PTH-IF 到 WDC-IF 映射等。

式中的 W_{Bi} 表示当前词对发音变化的影响，本文引入此项，是为了利用语言层的信息进行发音词典的剪枝，以降低多发音词典之间的混淆，其作用与其它上下文（如 Y_{Bi}）有所不同。

基于前面的讨论和以上的理论分析，本文提出了一种新的框架，尝试使用少量带口音数据，外加方言背景相关知识，将一个标准普通话识别器转换成为一个方言背景普通话识别器。如图 1.2 所示（以吴方言背景普通话为例）。

框架可以分为三部分，左边虚线框内的部分为“标准普通话识别器”，右边虚线框内为“方言背景普通话识别器”，而中间的部分表示识别器的转化过程。此框架的第一个研究重点是标准普通话识别器，它是此框架的研究基础，方言背景普通话要在此基础上建立。第二个研究重点是如何由标准普通话识别器构建方言背景普通话识别器——即识别器的转化过程。为了降低成本，提高扩展性，本文只使用少量方言背景普通话数据用于知识抽取和模型自适应。同时，针对特定的方言，需要收集整理的方言相关知识，进而将标准普通话识别器转换为一个方言背景普通话识别器。

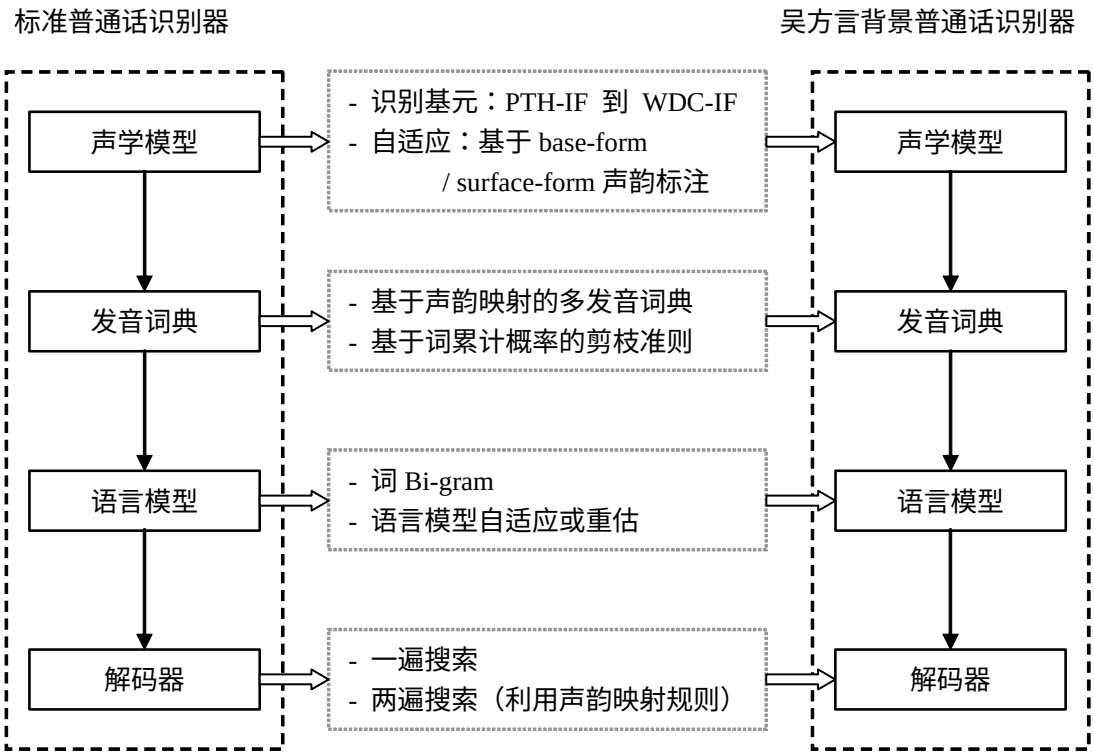


图 1.2　吴方言背景普通话识别框架

具体的转换可以在四个层面进行——声学模型层，发音词典层，语言模型层，解码器层。

(1)　声学模型

本层考虑两个方面的问题：识别基元定义和声学模型自适应。明显地，为了标注方言背景语音数据，仅仅使用标准声韵母（PTH-IF）集合是不够的，还需要引入一些方言相关的声韵母。同时，由于普通话的推广，越来越多的人可以较为流利地讲普通话，所以，并非所有的方言相关的声韵母都要被考虑进来。事实上，只有少量需要保留，而它们可以通过考察声韵层标注信息获得。因此，如公式(1-12)所示，可以采用扩展的声韵母集合来表示实际发音序列 **S**。

在进行声学模型自适应时，我们采用“有监督的 MLLR”方法。这种方法适用于数据量较少的情况，且 MLLR 本身是一种变换的方法，因此也符合本文“变换”的基本思想。本文采用两种标注信息来指导声学自适应：base-form（或 baseform）声韵标注，表示标准发音；surface-form 声韵标注，表示实际发音。而前者可以被认为是传统的方法。下面给出一个实际的例子来介绍这两层标注，其中 HZ 表示汉字层标注，PY 表示拼音层标注，BF 表示 base-form 标准发音声韵标注，SF 表示 surface-form 实际发音声韵标注：

HZ:	说		我	有	一	次	
PY：	shuo		wo	you	yi	ci	
BF：	sh	uo	uo	iou	i	c	ii
SF：	s	uo	uo	ieu	i	c	ii

(2)　发音词典

通过考察吴方言背景普通话语音库标注信息，我们发现，实际发音和标准发音之间在声韵层存在着 20～30%的差异。这种差异的存在提示我们，可以从声韵层的变化规律来考察方言背景对普通话的影响。我们可以考察以下声韵映射规则，如，上下文无关普通话声韵母映射规则，上下文无关方言背景普通话声韵母映射规则，音节相关或左相关声韵母映射规则，这些规则可以从专家知识或语音数据中获取，用于生成多发音词典。其使用方式如公式（1-14），（1-15）和（1-16）所示。

使用声韵映射规则得到的多发音词典可以覆盖多数可能的方言背景普通话发音，但发音的数目相对较大，因而也会引入更多的混淆。所以，需要对发音词典进行剪枝。本文提出了基于累积一元概率（AUP，Accumulated Uni-gram Probabilities）的准则对发音词典进行剪枝，从而降低多发音词典的长度。此方法通过公式（1-14）至（1-16）中的 W_{Bi} 项来体现。

(3)　语言模型

统计语言模型（如 Bi-gram 或 Tri-gram 等）的应用可以有效地提高识别器的性能，对于方言背景普通话也是如此。正如我们所知，这里存在一些方言相关的问题，如，标准普通话词表之外的方言相关的词汇，句子或片断中词序的不同，它们均会影响识别器的性能，尤其是那些口音很重的说话人。因此，有必要对普通话语言模型进行重新训练或自适应，以符合方言背景普通话识别器的需要。

(4)　解码器

一遍搜索和两遍搜索策略均可在解码时采用。本文中给出的所有结果都是基于一遍搜索策略的。但两遍搜索也是一种很好的选择，而且，可能会更适于方言背景普通话识别器。例如，类似于吴方言背景普通话，许多方言在发音时

都有其特定的韵律规律，在相应的方言普通话中也有所体现。还有，用标准普通话识别器识别方言背景普通话，通过数据驱动的方式，可以从识别结果中学习到一些知识。这些知识都可以方便地应用于第二遍搜索中。一种可行的两遍搜索的形式是，用标准普通话识别器进行第一遍搜索，生成单元（词/音节/声韵母）网格，然后再利用方言背景普通话知识进行第二遍解码。

1.4　论文结构安排

本文研究的总体思路和主要内容如图 1.3 所示。从图中可以看出，研究总体分为三部分内容：数据准备、标准普通话声学建模和吴方言背景普通话识别器构建。前两部分是吴方言背景普通话研究的基础，也是非常重要的。图中的阴影部分是本文工作涵盖的具体内容。

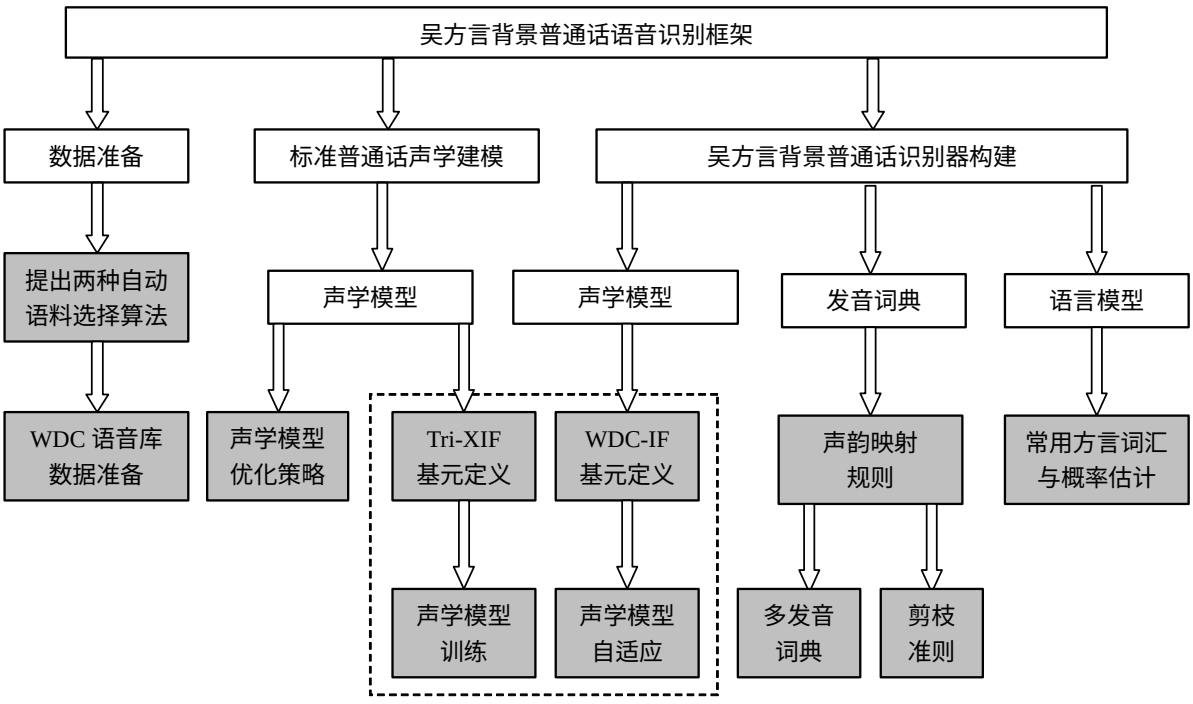


图 1.3　论文研究总体思路和主要内容

在数据准备部分，本文研究的是均衡语料选择算法和吴方言背景普通话（WDC）语音数据库准备工作。在这部分，本文提出了两种均衡语料自动选择算法，分别称为抑制高频单元（RHF，Restricting High-Frequency

units）和鼓励低频单元（ELF，Encouraging Low-Frequency units），并进行了对比，具体内容将在后面介绍。同时，还介绍了 WDC 语音库的设计、采集、标注和划分工作。

在标准普通话声学建模中，本文采用声韵母作为声学基元，进而提出了扩展声韵母基元（XIF，eXtended Initial/Final），进行基于状态共享的上下文相关扩展声韵母建模（Tri-XIF），并采用适当的策略优化模型。

吴方言背景普通话识别器构建部分是在图 1.2 所示的框架指导下进行的。在声学模型层定义了吴方言普通话识别基元，称为 WDC-IF，并分别利用声韵层标准发音（base-form）和实际发音（surface-form）来指导声学自适应。由于这部分内容与标准普通话声学建模关系十分密切，因此，本文在章节安排中将它们组织到了一起（图中虚线框）。在发音词典层，提出了基于声韵映射规则的多发音词典构建方法，并提出基于一元累积概率（AUP）的准则进行多发音词典的剪枝。在语言模型层研究并收集常用吴方言词汇，并给予适当的概率估计。在解码器层，本文使用一遍解码策略。对于多遍解码及重打分策略只进行了初步的研究，它将作为将来工作的重点之一，因而并未列入图 1.3 中（具体内容请参见“ 发表的学术论文 ”[8]）。

全文按如下方式进行组织：第 2 章介绍语音库的设计与采集方法，包括均衡语料自动设计算法和吴方言背景普通话语音库准备；第 3 章介绍标准普通话声学建模方法和吴方言背景普通话声学模型自适应方法，包括基于状态共享的上下文相关扩展声韵母建模方法及其优化策略，基于 base-form/surface-form 的 MLLR 自适应方法；第 4 章介绍基于声韵映射规则的多发音词典生成方法和基于累积一元概率（AUP）的剪枝准则，以及常用方言词汇的收集与概率估计；第 5 章给出相关的实验结果和分析；第 6 章是对工作的总结和对未来工作的展望。

第 2 章　语音库设计与采集

2.1　本章引论

声学模型训练和自适应都需要语音库的支持。近年来，越来越多的语音识别研究机构开始重视并逐渐加强了语音语料库的建设工作。当前，国内外已经成立了许多家组织，致力于语料库的收集、整理、发布工作，如 LDC（Linguistic Data Consortium）[53]，ELRA（European Language Resources Association）[54]，国际中文语言资源联盟（CCC，Chinese Corpus Consortium）[55][56]联盟等。我国的社科院语言所（CASS）专门从事汉语语音语料库的设计、采集、标注、研究等，也积累了很多的经验和成果。

语音库采集中的一个重要问题就是文本的设计。一般地，不同的应用对语料库有不同的要求，因此，语音文本要根据其用途进行设计。本章讨论的是用于语音识别的语音库设计，主要用途是声学模型训练和自适应。为此目的，语料库应该覆盖尽可能多的语音单元（如音节，声韵），协同发音现象，以及语言现象，同时，为了避免过多的冗余，各个语音单元出现次数应该尽可能地均衡。也就是说，为了让声学模型训练更为准确地描述各个基元及其协同发音现象，语音库中的各个语音单元要有一定的覆盖率和均衡度。

通常，语料库的设计是基于手工或半自动的方式[57][58][59]。在我国，现有应用最为广泛的汉语语音库就是 863 语音库，它是在国家 863 计划支持下，由 863 委员会组织，社科院语言所和中国科技大学合作完成的标准普通话连续语音数据库[58]。863 语音库的设计就是通过半自动的方式完成的，达到了声韵母单元的均衡。手工或半自动方法有如下优点：便于控制句子质量；易于覆盖高层知识信息；适用于中小规模的语料库设计。

设计大规模语料库时，使用手工或者半自动的方式就会带来很大的困难。因此，一般要采用自动抽取的方式，从一个真实的、大规模的候选语料库中自动抽取句子，生成所需的语料库[60][61]。根据语料库的用途，候选语料库中的句子可以从书籍、报刊、互联网等获取，因此可以建立很大的候选库。自动抽取方法优点是：便于建立大规模的语料库；候选语料库规模大，相对易于满足语音单元的覆盖和均衡等要求；灵活性高，可以根据需求随时重建语料库。但由

于是全自动的方式，也存在着一定的缺点，如候选语料库规模过大时，无法保证其质量，等等，也是值得关注的问题。

语料库设计算法的优劣常用语音单元的覆盖率和离散度来衡量。语音单元的覆盖率指达到一定次数的语音单元所占的比例，它对模型训练来讲是很有现实意义的。离散度常用方差来描述，用以表示各个单元的均衡度，而其数值常常很大，达到10^5以上，因而无法给人直观的感受，所以本文并不使用方差来衡量算法的好坏，而是给出直观的统计结果和分析。

区别于过去的方法，本文通过加强对高频单元的抑制或对低频单元的鼓励来达到最终的均衡。为此，本节提出了基于“抑制高频单元”和“鼓励低频单元”两种策略，同时列出了随机选择的方法作为参考，其物理意义会在后面具体介绍。虽然本文使用的是语音单元是右相关的声韵母（此处称为 Di-IF），但所提出的算法对所使用的语音单元并无限制。本节介绍的相关算法发表于 O-COCOSDA2003。

本章研究主要包括以下两个方面：（1）均衡语料库设计方法；（2）吴方言背景普通话数据准备。本章按照如下方式组织：2.1 节为本章引论；2.2 节提出两种均衡语料设计算法并进行比较；2.3 节介绍吴方言普通话数据准备工作；2.4 节为本章小结。

2.2　均衡语料设计算法

表 2.1 中列出算法所使用的符号表。

表2.1　均衡语料设计算法的符号表

符号	意义及说明
S	候选语料库
M	候选语料库 S 中的句子总数。即 S 中含有 M 个句子
n	语音单元总数。如，Di-IF 个数
s	S 中的一个句子
s_i	S 中第 i 个单元出现的次数。i =1,2,…,n
D	目标语料库。从 S 中抽取出句子，放入 D 中

续表2.1　均衡语料设计算法的符号表

符号	意义及说明
N	目标语料库 D 中最终的句子总数。即需要抽取的句子数目
d	D 中的一个句子
d_i	D 中第 i 个语音单元出现的次数。i =1,2,…,n
F	理想情况下 D 中各个语音单元数目。即平均数目
TH	D 中各个语音单元数目的上限阈值
$Score(D)$	D 的分数

2.2.1　随机选择

随机地从候选语料库中选择 N 个句子，可以得到目标语料库 D 的一个候选，则共有 C（M，N）种候选方案，而最佳方案就是其中之一。但是当 M 和 N 很大时，C（M，N）是非常大的，所以很难通过这种方式得到满足要求的语料库。

2.2.2　抑制高频单元

抑制高频单元（RHF，Restricting High-Frequency units）是通过对高频单元的抑制来达到均衡的。RHF 是一种**递增式**语料选择算法。开始时，所有句子都在候选语料库 S 中，目标语料库 D 为空。然后，对候选语料库 S 中每个句子进行打分，选出分数最高的句子，将其从 S 中抽取出来，加入 D 中。重复此过程，直至 D 中的句子数目达到 N。

RHF 方法与以往的方法比较，句子打分和选取的准则不同，主要体现在对高频单元的抑制方面。首先，根据候选库 S 和目标库 D 的规模大小，定义一个阈值 TH，D 中各个单元的数目不能超过此阈值，从而有效地起到抑制高频单元的作用。阈值 TH 只有在算法无法完成，即无法再继续选出句子满足阈值条件时，才会逐步放宽，直到算法完成。其次，对 S 中的句子要参照阈值 TH 进行打分，含有高频单元的句子分数相对较低，从而在一定程度上抑制了高频单元的出现。句子打分时考虑了句中每个单元的影响，并对分数进行了调整，从而更为有效地反映出此句对语音单元覆盖率和均衡度的贡献大小。简单地说，就是有一个阈值作为硬性指标来抑制高频单元，而让低频单元尽可能地被选入，从而保证

单元的均衡。

下面给出有关阈值设定和打分的详细讨论。理想情况下，目标语料库 D 中所有单元完全均衡，即各个单元数目相同（设为 F），但实际上，自然而有意义的句子中各个语音单元之间具有很强的相关性，所谓的理想情况实际上是无法达到的。为了抑制高频单元，我们根据理想均值 F 设定一个阈值 $TH = aF \quad (a \geq 1)$。若取 $a = 5$，则每种语音单元出现的最多次数为理想均值 F 的 5 倍。如果 D 中某种单元的总数已经超过此阈值，则含有此单元的句子将不会被选择进来，从而达到抑制 D 中高频单元的目的。另外，对 S 中的句子 s 进行打分时，将考虑 s 中所包含的所有单元。句子中所含的每个单元，如单元 i 的分数，按如下方式定义

$$Score_i = \frac{TH - d_i}{s_i} \qquad (d_i < TH) \tag{2-1}$$

（2-1）中，分子部分表示 D 中所需的单元 i 的数目，分母部分表示 S 可以提供的单元 i 的数目。此处利用每个单元的供求关系来表示其重要性。供不应求者会得到更高的分数，而供大于求者则得到较低的分数。

$$NormScore_i = \begin{cases} Score_i & (Score_i \geq 1) \\ \dfrac{1}{1 - \log(Score_i)} & (Score_i < 1) \end{cases} \tag{2-2}$$

（2-2）是为了避免对严重供大于求的单元（$Score_i < 1$）打分过低而进行的分数调整，其分数范围仍在（0，1）区间内。如果此单元分数过低，会影响其他单元发挥作用。换句话说，就是让严重供大于求的单元的影响适当降低。而对于 $Score_i \geq 1$ 的情况则不作处理。

RHF 算法流程如下所示：

步骤 1： 计算 s_i，即 S 中各个语音单元出现的次数， i =1,2,…,n

步骤 2： 计算 F，即理想情况下 D 中各个语音单元平均次数

$$F = \frac{N}{M} \sum_{i=1}^{n} s_i$$

步骤 3： 设置 D 为空集，初始化 d_i 为 0，　i =1,2,…,n

设置阈值 $TH = aF \quad (a \geq 1)$

步骤 4： 依次从 S 中取出一个句子，令句子分数 *SentScore*=0

for （句中每个语音单元）

{

设当前语音单元对应的序号为 k

if （$d_k \geq TH$）　　// 本语音单元在 D 中总数已经超出阈值

{ 句子平均分数 *AverageScore=MIN_SCORE*，并跳转到**步骤 5**}

else

{

$$Score_k = \frac{TH - d_k}{s_k}$$

if （$Score_k$<1）　$NormScore_k = \frac{1}{1 - \log(Score_k)}$

else　$NormScore_k = Score_k$

$$Score_k = \log(NormScore_k)$$

}

当前句子总分 $SentScore = SentScore + Score_k$

}

$AverageScore = \frac{SentScore}{num}$　// num 为当前句子包含语音单元总数

步骤 5： if （ 检查的句子总数 *CheckNum* 没有达到指定的阈值 ）

{ 跳转到**步骤 4** }

清除 *CheckNum*

if （ *S* 中全部句子的分数 *AverageScore* 均为 *MIN_SCORE* ）

{ 调整阈值 *TH* 为 *TH*+*bF*（*b*>0），跳转到**步骤 4** }

找出检查过的句子中得分最高的一个，从 *S* 中删除，并放入 *D* 中

重新调整 s_i 和 d_i， i =1,2,…,n

步骤 6： if （*D* 中句子总数未达到 *N*） { 跳转到**步骤 4** }

else { 输出 *D*，算法结束 }

此算法还可以根据需要采用以下策略加以调整和细化，如：

- 设置同一个句子可被重复选择的最大次数，默认为 1
- 设置不受阈值限制的单元，如“的”，出现比例极高，可以被忽略
- 是否将句子分数取对数或归一化，以降低分数间的差异，避免个别单元占主导作用
- 对于含有稀有单元的句子是否直接选取
- 设置最大步长，即，考察多少个句子后必须选出一句，主要用于加速

2.2.3　鼓励低频单元

鼓励低频单元（ELF，Encouraging Low-Frequency units）算法强调的是对候选数据库中低频单元的鼓励。在抽取句子时，我们一方面希望选出的句子中语音单元尽量均衡，另一方面又希望每种单元至少能够出现若干次以上，这一点对模型训练来说也是非常重要的。因此，本文提出一种鼓励低频的策略，能够尽可能地将含有低频单元的句子选进来。

这是一种**替换式**的语料选择方法。首先，从候选语料库 *S* 中随机选择 *N* 个句子，抽取出来并放入 *D* 中，作为 *D* 的初始集合。然后，找出 *D* 中对低频单元贡献最小的句子 *t*，再找出 *S* 中对 *D* 中低频单元贡献最大的句子 *s*，将它们进行替换。重复此过程，直至不再有替换发生。这样，含有 *D* 中所需的低频单元的句子会被优先选择出来并替换 *D* 中对低频单元贡献最小的句子，从而有效地提高了 *D* 中低频单元的数量。

集合 *D* 的分数 *Score*(*D*) 按如下方式计算：

1．计算 *D* 中所有语音单元出现的次数 d_i，i =1,2,…,n

2．将 d_i 按增序排列，即，$d_1 \le d_2 \le ... \le d_n$

3．令向量$[d_1, d_2, ..., d_n]$为 $Score(D)$

下面给出集合 D 和 D’的分数比较准则。

$Score(D) = [d_1, d_2, ..., d_n]$，$Score(D') = [d_1', d_2', ..., d_n']$。$Score(D) = Score(D')$，当且仅当，

$d_1 < d_1'$，或者

$d_1 = d_1'$，且 $d_2 < d_2'$，或者

…

$d_1 = d_1'$，…，$d_{K-1} = d_{K-1}'$，且 $d_K < d_K'$

具体算法如下所示：

步骤 1：抽取 S 中前 N 个句子，放入 D

$S \leftarrow S - D$

步骤 2：从 D 中选出句子 t，使 $Score(D - \{t\})$ 最大

步骤 3：for （S 中每个句子 s）

if （ $Score((D - \{t\}) \cup \{s\}) > Score(D)$ ）

$D \leftarrow (D - \{t\}) \cup \{s\}$，$S \leftarrow (S - \{s\}) \cup \{t\}$，即，$t$ 与 s 进行交换

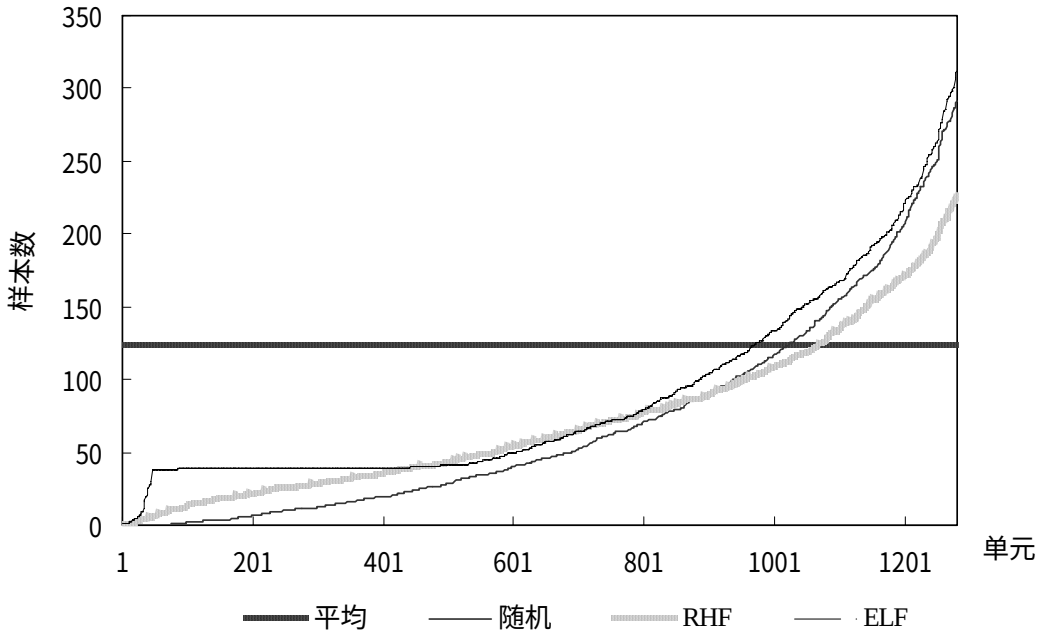
从 D 中选出句子 t，使 $Score(D - \{t\})$ 最大

如果有替换发生，则重复**步骤 3**；否则算法成功，结束。

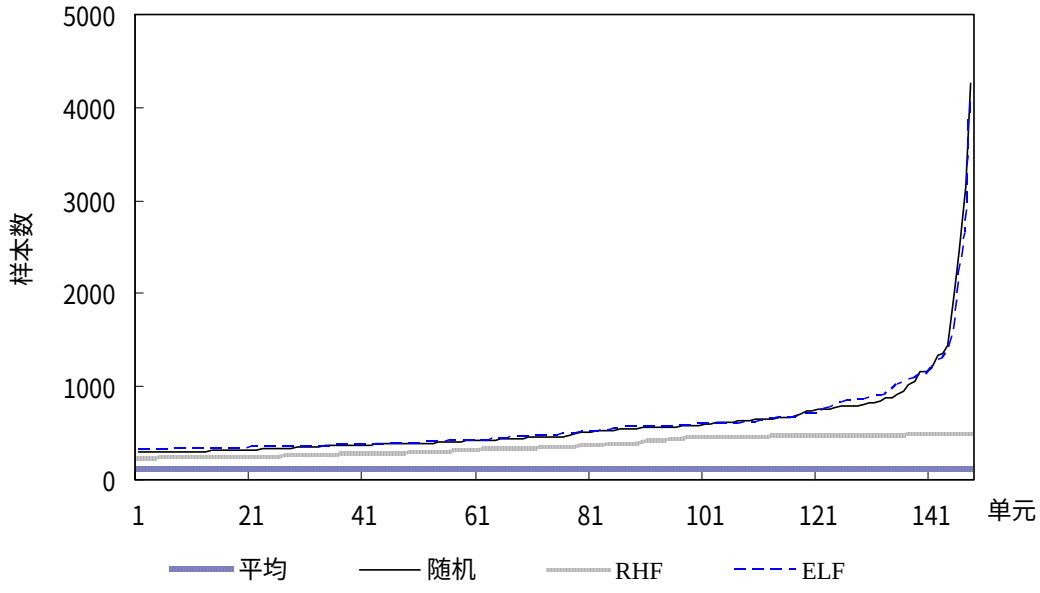
鼓励低频单元语料选择算法侧重将含有 D 中低频单元的句子选择进来，从而保证了 D 中包含更多的低频单元。为了保证一定的均衡度，也可以结合前面抑制高频单元策略中的方法，使用适当的阈值来对 D 中的高频单元进行适当的抑制。

2.2.4　结果与总结

本文实验中的候选语料库是从人民日报中选择出来的，共有 800,000 个句子，每个句子包含 15～20 个汉字。需要从中选择 6,000 个句子，尽量使得右相关声韵母基元均衡。图 2.1 中给出了实验统计结果。



（a）前 90%Di-IF 单元个数



（b）后 10%Di-IF 单元个数

图 2.1　目标语料库中各 Di-IF 单元个数分布

（平均表示理想均值；随机表示随机选择；

RHF 表示抑制高频单元；ELF 表示鼓励低频单元）

图 2.1 中给出了采用随机选择、理想情况（平均）、RHF 方法和 ELF 方法后，目标语料库中各个基元的个数分布情况。图中，横坐标表示的是不同的 Di-IF 单元，是按照它在目标集合中的个数从小到大排序的，纵坐标表示当前单元出现的总个数。（a）中显示的是前 90%的单元分布情况，从图中可以看出，随机选择既无法保证均衡度，也无法保证覆盖率，约有 70 个单元完全没有出现，而 RHF 和 ELF 方法则覆盖了所有的单元。还可以看出，ELF 在此时的优势比较明显，很多低频单元都满足三、四十以上的出现次数，达到了算法预期的目的。RHF 则要相对欠缺一点，但也在一定程度上保证了低频单元能够出现一定的次数。（b）中显示的是后 10%的单元分布情况，从中可以看到，RHF 算法可以有效地抑制高频单元，数目最多的单元也没有超过 500 个。而 ELF 方法和随机选择比较接近，不少高频单元的数目都达到了 1,000～4,500。因此，RHF 对高频单元的抑制，也达到了算法预期的目的，使得各个单元的均衡度大大加强了，同时也避免很多冗余单元被选进来。在声学模型训练中，单元覆盖率比均衡度更为重要，二者无法折衷时，优先考虑覆盖率。两种算法对比，ELF 更能保证低频单元的覆盖率，因此更适合于模型训练和自适应任务。

2.3　WDC语音库准备

吴方言背景普通话语料库包括两部分内容，一部分是朗读式语音库，一部分是自然发音式语音库。本节介绍此库的设计与采集。

2.3.1　语音库概况

吴方言背景普通话语音库包括 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。其口音等级、年龄大小、教育程度等都是按照预先的要求选择的。表 2.2 给出了库中说话人的年龄分布，40 岁以下与 41 岁以上的说话人比例基本相同。表 2.3 给出了库中说话人的教育程度分布。80%以上的说话人具有高中以上教育程度，近 20%说话人具有较低的教育程度，男女声采用了相同的比例。通过统计我们发现，教育程度较低的说话人往往口音很重，基本上，说话人的口音轻重与教育程度有着基本一致的关系。

表 2.2　吴方言背景普通话语音库说话人年龄分布

年 龄	男 声	女 声	总 计
26-40	27%	25%	52%
41-50	23%	25%	48%

表 2.3　吴方言背景普通话语音库说话人教育程度分布

教育程度	男 声	女 声	总 计
好（高中及以上）	41%	41%	82%
低（高中以下）	9%	9%	18%

图 2.2 列出了库中说话人普通话等级分布情况。等级的评定是由上海师范大学吴语专家给出的。按照普通话等级标准，分为三级六等，口音从轻到重依次为 1A、1B、2A、2B、3A 和 3B。

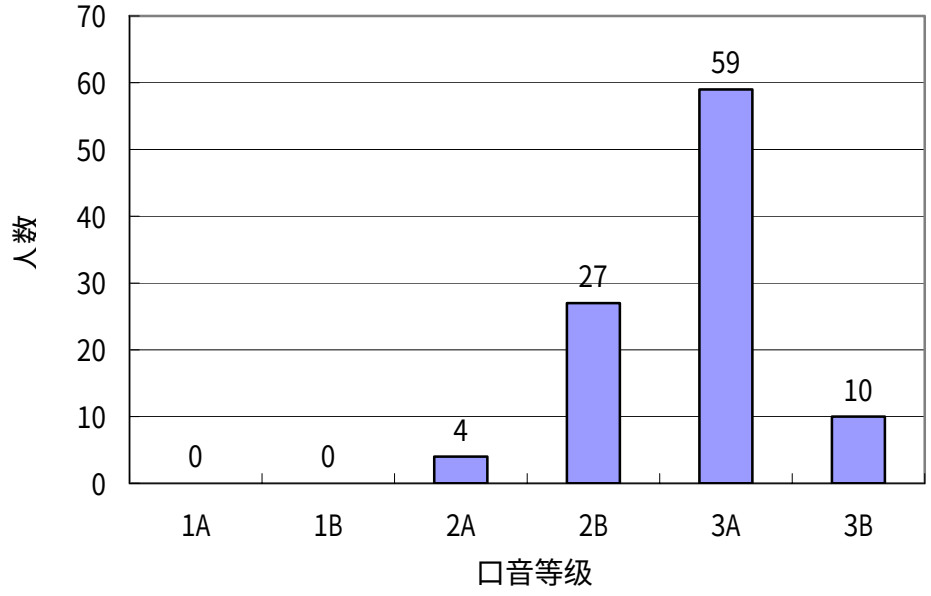


图 2.2　吴方言背景普通话语音库普通话等级

如图 2.2 所示，说话人口音大部分分布于 2B 和 3A。1A 和 1B 等级的说话人口音太轻，其发音相当标准，不在本文讨论的范围之内，因此没有选择这两个等级的说话人进行录音。

对于朗读式语音，每个说话人读 65 句话，共有 6,500 个句子。每人前 60 句为标准普通话语句，通过本文提出的均衡语料设计算法——鼓励低频单元（ELF）方法——从人民日报 800,000 个句子中自动抽取出来的。后 5 句是从网络、报刊上手工选择出来的，每个句子包含一个或多个在普通话中常用的吴方言词汇，如“晓得”（意同“知道”）等常用吴方言词汇。

对于自然发音式语音，我们采用对话的形式进行录制，一个标准普通话说话人与吴方言背景普通话说话人进行对话，只取后者的语音。为了使谈话内容更为丰富，覆盖更多的词汇和语言现象，本文设计了 5 个大的主题，以及若干小主题，让说话人自行选择，进行对话和录音。如表 2.4 所示。

表 2.4　自然发音式语音库对话主题设计

主题（Topic）	**子主题（Sub-Topics）**
体 育	泛谈，篮球，足球，乒乓球，网球，奥林匹克，……
政 经	泛谈，国际关系，恐怖活动，经济全球化，……
娱 乐	泛谈，电影，电视剧，喜剧，明星，春节联欢晚会，……
生 活	泛谈，家庭，工作，学习，教育，交通，出国，饮酒，吸烟，语言，友谊，爱情，购物，……
科 技	泛谈，计算机，网络，公司，克隆，……

每个说话人采集约 6～12 分钟语音，但仅选用前 3 分钟语音数据，其它剩余部分未作处理。

2.3.2　语音库标注方法

吴方言背景普通话语音库由社科院语言所负责手工标注和校对，包括汉字、拼音、声韵和杂项四层标注。如表 2.5 所示。

表 2.5　语音库四层标注信息

层 次	说　明
汉 字（HZ）	汉字标注
拼 音（PY）	拼音标注，包括标准发音的声调
声 韵（IF）	声韵标注，使用标准声韵母（PTH-IF）和吴方言声韵母（Wu-IF）集合进行标注，包括时间点和发音变化
杂 项（Misc）	非语音信息，如呼吸、咳嗽、不流畅等

此处着重介绍以下声韵层标注。SAMPA 是目前国际上通行的可机读语音键盘符号系统，它在语料库的标注中被广泛应用。在 SAMPA 基础上制定出了一套针对汉语普通话的语音标注符号系统，称为 SAMPA-C。进而将这个系统扩大，可以包括汉语各地方言[62][63][64]。有了声韵母集合与 SAMPA-C 符号集合的对照关系，就很容易将声韵母标注转换为 SAMPA-C 符号集合。

2.3.3　少量语音选取

从吴方言背景普通话语音库中分别选出 20 人数据，得到训练集（TrainSet）和测试集（TestSet）。训练集中的数据主要用于声学自适应、声韵映射规则学习等，因此，在选择这部分数据时，主要选择的是口音较重的说话人，而性别、年龄比例与原库基本保持一致。训练集和测试集中说话人基本情况分别如表 2.6 和表 2.7 所示。

表 2.6　训练集合（TrainSet）说话人信息（20 人）

编号	性别	普通话等级	编号	性别	普通话等级
001	男	3A	055	女	3A
003	男	3A	060	女	2B
004	男	3A	063	女	2B
013	男	3A	065	女	3A

续表 2.6　训练集合（TrainSet）说话人信息（20 人）

编号	性别	普通话等级	编号	性别	普通话等级
017	男	3A	069	女	2A
018	男	2B	081	女	3A
026	男	3A	085	女	3A
036	男	3A	090	女	2B
037	男	3A	094	女	3A
050	男	3A	095	女	2B

表 2.7　测试集合（TestSet）说话人信息（20 人）

编号	性别	普通话等级	编号	性别	普通话等级
032	男	3B	008	男	2A
035	男	3B	009	男	2B
043	男	3A	011	男	2B
046	男	3A	012	男	2B
047	男	3A	016	男	2B
053	女	3B	054	女	2B
059	女	3B	061	女	2B
076	女	3B	064	女	2B
098	女	3A	066	女	2A
099	女	3B	067	女	2A

表 2.8 中列出了训练集和测试集中的数据量。

表 2.8　训练集与测试集信息

说话方式	集　合	人　数	小时数	句子数
朗读式	训练集	20	1.4	970
	测试集	20	1.8	1,300
自然发音式	训练集	20	1.3	739
	测试集	20	1.7	1,100

从表中可以看出，对于朗读式和自然发音式语音，其训练集分别含有 1.4 小时和 1.3 小时的数据，不到 1,000 个句子，采集和标注这样量级的数据库，其成本是比较低而可接受的。这也是研究框架得以扩展应用于其它方言的基础之一。

2.4　本章小结

语音库在声学模型训练和自适应中占有十分重要的地位，因此，本章对语音库语料设计进行了研究，并提出了两种自动语料算法：抑制高频单元算法和鼓励低频单元算法。前者通过设置和调整阈值，并利用需求关系来对句子打分，从而有效地抑制了高频单元，尽可能地达到语音单元的均衡。而后者侧重鼓励低频单元，通过替换的方式，不断将含有低频单元的句子替换进来，从而有效地保证了目标语料库的覆盖率。其均衡度还可以通过加入其它策略，如阈值限制，在一定程度上得到保证。语料库设计中，语音单元的均衡度和覆盖率是两个重要的方面。事实上，由于句子中各个单元的相关性，这两者也是很难同时保证的，这就需要根据实际情况折衷考虑。本章提出的自动设计算法既可以用于标准普通话建模所需的大规模语料库设计，也可以用于方言背景普通话自适应所需的少量语音选取。二者对比，ELF 算法更适合于模型训练和自适应语料库设计。

本章还介绍了 WDC 语音库的设计、采集、标注与划分。为了满足相关研究的需要，WDC 语音库是按照一定规模来设计和采集的，采用的是本章提出的抑制低频单元算法（ELF）。同时，在社科院语言所指导下，此库由专业人员进行了精心的手工标注，可以方便地应用于吴方言背景普通话的各种相关研究中。

第 3 章　声学模型训练与自适应

3.1　本章引论

本章研究的是声学建模相关问题，包括标准普通话声学建模和方言背景普通话声学自适应。前者是基准模型，可被各种方言共享使用；后者是特定方言背景普通话声学模型，是前者经自适应得到的。

声学模型可以认为是语音识别的匹配模板，如果没有好的模板，便很难得到满意的识别结果，因此，它在语音识别系统中占有十分重要地位。从语音识别研究开始到现在，人们对声学建模的研究一直没有停止过，在每年的重要会议（如 ICASSP 等），以及重要期刊上，都会看到声学建模相关的专题或文章。同时，方言背景对发音的影响也主要表现在声学层，对声学模型训练和自适应的深入研究，将有助于建立高质量的标准普通话声学模型，进而得到高性能的方言背景普通话识别器。

在标准普通话声学建模研究中，本文采用的是基于决策树的上下文相关扩展声韵母建模方法，并提出了三种策略进一步优化模型；在吴方言背景普通话声学模型自适应中，本文研究了基于 base-form/surface-form 指导的 MLLR 自适应方法，并进行对比。

本章按照如下方式组织：3.1 节为本章引论；3.2 节介绍标准普通话声学模型训练方法和优化策略；3.3 节介绍吴方言背景普通话声学模型自适应方法；3.4 节为本章小结。

3.2　标准普通话声学模型训练

本节讨论标准普通话声学建模训练问题，它是汉语连续语音识别中的关键步骤之一。主要包括以下内容：（1）基于扩展声韵母的上下文相关声学建模方法；（2）声学模型优化策略。

用于非特定人、大词表连续语音识别的声学建模方法主要可以分为两类：基于概率统计模型的方法和基于人工神经元网络（ANN，Artificial Neural Network）的方法[65]。基于概率统计的建模方法在当前研究中仍占据着主导地位，

而其主流仍是 HMM 及其各种变形，如，链式隐马尔可夫模型，高斯混合模型，中心距离连续概率模型（CDCPM）[66]等。本文采用的就是连续概率密度 HMM 来描述声学模型。

根据汉语语音的特点，本文提出了扩展声韵母（XIF）识别基元，并针对上下文设计了相应的问题集，利用基于决策树的状态共享策略建立上下文相关声韵模型（Tri-XIF）。为了优化模型，本节还提出了三种策略用于改善标注、改进问题集和降低模型规模。

3.2.1　汉语常用识别基元

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）[67]，也可以是基于数据驱动方式的（例如，状态级基元 senone[68]）。同时，基元的选择也依赖于具体的应用，对于小型系统，如孤立词、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；而对于大型系统，如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，通常还要考虑上下文相关性，如常用的上下文相关音素基元（Triphone）。本文研究的是大词表连续语音识别，在选择基元时，考虑如下因素：

（1）**基元数目：**规模是否适当，是否便于建立上下文相关模型，来描述连续语音中的协同发音现象，等等。声学模型的规模与基元数目紧密相关，基元数目过大，则一般需要更多训练数据，且占用更多存储空间和计算时间，从而影响识别效率；基元数目太少，则不利于区分不同基元，从而影响识别性能。

（2）**基元的灵活性：**一个句子由一系列的词组成，而一个词又由一系列基元组成。不同的句子可以共享词，而不同的词又可以共享基元，同时，同一个词又可以由多个不同的基元序列组成（多发音词典），用以表征音变现象。因此，所选择的基元要有一定的灵活性，可以表示出这些变化。

（3）**相关语言学知识：**相关的语言学知识可以帮助我们在模型训练时进行参数共享，不仅可以对训练数据较少的基元给出适当的估计，同时也可以利用这些知识对没有训练数据的基元给出适当的参数。

基于以上因素，英语连续语音识别中常用的识别基元包括：词（Word）、音节（Syllable）、音素（Phone，Diphone，Triphone）[67]，以及状态级基元，如 senone[68]，其中最为常用的是 Triphone 基元。汉语连续语音识别中，常用的

基元包括：词、音节、声韵母（IF，Di-IF，Tri-IF）和音素（Phone，Diphone，Triphone）[67][69][70]等。基元对应的语音段长度从长到短依次为：词、音节、声韵、音素和 senone。一般地，基元对应的语音段越长，基元越稳定，但灵活性越差；对应的语音段越短，基元越灵活，但稳定性越差。

本文对于识别基元的研究，主要是基于“扩展”的思想。为了表示自然发音中的轻化、浊化等现象，有研究者提出了 GIF 基元[71][72]，进行精细建模，其本质上也是基于一种扩展的思想。这种扩展的思想的产生，主要源于当前基元对数据描述的不足。对于标准普通话，本文通过分析对比，选用声韵母作为识别基元，并针对标准声韵母基元的不足——上下文相关基元数目过大和插入错误过多——加入了零声母基元，提出了扩展声韵母基元（XIF，eXtended Initial/Final）。而对于吴方言背景普通话，由于同时受到标准普通话和方言背景的双重影响，仅用标准声韵母不足以表示其发音变化，因此，在标准声韵母基础上，引入了部分吴方言声韵母（Wu-IF），从而扩展成为吴方言背景普通话声韵母（WDC-IF）基元。这种扩展的方法同样可以应用于其它方言背景普通话。

汉语语音识别中常用的几种识别基元有：音节、音素和声韵，这包括上下文无关和相关建模。本节中，我们要对这几种常用基元进行分析对比。

（1）音节

汉语标准普通话中约有 400 多个无调音节和 1,300 多个有调音节[69]。在进行上下文无关的声学建模时，选用音节作为基元可以取得比较好的性能。但在连续语音识别中，音节间的协同发音现象比较严重，为了描述这种现象，需要进行上下文相关建模，而建立上下文相关模型对音节来说是比较困难的，因为基元数目会变得非常庞大，对训练和识别来说都是不适宜的。因此，一般使用音节作为基元时，大都进行上下文无关建模。也有一些研究者在研究音节的部分相关性，如，将音节模型划分为三部分，头、中、尾，头部考虑左相关，尾部考虑右相关，中部则认为受上下文影响较小，从而不考虑上下文相关性。同时，对上下文也进行适当的分类，从而减少了上下文相关基元总数。这种做法虽然在一定程度上解决了上下文相关问题，但不同音节之间的共享则被大大削弱了。

（2）音素

汉语标准普通话中有 35 个音素，如表 3.1 所示。音素基元在英语连续语音识别中得到了广泛的应用，并取得了很好的识别性能[73][74]。对于汉语，音素也是一个很好的选择。它的基元数目较少，便于建立上下文相关模型。但音素并

没有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。

表3.1　音素基元列表

辅音基元（22）	元音基元（13）
b, c, ch, d, f, g, h, j, k, l, m, n, ng, p, q, r, s, sh, t, x, z, zh	aI, a, Ie, eI, eN, e, Ci, CHi, Bi, oU, o, u, v

（3）声韵母

标准普通话中有约 59 个无调声韵母，如表 3.2 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：第一、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，因此，声韵更能反映汉语的特点；第二、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；第三、基元数目和语音段长度比较恰当。与音节比，便于建立上下文相关模型；与音素比，稳定性好。

表3.2　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

通过上述分析比较，我们可以看出，声韵母是汉语连续语音识别基元的最佳选择之一。事实上，在汉语连续语音识别中，标准声韵母基元也被很多研究者所采用[69][70][75]。

3.2.2　扩展声韵母基元

上节中对比分析了标准普通话连续语音识别中常用的三种基元——音节、音素、声韵母。通过分析我们可以得到这样的一个结论：在这三种基元中，声韵母是这三种基元中最适合汉语连续语音识别的基元。在 863 数据库上的实验结果也可以验证这一点（实验结果参见第 5 章）。

但是，标准的声韵母基元集合仍有其不足之处，从而影响了系统的整体识

别率。汉语中，有许多音节只有韵母部分，而没有声母部分，我们称其开始部分为零声母，如，“ a ”，“ an ”，“ wu ”，等。由于这些不带声母的音节的存在，导致了如下两方面的问题：

（1）由于这些音节没有声母，在进行上下文相关建模时，其上下文既可以是声母，也可以是韵母，因此，上下文相关基元数目会很大，约为 12 万个。基元数目庞大，则容易引起训练数据稀疏、基元混淆多、搜索网络大、搜索效率低等问题。

（2）在进行参数共享时，这些不带声母的音节，将被作为韵母与其它音节的韵母部分进行共享。例如，音节“ wu ”会与音节“ du ”的韵母/u/进行共享。而实际上，它们的语音段长度存在很大差异，发音有所不同。它们常被强制共享在一起，从而导致识别中不带声母的音节的插入和替代错误非常明显。如，“ yang ”常被识别为“ yi yang ”，“ du ”常被识别为“ du wu ”，等。

为了解决上述问题，本文将不带声母的音节的开头部分（零声母），定义为单独的基元，即零声母基元，提出了扩展声韵母基元（XIF）。本文共引入了 6 个零声母，它们是{_a，_o，_e，_y，_w，_v}。扩展的声韵母基元列表如表 3.3 所示，包括 27 个声母和 38 个韵母。表中，ii 和 iii 分别表示与{z，c，s}和{zh，ch，sh，r}相接的韵母/i/。加入零声母基元以后，声韵母的上下文关系变得更加规范，声母左边只能接韵母或静音，右边只能接韵母；韵母左边只能接声母，右边只能接声母或静音。这样，上下文相关基元数目从 12 万减少为 3 万左右，有效地缓解了训练过程中的数据稀疏等问题。同时，使用扩展的声韵母基元也有效地减少识别中的插入和替代错误，其性能也优于标准声韵母基元。

表3.3　扩展的声韵母基元列表

声母基元（27）	**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, _o, _e, _y, _w, _v	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

3.2.3　基于状态共享的上下文相关声韵母建模

在连续语音中，协同发音现象十分严重，因此，建立上下文相关模型来刻画协同发音现象是非常必要的。为解决上下文相关建模时的数据稀疏问题，本文采用基于决策树的状态共享策略。

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[76][77]。决策树是一个二叉树，每个结点都绑定着一个“ Yes/No ”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，其参数将被共享起来。它是基于数据驱动方法和基于知识方法的结合。与基于数据驱动方法相比，它能够对训练数据稀少的基元和没有训练样本的基元给出适当的参数估计。与基于知识的方法相比，它能够弥补专家知识不足带来的缺陷。

本节讨论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的几个主要问题：问题集的设计，状态共享策略，结点分裂和停止分裂准则，模型训练流程。

（1）问题集的设计

问题集就是供决策树构造使用的问题的集合。结点分裂时选中的那个问题，就与此结点绑定，从而决定哪些基元的哪些状态被共享起来。问题集的好坏会影响到上下文相关模型的性能。本文中使用的问题集是基于汉语语音学知识的[72][78][79][80]。根据这些先验知识，中心基元的上下文被划分为若干类，每一类作为一个问题。针对扩展声韵基元，设计了基于上下文分类的问题集。其中，作为问题集的声母基元类有 22 个，韵母基元类的问题有 39 个，分别如表 3.4 和表 3.5 所示。

表3.4　用于构建问题集的声母基元类

声母基元类	类中所包含的基元
响音（Sonorant）	{m, n, l}
塞音（Stop）	{b, d, g, p, t, k}
唇音（Labial）	{b, p, m}

续表3.4　用于构建问题集的声母基元类

声母基元类	类中所包含的基元
唇音（Labial2）	{b, p, m, f}
塞擦音（Affricate）	{z, zh, j, c, ch, q}
零声母（Zero1）	{_a, _o, _e, _y, _w, _v}
零声母（Zero2）	{_y, _w, _v}
零声母（Zero3）	{_a, _o, _e}
……	…….

表3.5　用于构建问题集的韵母基元类

韵母基元类	类中所包含的基元
前高（HighFront）	{i, u, v}
开口 n（Open_n）	{an, en}
开口 ng（Open_ng）	{ang, eng}
……	……

（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。通常，声韵母基元采用 3 个连续的状态进行描述，这三个状态分别反映了基元对应的语音段的起始部分、中间部分、结尾部分的情况。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，本文采用的是状态相关的共享策略。对于 Tri-XIF，共有 66 个中心基元，每个基元模型含有三个状态，要对每个中心基元的每个状态建立一棵独立的决策树，因此，共需要生成 66×3=198 棵状态树。

（3）结点分裂和停止分裂准则

在构建决策树时，首先要将所有可能共享的 HMM 状态放入一个状态共享池（State Pool）中，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。如图 3.1 所示。这是对中心基元为/an/的扩展声韵母基元的各状态建立决策树的示例。

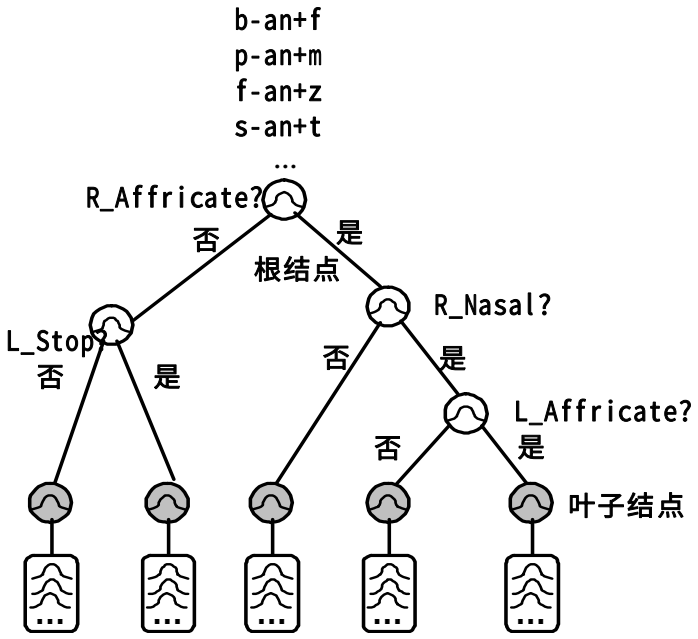


图 3.1 决策树结构图

本文中使用的分裂准则是最大似然准则，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[76]。决策树的停止分裂采用阈值进行控制。当分裂后的结点中训练样本数目少于一定数量时，或者，当本结点分裂后对数似然分数的增加小于一定的阈值时，停止分裂。

我们定义对数似然概率 $L(S)=\log P(X\mid S)$ 为结点 S 分裂的评估函数。其中 $X=\{X_1,X_2,\cdots,X_N\}$ 表示一个父结点总共包含 N 个样本。设 $X^1=\{X_1^1,X_2^1,\cdots,X_N^1\}$ 和 $X^2=\{X_1^2,X_2^2,\cdots,X_N^2\}$ 表示由父结点 X 分裂得到的两个子结点 X^1 和 X^2 所包含的样本，满足 $X=X^1\cup X^2$， $X^1\cap X^2=\varnothing$。父结点和两个子结点的评估函数的值分别表示为 L，L^1 和 L^2。则结点分裂后，似然分的增量为 $\Delta=L^1+L^2-L$。在每个叶子结点进行分裂的时候，我们从问题集中选择一个问题，然后根据此问题把结点分成两个子结点并计算增量 Δ，然后选择产生最大增量的问题作为此结

点绑定的问题，并根据此问题把结点分裂为两个子结点。当所有问题的增量都低于某个阈值的时候，即满足停止分裂准则时，停止分裂。

在具体的实现中，由于 $L(S)$ 不便于直接计算，一般采用如下的辅助函数作替换[77]：

$$Q(S)=\sum_{x_t}\sum_{s\in S}\gamma_s(x_t)\log N(x_t\mid\mu(S),\sum(S)) \qquad (3\text{-}1)$$

其中 $\gamma_s(x_t)$ 是观察矢量 x_t 在结点 S 上的后验概率。$N(\bullet|\mu,\Sigma)$是均值为 μ 和协方差矩阵为 Σ 的高斯密度函数。由于 $Q(S)$ 和 $L(S)$ 具有相同的单调性，也就是

$$Q(\hat{S})\geq Q(S)\Rightarrow L(\hat{S})\geq L(S) \qquad (3\text{-}2)$$

因此我们可以使用 $Q(S)$ 来作为评估函数。为了减少决策树分裂过程中的计算复杂度，分裂过程中每个结点上的样本分布都采用单高斯分布来描述。待决策树分裂结束后，再对每个叶子结点采用更加精确的混合高斯分布来描述。一般采用混合分裂的方式来增加混合数目，而最终的混合数目是根据模型的性能和规模综合决定的，且不同的基元可以使用不同的混合数目。

（4）模型训练流程

对于音节模型，每个基元使用自左往右的可单步跳转的 6 状态 HMM 来描述，即每个状态只能驻留或跳转到相邻的下一个状态。对于音素和声韵基元，使用 3 状态 HMM 来描述，如图 3.2 所示。

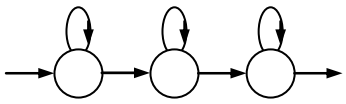


图 3.2 Triphone/Tri-XIF 基元模型拓扑结构

为了增加静音模型的灵活性，对其增加了两个弧，使其可以跳过第二个状态，且可以形成环路，从而可以表示长静音。如图 3.3 所示。

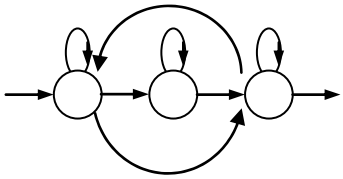


图 3.3 静音模型拓扑结构

本文采用的是中心分裂准则，由单混合依次分裂为 2 个、4 个、6 个、8 个或更多混合，可以根据训练数据的规模和任务需求来确定最终的状态数和混合数。通常地，每个状态含有 8～40 个混合。

上下文相关扩展声韵母模型（Tri-XIF）训练过程如下所述：

- 上下文无关声学模型训练（单混合 XIF）
- 通过复制中心基元模型，得到上下文相关声学模型，并进行迭代训练（单混合，非共享的 Tri-XIF）
- 基于决策树的状态共享（单混合，状态共享的 Tri-XIF）
- 混合分裂并迭代（多混合，状态共享的 Tri-XIF）

3.2.4　模型优化策略

基于决策树的状态共享策略利用了很多语音学知识，可以有效地将基于知识的方法和基于数据驱动的方法结合起来，因而能够较好地描述音节内及音节间的协同发音现象。同时，对于训练数据不充分，甚至没有训练数据的基元也可以利用生成的决策树给予合理的参数估计。因此，利用这种方法可以很好地解决数据稀疏的问题，从而得到高性能的声学模型。

但基于决策树的方法仍旧存在一些问题尚未解决，例如：

（1）初始模型粗糙

由于多混合的初始模型在进行决策树构建时计算量太大，所以采用了单混合的初始模型。单混合的初始模型比较粗糙，这就对决策树的生成造成了很大的影响，也就直接影响到了最终的模型性能。有的研究就是针对这种问题提出的，例如，使用多混合的初始模型，在结点分裂时，使用近似的方法来重估参数。但实验结果并不是很理想[81][82]。

（2）停止分裂准则的选取

现在常用的停止分裂准则是阈值方式，即分裂后样本数目小于一定的阈值或者对数似然分的增加小于一定的阈值时停止分裂。这种分裂方法需要确定适

当的阈值。有的研究针对这种问题提出一定的解决方案，例如，先进行分裂，直至分裂到每个状态占用一个叶子结点，然后采用回归的方式来选定最后的决策树。这种方法可以在一定程度上提高模型性能。

（3）局部最优

这里我们考察的是决策树的构建中存在的局部最优问题。在建立决策树时，虽然选定的问题是最大似然准则下的最优问题，但只保证了本层结点分裂后的似然分增加最大，而并没有考虑这些问题对其子结点分裂的影响。因此，采用这种方式只能达到局部最优的结果。针对这个问题，有的研究者提出了一些方法，尝试使问题的选择更加合理。例如，多层决策树的方法，即，多分裂几层，对比叶子结点的似然分增加情况，再选择要绑定的问题[77]。当然，要达到真正的全局最优几乎是不可能的。

针对决策树方法仍存在的一些问题，本文也提出了一些优化策略来改善模型的性能。如下：

（1）改善标注，提高初始模型性能

很多情况下，我们并没有完整而准确的基元标注信息，如，只有少量数据带有较准确的时间点信息；标注中常存在着由多音字或手误等造成的错误，使标注与发音不一致。标注问题会给模型训练带来一定的混淆。为此，我们利用训练好的 Tri-XIF 模型对训练数据进行强制对准（FA，Forced Alignment），从而获得新的声韵标注信息。用 **A** 表示对观察矢量 **X** 可能的时间划分（如状态划分，或对应于序列 **B** 中各基元起止时间划分），有

$$\begin{aligned}\hat{\mathbf{B}} &= \arg\max_{\mathbf{B}} P(\mathbf{X}\,|\,\mathbf{B})P(\mathbf{B}\,|\,\mathbf{W}) \\ &= \arg\max_{\mathbf{B}} \sum_{\mathbf{A}} P(\mathbf{X},\mathbf{A}\,|\,\mathbf{B})P(\mathbf{B}\,|\,\mathbf{W}) \\ &\approx \arg\max_{\mathbf{B},\mathbf{A}} P(\mathbf{X},\mathbf{A}\,|\,\mathbf{B})P(\mathbf{B}\,|\,\mathbf{W})\end{aligned} \tag{3-3}$$

对于给定的词序列 **W**（词标注信息），可以通过强制对准获得最佳声韵母序列 $\hat{\mathbf{B}}$，以同样的方式还可以得到序列 $\hat{\mathbf{B}}$ 对应的最佳时间划分 $\hat{\mathbf{A}}$。通过比较最佳解码序列 $\hat{\mathbf{B}}$ 和原始声韵标注信息，可以发现标注错误，如多音字相关的错误，其它手误等，还可以用来检查标注中遗漏的短静音等。同样地，我们得到最佳时间划分 $\hat{\mathbf{A}}$ 后，可以将其作为新的时间点标注，并用于初始模型训练。

为了检查标注中遗漏的短静音，本文在强制对准时修改了发音词典，为每个词条增加一个发音入口（即，在每个音节之后强制加入静音）。以音节发音词典为例，修改后的形式如下所示，

a　_a　a　　（标准发音入口）
a　_a　a　sil　　（**新增发音入口**）
b　b　a　　（标准发音入口）
b　b　a　sil　　（**新增发音入口**）
……

每行中，第一列为音节，后面为对应的扩展声韵母基元序列。在强制对准时，如果实际语音中含有静音段而没有被标记出来，则可以从解码结果发现被遗漏的静音并加以修正。

另外，也可以不修改发音词典，而直接在原始标注中的音节边界处强制插入静音标记，再进行强制对准。然后，根据解码结果中静音的时间长度，设定阈值，来确定是否为有效静音。这种方法对发现未标注的静音同样有效。

利用改善后的标注信息重新训练模型，从而为状态共享提供更好的初始模型，进而改善最终模型的性能。可以重复此过程，以得到更好的效果。这样就可以在解码结果中，从而减少标注错误对模型参数估计带来的负面影响。

（2）加强中间状态的共享程度

“停止分裂准则”决定了状态共享的方式和程度，一般都是采用阈值的方式。直观地，基元的不同状态受左右两边上下文影响程度是不同的。首状态受左边上下文影响较大，末状态受右边上下文影响大，中间状态要相对稳定。为了验证这个结论，我们对决策树中根结点绑定的问题（即，对分裂影响最大的问题）进行了统计分析，统计结果见图 3.4 和图 3.5。从图中可以看出，对于声母，三个状态都受韵母影响较大，但首状态受左边上下文影响较为明显，而中间状态和末状态受右边上下文影响较为显著。对于韵母，其首状态受左边上下文影响较大，末状态受右边上下文影响较大，中间状态居于两者之间。由此，根据上下文影响程度的不同，可以对不同状态选用不同的阈值，来降低模型的规模。

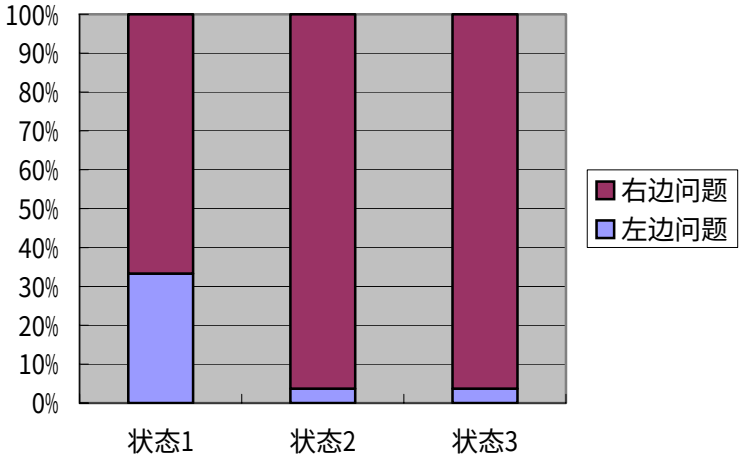


图 3.4 声母基元决策树根结点绑定问题比例

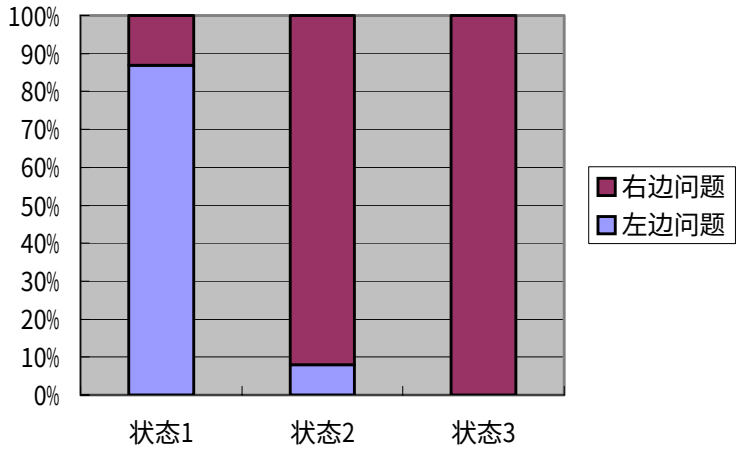


图 3.5 韵母基元决策树根结点绑定问题比例

（3）针对局部最优问题，提出了双向问题集，将原来的问题集加以扩展，用以改善决策树

问题集中的问题一般只提问单边的相关性，并且是基于已有的语音学知识的，其数量也不是很多。比如，“左边是否为响音{m，n，l}？”，或“右边是否开口韵母{an，en}？”，等，而并不同时考察两边的相关性。实际上，虽然基元两边的上下文对同一个状态的影响程度会有所不同，但均会对其产生一定的影响，所以，在进行结点分裂时同时考虑两边的相关性是合理的。另外，当决策树生成以后，从根结点出发，经过一系列非叶子结点，最终到达某个叶子结点。期间经过的各个结点上绑定的问题的交集，即决定了落入此叶子结点的基

元的上下文，对应的问题一般均涉及左右两边的问题。

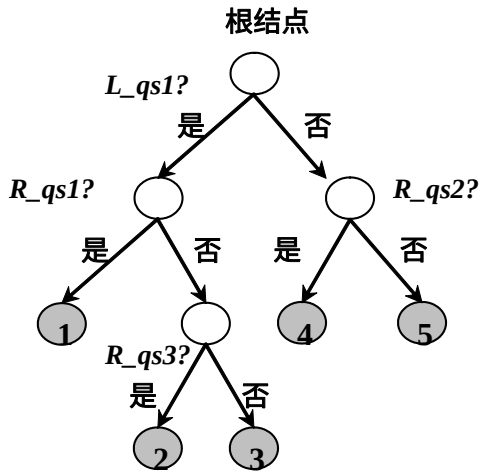


图 3.6 采用双向问题集的决策树示例

如图 3.6 所示，结点 1 对应的问题是$(L_qs1)\wedge(R_qs1)$，结点 2 对应的问题为$(L_qs1)\wedge(!R_qs1)\ \wedge(R_qs3)$，等。可以看出，这些问题就是左右两边问题的组合。由此，本文提出双向问题，首先将左边和右边的问题分别进行扩展，然后将它们进行组合，从而得到双向问题，并将其加入原来的问题集中。这样，决策树结点分裂时，可以同时考察中心基元两边的上下文。下面给出形式化描述。

定义问题集为集合**QS**，它包含所有可能的问题（每个问题是用分类上下文表示的，参见表 3.4 和表 3.5）。定义所有左相关的问题构成的集合为**LQS**，所有右相关的问题构成的集合为**RQS**。则有，

$$\mathbf{QS}=\mathbf{LQS}\cup\mathbf{RQS} \tag{3-4}$$

分别对集合 **LQS**，**RQS** 中的元素进行扩充，加入每个问题的补集，得到新的问题集**LQS'** 和**RQS'**。

$$\begin{cases}\mathbf{LQS'}=\mathbf{LQS}\cup\{\overline{qs}\mid qs\in\mathbf{LQS}\} & (3\text{-}5)\\ \mathbf{RQS'}=\mathbf{RQS}\cup\{\overline{qs}\mid qs\in\mathbf{RQS}\} & (3\text{-}6)\end{cases}$$

本文中，声母和韵母上下文是分别进行分类的，声母和韵母的上下文分类

没有交集。因此，此处 $\overline{qs}$ 的定义依赖于 qs 为声母还是韵母，为声母，则以全部声母上下文构成的左（或右）相关问题集为全集求补，为韵母，则以全部韵母上下文构成的左（或右）相关问题集为全集求补。

则双向问题集 **DQS** 定义为

$$\mathbf{DQS} = \{L\_qs \,\&\, R\_qs \mid L\_qs \in \mathbf{LQS}', R\_qs \in \mathbf{RQS}'\} \qquad (3\text{-}7)$$

表示同时考虑左右两边的问题 L_qs 和 R_qs。再加入原来的单向问题集 **QS**，便得到新的问题集 **QS**'，

$$\mathbf{QS}' = \mathbf{QS} \cup \mathbf{DQS} \qquad (3\text{-}8)$$

这样，新的问题集 **QS**' 包含了传统的单向问题，还包含了新引入的双向问题。因此可以在对结点提问时同时考察两边上下文的影响，而不只是单边的影响。

本文中，问题集中的问题是基于声韵母分类的，单向问题集中的问题总数约有 200 个（同一上下文分类既可以作为左边相关性问题，也可以作为右边相关性问题）。而以此方式生成的问题集 **QS**' 会很大，含有近万个问题。但实际使用时，其时间开销是完全可以接受的。

3.3　吴方言背景普通话声学模型自适应

3.3.1　WDC识别基元定义

方言背景普通话既不是标准普通话，也不是方言，而是介于两者之间，同时受到两者的影响。因此，其发音既有标准普通话的特点，又有方言的特点。以标准普通话为基准，受方言背景影响而产生的发音变化可以分为两类，一类是音节或声韵的一部分发生变化，一般称为“ sound change ”；一类是整个音节或声韵母发生替换，完全变为另一个音节或声韵母，类似于“ phone change ”，例如，受方言的影响，很多地方的人在说普通话时常将/n/读作/l/。

在自然发音的连续语音识别研究中，通常要考虑这两种发音变化。此时，发音变化多是由上下文、说话人的说话习惯等引起的。这时，仅仅使用标准普通话识别基元来描述就显得不够了，因此，就有研究[72]在进行发音变化建模时，针对发音变化将识别基元进行了扩展，进行精细建模研究。对于方言背景普通话，受同一种方言背景的影响，音节或声韵母的发音有着较为一致性的变化规

律。同样地，仅使用标准普通话声韵母或吴方言声韵母来表示其发音也是不够的。为此，本文将两个基元集合进行合并，并加以筛选，从而定义出吴方言背景普通话声韵基元（WDC-IF）。

见表 3.6。

表3.6　PTH-IF基元与WDC-IF基元定义

WDC-IF 基元	
PTH-IF 基元	**Wu-IF 集合**
a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, ii,iii, in, ing, iong, iou, o, ong, ou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn, j, k, l, m, n, b, c, ch, d, f, g, h, p, q, r, s, sh, t, x, z, zh	e>, eer, ie<, ieu, eu, io^, ioong, iuu, ni, o^, oong, voe, voong

表中，标准普通话声韵母集合简称为 PTH-IF，其中包含 21 个声母和 38 个韵母，共 59 个基元（参见表 3.2）。吴方言声韵集合简称为 Wu-IF，很多声韵与标准普通话发音相近，但有些是新的发音。本文根据吴方言背景普通话的发音特点和数据库中的声韵层标注，为吴方言背景普通话定义了新的声韵集合，称之为 WDC-IF，包括全部 PTH-IF 及 13 个常用的 Wu-IF。其中，新引入的 Wu-IF 基元是通过考察吴方言背景普通话语音标注而得到的，它们仍显著地保留在吴方言背景普通话发音中。WDC-IF 用以表示吴方言背景普通话实际发音的情况，而公式（1-12）中的序列 $\mathbf{S}^{\mathbf{WDC}}$ 便可以用 WDC-IF 序列来表示。

借鉴标准普通话的研究成果，我们也可以将 WDC-IF 集合进一步扩展，得到 WDC-XIF 集合。虽然语言学中也将零声母作为一种重要的语言现象来加以研究，但从一定程度上说，扩展声韵母主要应归于一种工程方法，而非语言学方法。因此，在本节实验中，我们主要使用标准声韵母进行研究，最后再给出基于扩展声韵母的实验结果。

3.3.2　声学模型自适应

声学模型自适应方法主要用于说话人自适应，是用少量语音去修正原有的说话人无关模型（SI），从而得到说话人相关模型（SD）。如前所述，受方言背景的影响，发音变化主要表现在声韵层。因此，本文中声学自适应是按照声

韵基元来进行的。自适应方法带来的另一好处是，可以在一定程度上解决信道带来的影响。

自适应方法研究在九十年代中期取得了很大的进展，这些方法被证明是非常有效的。自适应方法中应用最广的可以分为两类[83][84]：贝叶斯方法和基于变换的自适应方法。贝叶斯方法，即最大后验（MAP）估计方法[85][86][87]，其基本思路是直接根据贝叶斯准则，将初始模型信息作为先验分布的估计值纳入估计式中，与自适应数据中蕴含的依赖于说话人的信息结合，进而得到在形式上基于模型组合的自适应结果。变换自适应的基本思路是假设初始模型和当前说话人之间的差异可以用一组变换函数来描述，因此自适应过程即是利用自适应数据来估计对应的变换参数的过程。目前最为广泛使用的是：最大似然线性回归（MLLR，Maximum Likelihood Linear Regression）方法[88][89][90][91]。

对比两类方法，MLLR 更适合于数据量较少的情况，同时，MLLR 是一种变换方法，也符合本文研究声韵母变换的基本思想。因此，本文选用 MLLR 方法进行声学模型自适应。本文中自适应的目的是由标准普通话声学模型得到吴方言背景普通话声学模型，以便进行声学打分并解码得到最终的汉字序列。这里所谓的声学模型，就是公式（1-11）和（1-12）中声韵母序列 $\mathbf{S}^{\mathbf{PTH}}$ 和 $\mathbf{S}^{\mathbf{WDC}}$ 对应的模型。本文采用两种方式来监督自适应过程，分别基于 base-form 和 surface-form 声韵母标注，前者用于生成公式（1-11）中 $\mathbf{S}^{\mathbf{PTH}}$ 对应的模型，而后者用于生成公式（1-12）中 $\mathbf{S}^{\mathbf{WDC}}$ 对应的模型。

下面分别介绍这两种方式。首先，base-form 指的是标准发音，即，应该读作什么音。而 surface-form 表示的是实际发音，即，说话人实际读作了什么音。声韵母中，同一个 base-form 形式，可能会对应多个 surface-form 形式，同样地，同一个 surface-form 也可能来自不同的 base-form。如，标准发音/sh/可能发为/sh/或/s/两种不同的形式（“ 上海 ”，读为“ shang4 hai3 ”或“ sang4 hai3 ”），同样，实际发音/s/又可能来自两种不同的标准发音/sh/和/s/（如，实际发音“ si4 si2 ”，对应的标准发音可能是“ shi4 shi2 ”，即“ 事实 ”，也可能是“ si4 shi2 ”，即“ 四十 ”）。图 3.7 以/s/和/sh/为例，给出了标准发音与实际发音的关系。本文中，base-form 声韵标注由拼音层标注使用标准发音词典转换得到，表示的是声韵层的标准发音，也就是标准普通话中应该读作什么音；surface-form 声韵标注来自于声韵层手工标注，表示的是声韵层的实际发音。

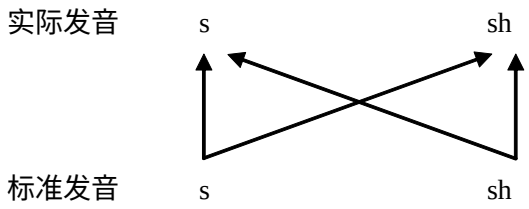
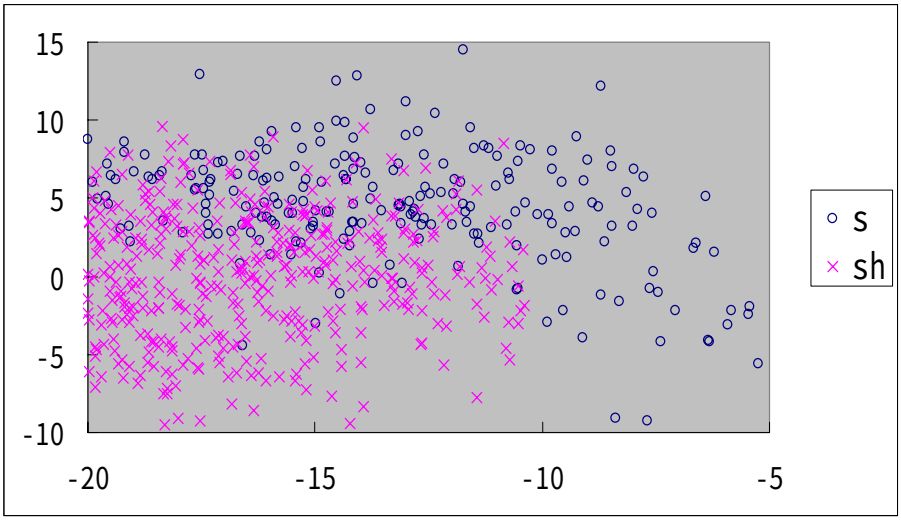
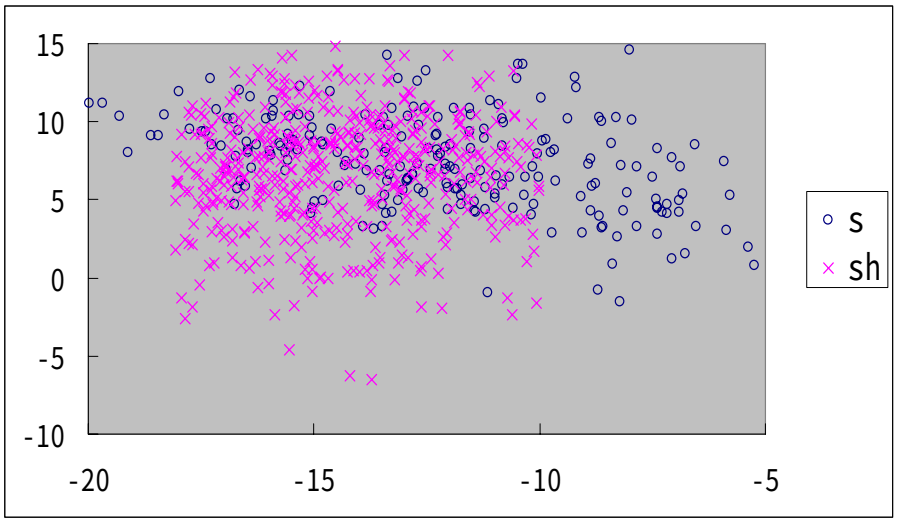


图 3.7 实际发音和标准发音对照关系

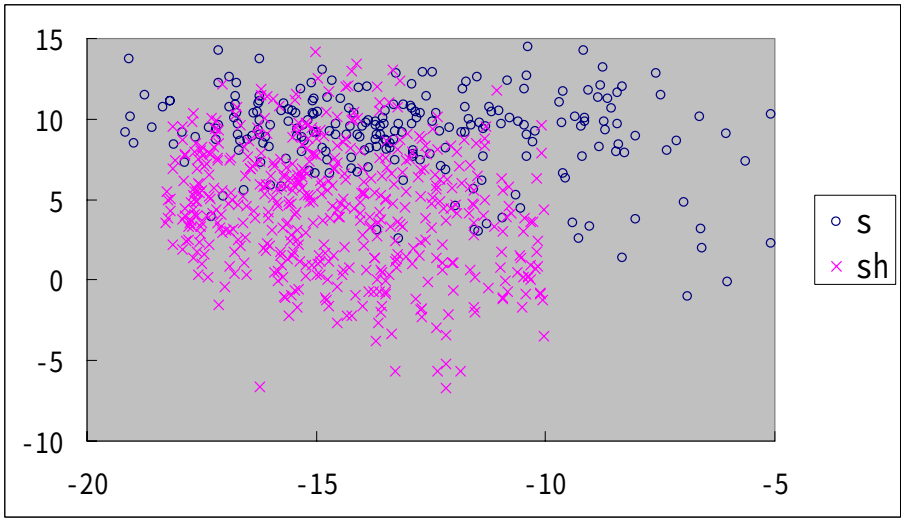
Base-form 监督下的自适应按照 base-form 标注来组织数据，即标准发音相同的数据将会被组织到一起，通过对模型参数的修改来涵盖发音变化。如，标准发音为/sh/，则无论实际读为/s/还是/sh/，其数据都会用来对声母/sh/进行自适应训练。这种情况下，声母/s/和/sh/的模型会变得更加相似。因此，在识别标准普通话或者口音较轻的吴方言背景普通话时可能带来更多错误。但在数据量充足的情况下，这种方式也许更为有效。与此相反，surface-form 监督下的自适应使用实际发音来组织数据，实际发音相同或相近的声韵母才会放到一起进行自适应，而不考虑其标准发音是否相同。如，不管标准发音为/sh/还是/s/，只要实际发音为/s/，它们将会被组织到一起进行自适应。此时，还要与多发音词典配合才能得到真正的结果。与 base-form 标注相比，由于实际发音相同或相近，用于自适应的数据内聚性好，更容易得到准确的声韵层的结果。图 3.8 展示了这两种方式的对模型参数的影响。



a） 普通话模型



b）基于 base-form 的 MLLR



c）基于 surface-form 的 MLLR

图 3.8 声母/s/和/sh/的声学模型第二个状态前两维均值向量分布图

a)普通话模型； b)基于 base-form 的 MLLR； c)基于 surface-form 的 MLLR

图 3.8 表示的是声母/s/和/sh/的声学模型第二个状态均值矢量分布情况，这里取的是前两维特征，横坐标和纵坐标分别表示第一和第二维特征。从中可以看出，基于 base-form 的 MLLR 中，两个基元的混淆度要高于基于 surface-form 的 MLLR。

进一步探讨一下这两种方法。对于基于 base-form 的自适应方法，其出发点是，无论说话人实际上读成了什么音，他们都是在试图读同一个标准发音，同

时，由于受到同种方言的影响，即使实际发音有所不同，也会体现出相当的一致性规律。因此，标准发音作为实际发音的纽带，可以用来组织数据，进行自适应。当数据量充足时，基于统计的方法可以较好地描述同一标准发音下的发音变化。而在数据量不足的情况下，则不易取得好的效果。因此，从某种程度上说，这是一种数据驱动的方式。另外，基于 base-form 的方法给出的是标准发音的识别结果，所以，即使只有单发音的词典，也可以得到最终的识别结果。而对于 surface-form 指导下的自适应方法，则使用实际发音来组织数据，即，无论对应的标准发音是什么，只要实际发音相同，就放到一起来进行自适应。很明显，这种方式下，数据的内聚性好，对数据量要求相对较小。但这种方式并不能直接得到识别结果，只能得到实际发音的结果。而同一个实际发音可能对应不同的标准发音，因此，还需要有多发音词典的配合，才能够得到最终的识别结果。事实上，这种方式更接近人耳识别。人耳首先识别的是实际发音，再利用上下文、声调等猜出说话人所说的是什么。比如，人们听上海人说“ zi1 si0 ”，实际听到的声母就是/z/和/s/，而不是/zh/和/sh/，但却可以猜出来是“ 知识 ”，而不是“ 自私 ”或“ 只是 ”。因此，这两种方式各有其道理所在。

3.4　本章小结

本章研究声学模型相关的问题，包括标准普通话声学模型训练和吴方言背景普通话声学模型自适应。

对于标准普通话声学建模，本文提出了基于扩展声韵母的上下文相关建模方法，并提出几种优化策略进一步优化模型。语音识别中，识别基元的选择是一个基本而重要的问题。本文根据汉语发音特点和方言背景普通话的需要，选用声韵母作为识别基元。对于标准普通话，通过分析标准声韵母基元的不足，将标准声韵母进行了扩展，提出了扩展声韵母基元 XIF。在进行上下文相关建模时，Tri-XIF 基元数据可以从 12 万左右降低至 3 万左右，对于模型训练和识别搜索都会带来好处。对于模型训练过程中存在的问题：（1）初始模型粗糙问题；（2）停止分裂准则的选取问题；（3）局部最优问题，采用了三种模型优化策略：（1）利用强制对准改善标注，提高初始模型性能；（2）加强中间状态的共享程度，降低模型规模；（3）针对局部最优问题，提出了双向问题集，将原来的问题集加以扩展，用以改善决策树。

而对于吴方言背景普通话，在标准普通话声韵母集合的基础上加入了少量吴方言声韵母，从而扩展成为 WDC-IF 基元，用以描述标准普通话和方言的双重影响。新引入的 Wu-IF 基元是从实际语音数据中统计得到的，此方法可以很方便地应用于其它方言。为了获得方言背景普通话声学模型，分别采用基于 base-form 和 surface-form 的 MLLR 自适应方法，将标准普通话声学模型变换为方言背景普通话声学模型，并对两种方法进行分析对比。

具体的实验结果将在第 5 章中给出。

第 4 章　多发音词典与常用方言词汇

4.1　本章引论

与方言背景的影响相比，发音变化（PV，Pronunciation Variation）是个更为宽泛的概念。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，如音素或声韵的替代、插入、删除错误等。受方言背景影响而产生的变化也主要发生在声学层，属于发音变化范畴，因此，许多发音建模（PM）的思想和方法可以用来解决方言带来的问题。但另一方面，由于方言的独特之处，其处理方式又有所不同。发音变化建模一般仅考虑声学层的变化，而方言的影响不仅体现在声学层，还发生在语言层，如用词的不同，语序的不同等；发音变化可能由多种原因造成，可能有规律，也可能没有规律，因此对发音变化规则的获取有一定的困难和盲目性，而在特定的方言背景影响下，就会有许多方言相关的知识可以使用，尽量避免规则获取的盲目性。

为了描述方言背景对声学层带来的影响，在上一章中介绍了基于 base-form 和 surface-form 的声学模型自适应方法，但对声学层的改进并不足以反映这些变化，还要有多发音词典的配合。通常，发音词典中的每个词条只有一个标准发音入口，多发音词典就是要为词条增加可能的发音入口，从而由单发音词典变为多发音词典。一方面，我们要尽可能地增加词条的发音入口，以覆盖更多的音变现象，另一方面，过多的发音入口又会引起更大的混淆，带来新的错误。因此，对多发音词典的研究主要涉及两个方面：如何产生多发音词典；如何对多发音词典进行剪枝。

多发音词典的产生可以是基于知识或数据驱动的方式。基于知识的方法多是利用音韵学知识来产生多发音词典[92][93]，而基于数据驱动的方式是从数据和标注中得到多发音词典[94][95][96][97]，其中最常用的方式就是通过将解码器识别结果与标注进行强制对准，进而研究音素的插入、删除、替代等发音变化规律。而在进行剪枝时，一般考虑发音入口在实际的语音数据中出现的频率[98][99][100]和发音之间的混淆程度[72][101]。

汉语中的音节具有独特的声韵母结构，这是区别于其它语言的，因此，本

文采用声韵母基元进行标准普通话声学建模。同时，受方言背景影响，声学层的发音变化主要表现在声韵层，且对于多数方言背景都具有这种特点。因此，在构建方言背景普通话识别器时，本文重点考察声韵层的变化规律，提出了基于声韵映射规则（IF-Mapping Rules）的多发音词典产生方式。这些映射规则可以从方言相关知识和少量数据中获取。在进行剪枝时，本文使用 Uni-gram 概率来衡量词条本身的重要程度，提出了基于累积一元概率（AUP）的剪枝准则。在此准则下，仅对满足 AUP 阈值的词条进行基于声韵映射规则的多发音扩展。

另外，除了声学层变化之外，方言背景普通话中还保留着许多常用方言词汇，在构建方言背景普通话识别器时，也要将其考虑进来。对于吴方言背景普通话，本文收集了约 200 个常用方言词汇，并赋予了适当的语言模型概率。

本章按照如下方式组织：4.1 节为本章引论；4.2 节介绍基于声韵映射的多发音词典；4.3 节介绍 AUP 剪枝准则；4.4 节为本章小结。

4.2　基于声韵映射的多发音词典

在普通话中，受方言背景影响而产生的发音变化主要表现在声韵层，因此，本文在生成多发音词典时所使用的方言相关知识是“声韵母映射规则”，简称为“声韵映射规则”。包括：

（1）上下文无关的普通话声韵母映射规则（PTH-IF Mapping）；

（2）上下文无关吴方言背景普通话声韵母映射规则（WDC-IF Mapping）；

（3）音节相关声韵母映射规则（Syllable Mapping）。

本文中的声韵映射规则可有两个知识源：一是专家知识，二是实际的语音数据。获得声韵映射规则以后，可以将标准普通话发音词典中对应的声韵母按照可能的映射规则进行变换，从而得到所有可能的发音序列。如，“知识”，在标准发音词典中表示为

知识　zh iii sh iii

而根据 PTH-IF 映射规则（zh->z；sh->s；iii->ii）得到的多发音词典为：

知识　zh iii sh iii

知识　z ii s ii

知识　zh iii s ii

知识　z ii sh iii

在构建多发音词典时，不合理的发音组合，如，/zh/和/ii/的组合等，已经被去掉了（这与/i/，/i/i 和/iii/的定义相关）。不同发音入口还可以带有一定的概率值，以表征此发音入口的重要程度。当不使用语言模型时，多发音词典中概率的应用是非常有效的[72]。但本文中加入了语言模型以后，实验表明，发音概率并没有真正发挥作用，而且在很大范围内波动都不会影响到实验结果，这主要是因为语言模型中的概率占据了主导作用。

本章后面各节中列出的映射规则是基于自然发音式语音、20 人训练集合的统计结果，对于朗读式语音，可以得到一个非常相似的结果，没有在此列出。

4.2.1　标准普通话声韵母（PTH-IF）映射规则

受方言背景影响而产生的声韵变化可以很容易地从实际语音中得到佐证，如，吴方言区来的同学，他们在说普通话时，经常区分不出前鼻音和后鼻音（如/en/和/eng/），卷舌音和非卷舌音（如/zh/和/z/）。本节研究上下文无关的标准声韵母映射规则，即，标准声韵母基元集合内的映射关系，不考虑吴方言相关的基元。

不考虑音节序列 **Y** 的影响，由公式（1-11）和（1-14）可得

$$\begin{aligned} P(\mathbf{S}\,|\,\mathbf{B},\mathbf{Y},\mathbf{W}) &\approx \prod_{i=1}^{ns} P(S_i^{\mathbf{PTH}}\,|\,B_i, W_{B_i}) \\ &= \prod_{i=1}^{ns} P_{W_{B_i}}(S_i^{\mathbf{PTH}}\,|\,B_i) \end{aligned} \tag{4-1}$$

其中 B_i 表示当前声韵母的标准发音（PTH-IF），$S_i^{\mathbf{PTH}}$ 表示 B_i 对应的实际发音（PTH-IF），W_{B_i} 为 B_i 所在的词。概率 $P_{W_{B_i}}(S_i^{\mathbf{PTH}}\,|\,B_i)$ 表示的就是 PTH-IF 映射规则 $B_i \Rightarrow S_i^{\mathbf{PTH}}$ 的概率，而 $P(\mathbf{S}\,|\,\mathbf{B},\mathbf{Y},\mathbf{W})$ 表示的是产生整个 **S** 序列的概率。W_{B_i} 对多发音词典的影响在本节暂时不予考虑，它的作用将在 4.3 节中进行讨论。

表 4.1 列出了由专家知识提供的标准普通话声韵母映射规则[102][103][104]。这些规则对应于公式中的 $P_{W_{B_i}}(S_i^{\mathbf{PTH}}\,|\,B_i)$，在没有给定概率的情况下，默认为 1。

表4.1　基于语言学专家知识的PTH-IF映射规则

标准发音	带口音的发音	标准发音	带口音的发音
zh	z	eng	en
z	zh	en	eng
ch	c	ing	in
c	ch	in	ing
sh	s	r	l
s	sh		

微软在此方面进行了研究，其中有关发音变化规则方面的工作给出的是几十组音节映射关系，实际上与表 4.1 中列出的映射规则是一致的[31]。

上下文无关 PTH-IF 映射规则还可以从数据标注中进行学习。表 4.2 中给出了从自然发音训练集合标注中学习得到的规则，其中包括到自身的映射。

表4.2　基于语言学专家知识的PTH-IF映射规则

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	88.57	iao	iao	87.69	r	l	22.05
ai	ai	91.29	iao	e	5.03	s	s	84.38
an	an	91.73	ie	ie	94.16	s	sh	10.00
ang	ang	91.58	ii	ii	83.33	sh	s	54.17
ao	ao	94.09	iii	ii	50.69	sh	sh	32.18
b	b	91.75	iii	iii	35.10	t	t	90.64
c	c	85.83	in	in	57.85	u	u	92.72
ch	c	81.63	in	ing	35.12	ua	ua	88.98
ch	ch	11.56	ing	ing	80.33	uai	uai	78.72
d	d	91.37	ing	in	10.77	uan	uan	91.40
e	e	91.99	iong	iong	94.00	uang	uang	86.21
ei	ei	94.12	iou	iou	90.44	uei	uei	92.43

续表4.2　基于语言学专家知识的PTH-IF映射规则

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(%)
en	en	73.90	j	j	91.38	uen	uen	73.33
en	eng	17.63	k	k	94.55	uo	uo	92.90
eng	eng	64.23	l	l	93.32	v	v	90.20
eng	en	29.67	m	m	96.19	van	van	84.47
er	er	96.36	n	n	93.85	ve	ve	90.08
f	f	90.69	ng	ng	93.02	vn	vn	74.29
g	g	91.39	o	o	84.00	x	x	92.22
h	h	86.23	ong	ong	91.20	z	z	86.50
i	i	92.17	ou	ou	84.20	zh	z	73.21
ia	ia	93.94	p	p	93.99	zh	zh	16.60
ian	ian	94.01	q	q	94.53			
iang	iang	89.66	r	r	65.75			

表 4.2 中，每三列为一组，表示了标准声韵母集合内的映射关系。第一列为标准声韵母，第二列为其对应的实际发音，第三列给出这种对应关系对于特定的标准声韵母的概率。这里的映射规则是通过统计自然发音的 20 人训练集合的标注得到的。标准声韵母标注表示的是语音数据对应的标准声韵母发音，而实际发音表示的是说话人实际上的读音（原始标注中包含的吴语相关的声韵母被映射到了对应的标准声韵母中，因此这里给出的实际发音全部在标准声韵母集合内），通过对这两种标注层的强制对准（最大匹配准则），可以得到以上映射规则。这里的概率表示同一个标准声韵母对应的各种发音变化的比例，可以通过公式（4-2）来计算，而式中的 $Prob_{S|B}$ 可以替代公式（4-1）中的 $P_{w_{B_i}}(S_i^{\mathbf{PTH}} \mid B_i)$，此处没有考虑所在词的影响。低于一定次数或比例的映射规则可能是由于干扰或标注错误引起的，它们是不可靠的，因此在选取规则时不予考虑。也正是由于一些低概率的映射被忽略，所以上表中同一个标准发音对应的各个实际发音的概率之和并不严格为 1。

$$Prob_{S|B} = P(S \mid B) = \frac{\#(S,B)}{\#(B)} \times 100\% \qquad (4\text{-}2)$$

公式（4-2）中，S 表示实际发音，B 表示标准发音。可以看出，表 4.1 中专家知识提供的映射规则是表 4.2 的一个子集。也就是说，从数据中统计得到的映射关系已经涵盖了表 4.1 中所有的专家知识。

4.2.2　方言背景普通话声韵母（WDC-IF）映射规则

说话人的实际发音中，除了包含标准声韵母外，还有一些吴语相关的声韵母，如/io/（接近标准普通话中的/iao/）。因此，使用标准声韵母集合来表示吴方言背景普通话发音是不够的。为此，本文引入了 WDC-IF 集合，并考察标准普通话声韵母到吴方言背景普通话声韵母集合映射规则，即，PTH-IF 到 WDC-IF 的映射规则。经统计，共有 13 个吴语相关的声韵母（Wu-IF）在训练集标注中达到一定的数量而被保留下来。

表 4.3 给出了从自然发音训练集合中统计得到的 PTH-IF 到 WDC-IF 的映射关系，包括 PTH-IF 到 PTH-IF 的映射，以及 PTH-IF 到部分 Wu-IF 的映射。而 WDC-IF 正是由 PTH-IF 和少量 Wu-IF 组成的。

表4.3　从自然发音训练集合中学习得到的PTH-IF到WDC-IF的映射关系

集内/集外映射	PTH-IF	WDC-IF	概率(%)
PTH-IF 集内映射（10 个）	en	eng	18.31
	eng	en	30.08
	iii	ii	51.67
	in	ing	37.19
	ing	in	11.24
	r	l	22.83
	s	sh	10.00
	ch	c	81.29
	sh	s	50.38
	zh	z	74.59

续表4.3　从自然发音训练集合中学习得到的PTH-IF到WDC-IF的映射关系

集内/集外映射	PTH-IF	WDC-IF	概率(%)
PTH-IF 集外映射（13 个）	ai	e>	8.51
	ao	o^	39.48
	er	eer	65.45
	iao	io^	35.18
	ie	ie<	26.80
	iong	ioong	48.00
	iou	iuu	30.48
	iou	ieu	24.94
	n	ni	9.33
	ong	oong	46.67
	ou	eu	51.53
	ve	voe	47.93
	vn	voong	22.86

总的来说，不包括自身映射，一共有 10 个 PTH-IF 集内映射（从 PTH-IF 到 PTH-IF 的映射）和 13 个集外映射（从 PTH-IF 到 Wu-IF 的映射），包括 PTH-IF 到{e>, o^, eer, io^, ie<, ioong, iuu, ieu, ni, oong, eu, voe，voong} 的映射，其中/voong/因为数据量不足而从朗读式识别器中删除了。

类似于公式（4-1），忽略音节对发音变化的影响，可得

$$\begin{aligned} P(\mathbf{S}\,|\,\mathbf{B},\mathbf{Y},\mathbf{W}) &\approx \prod_{i=1}^{ns} P(S_i^{\mathbf{WDC}}\,|\,B_i, W_{B_i}) \\ &= \prod_{i=1}^{ns} P_{W_{B_i}}(S_i^{\mathbf{WDC}}\,|\,B_i) \end{aligned} \tag{4-3}$$

上式表示的就是上下文无关 WDC-IF 映射规则产生的多发音序列 **S** 的概率。在进行模型自适应时，由于新引入的 WDC-IF 基元没有初始模型，其模型参数通过复制对应的 PTH-IF 参数而获得。如/e>/在不同上下文下的模型，是通过复

制/ai/对应的上下文的模型而得到的。然后，再使用相应的语音数据进行自适应。

4.2.3　音节相关声韵母映射规则

本节讨论上下文相关的吴方言背景普通话声韵（WDC-IF）映射规则，这里考虑的上下文为“音节”。实际上，我们在讨论上下文无关的映射规则时，也部分地考虑了规则的上下文相关性。例如，音节“shi”的声母/sh/映射为/s/时，其对应的韵母必须由/iii/变为/ii/。我们希望通过统计音节相关的映射关系，能够得到更为精确的规则。

类似于公式（4-1），但考虑音节对发音变化的影响，可得

$$\begin{aligned} P(\mathbf{S}\mid\mathbf{B},\mathbf{Y},\mathbf{W}) &\approx \prod_{i=1}^{ns} P(S_i^{\mathbf{WDC}}\mid B_i, Y_{B_i}, W_{B_i}) \\ &= \prod_{i=1}^{ns} P_{w_{B_i}}(S_i^{\mathbf{WDC}}\mid B_i, Y_{B_i}) \end{aligned} \tag{4-4}$$

其中音节相关的映射规则 $P_{w_{B_i}}(S_i^{\mathbf{WDC}}\mid B_i, Y_{B_i})$ 可以通过公式（4-5）中的 $Prob_{S|B,Y_B}$ 进行估算，即

$$Prob_{S|B,Y_B} = Prob(S\mid B,Y_B) = \frac{\#(S,B,Y_B)}{\#(B,Y_B)} \times 100\% \tag{4-5}$$

公式（4-5）中，只有属于同一个音节 Y_B 的映射才被统计进来。我们的实验结果表明，当使用音节相关的映射规则时，识别结果明显变差了。分析原因，主要是因为使用的数据量较少，统计数据变得不可靠了。但在一些方言中，确实存在这种词/音节相关的发音变化规律，如四川方言中，拼音“guo”在不同的上下文中发音是不同的，如“锅”中读做“guo”，但在“中国”中，读作“guai”。这样的问题仍是值得关注的。

4.3　发音词典剪枝准则

我们知道，单发音词典无法表示发音变化现象，尤其对于自然发音的语音。但另一方面，一个完全扩展的多发音词典虽然会对发音变化的描述有所帮助，但也会带来更多的混淆。为此，本文提出基于累积一元概率（AUP）的多发

音剪枝准则，来平衡这种矛盾。这里的一元概率指的是词在语言模型训练文本中出现的概率，对应语言模型中的 Uni-gram，它表征了词条本身的重要程度。而前面公式中的W_{B_i}项正是体现了这一点。如，公式（4-1）中的$P_{w_{B_i}}(S_i^{\mathrm{PTH}} \mid B_i)$，其中$W_{B_i}$的作用就是，依据 AUP 准则确定是否应用此声韵映射规则。

AUP 剪枝准则如下：

（1）只有 Uni-gram 概率较高的词才会给出多发音；

（2）对于 Uni-gram 概率较低的词，将选用单发音，即，标准普通话发音。

首先将词条按 Uni-gram 概率降序排序，然后将它们依次累加，得到累积概率值。再根据事先设定的阈值，将词表分为两部分。超过阈值的词条属于高频词，采用声韵映射规则生成多发音入口；没有超过阈值的词条则不进行多发音扩展，仅保留单发音入口。

词	概率(降序)	累积概率	
		0.000	
</s>	0.10782136	0.108	
的	0.03608752	0.144	
你	0.02161165	0.194	
是	0.01907339	0.213	
…			
标准	**0.00005742**	0.899	← 实际累积概率最小值
…	**0.00005742**		
团	**0.00005742**	**0.900**	← 需要的累积概率
…	**0.00005742**		
最多	**0.00005742**	0.901	← 实际累积概率最大值
…			
鲫鱼	0.00000124	1.000-	
黛	0.00000124	1.000-	

多发音扩展

标准单发音

图 4.1 多发音扩展准则

图 4.1 中对基于 AUP 的多发音扩展准则给出了详细的说明。图中表格里，第一列为词，是按照其一元概率降序排列的。</s>是句子结束符号，这里可以忽略它。第二列为本词所对应的一元概率，第三列为到当前词为止的累积概率。假定我们选定 0.90 为 AUP 的阈值，即图中“需要的累积概率”，一般情况下，此概率对应的词存在很多同概率的词（如图中阴影区域所示），所以可以选择实际累积概率的最小值或者最大值作为实际阈值。本文中选用最大值作为阈值，则一元概率超过此阈值的词进行多发音扩展，而其它词则仅使用单发音。

表 4.4 给出了自然发音式累积一元概率分布情况。

表4.4　自然发音式累积一元概率分布情况

需要的累积概率	实　际		
	累积概率	一元对数概率阈值	涵盖的词数
0.80	0.80018502	-3.594763	416
0.90	0.90050580	-4.240960	1,292
0.92	0.92081776	-4.419749	1,735
0.94	0.94120825	-4.624082	2,427
0.96	0.96064531	-4.867548	3,543
0.98	0.98000226	-5.250755	5,838
1.00	1.00000000	无	15,724

表 4.4 中，第一列为“需要的累积概率”，其它列为实际的分布信息，包括“累积概率”，“对数概率阈值”，“涵盖的词数”。以“需要的累积概率”等于 0.90 为为例，则对应的“一元对数概率阈值”为-3.594763（与图 4.1 中的一元概率值是对应的），超过此阈值的词有 1,292 个，它们的“实际累积概率”为 0.90050580。则有 1,292 个词将进行多发音扩展，而剩余的词则只有标准单发音。

本文还会与另外一些单发音或多发音词典进行对比，如标准单发音词典（STD，等价于 AUP=0%），最可能发音（MLP，Most Likely Pronunciation）

[106][107]，以及它们的组合，STD＋MLP。一般地，MLP 已经可以取得与多发音词典非常相近的结果，也被广泛采用。

4.4　常用方言词汇和语言模型

本文采用基于词的 Bi-gram 的统计语言模型[52][108][109][110][111]。首先，本文收集了约 200 个常用方言词汇，这些词汇属于吴方言词汇，但仍广泛应用于吴方言背景普通话中。然后，对选出的常用吴方言词汇给出适当的语言模型概率。这些方法都可以应用于其它方言。

4.4.1　常用吴方言词汇的收集

有吴语专家提到，标准普通话和吴方言之间词汇的相似度大约为 60～70%，且有不少吴方言词汇仍被广泛应用于吴方言背景普通话中，尤其是对于口音较重的说话人。这些信息提示我们，应该收集吴方言背景普通话中常用的吴方言词汇，并将其加入词表。本文根据一些对吴方言词汇的研究著作，以及在吴语专家的帮助下，收集了约 200 个常用吴方言词汇[1][2][103][104][105]，比如：“ 晓得 ”（知道），“ 两样 ”（不同）等。

常用吴方言词汇的**选择准则**是：词汇要在典型的吴语区（如上海，温州，苏州等）通用，且不同于对应的标准普通话词汇。然后再由吴语专家进行确认。这些词汇基本上都有同义的标准普通话词汇相对应，但也有个别词汇没有特别明确的意思，如吴方言背景普通话口语中常用到的“ 一般性 ”，可以近似认为是“ 一般来说 ”的意思，也可以认为仅仅作为口语插入语。

表4.5　常用吴方言词汇举例

吴方言词汇	标准普通话词汇	吴方言词汇	标准普通话词汇
晓得	知道	莲花	荷花
日头	太阳	横竖	反正
以前	从前	自来火	火柴
雄马	公马	鼻头	鼻子
好天	晴天	……	……

4.4.2　常用吴方言词汇的概率估计

由于朗读式与自然发音式语音差异很大，一般需要为它们分别定义词表，选择重新训练或自适应语料。本文为自然发音式和朗读式识别器分别定义了 15k 大小的词表。语言模型的训练或自适应可以通过收集方言背景普通话相关的文本语料来进行[52]。对于自然发音式语音，直接将常用吴方言词汇加入词表，并利用一些自然发音语料文本进行模型的重新训练。对于朗读式语音，首先使用标准词表和标准普通话文本训练语言模型，再加入常用吴方言词汇，并对这些词的 Uni-gram 和 Bi-gram 概率进行重估。重估时，如果方言词汇的同义词在词表中，则认为它们是等价的，并使用其对应的概率；如果对应的同义词不在词表中，则选择同类词进行概率的估计。之后，再对语言模型的概率进行归一化处理。

4.5　本章小结

依据汉语方言特点，本章提出并介绍了基于声韵映射规则的多发音扩展方法，基于 AUP 的多发音词典剪枝准则，以及常用吴方言词汇的收集方法和概率估计等。

这里没有进行解码器相关的讨论。在本文中采用的是一遍解码的方式，对于两遍解码策略，绪论中曾有所介绍。在进行两遍解码时，首先可以得到一个音节/声韵网络，然后再利用方言相关知识进行重打分。我们在关键词识别系统中进行了尝试（参见作者在 ISCSLP2004 中的文章），在未来的工作中，会进一步研究方言背景普通话语音识别中的两遍解码和重打分策略。

第 5 章　实验结果与分析

5.1　本章引论

本章介绍实验结果，主要包括两部分内容，一是标准普通话识别器的构建，主要介绍声学建模相关的实验结果，包括基于扩展声韵母的上下文相关模型性能，与其它常用识别基元的对比，以及模型优化策略等，二是所提出的框架应用于吴方言背景普通话的实验结果，用于验证本文提出的研究方法。

本文采用了一些较为通用的建模工具，如 HTK 工具包[112]，SRILM 工具包[113]等来辅助实验，以验证本文的研究思路。这些工具在国际上得到普遍认可与广泛应用，已经成为语音识别研究的标准平台之一。同时，本文研究的方法与思路是独立于工具的，其有效性和可扩展性不会受到它们的影响。

本章按照如下方式安排。本节为引论，5.2 节介绍标准普通话识别器相关结果，5.3 节介绍吴方言背景普通话识别器实验结果。

5.2　标准普通话声学建模

本节介绍标准普通话识别器的构建，主要是声学建模相关的实验结果。本文提出了基于扩展声韵母的上下文相关建模方法，并提出若干优化策略来改善模型。下面将给出这些方法的实验结果。

5.2.1　识别基元对比实验

本节给出了在 863 标准语音数据库下几种识别基元的测试结果。863 语音库是由我国 863 委员会组织，国家 863 计划支持，社科院语言所和中国科技大学合作完成的朗读式连续语音数据库[58]，是一个经过精心设计、采集、标注，并在我国语音识别、合成研究领域广泛应用的标准语音库。库中文本共有 1,560 个不同的句子及若干词汇，分为 A，B，C，D 四组，其中 A、B、C 三组为长句，每组约含 520 个句子，D 组为词汇。本文使用 863 数据库中的男声数据进行实验，其中 70 人男声数据定义为训练集合，剩余 10 人男声数据定义为测试集合。在保证各组句子基本均衡的前提下，数据库的划分是随机进行的。

声学模型使用连续 HMM 来描述，其中，音素和声韵基元采用 3 个连续的状态，而音节基元时间跨度相对较长，使用 6 个连续的状态。每个状态只能驻留或跳转到相邻的下一个状态。为了使静音模型更加灵活，增加了 1 至 3 状态之间的跳转弧，因而可以很好地匹配长短静音。声学特征采用的是 MFCC[114][115]，包括 14 维倒谱特征，以及一阶差分Δ和二阶差分ΔΔ，共 42 维。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，Cepstral Mean Normalization）[116][117]。识别时不使用语言模型，仅对比声学模型的性能，搜索网络为 400 多个音节的平行循环网络，评测结果用音节错误率（SER，Syllable Error Rate）来表示。上下文无关、相关模型的测试结果分别见表 5.1 和 5.2。表 5.1 列出了上下文无关模型的识别结果。

表 5.1　上下文无关基元音节错误率（SER%）

基元类别	1 混合	2 混合	4 混合	8 混合
音素（phone）	74.93	66.17	59.65	53.75
标准声韵母（IF）	61.86	54.59	48.26	43.14
扩展声韵母（XIF）	60.44	52.84	46.43	41.85
音节（Syllable）	44.30	38.14	32.71	29.07

从表 5.1 中可以看出，音节模型的识别率远高于音素和声韵基元模型，这是因为音节模型使用了更多的参数来描述模型，音节内部的相关性已经得到了很好的描述。同时，我们可以看出，扩展的声韵母基元的性能也明显优于标准声韵母和音素基元。

表 5.2 表示的是上下文相关模型性能。为了便于对比，表中也加入了上下文无关音节（Syllable）的结果。从表中可以看出，对于上下文相关模型，其性能要远远好于对应的上下文无关模型，也明显优于无关模型中性能最好的音节模型。对于 8 混合模型，Tri-XIF、Tri-IF、Triphone 与音节模型相比，其音节误识率分别降低了 39.04%，30.07%和 24.42%。这主要是由于引入了上下文相关建模和状态共享策略的缘故。同时，在模型规模相当的情况下，本文定义的扩展声韵母基元也优于其它常用基元。对于 8 混合模型，Tri-XIF 与 Tri-IF 和 Triphone

相比，其音节误识率分别降低了 12.84%和 19.34%。

表 5.2　上下文相关基元音节错误率（SER%）

基元类别	1 混合	2 混合	4 混合	8 混合
音节（Syllable）	44.30	38.14	32.71	29.07
音素（Triphone）	31.73	27.18	23.39	21.97
声韵母（Tri-IF）	30.25	26.08	22.51	20.33
扩展声韵母（Tri-XIF）	26.19	23.11	19.64	17.72

表 5.3 列出了几种声学模型的状态数目。可以看出，上下文相关音素、声韵母模型的状态数目是相当的。也就是说，表 5.2 中对于上下文相关基元性能的对比，是在模型规模相当的情况下作出的，也更能说明扩展声韵母基元的优点。

表 5.3　模型规模

基元类别	状态数目
音节（Syllable）	2,412
音素（Triphone）	10,851
声韵母（Tri-IF）	11,452
扩展声韵母（Tri-XIF）	11,708

总之，在几种常用的汉语连续语音识别基元中，上下文相关扩展声韵母（Tri-XIF）是最佳选择。它的优势主要体现在两个方面，一是基元数目，一是识别性能。在进行上下文相关建模时，标准声韵母基元约有 12 万个，这会导致识别网络规模过大，而扩展声韵母基元仅有约 3 万个，基元数目降低了 75%。同时，由于有些音节只有韵母部分，因此，整个音节需要与其它音节的韵母部分进行共享，这也导致了插入和替代错误较多。另外，在同等模型规模下，扩展声韵母模型也明显好于标准声韵母和音素模型。基于以上原因，在几种常用

基元中，上下文相关扩展声韵母（Tri-XIF）是最佳选择。

5.2.2　模型优化策略

本文提出和采用了三种策略来优化声学模型，这里讨论这些策略应用于 Tri-XIF 模型后的结果。

（1）得到 Tri-XIF 模型后，通过强制解码得到新的标注，以改善原始标注。此过程可以反复进行。当使用修正后的标注重新训练模型时，上下文无关模型的音节正确率提高了约 2 个百分点，进而，单混合的相关模型也有 1～2 个百分点的提高。但随着每个状态混合数目的增加，模型性能的提高变得不太明显，最终有不到 1 个百分点的提高，误识率下降约 5%。这主要是因为增加混合数以后，模型的描述能力已经大大增加，初始标注带来的影响已大大减小了。

（2）问题集的优化。本文考察了单问题集（即，未分类的问题集）、分类问题集、分类问题集＋双向问题集等。通过实验发现，使用单问题集已经可以取得较好的性能。将单问题集作为基线系统（Baseline），则使用分类问题集可以使音节误识率下降约 6%，进而加入双向问题集，音节误识率下降约 8%。与分类问题集相比，加入双向问题集以后，音节误识率可以下降 2～4%。

（3）加强中间状态的共享程度。通过调整分裂阈值，使中间状态加强共享，可以使模型规模（总状态数）降低 20%以上，而音节识别率只降低约 0.6 个百分点。当对模型规模有一定要求，又不希望性能有太大损失时，可以采用这种策略来进行优化。

5.2.3　标准普通话识别器

基于已有的标准普通话声学建模研究成果，本节将讨论如何构建标准普通话识别器，以便用于方言背景普通话识别框架。本节将介绍标准普通话构建中的主要方面，（1）声学模型（选用的语音库、识别基元、模型拓扑结构）；（2）发音词典；（3）语言模型（训练语料库、训练方法）；（4）实验结果：标准普通话识别器的性能。

虽然 863 语音库是一个精心设计的高质量语音库，语速缓慢，语料总数较少，不能涵盖更多的上下文，因此并不是训练基准模型的最佳选择，尤其是对于自然发音式语音。由 JHU 大学 CLSP 研究中心提供的朗读式标准普通话数据库——MBN（Mandarin Broadcast News）数据库[53]，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，发音更自然，内容更丰富，是一个更好的朗读式语音库。我们获准使用该语音库的标注文件和特征文件进行相关研究。JHU 提供的声学特征为 39 维的 MFCC，包括 13 维倒谱特征，以及一阶差分Δ和二阶差分ΔΔ，与我们的 42 维特征相比，没有加入频带能量，差别不大，不会影响实验结果的可靠性。而在 863 语音库上得到的一些主要结论，也在本库上也得以验证。经过与 863 语音库的对比，本文最终采用 MBN 语音库来建立标准普通话基准模型。

由于原始的标注、专家知识、方言背景普通话标注等都没有涉及零声母的问题，因此，本文首先使用标准声韵母基元建立基准声学模型。同时，为了验证扩展声韵母基元的有效性，本文也将提供基于扩展声韵母的实验结果。实验仍采用连续 HMM 来描述声学模型，拓扑结构等均与基于 863 语音库的实验类似，不同之处在于，状态共享后的总状态数为 3,000，每个状态的混合数增加为 14 个。

本文使用 Bi-gram 统计语言模型，用来进行语言模型的训练和平滑。语言模型的训练数据为 MBN 语音库中的文本标注。由此，可以构建标准普通话识别器。当用此系统测试 NIST 1997 广播新闻评测任务（NIST 1997 Broadcast News evaluation task[15]）时，其字错误率为 21.38%。这表明，标准普通话对于相同信道、标准普通话连续语音识别任务来说，是很好的。但是，当我们用此标准普通话识别器去识别吴方言背景普通话时，其字错误率则急剧升高。对朗读式和自然发音式吴方言背景普通话测试集（Test set），其字错误率分别升高为 61.89%和 72.17%。

5.3　吴方言背景普通话识别器

5.2.3 节中，我们给出了标准普通话声学建模的研究成果，本节中将给出应用新的识别框架后，方言背景普通话识别器的识别结果。采用上下文相关的映射规则后，并没有得到性能的提升，反而会下降，这主要是由于用于统计的数据量不足，使得统计结果不够准确。因此，此处给出的是基于上下文无关声韵映射规则的结果，包括 PTH-IF 和 WDC-IF 映射规则。我们使用字错误率/音节错误率（CER/SER）来表示识别结果。自然发音式和朗读式语音测试集上的测试结果将在后面给出。

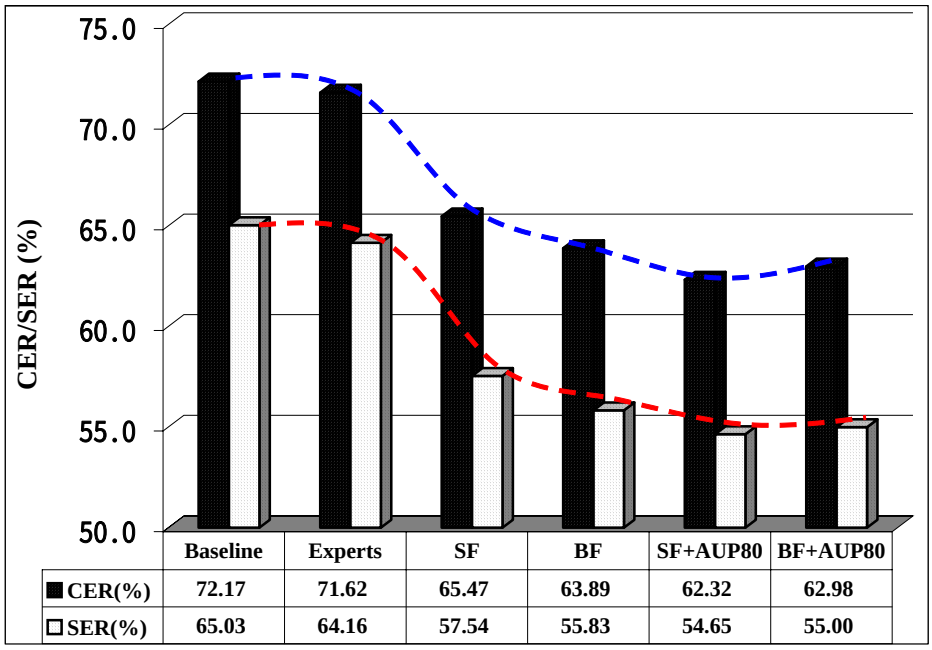
5.3.1　实验条件

在本文提出的方言背景普通话识别框架中，需要少量方言背景数据，所收集的方言相关知识，从而将一个标准普通话识别器转化为特定的方言背景普通话识别器。在本文中，吴方言背景普通话被选定为研究示例。

5.3.2　自然发音式语音识别结果

图 5.1 给出了各种方法在自然发音式语音测试集上的识别结果，包括，

1) 基线系统（Baseline）：标准普通话识别器
2) 基于专家知识（Experts）的多发音词典（不进行声学自适应）
3) 基于 surface-form 的 MLLR＋单发音词典（SF）
4) 基于 base-form 的 MLLR＋单发音词典（BF）
5) 基于 surface-form 的 MLLR＋WDC-IF 映射规则＋AUP（＝80%）剪枝准则（SF＋AUP80）
6) 基于 base-form 的 MLLR＋PTH-IF 映射规则＋AUP（＝80%）剪枝准则（BF＋AUP80）



	Baseline	Experts	SF	BF	SF+AUP80	BF+AUP80
■CER(%)	72.17	71.62	65.47	63.89	62.32	62.98
□SER(%)	65.03	64.16	57.54	55.83	54.65	55.00

图 5.1　各种方法在自然发音式测试集上的
字错误率/音节错误率（CER/SER）

声学自适应中，首先进行利用 MLLR 进行全局自适应，然后使用一个基于上下文无关声韵的回归树进行有监督的 MLLR。所有标准普通话声韵基元各占一类，吴方言背景普通话声韵则对应的标准普通话声韵母放在临近的位置。本文尝试了两种回归树的拓扑结构，一种是单右枝二叉树，一种是基于手工分类的二叉树。实验结果表明，在给定的数据量下，它们性能是相当的，因此我们直接采用前者进行实验。

从结果中可以看出，所提出的方法能够有效地降低系统的错误率。其中最好的结果是 surface-form 指导下的 MLLR 自适应，加上基于 AUP 的多发音词典。图 5.3 给出了 CER 在不同 AUP 阈值下的变化曲线。

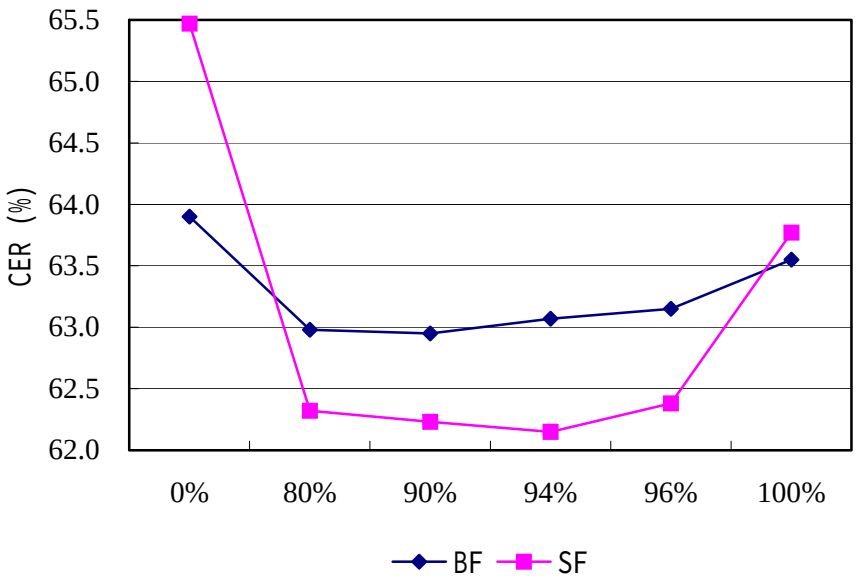
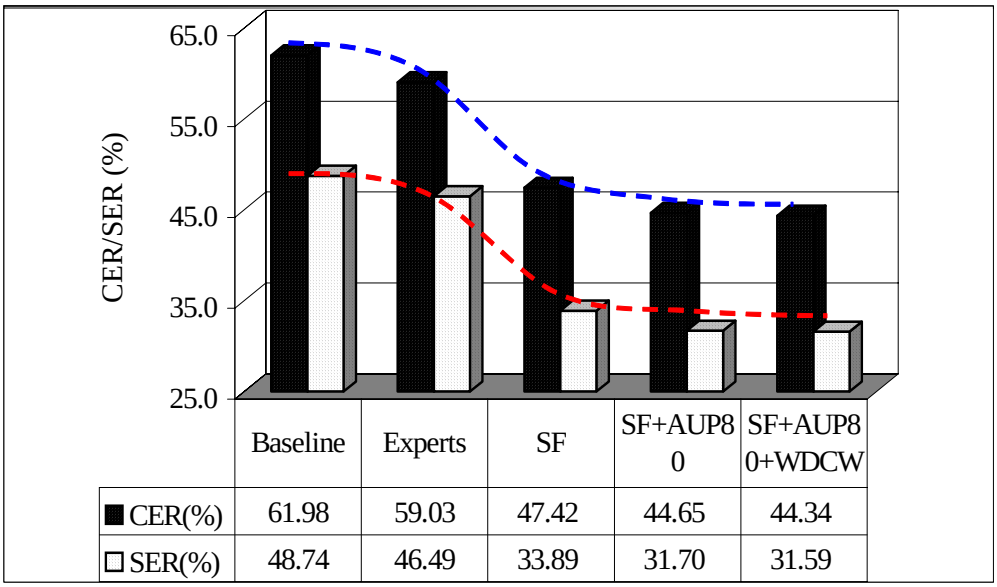


图 5.2　AUP 方法在自然发音式测试集上的字错误率（CER）曲线：基于 surface-form（SF）和 base-form（BF）的 MLLR 方法的对比

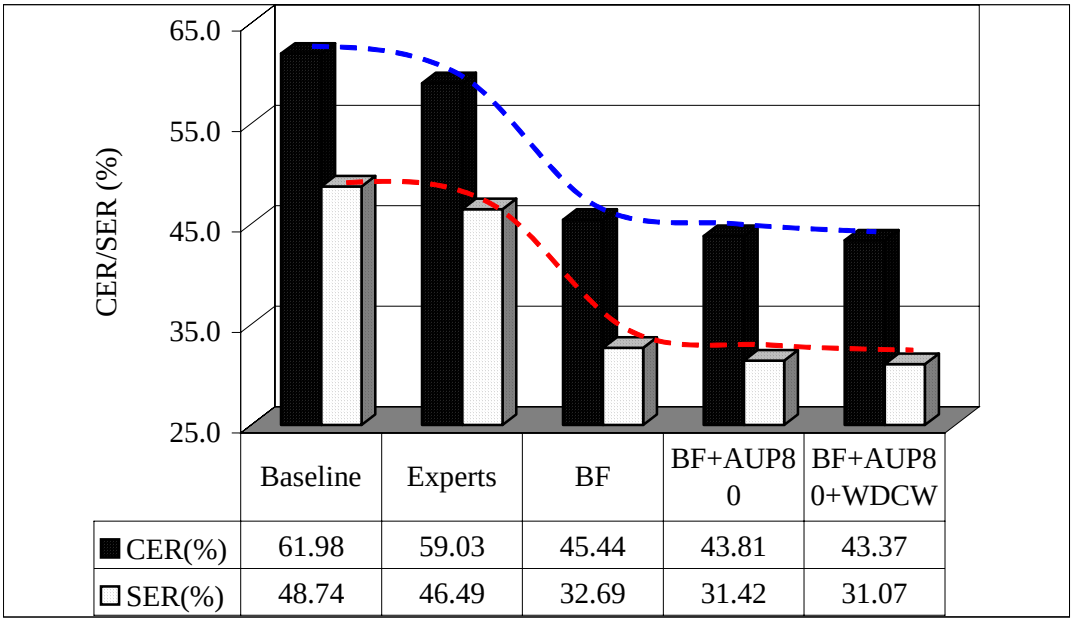
图 5.2 中，横轴表示 AUP 的阈值，0%表示所有词均使用单发音词典（即，标准单发音词典），而 100%表示所有词都使用多发音词典。对于 p%（0<p<100），只有累积概率大于 p%的词才进行多发音扩展，而其余的词仅使用单发音词典。p 越大，拥有多发音的词就越多，词典的条目也就越多。可以看出基于 AUP 的多发音扩展策略对于基于 base-form 和 surface-form 的 MLLR 都是有效的。单发音或全部多发音都不能得到最好的识别性能。我们需要根据识别性能和词典中条目的多少来决定使用什么样的阈值。

5.3.3　朗读式语音识别结果



	Baseline	Experts	SF	SF+AUP80	SF+AUP80+WDCW
CER(%)	61.98	59.03	47.42	44.65	44.34
SER(%)	48.74	46.49	33.89	31.70	31.59

a）基于 base-form 的 MLLR



	Baseline	Experts	BF	BF+AUP80	BF+AUP80+WDCW
CER(%)	61.98	59.03	45.44	43.81	43.37
SER(%)	48.74	46.49	32.69	31.42	31.07

b）基于 surface-form 的 MLLR

图 5.3 所采用的方法在朗读式测试集上的测试结果：
字错误率/音节错误率（CER/SER）

图 5.3 和图 5.4 分别给出了所提出的方法在朗读式语音测试集上的结果。从图 5.3 结果中可以看出，对基于 base-form 和 surface-form 的 MLLR，以

及基于 AUP 的多发音扩展策略是有效的。当加入方言词汇并重估其语言模型概率后（WDCW），总体误识率也有所下降，虽然只有约 0.5 个点的下降，但考虑到常用方言词汇在测试数据中仅占约 1%，这种下降也是比较明显的。

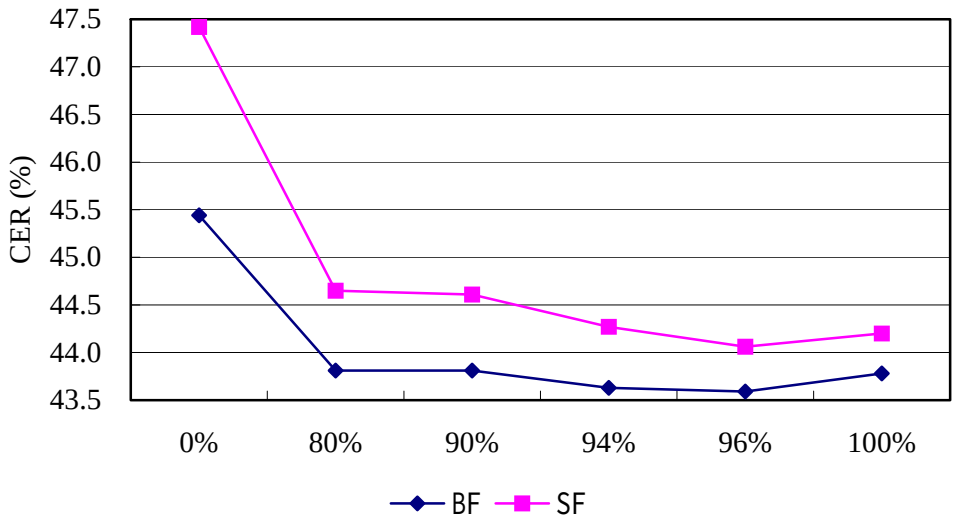


图 5.4 AUP 方法在朗读式测试集上的字错误率（CER）曲线：
基于 surface-form（SF）和 base-form（BF）的 MLLR 方法的对比

图 5.4 给出了 AUP 方法在朗读式测试集上的字错误率（CER）曲线。可以看出，base-form 指导下的 MLLR 方法的性能要略好于 surface-form 指导下的 MLLR。同时，区别于自然发音式语音，当 AUP 阈值接近 100%时，曲线也没有明显上升的趋势。主要原因是，朗读式语音与自然发音式语音相比，发音较清晰，且受方言背景影响要小，因此混淆度相对较小。

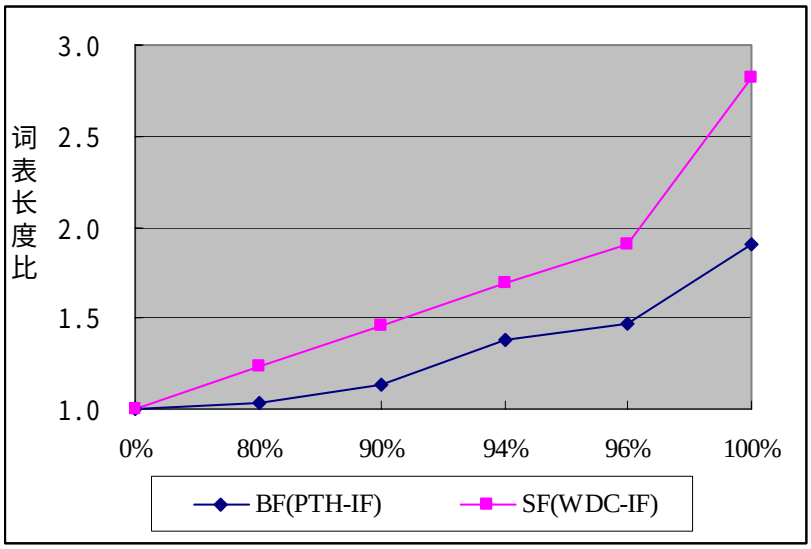


图 5.5 基于 AUP 准则的多发音词典与标准发音词典
词表长度比随阈值变化的曲线（朗读式）

图 5.5 给出了基于 AUP 准则的多发音词典与标准发音词典词表长度比随阈值变化的曲线（朗读式）。图中，BF/SF 分别表示与 base-form/surface-form 自适应模型对应的多发音词典，即，基于 PTH-IF/WDC-IF 声韵映射规则的多发音词典。当 AUP＝80%时，对于 BF 和 SF，其对应的词长比分别为 1.03 和 1.23，而从图 5.5 的结果中可以看出，此时已经取得了很明显的性能提升。当 AUP＝94%时，其词长比分别为 1.38 和 1.70，性能已经比较稳定了。而在发音变化建模中，这个比例一般不超过 2.5，过多则带来更多混淆，从而影响识别率和识别速度。这与本文的结论是一致的。

图 5.6 给出了采用几种不同的单发音、多发音词典后，在朗读式测试集上的性能对比，包括最可能发音（MLP），标准单发音（STD），以及它们的组合：STD＋MLP。STD＋MLP 的意义在于，它一方面具有 MLP 的优点，涵盖了最可能发音；另一方面，它既包含标准普通话发音，也包含吴方言发音，也与“方言背景普通话受标准普通话和方言背景双重影响”这一事实相一致，所以可以取得优于 STD 和 MLP 的性能。AUP 准则获得了最好的性能，优于其它几种单发音和多发音方法。

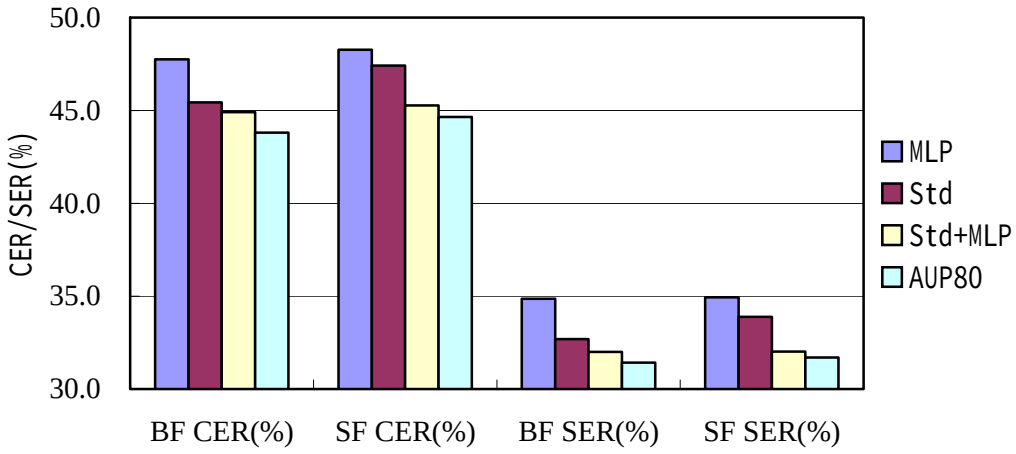


图 5.6 几种发音发音词典在朗读式测试集上的性能比较

5.3.4　扩展声韵母

我们将扩展声韵母基元应用于吴方言背景普通话识别，可以得到基于扩展声韵母的吴方言背景普通话识别器。首先要更新标注文件，按照扩展声韵母声学模型训练方法得到标准普通话声学模型。然后，将零声母相关的规则加入声韵母映射规则中，并重新生成 15k 词表的多发音词典。修改吴方言背景普通话小语料库的标注文件，进行基于 surface-form 和 base-form 的 MLLR 自适应。而语言模型即解码器部分与基元无关，不用进行修改。由此，可以得到基于扩展声韵母的吴方言背景普通话识别器。为了表示清楚，本文在吴方言背景普通话识别中，使用 WDC-XIF 表示扩展声韵母，以区别于 WDC-IF。表 5.4 给出了同等实验条件下，base-form 指导下的 MLLR 的系统测试结果，表中给出的是字错误率（CER）。

表 5.4　吴方言背景普通话识别器性能对比（CER%）：标准声韵母与扩展声韵母

说话方式	WDC-IF	WDC-XIF	误识率降低
自然发音式	62.98	60.39	4.11
朗读式	43.37	39.31	9.36

表 5.4 中的结果是在完全相同的条件下得到的对比结果，因此可以通过对比得到结论：扩展声韵母应用于方言背景普通话识别时，无论对于自然发音方式还是对于朗读方式，都可以取得性能上的提高。

在解码时间可以接受的情况下，适当优化识别器的参数设置，可以获得更好的识别性能。包括，适当放宽搜索阈值宽度，（如，Beam 宽度，即，用于剪枝的相对分数宽度，一般设为某一固定值，是各种剪枝策略中最有效的方式之一），调整词插入折扣等，可以得到更好的系统性能。实验表明，优化参数设置以后，时间开销上由原来的约 2 倍实时增加到 3 倍实时，但性能上有了进一步的提高。而各实验中总体变化趋势并没有发生变化。表 5.5 给出了在新的解码参数下新的对比结果。

表 5.5　优化参数后的吴方言背景普通话识别器性能对比（CER%）：标准声韵母与扩展声韵母

说话方式	WDC-IF	WDC-XIF	误识率降低
自然发音式	56.00	54.53	2.63
朗读式	38.65	34.67	10.30

从表 5.5 中可以看出，适当调整参数以后，可以获得更好的识别性能，而扩展声韵母性能一直优于标准声韵母的结果。在朗读式语音上，这种优势要更为明显。而自然发音式语音受发音习惯等的影响较大，在很大程度上影响了识别率。

第 6 章　总结与展望

6.1　总结

本文研究的是一种可扩展的方言背景普通话识别方法。借鉴以往的研究成果，本文提出了一种可扩展的方言背景普通话识别框架，并应用于吴方言背景普通话识别研究。在此框架下，本文研究内容涉及以下三个方面：均衡语料设计；标准普通话声学建模；吴方言背景普通话识别器构建。总结起来，本文有如下贡献：

（1）本文提出了一种可扩展的方言背景普通话识别框架，并加以研究。首先，建立一个标准普通话识别器，对于一种特定的方言，采集少量方言背景普通话语音（如 1 小时左右），并收集方言相关知识，将标准普通话识别器变换为方言背景普通话识别器。转换工作可以在声学模型、发音词典、语言模型、解码器等四个层面进行。在声学层，设计和选择方言背景普通话特定基元，并利用少量数据进行基于 surface-form/base-form 的 MLLR 自适应；在发音词典层，利用专家知识和标注文件中统计得到的声韵映射规则生成多发音词典，并进行词典的剪枝；在语言模型层，收集常用方言词汇，并通过重估或自适应方式重估语言模型；在解码器层，可以利用方言相关知识来优化解码器。与以往的研究相比，更注重框架整体性，方法的可移植性，以及方言相关知识的收集与利用。

（2）提出了两种均衡语料自动选择算法，可以从大量文本语料中自动抽取语句，使得语料库文本能够尽可能地均衡。区别于以往的方法，本文提出的两种方法分别从两个角度出发，一是抑制高频单元，一是鼓励低频单元，从而尽可能达到各个单元的均衡。二者相比，ELF 算法对训练用语料设计更为有效，因为它很好地保证了低频单元的覆盖率。

（3）提出了基于扩展声韵母的上下文相关声学建模方法，用于标准普通话和方言背景普通话声学建模。为了规范上下文以减少相关基元数目，同时减少插入和替代错误，本文定义了扩展声韵母基元，并设计了状态共享的问题集，建立基于扩展声韵母的上下文相关模型，并与其它几种常用基元进行了对比和分析。在同等条件下，扩展声韵母模型的性能要明显优于音节模型、音素模型、

标准声韵母模型。对于 8 混合模型，Tri-XIF 与 Tri-IF 和 Triphone 相比，其音节误识率分别降低了 12.84%和 19.34%。同时，当扩展声韵母应用于吴方言背景普通话识别时，其音节误识率与标准声韵母相比也可以降低 2～10%。同时，本文还提出了三种模型优化策略，实验表明，这些策略在一定程度上是有效的，可以根据需要进行使用。

（4）在吴方言背景普通话识别研究中，提出了基于 surface-form/baseform 的自适应方式与声韵母映射规则的多发音词典结合的方法。提出了基于累积一元概率（AUP）的剪枝策略，可以有效地降低词典长度，而几乎不影响识别率。与标准普通话识别器相比，仅使用约 1 小时语音数据和相关知识得到的吴方言背景普通话识别器，可以将自然发音式和朗读式语音的字错误率分别降低 13.65%和 30.03%，再考虑本文提出的扩展声韵母带来的好处，则字错误率降低更为明显，分别为 16.32%和 36.58%。如果将基于 base-form 的 MLLR 自适应视为较传统的方法，则采用本文所提出的方法后，自然发音式和朗读式语音的字错误率分别降低了 5.49%和 13.49%。

总的来说，对于朗读式语音，使用约 1 个小时的 WDC 语音数据和本文提出的方法，可以将字错误率从 61.89%降低到 34.67%，对于自然发音式语音，可以将字错误率从 72.17%降低到 54.53%。错误率的降低还是很明显的。在绪论中也给出过其它口音的识别结果（表 1.1），其改进后的错误率也都在 30%~60%之间。虽然不同语言复杂程度不同，但这些结果也说明了口音问题的困难程度。

再以朗读式语音为例，作为上限，标准普通话识别器的字错误率为 21.38%（国际上最好的结果在 20%左右），而采用本文提出的方法后，吴方言背景普通话识别器的字错误率为 34.67%，它们之间存在约十几个百分点的差距。虽然前者是在标准发音、同信道情况下得到的，占有一定的优势，但是，这种差距的存在正说明了问题的存在，而尽量缩短这种差距，就是我们继续研究的目标。

本文的出发点是，利用少量数据和方言相关知识，以较低的成本将标准普通话识别器变换为特定的方言背景普通话识别器。同时，希望方法本身可以扩展应用于其它方言背景。因此，本文充分考虑了框架的通用性和可扩展性，采用变换的方式，仅选用 1 小时左右的语音数据，利用专家知识和少量数据来获取声韵映射规则，采集常用方言词汇等，这些方法都可以方便地扩展到其它方言背景普通话识别中。当然，本研究仍处于起步阶段，方言相关知识的采集与应用还比较欠缺，对多遍解码等策略也需要进行更为深入、细致的研究。

6.2　展望

方言背景问题是汉语语音识别研究与应用中的一个重要课题，因此，对于方言背景普通话语音识别的研究是具有很强的现实意义的。在国际上，与此相关的口音问题、说话方式问题等，已经成为近几年来的研究热点，在 JHU 大学举行的 Workshop 在 2000 年和 2004 年都涉及到相关的研究方向。而我们国内在方言背景普通话识别方面的研究还相对欠缺。

本文在前人研究的基础上，提出了一个相对完整的、可扩展的方言背景普通话语音识别框架，并基于此框架进行研究。本文的研究工作只是一个开始，还有许多内容都需要进行深入细致的研究。总结起来，有如下几个重要问题值得考虑：

（1）**如何更好地收集方言相关知识。**知识源包括专家知识和少量实际数据。汉语中，对方言语言学、语音学的研究要比语音识别广泛而久远，因此，会有许多研究成果可以被借用过来，指导语音识别工作。针对要研究的特定方言背景普通话，我们需要更全面地收集方言相关的知识。包括，发音、声调、语调、词法、语法等方面的知识。另外，从实际采集的语音数据中，也可以获得许多知识，包括发音变化、音调变化、用词等。

（2）**如何更好地利用方言相关知识。**本文中，我们收集了上下文相关声韵映射规则，还给出了多发音词典的发音概率，但实验中，这些信息的应用却没有带来系统性能的改善。当然，其中一部分原因是由于语言模型发挥了很大的作用，从而使得发音词典层的概率变得影响很小。但从物理意义上讲，发音概率还是应该有作用的。因此，如何利用发音概率等信息，也是一个重要的问题。类似地，还有声调、语调变化等问题，对于特定的方言，都会有其特定的变化规律。本文研究中没有考虑声调、语调的变化，但实际上，这也是非常重要的知识，值得进一步研究。总的来说，当前语音识别研究中，知识的应用越来越受到重视，李锦辉在 ICSLP2004 大会报告上做得题为“ From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition ”的报告[118]，强调了知识在语音识别中的重要作用。

（3）**语言模型自适应。**本文对方言背景普通话识别的研究主要还是集中于声学层面，在语言模型方面进行的工作比较有限，主要是常用方言词汇收集和概率估计方面。由于方言背景的影响，还存在词法、语法上问题需要解决。在

有些方言中，语序与标准普通话相比会有所不同，如“你先走”，在一些方言中为“你走先”，而且常被保留在普通话中。因此，需要进一步加强对语言模型自适应的研究，以反映词法、语法差异对识别带来的影响。

（4）**解码策略。**本文报告的结果是基于一遍解码策略的。同样，我们可以采用两遍解码等策略，如 Word-Graph 搜索。首先，不使用语言模型，或使用简单的语言模型，通过解码得到声韵或音节网络（Lattice），然后进行重打分，更新网络和概率分，再加载更精细的语言模型，如 Bi-gram 或 Tri-gram，得到最终的识别结果。这种策略，我们在关键词识别中加以应用，可以取得一定的效果。在后面的研究中，可以考虑应用于方言背景普通话识别中。

总之，方言背景普通话语音识别研究是一个充满了挑战的课题，具有很强的现实意义，需要继续加强研究。

参考文献

[1] 钱乃荣．当代吴语研究．上海：上海教育出版社，1992

[2] 受鸣．上海方言十夜谈．上海：华东师范大学出版社，1992

[3] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989, 77(2): 257-285

[4] 剑桥大学．http://htk.eng.cam.ac.uk/

[5] IBM. http://www.watson.ibm.com/

[6] CMU. http://www.speech.cs.cmu.edu/

[7] Microsoft. http://www.microsoft.com/speech/

[8] ATT. http://public.research.att.com/

[9] Huang X-D, Alleva F, Hon H-W, et al. The Sphinx-II speech recognition system: an overview. Computer Speech and Language. 1993, 7(2): 137-148

[10] Viavoice. http://www-3.ibm.com/software/speech/

[11] Dragon Systems. Dragon Dictate, 2000. http://www.dragonsys.com/

[12] Yan P-J, Zheng F. Context directed speech recognition in dialogue systems. International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages. 2004, 225~228

[13] Liu Z, Wang Y. Audio indexing and retrieval. Handbook of Video Databases Design and Applications, CRC Press. 2003. 483~510

[14] Gu L, Gao Y-Q. Use of maximum entropy in natural word generation for statistical concept-based speech-to-speech translation. In: Interspeech'2005. 2005. 3189~3192

[15] NIST. 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, 1997

[16] http://www.clsp.jhu.edu/workshops/

[17] Sproad R, Zheng F, Gu L, et al. Dialectal Chinese speech recognition: Final Report. CLSP Summer Workshop. 2004. 1~82

[18] Wang H-Y, Heuven V J van. Mutual intelligibility of american, Chinese and dutch-accented speakers of English. In: Interspeech'2005. 2005. 2225~2228

[19] Hong Y, Abeer A, Abe K, et al. Pronunciation variations of Spanish-accented English spoken by young children. In: Interspeech'2005. 2005. 749~752

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

[21] Huang R, Hansen JHL. Dialect/accent classification via boosted word modeling. in: ICASSP'2005. 2005, 1:585-588.

[22] Chen T, Huang C, Chang E, et al. Automatic accent identification using Gaussian mixture model. IEEE workshop on ASRU'2001. 2001.

[23] Tobias C, Rainer G, Satoshi N. Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models. In: Interspeech'2004. 2004. 1509~1512

[24] Sun X. Pitch accent prediction using ensemble machine learning. In: ICSLP'2002. 2002. 561~564

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

[26] Raux A. Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition. In: Interspeech'2004. 2004. 613~616

[27] Zheng Y-L, Sproat R, Gu L, et al. Accent detection and speech recognition for Shanghai-accented Mandarin. In: Interspeech'2005. 2005. 217~220

[28] Ikeno A, Pellom B, Cer D, et al. Issues in recognition of Spanish-accented spontaneous English. In: Proceedings of IEEE/ISCA Workshop on Spontaneous Speech Processing and Recognition. Tokyo, Japan. 2003

[29] Tomokiyo L-M. Recognizing non-native speech: characterizing and adapting to non-native usage in LVCSR. PhD Thesis, Carnegie Mellon University, 2001

[30] Wang Z-R, Schultz T and Waibel A. Comparison of acoustic model adaptation techniques on non-native speech. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003). 2003, 1:540-543

[31] Huang C. Accent issue in large vocabulary continuous speech recognition. Microsoft Research Technical Report. MSR-TR-2001-69, 2001

[32] Huang C, Chen T, Chang E. Accent issue in large vocabulary continuous speech recognition. International Journal of Speech Technology, Kluwer Academic Publishers. 2004, 7(2):141-153

[33] Tjalve M, Huckvale M. Pronunciation variation modelling using accent features. In: Interspeech'2005. 2005. 1341~1344

[34] Aalburg S, Hoege H. Foreign-accented speaker-independent speech recognition. In: Interspeech'2004. 2004. 1465~1468

[35] Leggetter C J, woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language. 1995, 9:171-186

[36] Gales M J, Woodland P C. Mean and variance adaptation within the MLLR framework. Computer Speech and Language. 1996, 10: 249-264

[37] Lee C-H, Lin C-H, Juang B-H. A study on speaker adaptation of parameters of continuous density hidden Markov models. IEEE Trans. SP, 1991, 39(4): 806-814

[38] Lee C-H, Gauvain J L. Speaker adaptation based on MAP estimation of HMM parameters. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 1993, 2: 652-655

[39] Jurafsky D, et al. What kind of pronunciation variation is hard for triphones to model? In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'2001). 2001, 577~580.

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

[41] Byrne W, Finke M, Khudanpur S, et al. Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'1998). Settle, USA, 1998. 313~316

[42] Wester M. Pronunciation modeling for ASR - knowledge-based and data-derived methods. Computer Speech and Language. 2003, 17:69-85

[43] Bell A, Jurafsky D, Fosler-Lussier E, et al. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. Journal of the Acoustical Society of America. 2003, 113(2):1001-1024

[44] Fosler-Lussier E. A tutorial on pronunciation modeling for large vocabulary speech recognition. In: S. Renals and G. Grefenstette, Text and Speech Triggered Information Access, Springer Verlag, Berlin, 2003. (also available: http://www.cse.ohio-state.edu/~fosler/publications.html)

[45] Liu Y, Fung P. State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition. IEEE Transactions on Speech and Audio Processing. 2004, 14(4):351-364

[46] Liu Y, Fung P. Modeling partial pronunciation variations for spontaneous mandarin speech recognition. In Computer Speech & Language. 2003, 17(4):357-379

[47] Liu Y, Fung P. Partial change accent models for accented Mandarin speech recognition. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'2003). 2003.

[48] http://www.chineseldc.org/doc/CLDC-SPC-2004-005/intro.htm

[49] Liu M-K, Xu B. Accent-specific Mandarin adaptation based on pronunciation modeling technology. In: ICSLP'2000. 2000, 2:330-333

[50] Zhang H-Y. Xu B. Geometric constrained maximum likelihood linear regression on Mandarin dialect adaptation. In:Eurospeech'2003. 2003. 1465~1468

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

[52] Rosenfeld R. Two decades of statistical language modeling: Where do we go from here? In: Proceedings of the IEEE. 2000, 88:1270-1278

[53] LDC. http://www.ldc.upenn.edu/

[54] ELRA. http://www.elra.info/

[55] CCC. http://www.cccforum.org

[56] Zheng T F. Making full use of Chinese speech corpora. Invited Keynote Speech, Oriental-COCOSDA. 2003. 9~23

[57] Sun J-S, Wang Z-Y, Wang X. Construction of the lexicon for continuous acoustic model training. In: Proc. of the Improvement of Intelligence Computer Interface and Application. 2000. 161~121

[58] 祖漪清，李爱军．连续语音数据库设计的科学性问题．语音研究报告．中国社会科学院，1998

[59] 王天庆，李爱军．连续汉语语音识别语料库的设计．天津：第六届全国现代语音学学术会议．2003

[60] Li M, Junkawitsch J, Yun T. An incremental approach to selection of well balanced corpus. In: 8th Aust. Int. Conf. Speech Sci. & Tech.. 2000. 440~444

[61] 宁振江，杜利民．一种改进后的递增式语音语料抽选算法．中国科学院研究生院学报．2005, 22(2)：140-146

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

[63] Li A-J., Chen X-X., et al. The phonetic labeling on read and spontaneous discourse corpora. In: International Conference on Spoken Language Processing (ICSLP'2000). 2000, 4:724-727

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

[65] Wei W, Van Vuuren S. Improved neural network training of inter-word context units for connected digit recognition. In: ICASSP'98. 1998. 497~500

[66] Zheng F, Chai H-X, Shi Z-J, et al. A real-world speech recognition system based on CDCPMs. In: Int. Conf. On Computer Processing of Oriental Languages (ICCPOL'97). 1997, 1:204-207

[67] 杨行峻，迟惠生，等．语音信号数字处理. 电子工业出版社，1995

[68] Hwang M-Y, Huang X, Alleva F. Predicting unseen triphones with senones. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. Minneapolis. 1993. 311~314

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

[70] 张继勇．汉语语音识别中声学建模及参数共享策略的研究：[硕士学位论文] ．北京：清华大学计算机系，2001

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

[72] 宋战江．汉语自然语音识别中发音建模的研究：[博士学位论文] ．北京：清华大学计算机系，2001

[73] Lee C-H, Rabiner L, Pieraccini R, et al. Acoustic modeling for large vocabulary speech recognition. Computer Speech and Language. 1990, 4(2): 127-165

[74] Young S J, Woodland P C. Tree-based state tying for high accuracy acoustic modeling. In: Proc ARPA Human Language Tech Workshop. Plainsboro, NJ: Morgan Kaufmann Publisher. 1994. 307~312

[75] 时宇，张益肇．微软中国研究院在普通话语音识别领域的现况和展望．第六届全国人机语音通讯学术会议 (NCMMSC6)．2001．2799~2802

[76] Reichl W, Chou W. Decision trees state tying based on segmental clustering for acoustic modeling. In: Int. Conf. of Acoustics, Speech, Signal Processing(ICASSP'98). Seattle, Washington: IEEE Press. 1998, 801~804

[77] Reichl W, Chou W. Robust decision tree state tying for continuous speech recognition. IEEE Trans Speech and Audio Proc. 2000, 8(5): 555-566

[78] 曹剑芬．现代语音基础知识．北京：人民教育出版社，1990

[79] 吴宗济．试论人－机对话中的汉语语音学．北京：世界汉语教学，1997，42(4)：3-20

[80] 罗安源．田野语音学．北京：中央民族大学出版社，2000

[81] Gao S., Xu B, Huang T-Y. Class-triphone acoustic modeling based on decision tree for Mandarin continuous speech recognition. In: International Symposium on Chinese Spoken Language Processing (ISCSLP' 98). 1998. 44~48

[82] Duchateau J, Demuynck K, Van. Compernolle D. A novel node splitting criterion in decision tree construction for semi-continuous HMMs. In: Eurospeech'97. Rhodes, 1997. 1183~1186

[83] Digalakis V V, Rtischev D, Neumeyer L G. Speaker adaptation using constrained estimation of Gaussian mixture. IEEE Trans. on Speech and Audio Process. 1995, 3(5): 357-366

[84] 何磊．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]．北京：清华大学计算机系，2001

[85] Lee C-H, Lin C-H, Juang B-H. A study on speaker adaptation of parameters of continuous density hidden Markov models. IEEE Trans. SP. 1991, 39(4): 806-814

[86] Lee C-H, Gauvain J L. Speaker adaptation based on MAP estimation of HMM parameters. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing. 1993, 2: 652-655

[87] Huo Q, Chan C, Lee C-H. Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition. IEEE Trans. On Speech and Audio Processing. 1995, 3(5): 334-345

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

[89] Leggetter C J, Woodland P C. Flexible speaker adaptation for large vocabulary speech recognition. In: Proc. Eurospeech. 1995. 1155~1158

[90] Leggetter C J. Improved acoustic modeling for HMMs using linear transformations. Ph.D. thesis. Cambridge University. 1995

[91] Gales M J F. Maximum likelihood linear transformation for HMM-based speech recognition. Computer Speech and Language. 1998, 12:75-98

[92] Cohen P S, Mercer R L. The phonological component of an automatic speech recognition system. In: Reddy, D.R. (ed.), Speech Recognition. Academic Press, Inc., New York. 1975. 275~320

[93] Lamel L, Adda G. On designing pronunciation lexicons for large vocabulary continuous speech recognition. In: Proc. of ICSLP-96. Philadelphia, 1996. 6~9

[94] Amdal I, Korkmazskiy F, Surendran A C. Joint pronunciation modelling of non-native speakers using data-driven methods. In: Proc. of ICSLP'2000. 2000. 622~625

[95] Riley M, Byrne W, Finke M, et al. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Communication. 1999, 29(2-4):209-224

[96] Williams G, Renals S. Confidence measures for evaluating pronunciation models. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition'. 1998. 151~156

[97] Kessens J M, Cucchiarini C, Strik H. Data-driven method for modeling pronunciation variation. Speech communication. 2003, 40(4), 517-534

[98] Ravishankar M, Eskenazi M. Automatic generation of context-dependent pronunciations. In: Proc. of EuroSpeech'97. Rhodes, 1997. 2467~2470

[99] Kessens J M, Wester M, Strik H. Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. Speech Communication. 1999, 29(2-4):193-207

[100] Holter T. Maximum likelihood modelling of pronunciation in automatic speech recognition. Ph.D. thesis. Norwegian University of Science and Technology. 1997

[101] Torre D, Villarrubia L, Hernndez L, et al. Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In: Proc. of ICASSP'97. Munich, 1997. 1463~1466

[102] Zheng F, Wu J, Song Z-J. Improving the syllable-synchronous network search algorithm for word decoding in continuous Chinese speech recognition. J. Computer Science & Technology. 2000, 15(5):461-471

[103] 李荣，徐宝华，陶寰，等.上海方言词典．上海：江苏教育出版社，1993

[104] 颜逸明．吴语概说．上海：华东师范大学出版社，1994

[105] 王福堂．汉语方言词汇．北京：语文出版社，1995

[106] Schiel F, Kipp A, Tillmann H G. Statistical modelling of pronunciation: It's not the model, It's the data. ESCA Workshop on Pronunciation Variation for Automatic Speech Recognition. Kerkrade, 1998. 131~136

[107] Riley M D, Ljojle A. Automatic generation of detailed pronunciation lexicons. In: Automatic Speech and Speaker Recognition. Kluwer Academic Pubs. 1996. Ch. 12, 285-301

[108] Witten I H, Bell T C. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Trans. On Information Theory. 1991, 37(4):1085-1094

[109] Katz S M. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. On Acoustic, Speech and Signal Processing. 1987, 35(3):400-401

[110] Jelinek F, Mercer R L. Interpolated estimation of Markov source parameters from sparse data. In: Gelsema D and Kanal L, eds. North-Holland: Pattern Recognition in Practice. 1980

[111] 武健．汉语语音识别中统计语言模型的构建及其应用：[硕士学位论文]，北京： 清华大学计算机系，2000

[112] Young S, Evermann G, Hain T, et al. The HTK book (for HTK Version 3.2.1). http://htk.eng.cam.ac.uk, 2002

[113] Stolcke A. SRILM - an extensible language modeling toolkit. In: International Conference on Spoken Language Processing (ICSLP'2002). Denver, 2002, 2:901-904

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

[115] Zheng F, Zhang G-L. Integrating the energy information into MFCC. In: International Conference on Spoken Language Processing (ICSLP'2000). 2000, I:389-292

[116] Viikki O, Laurila K. Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization. In: ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels. 1997. 107~110

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

[118] Lee C-H. From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition. In: ICSLP'2004. Plenary Session 2. 2004

致　谢

衷心感谢导师吴文虎教授和郑方教授对本人的精心指导。吴文虎教授严谨求实，平易近人，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。郑方教授严谨治学，忘我工作，给我树立了很好的榜样，并引领我进入语音识别研究领域。他们的言传身教将使我终生受益。

感谢方棣棠教授、李树青教授和徐明星副教授的关心和帮助。方老师和李老师一直在关注着实验室的发展，让我为之感动。徐老师给我提了很多中肯的意见和建议，不断帮助我进步。同时感谢邬晓钧，熊振宇，孙辉，刘林泉，邓菁，刘建以及实验室全体老师和同窗的热情帮助和支持！

感谢我的父母和妻子王丽坤对我学业的支持。

本课题承蒙美国国防部和国家自然科学基金资助，特此致谢。

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：_____

附录 上海方言声母/韵母表

上海方言声母表

部位 Position / 方法 Method			双唇 Bilabial	唇齿 Labiodental	舌尖前 Dental	舌面前 Dorsal	舌根 Veolar	喉 Guttural
塞 Plosive	清 Voiceless	不送气 Unaspirated	b [p]		d [t]		g [k]	
		送气 Aspirated	p [pʰ]		t [tʰ]		k [kʰ]	
	浊 Voiced		bb [b]		dd [d]		gg [g]	[?]
塞擦 Affricate	清 Voiceless	不送气 Unaspirated			z [ts]	j [tɕ]		
		送气 Aspirated			c [tsʰ]	q [tɕʰ]		
	浊 Voiced					jj [dʑ]		
擦 Fricative	鼻 Nasal		m [m]		n [n]	ni [Â]		
	边 Lateral				l [l]		ng [Ð]	
	清 Voiceless			f [f]	s [s]	x [ɕ]		h^ [h]
	浊 Voiced			ff [v]	ss [z]	xx [ʑ]		hh [ä]

上海方言韵母表

单韵母 Single vowel			
	i [i]	u [u]	v [y]
ii [¡]			
iio [{]			
	复合韵母 Compound vowel		
a [A]	ia [iA]	ua [uA]	
a>[§]	ia>[i§]	ua>[u§]	
o^=ô [C]	io^=iô [iC]		
o [o]			
eu=e [°]/[°M]/[°u]	ieu [i°]/[i°M]		
e< [e]	ie< [ie]	ue< [ue]	
e^=ê [E]	ie^=ie=iê [iE]	ue^=uê [uE]	
e> [Q]	ie> [iQ]	ue> [uQ]	
oe [O]		uoe [uO]	voe [yO]
	iuu [iu]		
复合鼻韵母 Compound vowel followed by a nasal			
a<~ [a]	ia<~ [ia]	ua<~ [ua]	
a~ [A]	ia~ [iA]	ua~ [uA]	
a>~ [§]	ia>~ [i§]	ua>~ [u§]	
en [Èn]/[È´]	in=ien [in]	un=uen [uÈn]	vn= ven [yÈn]/[yn]
eng [ÈÐ]	ing [iÐ]	ueng [uÈÐ]	veng [yÈÐ]/[yÐ]
oong [oÐ]	ioong [ioÐ]		voong [yoÐ]
入声韵母 Vowel of Entering Sound			
a<k [a?]	ia<k [ia?]	ua<k [ua?]	
a>k [§?]	ia>k [i§?]	ua>k [u§?]	
ok [o?]	iok [io?]		vok [yo?]
o^k [C?]		u^k [uC?]	

oek [O?]			voek [yO?]
ek [È?]	iek [iÈ?]	uek [uÈ?]	
	ie<k [ie?]		
i>k [I?]	ii>k [iI?]		vi>k [yI?]
自成音节 Individual Syllable			
eer [Èr]/ [Èl]	m [m]	n [n]/[ˊ]	ng [Ð]

表中列出了上海方言声母/韵母，以及它们的 IPA 发音（置于中括号内，需要安装 ZCunsil4 字体）补充说明如下：

（1） 灰色背景的声母是标准普通话的声母，而其他为上海话特有的声母。

（2） 声母 /bb/, /dd/, /gg/, /jj/, /xx/, /ss/, /ff/, /hh/ 分别是对应于 /b/, /d/, /g/, /j/, /x/, /s/, /f/, /h/ 的浊声母。

（3） 声母 [?] 为入声声母。符号 /v/ 表示韵母 /ü/。符号 /ii/ 表示拼音 /zi/, /ci/, and /si/ 中的韵母，而 /iii/ 表示 /zhi/，/chi/，/shi/，/ri/ 中的韵母。

相关的参考文献：

[1] 许宝华，陶璟. 《上海方言词典》. 江苏教育出版社，1997

[2] 许宝华，汤珍珠.《上海市区方言志》. 上海：上海教育出版社. 1988

[3] 钱乃荣. 《跟我学上海话》. 上海：上海教育出版社. 2002

[4] 阮恒辉. 《自学上海话》. 上海：上海大学出版社. 2000

[5] 叶盼云 编著，范毓民 译.《学说上海话》. 上海：上海交通大学出版社. 2001

[6] 徐子亮. 《上海话三月通》. 上海：上海海文音像出版社. 2000

个人简历、在学期间发表的学术论文与研究成果

个人简历

1975 年 4 月 8 日出生于河北省赵县。

1994 年 9 月考入清华大学计算机系计算机应用专业，1999 年 7 月本科毕业并获得工学学士学位。

1999 年 9 月免试进入清华大学计算机系攻读计算机科学与技术学科工学博士至今。

发表的学术论文

[1] Li J, Zheng T F, Byrne W, Jurafsky D. A dialectal Chinese speech recognition framework. J. Computer Science & Technology (JCST). 2006, 21(1), 106-115（**SCI** 待检索，**EI** 待检索）

[2] 李净，郑方，张继勇，等．汉语连续语音识别中上下文相关的声韵母建模．清华学报（自然科学版），2004，24（1）：61~64（**Compendex：04218172339，INSPEC：8164092**）

[3] Li J, Zheng F, Xiong Z Y, et al., Construction of large-scale Shanghai Putonghua speech corpus for Chinese speech recognition. In: Oriental-COCOSDA, Singapore, 2003, 62-69.（**INSPEC：8151204**）

[4] Li J, Xu M X, Wu W H. Study on framework for Chinese pronunciation variation modeling. In: International Symposium on Chinese Spoken Language Processing (ISCSLP'02), 2002, 21-24.

[5] Li J, Zheng F, Zhang J Y, et al., The definition and extension of the question set for decision tree based state tying in Chinese speech recognition. In: International Conference on Chinese Computing (ICCC'2001), Singapore, 2001, 106-110.

[6] 李净，徐明星，张继勇等．汉语连续语音识别中声学模型基元比较：音节、音素、声韵母．第六届全国人机语音通信学术会议（NCMMSC6），2001，267-271．

[7] Li J, Zheng F, Wu W H. Context-independent Chinese initial-final acoustic modeling. In: International Symposium on Chinese Spoken Language Processing (ISCSLP'00). Beijing, 2000, 23-26.

[8] Zheng F, Li J, Song Z J, et al., A two-step keyword spotting method based on context-dependent a posteriori probability. In: International Symposium on Chinese Spoken Language Processing (ISCSLP'2004), Hong Kong, 2004, 281-284.

[9] 陈肖霞，郑方，李净，等．基于标注的上海口音普通话语音变化分析．第七届全国人机语音通信学术会议（NCMMSC7），2003，224-227

[10] Xiong Z Y, Zheng F, Li J, et al., An automatic prompting texts selecting algorithm for di-IFs balanced speech corpus. In: National Conference on Man-Machine Speech Communications (NCMMSC7), 2003, 252-256.

[11] 张国亮，徐明星，李净，等．语音识别中基于两层词法树的跨词搜索算法．清华学报（自然科学版），2003，43（7）：981-984（**Compendex：03507781447，INSPEC: 7932361**）

[12] Lu Z Z, Xu M X, Yan P J, Li J, Wu W H. A target-independent solution to Mandarin speech recognition engine of dialogue system on specified domain. The First International Conference on Machine Learning and Cybernetics, Prime Hotel, Beijing, China, 2002（**Compendex：03127405833**）

[13] Lu Z Z, Li J, Xu M X, et al., Mandarin keyword spotting based on syllabic fillers. In: International Conference on Chinese Computing (ICCC'2001), Singapore, 2001, 122-125.

[14] Zhang J Y, Zheng F, Li J, et al., Improved context-dependent acoustic modeling for continuous Chinese speech recognition. In: Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech'2001), Aalborg, Denmark, 2001, 3:1617~1620.

在读期间完成的其它研发工作

1. 改进用于连续语音识别的帧同步搜索算法，并搭建命令系统 EasyCmd
2. 设计开发关键词识别 API v1.0 和 v2.0，并应用于自动语音总机，现已投产（2 人合作完成）
3. 实验室大型数据库设计与采集，包括：TRSC（500 人朗读式电话语音库），WDC（吴方言背景普通话语音库）等