基于语音属性的带口音汉语语音识别研究

（申请清华大学工学硕士学位论文）

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计 算 机 科 学 与 技 术

研 究 生 ： 张 超

指 导 教 师 ： 刘 轶　　副研究员

二〇一二年五月

基于语音属性的带口音汉语语音识别研究

张超

Accented Chinese Speech Recognition Based on Speech Attributes

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Master of Science

in

Computer Science and Technology

by

Zhang Chao

Thesis Supervisor：Doctor Liu Yi

May, 2012

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

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

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

人类的语音包含了丰富的信息，如语言文本，说话人情绪、各种声学、语音学参数等，统称为语音属性。为提高 ASR 对由多种汉语方言口音所造成的发音变的鲁棒性，本文分别采用了两种方法进行带有多种方言口音的汉语语音识别研究：基于单一语音属性的统计 ASR 方法和自动语音属性转译方法。

对基于单一语音属性的统计 ASR 方法，本文在模型静态重构方法的基础上，提出了基于时间对准的音子识别以得到可靠的口音相关单元。本文还提出了动态高斯混合选择算法。在解码搜索中，对输入语音帧，它通过根据最近邻原则动态选取指定数目的高斯成分为每个状态分别构造动态观测密度，相当于对声学模型进行了动态重构。算法在每个状态中动态选取的高斯成分数目都是独立的离散变量，我们基于最小化分类错误准则和遗传算法对它们进行优化，相当于对算法参数进行离散区分性训练。得到的区分性动态高斯混合算法可以提高模型的鲁棒性和精度，从而改善系统在剪枝束搜索中的表现。联合使用这些算法使 ASR 在川、粤、吴 3 种口音普通话数据上比常用的 MAP 自适应算法音节准确率分别相对提高了 7.57%、3.69%和 4.19%，并保持了对标准普通话的识别率。

自动语音属性转译是近年来兴起的一种新兴的语音识别方法，它通过在 ASR 中导入更多的语音属性和先验知识来提高系统性能和鲁棒性，也称为基于多语音属性的探测式 ASR。我们首次将探测式 ASR 方法应用于带口音语音识别中。为使 ASR 能够解决上下文相关的发音变异，探测器采用基于发音特征的上下文相关 HMM。我们使用仅含状态特征函数的 CRF 融合探测器输出以解决音子欠生成问题，再利用基于 HMM 的有限状态网络验证 CRF 输出的所有假设序列并得到识别结果。探测式 ASR 的设计符合汉语非线性音系学知识，在官话、粤、吴口音中，音子准确率分别比 Monophone HMMs 系统相对提高了 5.04%、4.68%和 6.06%。我们还提出了语音属性区分单元以简洁高效地解决口音产生的发音变异而无需修改任何已有模型。本文最后还提出了对多种方言口音鲁棒的并行探测式 ASR，它与 Triphone HMMs 系统具有相近的识别率，但识别速度提高了 3.44 倍。

关键词：自动语音识别；多方言口音；语音属性；区分性动态高斯混合选择；自动语音属性转译

Abstract

There is a rich set of information embedded in the speech utterance: the content of the spoken language, the emotional status of the speaker, various acoustic and phonetic parameters, and *etc.*, which is jointly called the speech attribute. In order to improve the robustness in automatic speech recognition (ASR) for pronunciation variation caused by a diversity of Chinese regional accents, in this thesis two approaches are adopted separately for multi-accent speech recognition task: the state-of-the-art HMM-based ASR with single speech attribute, and a detection-based multi-attribute automatic speech attribute transcription (ASAT) approach.

For HMM-based ASR with cepstrum coefficients as the single attribute, we propose time-aligned phone recognition to generate reliable accent specific unit, which are used by static acoustic model reconstruction approach. We also propose dynamic Gaussian mixture selection (DGMS) scheme that customizes a dynamic observation density for each input speech frame on each state in decoding, by selecting a specified number of Gaussian-components nearest to that frame. DGMS is able to dynamically reconstruct the acoustic models. The selection number of Gaussian components in each HMM state is pre-defined individually based on the discriminative minimum classification error criteria with genetic algorithm, by which process can be regarded as discrete-variable discriminative training. The obtained discriminative DGMS is able to improve the robustness and performance for multi-accent variations, and alleviate the performance loss in pruned Beam search. Our system obtained 7.57%, 3.69%, and 4.19% relative phone accuracy improvement over a system with widely used MAP adaptation, and retained its performance on standard Putonghua.

ASAT is a novel speech recognition paradigm proposed recently. It aims at enhancing the performance and robustness by integrating more speech attribute and knowledge into ASR. An ASAT system is also called a detection-based ASR which is first applied to accented speech recognition task in this thesis. To capture a diversity of context-dependent pronunciation variations either from accents or co-articulations, context-dependent HMMs are adopted as the detectors for articulatory features. We use CRF with merely the state features as the attribute-to-phone event merger to avoid

phone under-generation problem. The hypothesis sequences embedded in the output probabilistic phone lattice by CRF are verified by searching through an HMM-based finite state network to generate the recognition result. Our systems coincide with the non-linear phonology, and obtained 5.04%, 4.68%, and 6.06% relative phone accuracy improvement on Guanhua, Yue, and Wu accents individually, compared to Monophone HMMs. Meanwhile, we propose accent-related attribute discrimination module to handle accent variations in high efficiency without retraining any model in the system. Furthermore, we propose a parallel detection-based ASR covering multiple accents, which obtained comparable phone accuracies at 3.44 times faster than Triphone HMMs on average.

Key words: Automatic speech recognition; Multiple accents; Speech attribute; Discriminative dynamic Gaussian mixture selection; Automatic speech attribute Transcription

目　录

第 1 章　绪论

1.1　自动语音识别研究的意义和重要性

语音是人类语言和思维最古老的载体之一[1]，也是目前人类最自然、最方便的交流方式。自动语音识别（Automatic Speech Recognition, ASR）是指计算机通过计算自动获得语音对应的文本信息，完成由语音到文字的转换，如图 1-1 所示。于是通过 ASR 可以实现自然的人机交互（Human Computer Interaction），改善用户体验。尤其是随着普适计算（Ubiquitous Computing）的快速发展，以智能手机、随身听、平板电脑等为代表的便携式产品受限于设备体积等原因，很难采用键盘等经典人机交互方式，给用户表达自然语言等复杂意图造成了很大的不便。采用 ASR 可以最自然的解决这些问题。而且使用 ASR 无需用户专门学习任何技能，门槛极低，易于扩展更广泛的用户群。ASR 与语音合成（Speech Synthesis）、说话人识别（Speaker Recognition）等其它语音技术共同构成了基于语音的人机交互技术的基石。

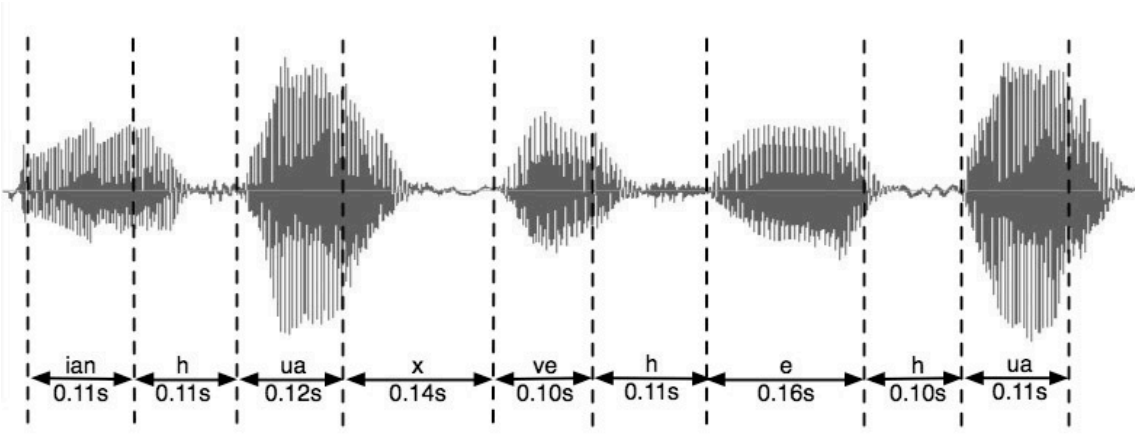


图 1-1　一段连续语音及其基于汉语声韵母的文本信息

同时，随着文明的高速发展，人类正面临着空前的信息爆炸，每时每刻都有以文本、音频、图像、视频等为载体的海量信息被创造出来。于是，如何在信息的汪洋中高效地检索到所需要的信息，即信息检索（Information Retrieval），也成为了日益重要的课题。现今，美国 Yahoo!、 Google 等公司开展的传统信息检索业务主要专注于文本信息（多媒体信息也多依赖标题、附近文字等文本信息进行

检索），而很少涉及同样具有很高价值的多媒体信息的内容检索。使用 ASR 能够自动获得音频、视频中的语音所对应的文本标注，很容易检索包含语音的音频、视频，方便用户快速获取所需信息。

ASR 还可以与对话系统（Dialogue System）、机器翻译（Machine Translation）等技术相结合，构成语音对话系统（Spoken Dialogue System，SDS）、语音翻译机（Spoken Language Translation）等重要应用。近来，随着苹果（Apple）公司的 Siri 等 SDS 产品的流行，兼具人类的智慧、计算机的高效和互联网的无所不知的电子秘书让人机交互接近于人与人之间的自然交互，为用户带来了革命性的体验作为 SDS 人机交流的媒介，ASR 正是其中的关键之一。Siri 的成功使得 SDS 受到了空前的关注，点燃了人机交互与移动互联网连接的热点，越来越多的用户希望将 SDS 应用于汽车、房屋、家电等设施。ASR 的进步可以提高 SDS 的易用性，使我们的生活更加方便。

最后，ASR 的发展还有助于增进我们对人类语音本身及其产生和接收机制的了解，从而增进我们对人类智能的产生、发展、表达的认识[1]，促进语音学（Phonology）、认知科学（Cognitive Science）和人工智能（Artificial Intelligence）的进步，对理论研究具有重要意义。

由于具有广泛的应用价值和理论意义，ASR 技术 60 年来[2]一直是学界、工业界研究的热点。包括贝尔（Bell）实验室、 麻省理工学院（Massachusetts Institute of Technology, MIT）、卡耐基梅隆大学（Carnegie-Mellon University, CMU）、剑桥大学（University of Cambridge）、国际商业机器 （International Business Machine, IBM）公司、微软公司研究院（Microsoft Research, MSR）等世界顶尖学术机构和信息技术公司都投入了大量精力进行 ASR 的研究，极大的推进了 ASR 技术的进步并产生了许多相关应用。目前，ASR 研究早已从早期的特定说话人、小词汇量、孤立字语音识别问题扩展到了非特定说话人的大词汇量连续语音识别（Large Vocabulary Continues Speech Recognition，LVCSR）[134]问题，产生了诸如 IBM ViaVoice 等成功的产品。目前国外的研究热点正集中于自然语音（Spontaneous Speech）和跨语言语音（Cross Lingual Speech）中广泛存在的发音变异（Pronunciation Variation）问题，希望通过使 ASR 对发音变异更加鲁棒（Robust）来提高系统的实用价值。

由于计算机中汉字的输入比英语更为困难，通过 ASR 降低汉语这一世界上使用者最多的语言的输入难度有着格外重要的价值。标准普通话（Standard Putonghua, PTH）是汉语的标准语音，但由于多样的汉语方言（Regional Dialect）的存在，普

通话语音中常带有不同程度的方言口音（Regional Accent）。统计表明，80%的普通话使用者带有方言口音，其中 55%的说话人带有严重的方言口音[3]，而实验表明方言口音通常会导致多数汉语 ASR 的性能严重下降[4]。所以研究带方言口音普通话的 ASR，使它对方言口音产生的发音变异具有更好的鲁棒性是汉语 ASR 应用进入成熟的一个关键问题，也是 MIT[108]、伊利诺伊大学香槟分校（University of Illinois at Urbana Ch 声学模型 paign）[86]、佐治亚理工学院（Georgia Institute of Technology, GaTech）[109]、MSR[59][87]、香港中文大学（Hong Kong University of Science and Technology）[4][13]、清华大学[78]等众多研究机构关注的热点。

本文研究带方言口音的普通话 ASR，从语音属性这一 ASR 的核心角度出发，分别使用传统的基于单一语音属性的ASR和新型的基于多语音属性的ASR两种方法，提高 ASR 对方言口音变异鲁棒性（Robustness）。本文提出的部分算法对机器学习（Machine Learning）中的其它问题同样具有应用价值。

本章剩余部分按照如下方式组织：1.2 节介绍了口音和方言的关系，并从声韵母的角度分析了汉语方言口音的特点；1.3 节介绍了隐式马尔科夫模型的经典统计 ASR 及带口音语音识别研究的现状；1.4 节详细介绍了语音属性和基于多语音属性的 ASR 及其研究现状；1.5 节介绍了本文的组织结构和创新点，最后是本章小结和实验数据库的说明。

1.2　口音与方言的关系

方言是指某个语言在发音（Pronunciation）、词汇（Vocabulary）、习语（Idiom）等方面具有地方特色的变种（Variety）[5]。例如，粤语（Yue）是汉语的一种变种语言，主要在广东和香港地区使用，它与作为标准汉语的 PTH 在发音和用语上都有显著差异，如粤语中词语“老窦”读作[lou dao][1]而非 PTH 话中的[lao dou]；同时“老窦”在粤语中表示“父亲”之意，而在 PTH 中则表示“一个姓窦的熟人”。

口音（Accent）是一种由某个特殊个体、地区或国家所特有的发音方式[1]。也就是说，带口音的语音与标准语音间的差异主要体现在发音方面，而不包括词汇等层面。例如一个母语是粤语的说话人，因粤语和 PTH 的发音差异，在学习 PTH 时其 PTH 发音往往会受到粤语的影响。如前述例子中“窦”的韵母’ou’[2]在粤语中读作’ao’，类似的情况还出现在“口”（粤音[hao]）、“九”（粤音[gao]）、“手”

1 中括号中为汉语拼音，下同。
2 单引号中为声韵母，下同。

（粤音[sao]）等汉字中。这样当粤语口音的说话人在日常交流中说 PTH“窦”、“口”、“九”、“手”时，很容易把韵母’ou’说成为’ao’。本文主要关注这种因地域原因造成的口音，即方言口音。后文简称方言口音为口音。

本节首先主要介绍汉语方言的产生和特点，并从传统语音学的角度介绍了本文使用的川语、粤语和吴语。

1.2.1　汉语方言和方言口音

汉语是指以汉民族语言为主构成的包含一系列可以在不同程度上互相理解的方言的一种有声调（Tone）语言[6]。汉语的书写单位是汉字。汉字是以象形文字（Pictograph）为基础发展而来的，东汉的许慎在其著作《说文解字》中将汉字的构字法分为象形、指事、形声、会意、转注、假借等 6 种，统称为六书。其中除使用形声及部分转注方法构造的汉字外，其它汉字的读音都很难从文字本身判断出来。这样当说话人因地理等原因出现长期分隔后，他们对汉字的读音很容易出现差异，而不同的生活环境和经验还会产生不同的词汇和习语，这样就产生了不同的汉语方言。汉语方言间的发音差异往往非常显著，例如统计表明，粤语和 PTH 之间有多达 60%的发音具有显著差异，而同一个汉字在粤语和 PTH 中的发音差异往往与同一个词在英语和法语中的读音差异一样大[7]。同时，由于中国历史悠久、幅员广大，汉语方言的种类也格外多。研究表明，汉语有官话（Guanhua）、粤语、吴语（Wu）、湘语（Xiang）、赣语（Xiang）、闽语（Gan）、客家话（Kejia）等 7 大方言，如图 1-2 所示[137]，它们还可以被进一步分为 30 多种子方言[6]。PTH 就是在北方官话的基础上通过规范发音和语法得到的。另外很有趣的是方言中包含了很多不同时期的古汉语发音和词汇[7]，如粤语的 9 种声调中就包含了唐代汉语的 8 个声调；前文例子中的粤语用“老窦”表示“父亲”就源自古汉语对《三字经》中“窦燕山，有义方；教五子，名俱扬。”一句的引申。

作为一种单音节（Monosyllable）语言，汉语的音节结构非常简单。汉代就发现可以将一个汉字的读音分成前后两半，利用一个汉字的前半部分读音和另一个汉字的后半部分读音拼成一个新的音节为其它汉字注音，即所谓“反切”注音法[9]。现代汉语的声韵母（Initial/Final, IF）系统和这种思路一脉相承，将一个音节（Syllable）拆分成“声母+韵母”的形式。普通话声韵母集中（Putonghua Initial/Final, PTH-IF）共有 22 个声母和 36 个韵母[9]，其它方言的 IF 数目和种类则各有不同[7][10]。IF 的差异导致了方言间音节的差异，故传统汉语语音学对方言发音差异的分析主要基于 IF[7][9][10]，下一小节主要从 IF 角度分析川、粤、吴 3 种典型汉语口音与 PTH

的发音差异。

图 1-2　汉语 7 大方言区及其使用范围

1.2.2　川语、粤语、吴语口音

川语（Chuan）属于西南官话的一种，是官话的一种代表性的子方言，主要在四川省和重庆市等地区使用。川语声韵母集有 21 个声母和 38 个韵母[10]，川语声母中缺少 PTH-IF 的’zh’、’ch’、’sh’、’r’、’l’，但多了一些浊辅音。于是川语中可能使用另外的声母来取代缺失的 PTH 声母，如 PTH 读音为[l-]的汉字在川语中通常读作[n-]。但并非所有 PTH-IF 中的’l’在川语中均用’n’替代，如“年”的声母为鼻音舌面前浊音。韵母方面，川语没有 PTH-IF 的’e’，相应汉字的韵母依据声母的不同可能分别替换为’o’或/ae/。川语还没有’an’，’an’在川语中经常对应于/ae/。另外值得一提的是虽然川渝中有韵母’uen’，但在声母为舌尖前音’z’、’c’、’s’或舌尖中音’d’、’t’、’n’时对应音节的发音为[-en]，在其余声母后仍为[-uen]。最后，川语

还缺少韵母’eng’和’ing’，PTH 中韵母为’eng’的汉字会被读作[-en]或[-ong]，韵母为’ing’的汉字则读作[-in]。

粤方言是广东省、香港特别行政区等地区主要使用的方言。粤语声韵母集有 20 个声母和 53 个韵母[10]。与川语相似，粤语的声母中没有’zh’、’ch’、’sh’、’r’。但粤语却有边音声母’l’和额外的鼻音声母/ng/。粤语的韵母结构与 PTH 有很大差异，如果将一个韵母顺序划分成韵首、韵腹和韵尾 3 部分[3]，PTH 仅有 2 种韵尾 ‘n’和/ng/，而鼻音’m’、’n’、/ng/和清塞音’d’、’t’、’k’在粤语中都可以作为作为韵母的结尾，这是导致粤语韵母数目多的主要原因。同时粤语没有 PTH 中的后响复韵母（’ia’、’ua’、’ie’、’uo’、’ue’），这些韵母在粤语中对应于只有相应韵腹的读音，这种现象被语言学家称作韵首的脱落[10]。此外前文中提到 PTH 中发音为[-ou]的汉字常被读为[-ao]也是粤语的发音特点之一。

吴方言主要在江苏省、浙江省和上海市等地使用。吴语声韵母集共有 26 个声母和 49 个韵母[10]。在所有主要汉语方言中，吴语具有最丰富的浊声母。PTH-IF 中的清塞音、清塞擦音和清擦音在吴语中均有对应的浊音。如吴语中有与清音舌尖中塞音’d’对应的浊音舌尖中塞音。与川语和粤语类似，吴语中也没有舌尖后音’zh’、’ch’、’sh’，并常代之以对应的舌尖前音’z’、’c’、’s’。吴语中还缺少 PTH 的浊音舌尖后擦音’r’，而在不同的上下文中分别代之以额外的浊音舌尖前擦音或浊音舌面前鼻音。韵母方面，吴语中单元音韵母很丰富，但却很少有 PTH 中的前响复韵母（’ai’、’ei’、’ao’、’ou’）。最后吴语仅有一种鼻音韵尾/ng/，故 PTH-IF 中以’n’结尾的韵母’an’、’en’、’in’ 等在吴语中常读作’ang’、’eng’、’ing’[10]。

于是母语为川语、粤语或吴语的说话人在学习使用 PTH 时拼读的音节常带有其母语方言 IF 的特点，分别产生带有川语口音的普通话（Chuan Accented Putonhua）、带有粤语口音的普通话（Yue Accented Putonghua）和带有吴语口音的普通话（Wu Accented Putonghua）[3]。例如粤语背景的说话人在说 PTH 时很难读准“张”字的声母’zh’，他可能使用粤语中与之读音最接近的声母’z’来替代’zh’[4]，从而发生了发音变异 [11]，本文用记号’zh’→‘z’的形式来表示’zh’被拼读成’z’的发音变异。这些口音相关的发音变异具有以下特点：

(1) 一个 PTH 发音可能变异为不同发音，如粤语口音中同时有发音变异’zh’→‘z’和’zh’→‘j’，这些发音变异通常发生在不同的上下文中，语音学中将它们总结为一些上下文相关（Context-Dependent）的变异规则。2) PTH 中的不同 IF

3 后文将详细介绍汉语韵母的结构。

也可能变异为同一个发音，如粤语中存在发音变异变体’j’→‘z’、’s’→‘z’和’x’→‘z’[10]。3) 不同说话人的发音变化的倾向性（Tendency）不同，例如口音较重的说话人使用’z’取代 PTH-IF 的’zh’，而口音较轻的说话人的发音却可能介于’zh’和’z’之间[4]。4) 口音产生的发音变异往往是单向的（Unidirectional）[12]，如粤语中只有’zh’→‘z’而没有’z’→‘zh’。当涉及到多方言口音时，发音变异的情况往往更加复杂：5) 一个在多个方言区长期居住过的说话人可能带有混和方言口音（Mixed Accent），即其口音可能同时带有两种方言的发音特点[13]。于是需要将不同方言产生的发音变异进行联系和比较。6) 不同方言口音中发音变异的趋向性可能不同，具有不同的声学特性。例如吴语中’p’、’k’或’t’的破裂性较强，而粤语中这 3 个音则要软一些[10]，故对不同口音中记号相同的发音变异不能简单的合并处理[13]。

值得说明的是，虽然汉语音节具有不同声调，但研究表明声调对汉语 ASR 并不是必要的[14][15]。由于汉语声调系统的变化有其复杂性[9][10]，声调探测准确率比较有限[15][16]。所以虽然口音也会带来声调上的发音变异[9][10]，但本文并不涉及这方面的讨论。本文统称因口音导致的各种发音变异现象为口音变异（Accented Pronunciation Variation）。最后有必要区分音素（Phoneme）和音子（Phone）的差别，音素通常用于指代说话人在发音前脑海中想表达的发音，也称为基准发音（Base Form）；说话人实际说出的语音称为音子，或表象发音（Surface Form）[4][17]

1.3　基于单一语音属性的统计ASR

60 年来，ASR 的发展经历了许多阶段，诞生了包括基于语音学知识和频谱分析（Spectrum Analysis）技术的声学-语音学方法（Acoustic-Phonetic Approach）、利用模式识别技术的模板匹配方法（Template Matching Approach）、基于规则的专家系统（Expert System）方法、基于隐式马尔科夫模型（Hidden Markov Model, HMM）和人工神经网络（Artificial Neural Network, ANN）等算法的统计机器学习方法（Stochastic Machine Learning Approach）[2][22]。

声学-语音学方法的核心思想是对各种时域（Temporal）、频域（Spectral）特征分别构造发音特征（Articulatory Feature, AF）[4]探测器，在探测到的发音特征中依据预定义的映射规则进行搜索以得到词语序列 [2]。声学-语音学方法完全基于实验语音学知识鉴别发音特征并合成词语，故声学-语音学方法通常不需要任何语音数据库。但受限于语音学研究，可用的先验知识并不完备，对如何构造 ASR 缺乏

4 本文第 3 章将详细介绍发音特征。

指导。

模板匹配方法通过收集数据中词语的语音片段构造 ASR，这些片段称为标本（Exemplar）；识别时利用统计模型比较输入语音与已有标本间的相似度，选择相似度最高的标本对应的词语作为语音的识别结果[2]。这种方法的时间、空间复杂度很高，并且对信道（Channel）、噪声（Noise）等方面的变化缺乏鲁棒性。

专家系统方法将 ASR 分为多个层次，按照自底向上（Bottom-Up）或自顶向下（Top-Down）等策略逐层集成声学（Acoustic）、词法（Lexical）、句法（Syntactic）、语义（Semantic）等方面的规则，基于规则进行语音识别[2]。与其它人工智能（Artificial Intelligence）问题一样，专家系统在 ASR 中也面临着如何恰当设计规则及规则增多后系统混淆增大的问题。

与以上各种方法相比，基于倒谱系数等单一语音属性和 HMM 等统计模型的 ASR 具有准确率高、数学背景可靠等优点[23]，关于口音和发音变异的主流研究都使用基于单一语音属性的 ASR。本节以 HMM 声学模型为主这种 ASR 方法。

1.3.1　统计语音识别的原理

统计 ASR 系统主要包括以下 3 部分：特征提取、模式分类、搜索解码，如图 1-3 所示。其中搜索解码过程可能包含对点阵进行重评分处理。

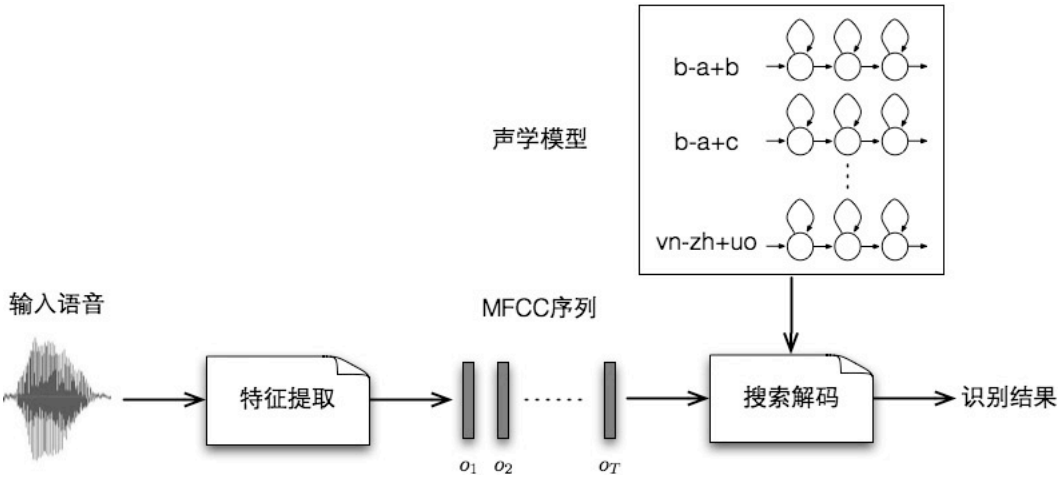


图 1-3　统计语音识别流程图

ASR 的任务是在输入一段语音的情况下得到相应的音节[5]序列 B。对一段输入

5 在汉语中，音节序列可以很容易转换为词语序列。本文中依据识别任务不同识别结果还可能为音子序列或发音特征序列等，下同。

语音，首先通过特征提取模块将其转换为更具有区分性且更容易处理的时序声学特征，也称为观测序列，记作O。后验概率（Posteriori）$P(B|O)$给出了已知O的情况下每种音节序列的可能性，可以合理假设可能性最高的音节序列$\hat{B}$为 ASR 的识别结果。即

$$P(\hat{B}|O)=\max_{B}P(B|O). \tag{1-1}$$

称为最大化后验概率（Maximum A Posteriori, MAP）。

将贝叶斯[25]公式

$$P(B|O)=\frac{P(O|B)P(B)}{P(O)} \tag{1-2}$$

代入公式(1-1)，$P(O|B)$为观测序列由指定音节序列生成的似然度（Likelihood）。观测序列确定时$P(O)$为常数，故(1-1)可改写为

$$P(\hat{B}|O)=\max_{B}P(O|B)P(B). \tag{1-3}$$

$P(B)$代表了在指定语言中每个音节出现的概率，称为语言模型（Language Model），$P(O|B)$描述了从音节序列B对应的模型生成观测序列O的概率，称为声学模型（Acoustic Model）[2]。声学模型和语言模型用于对观测序列进行模式分类。求解具有最大后验概率的音节序列$\hat{B}$的过程则称为搜索解码（Decoding），即

$$\hat{B}=\arg\max_{B}P(X|B)P(B). \tag{1-4}$$

1.3.2　特征提取

ASR 中通常使用倒谱系数作为声学特征。常用的倒谱系数包括线性预测倒谱系数（Linear Predictive Cepstrum Coefficients, LPCC）[23]、梅尔频率倒谱系数（Mel-Frequency Cepstrum Coefficients, MFCC）[26]、感知线性预测（Perceptual Linear Prediective, PLP）[27]等。下面以最常见的 MFCC 为例简介倒谱系数，图 1-4 是 MFCC 的提取流程。

原始语音点序列$s_t,t=1,\ldots,T_0$首先要经过公式(1-5)的预加重（Pre-emphasis）处理。$0<\alpha<1$，称为预加重系数。

$$s_t' = s_t - \alpha \cdot s_{t-1} \tag{1-5}$$

公式(1-5)相当于将 s_t 序列通过 1 个高通滤波器以加强高频其信息。这是因为说话人的声道具有低通滤波器的物理特性，会导致语音中声门产生的高频信息被削弱。

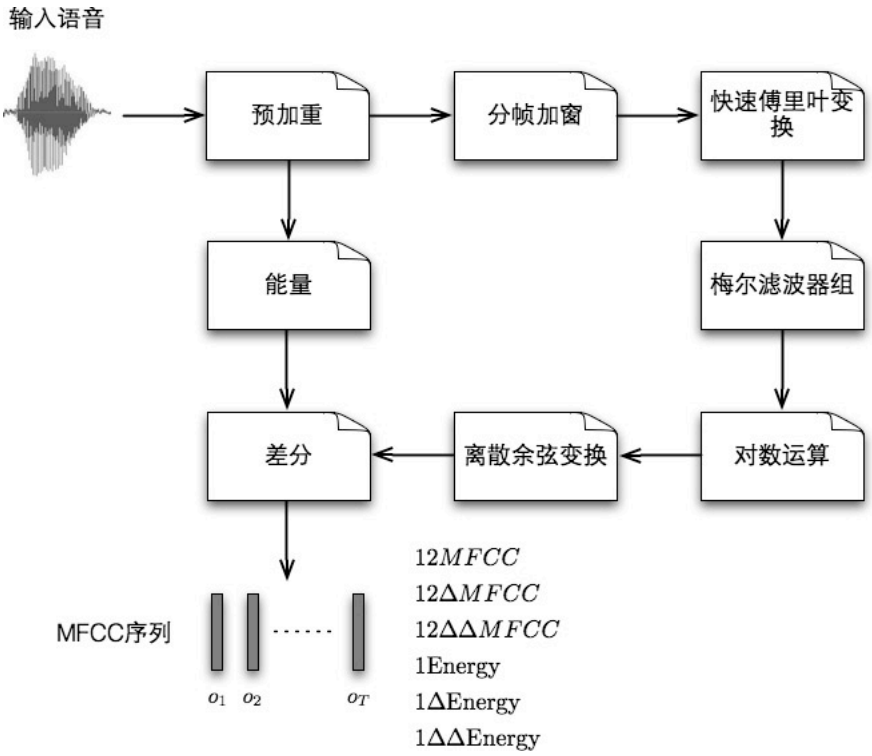


图 1-4　本文使用的 39 维 MFCC 特征提取流程

虽然语音波形是随时间连续变动的，且每个音段的起止时间点未知，但语音具有短时平稳（Short-Time Stationary）的特征，即 10~30ms 的语音的片段是近似平稳不变的[22]。于是可以将语音切分成等长的平稳片段，每个片段称为一个语音帧（Speech Frame）。为避免使用矩形窗（Rectangular Window）直接对语音切片导致的频谱泄露（Spectral Leakage）效应[28]，通常使用汉明窗（Hamming Window）[22]进行分帧，

$$s_t'' = \left\{0.54 - 0.46\cos\left(\frac{2\pi(t-1)}{T_0 - 1}\right)\right\} s_t'. \tag{1-6}$$

语音是连续变化的，如图 1-1 所示，所通常取汉明窗的长度大于帧长以使相邻窗之

间具有一定的重叠区域，减弱分帧给窗边缘带来的不连续性。记得到语音帧序列为 $x[n],n=1,\ldots,N$ 。

将语音帧序列使用公式(1-7)所示离散傅里叶变换（Discrete Fourier Transform, DFT）[28]转换到频域进行处理。设对应的频谱序列为 $X[k],k=1,\ldots,N$ 。

$$X[k]=\sum_{n=1}^{N}x[n]\cdot e^{-i2\pi\frac{k}{N}n} \tag{1-7}$$

工程中经常使用快速傅里叶变换（Fast Fourier Transform, FFT）取代 DFT，但 FFT 只能处理样本数为 2 的整数次幂的情况，故需要先对每帧中的语音样本点补零。

研究表明，人耳耳蜗（Cochlear）的感知能力存在许多临界频带（Critical Band），它们对于感受音质（Timbre）、音强（Loudness）等信息非常关键[22]。在信号技术中，可以利用带宽等于临界频带带宽且相互重叠的带通滤波器（Band-Pass Filter）组模拟耳蜗的特性。实验证实人的听觉系统对高频信息不敏感：临界频率在 1000Hz 以下的临界频带等间距分布，1000Hz 以上的临界频带的间距不断增大，频谱分辨率下降，通常在梅尔频率中利用等间隔分布的三角滤波器模拟这种特性[26]。梅尔频率的定义为，

$$Mel(f)=1125\ln(\frac{f}{700}+1). \tag{1-8}$$

梅尔频率下的三角滤波器的频域表达式可写为[22]，

$$H_m[k]=\begin{cases}0 & f_X[k]<f[m-1]\\ \dfrac{(f_X[k]-f[m-1])}{(f[m]-f[m-1])} & f[m-1]\le f_X[k]\le f[m]\\ \dfrac{(f[m+1]-f_X[k])}{(f[m+1]-f[m])} & f[m]\le f_X[k]\le f[m+1]\\ 0 & f_X[k]>f[m+1]\end{cases}. \tag{1-9}$$

其中 f_s 为语音的采样率，$f_X[k]=\frac{f_s}{N}k$ 为第 k 个语音点的频率；$f[m]$ 为第 m 个滤波器对应的临界频率。本文选择 22 个三角滤波器。

研究还发现，人耳对较低能量产生的变化更敏感，这与底数大于 1 的对数特性吻合，故对每个滤波器输出的能量求对数，有

$$S[m]=\ln\left[\sum_{k=1}^{N}|X[k]|^2 H_m[k]\right],0\le m<M. \tag{1-10}$$

最后使用第 2 型离散余弦变换（Discrete Cosin Transform, DCT）[28]求出每帧语音的倒谱系数，

$$c[n]=\sum_{m=0}^{M-1}S[m]\cos(\frac{\pi n}{M}(m+0.5)),0\le m<M. \tag{1-11}$$

倒谱系数中，声门特性处于较低维度，声道特性处于较高维度[17]。由于人类的声道特性存在一定差异，去除声道特性有助于降低说话人特性对语音的影响，故仅保留最低维度的倒谱系数。另外，倒谱系数的各维间的相互独立[28]，有助于降低进行模式分类的难度。

倒谱系数中通常还常使用倒谱均值归零化（Cepstral Mean Normalisation, CMN）技术[29]。语音通过不同的信道通常可以看做在频域乘以相应的传递函数， 取对数并进行 DCT 变换后相当于在每一维倒谱系数中分别偏移一个常数。假定原始语音中每一维倒谱系数的均值都为 0，那么此时的均值反应了信道特性。将倒谱系数中的每一维的均值归 0 可以去除信道的差异造成的影响。

MFCC 中通常同时包含能量与倒谱系数。每一帧能量为该帧中各样本的能量之和，再取对数；对每一帧能量分别减去输入语音中能量最大的帧的能量并加 1，可以将帧能量的范围进行归一化。

最后，通过在特征中包含一阶差分（1st Order Difference）和二阶差分（2n Order Difference）参数等动态参数 [29]可以在当前帧中引入附近帧的性质，一阶差分的计算公式为

$$d[n]=\frac{\sum_{\theta=1}^{2}\theta\cdot(c[n+\theta]-c[n-\theta])}{2\sum_{\theta=1}^{2}\theta^2}. \tag{1-12}$$

在公式(1-11)代入一阶差分可以得到二阶差分。使用一阶差分和二阶差分可以在 1 帧中最多增加它前后 8 帧的信息。这样就得到了 MFCC 特征，也将每个语音帧称作一个观测（Observation）。

可以看出，MFCC 降低了特征的规模，减少了说话人生理特征所带来的干扰，并去除了特征各维间的相关性。另外 MFCC 还通过模拟人类的听觉特性来增强与语音内容相关的区分性。需要注意的是 MFCC 没有直接触及音段等构成语音的核心要素，还保留了与语音识别无关的说话人情感等信息。

1.3.3　模式分类

1.3.3.1　声学模型

一个 HMM 可以看做一个包含有限状态（State）的机器，在每个时间点 t，HMM 都会从它的前一状态 i 跳转到当前状态 j，并根据状态 j 的观测密度（Observation Density）得到它生成当前观测 o_t 的概率 $b_j(o_t)$。观测密度描述了观测服从的分布，故也称为输出分布（Output Distribution）。HMM 中的一个重要假设是每个观测都是独立产生的随机信号[23][28]；观测服从 1 阶马尔科夫过程（1st Order Markov Process），即它的状态转移概率只和当前状态有关，与更早的历史无关[25]。通常使用离散常量 $a_{i,j}$ 作为转移概率，即状态转移服从几何分布（Geometric Distribution）[25]。图 1-5 为本文使用的 HMM 拓扑结构，每个 HMM 包含 5 个状态，状态 1 和状态 5 是用于简化公式形式的入口和出口虚状态，只有转移概率而没有观测密度[29]。状态间的转移都按照从左到右的顺序，且不存在状态被跳过的情况。令模型 Λ 按状态序列 $Q=q_1,q_2,\ldots,q_T$ 生成帧序列 $O=o_1,o_2,\ldots,o_T$，有

$$P(O,Q\mid\Lambda)=\prod_{t=1}^{T}P(o_t,q_t\mid\Lambda)=a_{q_0,q_1}\prod_{t=1}^{T}b_{q_t}(o_t)a_{q_t,q_{t+1}}. \tag{1-13}$$

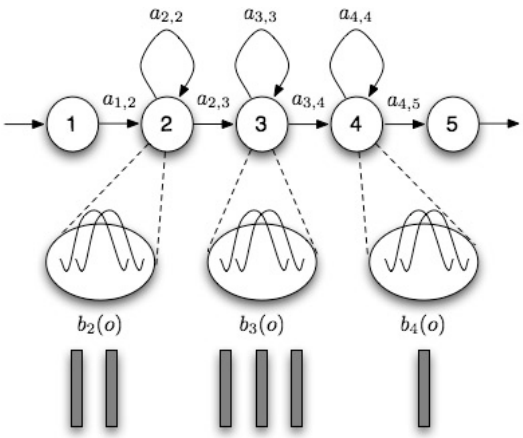


图 1-5　本文使用的 5 状态 HMM 拓扑结构

q_0 和 q_{T+1} 分别为入口和出口状态。HMM 称为“隐式”模型是因为在使用模型时仅已知观测 O，而状态序列 Q 未知，Q 为隐变量（Hidden Variable）。

声学模型中 HMM 通常对应一个 IF，故

$$P(O,Q \mid W) = P(O,Q \mid \Lambda). \tag{1-14}$$

综合公式(1-13)和(1-14)，可以由加法公式[26]可求得似然度

$$P(O \mid B) = \sum_{Q} P(O,Q \mid B) = \sum_{Q} a_{q_0,q_1} \prod_{t=1}^{T} b_{q_t}(o_t) a_{q_t,q_{t+1}}. \tag{1-15}$$

通常根据观测密度的类型对 HMM 分类。ASR 中以连续观测密度最常见，包括高斯混合模型（Gaussian Mixture Model, GMM）[32]、迪利克雷混合模型（Dirichlet Mixture Model）[33]等。本文使用 GMM 模型作为观测密度[34]，这是因为随着高斯成分的增加，GMM 在理论上可以按任意精度逼近任何随机分布[32]。状态 q_t 的 GMM 观测密度的定义为

$$b_{q_t}(o_t) = \sum_{m=1}^{M_{q_t}} c_{q_t m} \mathrm{N}(o_t; \mu_{q_t m}, \Sigma_{q_t m}), \tag{1-16}$$

其中 M_{q_t} 是观测密度中高斯成分（Gaussian Component）的数目，$c_{q_t m}$ 是第 m 个高斯成分的权值，$\mathrm{N}(o;\mu,\Sigma)$ 是均值向量为 μ、协方差矩阵为 Σ 的 d 维多变元高斯分布（Multivariate Gaussian Component），

$$\mathrm{N}(o;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)}. \tag{1-17}$$

HMM 的成功应用推动了 ASR 的进步，但也具有一些局限性[23]：

(1) 首先，HMM 假定了每个输入观测都是统计独立（Statistical Independent）的，这样才能将公式(1-13)中的联合概率写为连乘积的形式。但从图 1-1 可以看出发现相邻的观测间显然是相关的。

(2) 其次，使用 GMM 很难准确近似假设中的真实分布。由于缺乏对真实分布的先验知识，无法确定 GMM 中应包含多少高斯成分。另外训练 GMM 常用的最大期望（Expectation Maximization, EM）算法也很难收敛到目标函数的全局最优解[31]。

(3) 1 阶马尔科夫性本身限制了 HMM 的下一状态的变化仅取决于当前状态，但语音学分析表明这一假设在很多情况下并不合理[35]。

1.3.3.2　HMM 的替代模型

随着机器学习和模式识别的不断进步，越来越多的统计模型被应用到 ASR 中，可以大致分为以下 3 类。

第一种模型对 HMM 的特征限制进行扩展。HMM 要求输入观测都是统计独立的，但 MFCC 等往往不能满足这种限制。取消这一限制的方法之一是使用区分性（Discriminative）的模型，与生成性（Generative）的 HMM 不同，区分性模型直接优化后验概率，无需为求解似然度而强加观测的独立性假设[36][37]。典型的区分性模型包括最大熵马尔科夫模型（Maximum Entropy Markov Model, MEMM）[38]、条件随机场（Conditional Random Field, CRF）[39]、隐式条件随机场（Hidden Conditional Random Field, HCRF）等[37]。另外，由于最大化区分性模型的目标函数 $P(B|O)$ 的过程相当于直接求解统计 ASR 的基本公式(1-1)，比计算公式(1-3)更直接，故称之为直接模型（Direct Model）[38]。

第二类模型通过修改模型结构以降低马尔科夫性给 HMM 带来的约束。这类模型又可以分为两种类型。第 1 种类型使用跳转概率与时间有关的半马尔科夫（Semi-Markov）假设取代 1 阶马尔科夫假设，得到的模型通常具有更好的转移特性，如隐式半马尔科夫模型（Hidden Semi-Markov Model）[40]、半马尔科夫条件随机场（Semi-Markov Conditional Random Fields, SCRF）等[41][42][43]。第 2 种类型将 1 阶马尔科夫性从 1 维线性结构中前后相连的 2 个状态扩展到 2 维图中以任意方式连接的多个邻接状态，如动态贝叶斯网络（Dynamic Bayesian Network, DBN）就是 HMM 在这种思想下得到的扩展形式。DBN 中状态间的复杂约束常用来描述发音时不同器官间的关系[35]。

第三类模型包括 ANN、深层（Deep-Structured）模型等具有认知科学背景的模型。ANN 模拟了人脑神经元（Neuron）构成神经网络（Neural Network）的连接和结构，可以近似任意函数并使用任意特征[31]。利用 ANN 进行模式划分 20 多年来一直是 ASR 的重要研究方向之一[44][45]。深层模型通常指由 3 层以上信息处理系统构成的一种层次结构的模型[46]。它模拟了人脑包含多层神经网络的深层结构，并可以利用未标注数据进行训练，而仅需要少量的标注数据进行调整[46]。基于深层模型进行模式划分的 ASR 是近年来兴起的研究热点，包括使用包含 4 层受限玻尔兹曼机（Restricted Boltzmann Machine）的深层置信度网络（Deep Belief Network）进行音子识别、使用包含 7 层上下文相关深层神经网络（Deep Neural Network, DNN）进行自然语音识别的工作等，都取得了显著的进展[47][48][49]。

1.3.3.3　语言模型

语言模型将高层语言信息转换为概率以修正声学模型生成结果[50]。常见的语言模型包括 n-gram 统计模型[50]、基于移进归约句法分析器（Shift-Reduced Parser）的结构化语言模型[101]、基于 CRF 和 MEMM 的语言模型[51][52]、基于 ANN 的语言模型等[53]。语言模型在训练语音与测试语音内容相关时可以显著提高 ASR 的准确率，但反之也可能降低识别率[50]。

为更准确的评估算法对发音变异造成的影响，本文不使用语言模型以避免语言层信息的影响。这时公式(1-4)退化为

$$\hat{B} = \arg\max_{B} P(X \mid B). \tag{1-18}$$

1.3.4　搜索解码

许多经典算法可以求得公式(1-18)中的参数最大化过程，如 A*算法、前向后向（Forward-Backward）算法等。下面简介本文使用的维特比（Viterbi）算法[54]。

设一个 HMM 共有 N 个状态，令 $\phi_j(t)$ 为 t 时刻处于第 j 个状态的所有状态序列中似然度最大的序列，$1 \le j \le N$，有

$$\phi_j(t) = \max_i \{\phi_i(t-1)a_{i,j}\} b_j(o_t), \tag{1-19}$$

且

$$\phi_j(1) = \begin{cases} 1 & j = 1 \\ a_{1j} b_j(o_1). & \text{otherwise} \end{cases}. \tag{1-20}$$

搜索得到的最大似然度为

$$\hat{P}(O \mid W) = \max_i \{\phi_i(T) a_{i,N}\}. \tag{1-21}$$

实际应用中，为避免因数值范围过大而导致计算溢出（Overflow），对公式(1-19)-公式(1-21)常取以自然数 e 为底的对数，搜索目标变为 $\ln(\phi_j(t))$，得分称为对数似然度（Log-Likelihood）。

ASR 中通常使用有限状态网络（Finite State Network, FSN）对识别结果 $\hat{B}$ 进行一定的约束。搜索中遍历的每个可能状态序列都是 1 条从网络的起始节点到终止节点的路径（Path）。图 1-6 中给出了一种简单的网络结构，每个节点对应于 1 个音节等单元。由于存在反向边，路径可能包含任意节点的任意多次出现。这种网

络结构的通常称为自由语法（Free Grammar）[29]。另外值得注意的是，当使用上下文相关的声学模型时，需要将网络中每个节点用符合它在当前上下文的特定 HMM 代替。

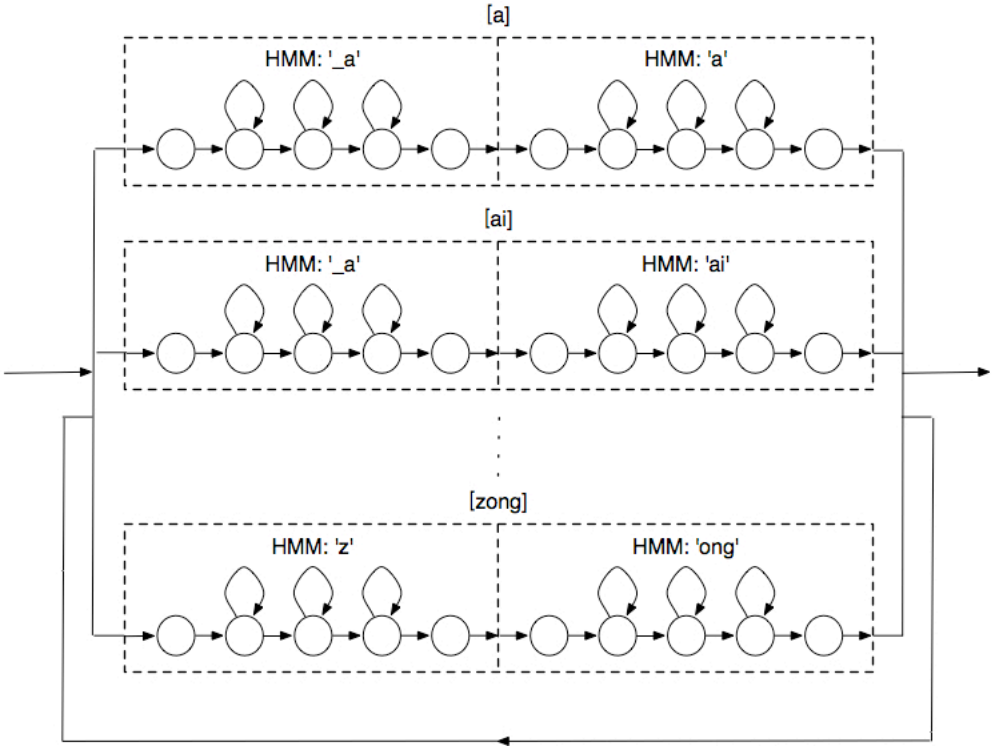


图 1-6　自由语法的有限状态网络

由公式(1-19)，维特比算法在搜索中需要保留所有路径，称包含所有路径的数据结构为点阵（Lattice）[29]。图 1-7 给出了识别音节[chao]的部分点阵，图中所有路径产生的 HMM 序列相同（即识别结果 $\hat{B}$ 相同），但状态序列不同。将解码产生的每个音节序列 $\hat{B}$ 称作一个假设序列（Hypothesis），对应的状态序列为网络中所有产生这一假设序列的路径中对数似然度最大的一条。相应的，称输入语音的人工标注 B 为参考标注序列（Reference）[29]。点阵中对数似然越高的假设序列在识别结果中排序越靠前，前 n 个序列称为 n 最优结果（n Best Results）[6]。另外，还可以在参考序列 B 已知的情况下构造识别结果只可能为 B 的 FSN，其中具有最高对数似然度的状态序列具有每个状态在语音中的起止时间，称之为强制对准（Forced Alignment, FA）[29]。

6 如无特殊说明后文中的识别结果通常值 1 最优结果。

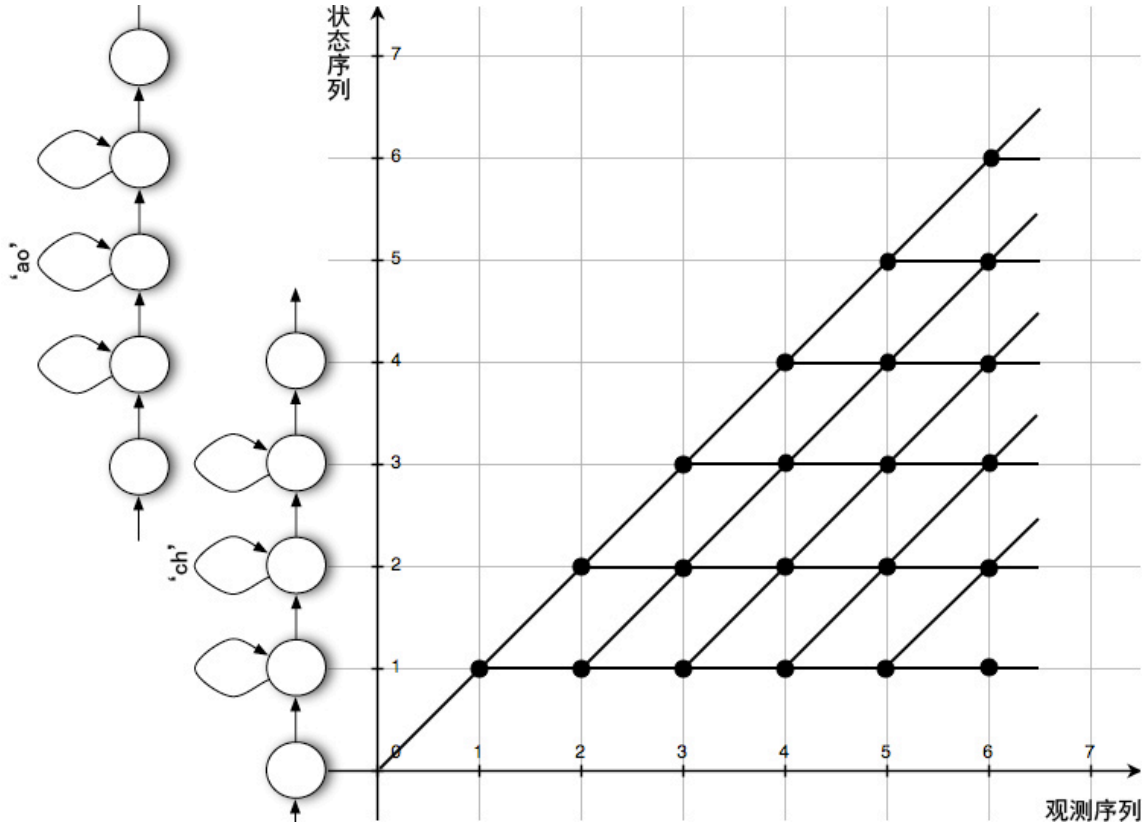


图 1-7　识别[chao]时的部分点阵

LVCSR 任务中的点阵在每个时间点可能包含上百万条路径，对所有路径都扩展的代价非常高。为提高搜索效率通常采用路径合并和剪枝。路径合并策略是在每一时刻将处于同一状态且已产生相同假设序列的所有路径中仅保留具有最大对数似然的 1 条。图 1-7 中经过每点的各条路径都进行了状态合并。本文在剪枝时采用基于贪心（Greedy）策略的束搜索（Beam Search）算法[50]。假设在每一时刻对数似然度很低的路径在后续的搜索中很难成为最优路径，将这些路径提前排除可以降低搜索的复杂度。本文采用 1 个预先设定的分数阈值 t_0 控制剪枝[29]：在每一时刻仅对满足以下条件的路径 j 进行扩展，

$$\max_i\{\ln(\phi_i(t))\}-\ln(\phi_j(t))\le t_0, \tag{1-22}$$

即丢弃与当前最优路径间对数似然度相差 t_0 以上的路径。这种剪枝方法相当于在搜索中维护了宽度为 t_0 的搜索束（Beam）。

经验表明，通常 ASR 在参考序列和假设序列音节数目相近的时候识别率最高[109]。为达到相近的音节数，在点阵中的每个音节后都为当前路径的对数似然度加

penalty 作为对新生成一个音节的鼓励或惩罚。*penalty* 预先设定，称为词惩罚分（Word Penalty Score）[29]。类似这样在搜索中或搜索后修正路径的对数似然度的方法称为点阵重评分（Lattice Rescoring），使用这种方法可以方便利用语言模型等修正识别结果[29] [55]。

本文采用正确率（Correct%）和准确率（Accuracy%）评估 ASR 的识别结果。对参考标注序列和识别结果使用动态规划（Dynamic Programming, DP）算法进行对齐，当两个序列中正确匹配的音节对数目最多时。设参考序列中有 N 个音节，正确匹配的音节数目为 H。识别结果中未匹配的音节称为插入错误（Insertion Error），记为 I；参考标注序列中未匹配的音节称为删除错误（Deletion Error），记为 D [29]；其它的称为替换错误。定义正确率为

$$\frac{H}{N}\times 100\%. \tag{1-23}$$

准确率为，

$$\frac{H-I}{N}\times 100\%. \tag{1-24}$$

对识别结果做深入分析时通常使用混淆矩阵（Confusion Matrix），定义为每个音节被错误识别为其它所有音节的概率所构成的矩阵，矩阵的对角线元素为音节的正确率。也就是说，混淆矩阵反映了每个音节与其它所有音节的混淆程度。

1.4 基于语音属性的ASR和带口音语音识别的研究现状

语音属性（Speech Attribute）是指表征语音性质的声学、语音学、语言学等方面的特征[88]。从特征提取技术的角度，语音属性又可以分为时域特征、频域特征、仿生学（Bio-Inspired）特征、利用自然语言处理（Natural Language Processing, NLP）技术获得的语法、语义相关的特征，反应说话场所特性的先验特征等。ASR 研究中常用的 MFCC、PLP 等特征也都属于语音属性。

使用 MFCC 等特征的传统统计 ASR 可以视作基于单一语音属性的 ASR，近年还兴起了通过整合多种语音属性进行识别的基于多语音属性的 ASR。由于统计 ASR 中通常使用 HMM 作为声学模型，故若无特殊说明，后文用基于单一语音属性的 ASR 指代使用基于单一语音属性的 HMM 作为声学模型的 ASR；通常利用探测器从语音中探测语音属性，故基于多语音属性 ASR 又称作探测式 ASR。

特征是模式识别的核心问题[22][24][89]，本节从作为 ASR 特征的语音属性出发，介绍基于单一语音属性 ASR 方法的主流带口音语音识别研究的现状。最后介绍了语音属性、基于多语音属性的探测式 ASR 及其研究现状。

1.4.1　基于单一语音属性的带口音语音识别研究现状

1.3.2 节以 MFCC 为例详细介绍了倒谱系数特征的提取方法及其性质。事实上，MFCC、PLP 等倒谱系数特征在近 20 年的 ASR 研究中占据了支配地位[90]，本小结综述的几乎所有主流带口音语音识别研究都只使用 MFCC 或 PLP 一种语音属性。

由于标准语音的采集和标注难度较小，容易获得训练统计模型所需要的大量数据，通常基于标准语音构造 ASR。当说话人带有某种口音时，输入语音中含有的口音变异会导致测试数据的声学特性与标准语音训练的 ASR 失配（Mismatch），导致 ASR 的性能急剧下降[4][17]。

ASR 中，口音变异通常可以分为发音层变异（Phonetic Level Variation）和声学层变异（Acoustic Level Variation）两种[13]。发音层变异是指说话人把一个语音完全错发成了另外一个发音，如一个粤语口音的说话人可能把基准发音 ‘r’说成表象发音’l’。由 1.2.2 节，发音层变异通常可以视作音素间上下文相关的概率转移[56]，当遇到发音层变异时 ASR 就会产生把基准发音识别为表象发音的识别错误[13]。声学层变异指发音的声学特征位于空间中基准发音模型和表象发音模型之间[12]所产生的识别错误，例如粤语口音说话人的拼读 PTH 韵母’ou’时他的发音可能介于’ou’和’ao’之间。需要注意的是，口音变异只是声学层变异的一种来源，语音数据库的信道和质量、特征的区分性、算法性能等原因都可能导致声学模型因对声学空间划分不当而具有一定的混淆性，产生声学层变异[57]。

ASR 中常通过分别修正由发音层变异或声学层变异导致的失配来提高模型对口音变异的鲁棒性。可以通过数据驱动（Data Driven）的方法或基于语音学知识得到相应的口音变异。当口音变异的表象发音所对应的音素在声学模型中已有对应模型时，带有多种发音候选的多发音字典（Multiple Pronunciation Dictionary）是一种常用方法[58][59][60][61]。例如对口音变异 ‘zh’→‘z’，表象发音’z’在字典中对应于基准发音’zh’或’z’的概率会作为路径得分的一部分影响 ASR 的识别结果。研究表明，利用数据驱动[58][59]或语音学知识[60][61]得到的发音层变异都可以在一定程度上提升 ASR 对带口音的汉语、英语的识别率。类似的，利用数据驱动方法获得的发音层变异可以用于在搜索解码时扩充 FSN 以解决口音变异[62]。当发音层变异的表象发音所对应的音素在原声学模型中不存在时，通常使用扩展发音基元（Phone Set

Extension）的方法为该音素训练模型并将其作为新发音基元加入声学模型[63][64][65][66]，例如可以建立与清辅音’d’对应的浊辅音模型以解决吴语口音中相应的发音变异。论文[63][64]为基于语音学规则得到的 SAMPA-C[7]单元建立模型并使用语音学家人工标注的数据进行训练；论文[65]则从吴语口音数据中自动获得与发音层变异直接相关的 ASR 识别错误，并将其作为新增发音基元；进一步，论文[66]将语音分割为频谱特征较为平稳（Stationary）的片段并使用 K 均值（K-Means）[31]算法进行聚类以得到全数据驱动生成的发音基元。含有扩展发音基元的声学模型可以使用动态 HMM 选择（Dynamic HMM Selection）算法解码以更好地结合口音变异的上下文条件[67]。

以上为口音变异建模的方法可以总结为用自动或人工方式得到的口音变异规则扩充字典、识别器或发音基元。由于通常存在大量的口音变异规则，且对如何准确利用这些规则缺乏指导，实践表明直接扩充会显著增加 ASR 的词法混淆度，从而导致新引入的规则在解决对应口音变异的同时也造成了新的识别错误[68]，制约了系统性能的提高。为解决这一问题，论文[4][58][59][60][64][65]分别利用语音学知识及对数似然率（Log-Likelihood Ration）、频率、卡方检验（Chi-Square Test）等统计假设检验方法对规则集合剪枝，只对存留的口音变异进行处理；论文[66]利用迭代的策略对在词典和发音基元中引入的口音变异规则进行联合优化。研究表明，这些数据驱动和基于先验知识的方法各有优劣，实用中通常将它们进行一定程度的结合：论文[65]通过对自动生成的口音变异进行人工筛选从而在数据驱动的方法中引入了先验知识；论文[60]把训练数据中的每一个基准发音替换为先验规则给出的所有表象发音，对得到的多组参考标注序列分别进行 FA，得分最高的序列是自动修正的基准发音序列，通过这种算法在基于知识的方法中加入了自动处理，以减轻人工标注数据的工作量。

提高 ASR 对声学层变异鲁棒性的主要方法是使用说话人自适应（Speaker Adaptation, SA）、状态级发音模型（State-Level Pronunciation Modeling, SLPM）、最小化错误率的区分性训练（Discriminative Training）等技术修正声学模型。SA 是指使用特定说话人的少量语音来调整非特定说话人（Speaker Independent, SI）的 ASR，使得原本反应大量说话人共性的声学模型更加适应该说话人的口音、语速等个性特征，使模型对该说话人具有更好的识别率[69]。在带口音语音识别中应用

7 国际音标（International Phonetic Alphabet）对应的计算机可读的符号集合称为 SAMPA（Speech Assessment Methods Phonetic Alphabet），将 SAMPA 修改为更适合汉语语音的形式所得到的符号集叫做 SAMPA-C（SAMPA-Chinese）。

SA 可以达到使用少量带口音数据让声学模型匹配带口音语音声学特性的目的[59][70]。论文[59][70]使用基于最大似然线性回归（Maximum Likelihood Linear Regression, MLLR）[135]的 SA 分别解决上海口音 PTH 和美式英语口音中的声学层变异；论文[71]联合使用 MAP[121]自适应和 MLLR 自适应解决 ASR 对韩语口音英文的失配。其它常见的 SA 还包括基于本征音（Eigenvoice）的 SA[73]、基于条件最大似然线性回归（Conditional Maximum Likelihood Linear Regression）的区分性 SA[74]等。使用 SA 技术进行带口音语音识别的主要问题是 SA 会给声学模型带来不可逆的变化，导致自适应后的模型不再匹配标准语音和其它口音[13]。

SLPM 是指通过调整高斯成分改变 HMM 状态的观测密度，从而在状态层级为发音变异进行建模[75]。SLPM 利用数据驱动[76]或人工标注[77]的方法获得口音变异，并训练相应的口音变异模型。利用口音变异模型中带口音的高斯成分扩充基准发音的观测密度，达到调整声学空间划分的目的。SLPM 可以在改善 ASR 对声学层变异鲁棒性的同时保持对标准语音的识别率。论文[78]使用少量带口音语音构造上下文无关的口音变异模型并通过 SLPM 解决吴语口音发音变异。论文[79]则构造了上下文相关的口音变异模型，再使用 SLPM 和决策树融合（Decision Tree Merge）算法扩充 PTH 声学模型[80]。以上两种方法都会引入额外带口音高斯成分，从而增加了模型的复杂度。论文[81]基于高斯成分共享技术，使用模型中已有的高斯成分代替带口音高斯成分，在提高 ASR 性能的同时保持模型原有的复杂度。另外，SLPM 算法还可以在提高 1 套 ASR 对多方言口音、混合口音[13]、混合语言[82]的鲁棒性，使系统具有广泛的适用性。

最小化错误率的区分性训练包括在语句（Utterance）、词语、音节、音素等不同层次最小化 ASR 的识别错误，包括最小化分类错误（Minimum Classification Error, MCE）[83]、最小化音子错误（Minimum Phone Error, MPE）[84]、最小化非均一错误率（Minimum Non-Uniform Error）[85]等。区分性训练通常需要 ASR 生成假设序列以计算错误率，由此产生的高计算复杂度是应用它们进行带口音语音识别的主要困难。另外，区分性训练在最小化由口音变异导致的识别错误的同时也会最小化因识别器等其它原因产生的声学层变异[4]。

另外，还有一些可以同时解决发音层变异和声学层变异的方法[86][87]。最直接的方法是使用带口音语音构造训练声学模型，但这样通常需要收集大量的数据并重新训练模型，成本很高；这种方法在面对多方言口音时还需要额外的口音辨识（Accent Identification）模块[86]，会产生额外的口音分类错误[88]。相反，论文[4]利用不同的距离度量把口音变异分类为发音层变异和声学层变异，并分别使用多发

音字典和 SLPM 算法分层处理。

1.4.2　基于多语音属性的ASR

尽管基于单一语音属性的统计 ASR 取得了巨大的成功[91]，但在自然语音等问题中 ASR 的准确率相比于人的语音识别 （Human Speech Recognition, HSR）仍有较大差距[88][92]。由于 MFCC 等倒谱特征只起到频谱压缩的作用而未触及与语言相关的本质特征[97]，噪声、信道差异、说话人口音等情况都会导致 MFCC 的变化，造成 ASR 识别率的下降[93]。另外 HMM 本身也存在一些缺陷，较难为由语音本身复杂的多线性结构产生的发音变异进行建模。第三，基于 FSN 或语言模型的搜索结构会导致 ASR 无法准确识别包含集外词（Out-of-Vocabulary）和不符合语法规范的语句（Out-of-Grammar） [92]限制了 ASR 的应用范围。

解决这些问题的思路之一是在 ASR 中模拟 HSR 的识别机制[88][93]。通常认为人类是通过自底向上综合所感知到的各种语音属性识别语音的[88][94][95][96][97]。许多时域、频域的语音属性都有助于提高 ASR 的性能：共振峰（Formant）可以区分元音并反映口音产生的发音变异[94]；浊音起始时间（Voice Onset Time）有助于对塞音的识别[98]；清浊音能量比可以有效区分清音和浊音[9]；语调（Intonation）能够提供语言层的信息[95]等。ASR 还经常使用描述发音器官状态的发音特征作为语音属性，以期从语音生成机制描述语音的深层结构，提高 ASR 的性能[99]。论文[100][101]使用 HMM 获得不同长度的发音特征语音段作为冗余特征来修正音素层级的识别结果；论文[35][102][103]使用 ANN 获得发音特征在语音帧上的得分，并证明使用它作为特征比 PLP 的区分度更好；一些最新的研究还使用了性能更强大 DNN 模型甚至额外的肌电（Electromyographic）设备[104][105]以更准确地得到发音特征[105]。研究发现使用发音特征可以提高 ASR 的抗噪能力[106]。于是，不同于基于单一语音属性的统计 ASR 中很少利用先验知识的情况，基于多种语音属性的 ASR 基于先验知识来选择多种有价值的语音属性，并在统计模型中结合更多的先验知识构造知识导向（Knowledge-Rich）的 ASR[93]。

另外，语言学研究还发现“语法是自立（Autonomous）并独立于语义之外的，而统计模型很难洞察句法结构的基本问题”[107]。这表明为了在 ASR 中更好的利用语言学信息改善识别结果，不可避免的要在语言模型中引入更多的语法知识[111]。将语法知识视作语音属性，利用基于多语音属性的 ASR 是实现这一目的的途径之一[93]。

基于多语音属性的 ASR 的这些优点吸引了 GaTech[45][55][88][93][104][109]、MIT[110]、

CMU[100]、华盛顿大学西雅图分校（University of Washington）[99]、加州大学伯克利分校（University of California Berkeley）[88][102]、新竹国立清华大学（National Tsinghua University）[101]、MSR[42][43][104]等学术机构的高度关注，产生了许多重要的研究成果。GaTech 的李锦辉（Chin-Hui Lee）教授提出的自动语音属性转译（Automatic Speech Attribute Transcription, ASAT）方法[93][104]是目前最新、设计最完善、最广受关注的基于多语音属性的 ASR 框架，具有重要的理论和应用价值[88][93]。

1.5　本文的组织结构和创新点

图 1-8 中给出了本文的基本研究思路，灰色表示创新点。

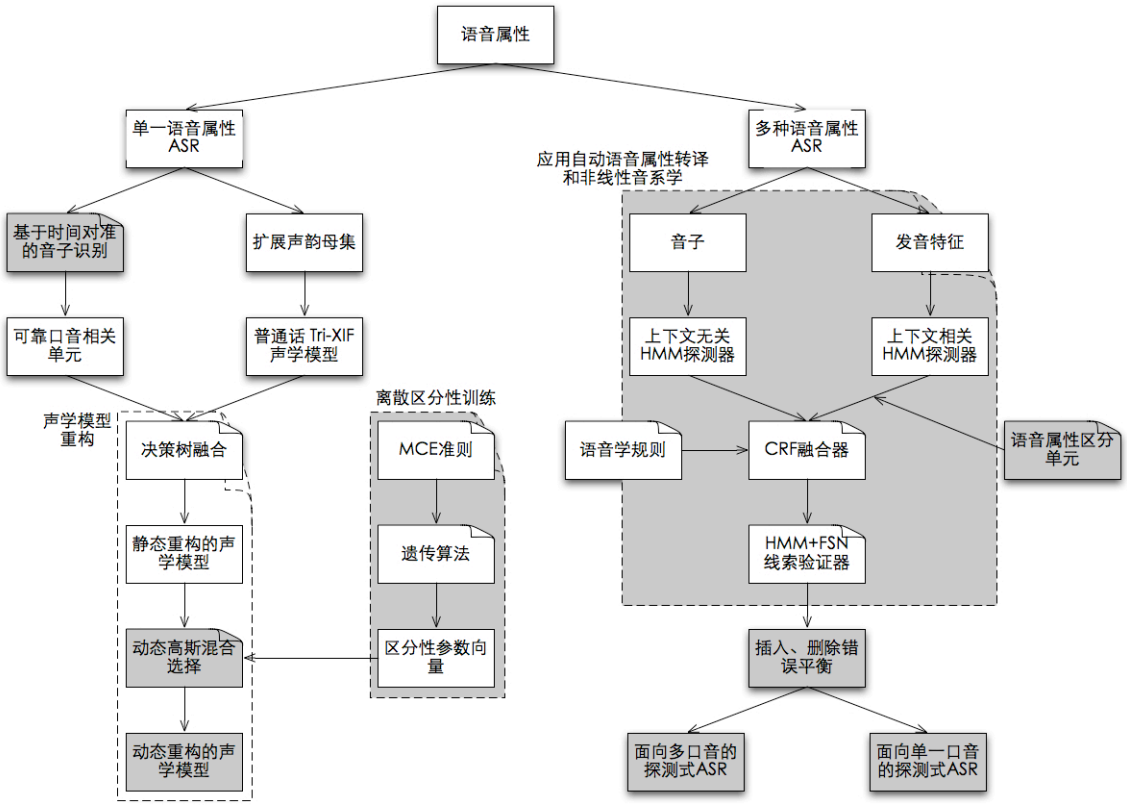


图 1-8　本文的研究思路和创新点

本文共分为 4 章，其各自内容如下，

第 2 章 提高基于单一语音属性的 ASR 对由多种口音产生的发音变异的鲁棒性。联合使用改进的 SLPM 方法和我们提出的动态高斯混合选择算法、离散区分性训练提升单一声学模型对 3 种方言口音的鲁棒性。

第 3 章 利用基于多语音属性的 ASR 方法进行多口音语音识别研究。我们使用发音特征作为语音属性，基于 ASAT 框架设计探测式 ASR，从而在发音特征层次解决口音变异。

第 4 章 对全文内容进行总结，并对相关研究进行展望。

各章的创新点如下，

(1) 第 2 章中针对 ASR 对汉语方言口音的鲁棒性，提出了时间对准的音子识别算法，通过消除帧错位（Frame Mismatch）现象改善使用数据驱动方式获得 ASU 的准确性。

(2) 第 2 章中提出了动态高斯混合选择算法，并使用基于最小化分类错误准则的遗传算法对动态高斯混合选择算法的参数进行离散区分性训练，通过提升使用 SLPM 方法得到的声学模型的精度，提高了系统的识别率和实用性。

(3) 第 3 章提出了基于 ASAT 的汉语带口音语音识别方法，和应用发音特征及非线性音系学描述口音发音变异的方式。为了提高系统的灵活性和实用性还提出了语音属性区分单元及并行探测式 ASR。实验表明这些方法有效地提升了 ASR 对多口音发音变异的鲁棒性。

(4) 第 3 章中提出了仅使用 CRF 的状态特征函数并使用基于 HMM 的 FSN 进行解码的方法以解决使用 CRF 作为融合器时产生的音子欠生成现象。

1.6 本章小结

本章首先介绍了 ASR 和汉语方言口音的概念，然后从 IF 的角度给出了不同汉语口音中发音变异的规律。接着对基于 HMM 的统计 ASR 及其在带口音语音识别中的应用进行了简介，引出了统计 ASR 的缺陷并给出了用基于多语音属性的 ASR 解决这些问题的思路和可行性。

1.7 语音数据库说明

本文所有实验中带口音数据都来自 RASC863 – 四大方言普通话语音语料库（863 Annotated 4 Regional Accent Speech Corpus, RASC863）[116]及其后续的 RASC863-G2 – 863 地方普通话语音语料库，第 2 批，6 地语音库（863 Regional Accent Speech Corpus, Group2, 6 Regional Corpus, RASC863-G2）[117]。RASC863 中包含广州口音（粤语）、上海口音（吴语）、厦门口音（闽语）和重庆口音（川

语）4 种不同口音数据，RASC863-G2 中则包含长沙口音、太原口音（晋语）、洛阳口音、南京口音、南昌口音、温州口音 6 种不同的口音数据。每种方言口音中包含 100 男说话人、100 女说话人的约 60 小时的朗读语音、方言常用词汇等，并带有语音学家评估的每个说话人的口音程度分类。这套数据库是目前汉语方言口音普通话研究中最大、最权威的语音数据库。

本文第 2 章中使用的普通话语音来自“SONY 普通话数据库”，数据库中包含来自 100 男说话人、100 女说话人的共约 60 小时标准普通话朗读语音。这套数据库在汉语 ASR 研究中被广泛应用[109][118][119]。

以上 3 个语音数据库均使用 16k、16b 的采样精度进行录制。

第 2 章　基于单一语音属性 ASR 的声学模型重构

2.1　本章引论

使用 MFCC 等单一语音属性作为特征的 ASR 系统结构简单、数学背景完整，对噪声、跨信道、发音变异等问题已有很多相对成熟的解决方案，故仍然是时下最流行的 ASR 方法[17][22]。同时，隐式马尔科夫模型工具包（Hidden Markov Model Toolkit, HTK）[29]等开发工具的出现也极大的简化了构建高性能 ASR 系统的工作，推动了基于单一语音属性的 ASR 的广泛应用。

本章使用基于单一语音属性的ASR方法进行带多方言口音PTH的语音识别研究[1]。这类研究通常有两种不同策略：(i) 为每种口音分别建立 1 套 ASR，应用时先利用口音辨识模块判断出说话人口音的类型，再用对应口音的系统进行识别[13]；(ii) 不辨明口音的种类，使用 1 套 ASR 对多种口音的语音进行识别[86]。使用第一种策略需要存储多套 ASR，空间复杂度很高；同时口音辨识模块会在系统中引入一定的辨识错误，产生错误传播（Error Propagation）影响系统整体性能[87]。所以本章研究着眼于改善 1 套 ASR 对多方言口音发音变异的鲁棒性。

带口音语音识别研究中常使用自动生成的口音相关单元（Accent-Specific Unit, ASU）代表相应的口音变异。ASU $b \rightarrow s$ 代表基准发音 b 变异为表象发音 s。使用 SLPM 算法构造 ASU 模型并对声学模型进行静态重构（Static Reconstruction）可以在保持对 PTH 识别率的基础上同时解决多种口音产生的发音层变异和声学层变异，还可以解决混合口音变异[13]。本章使用 SLPM 算法对川、粤、吴 3 种口音的口音变异进行建模。但是，静态重构通常会增加声学模型的大小，降低了使用高斯成分的效率，导致模型的精度（Model Resolution）下降。而模型精度的降低会造成剪枝束搜索算法性能的急剧下降。

为解决 SLPM 算法的这些问题，本章从获取发音变异的角度提出了基于时间对准的音子识别（Time-Aligned Phone Recognition, TAPR）算法及可靠口音相关单元（Reliable Accent-Specific Unit, RASU）的生成流程；从模型重构和搜索解码的角度提出了对声学模型进行动态重构（Dynamical Reconstruction）的动态高斯混合选择算法（Dynamic Gaussian Mixture Selection, DGMS）和确定它参数的离散区分

1 如无特殊说明，本章中 ASR 均指基于单一语音属性的 ASR。

性训练（Discrete-Variable Discriminative Training）。本章的贡献小结如下：

(1) 提出了 TAPR 算法以准确、高效的获得口音变异。TAPR 算法可以通过消除使用传统数据驱动方式生成 ASU 时存在的帧错位现象，使自动获得的 ASU 候选及其对应样本更可靠，从而更准确地进行发音变异建模。

(2) 静态模型重构会对观测密度进行扩充而导致声学模型精度的下降。为提升重构后模型的精度，本章提出了 DGMS 算法。在搜索解码中，对每个 HMM 状态，DMGS 算法在该状态的观测密度函数中选择指定数目的高斯成分为当前语音帧定制动态观测密度，算法选取到当前帧最近且最具有代表性的高斯成分。动态观测密度中高斯成分的数目在同一个 HMM 状态中固定不变，相当于一系列离散变量。我们提出使用基于 MCE 准则的遗传算法（Genetic Algorithm，GA）高效地对最优解进行启发式搜索，这一过程可以视作对离散变量进行区分性训练。利用区分性的 DGMS 算法可以同时增强模型的性能和鲁棒性，并在保持声学模型大小不变的情况下增加模型精度，从而降低了剪枝束搜索中剪枝错误造成的识别率下降。

本章剩余部分按照如下方式组织：2.2 节给出本文使用的 HMM 模型的训练算法和设置；2.3 节提出了 TAPR 算法以及 RASU 的生成流程和优越性；2.4 节介绍了利用 SLPM 算法对声学模型进行静态重构的原理及存在的问题；2.5 节提出了使用 DGMS 算法对声学模型进行动态重构；2.6 节提出了利用 GA 和 MCE 准则对 DGMS 算法进行参数的区分性优化；2.7 节给出了实验结果，最后是本章小结。

2.2　基于HMM的声学模型及训练

对 PTH 中的 416 个音节，在训练数据集中某些音节在特定上下文中可能只有很少的实例[30]，如果基于音节或更高层单元建立 HMM 可能出现因数据不足而导致模型参数过拟合的（Over-Fitting）现象，严重影响模型性能[31]。为解决这一问题，通常基于扩展声韵母集（Extended Initial/Final, XIF）构造 HMM[2]。XIF 是在 IF 中为’a’、’e’、’i’、’o’、’u’、’v’ 6 个单元音各补充 1 个零声母（Zero Initial）所得到的集合，零声母分别记为’_a’、’_e’、’_i’、’_o’、’_u’、’_v’。使用 XIF 可以维持 PTH 中音节严格的“声母+韵母”结构。使用 XIF 取代音节建立模型是因为 XIF 的数目比音节少得多，从而更容易获得充足的训练数据。当 1 个 HMM 对应 1 个 XIF 时声学模型称为 Mono-XIF。在连续语音中，发音受到它上下文的带来的协同发音（Co-Articulation）现象的影响，故 ASR 中常用上下文相关模型。当每个 XIF

2 本文还基于音素或发音特征构造 HMM。

都受到它 1 个和后 1 个 XIF 的限制时，称这样的 3 元组为 Tri-XIF 单元。如 Tri-XIF 单元'b'-'a'+'c'表示韵母'a'在上文为'b'、下文为'c'时的发音，'a'称为中心单元。当基于 Tri-XIF 单元建立 HMM 时，称声学模型为 Tri-XIF 模型。

本节将详细介绍如何从训练数据中获得基于 HMM 的声学模型，包括 Mono-XIF 和 Tri-XIF[3]模型的训练流程、基于决策树的状态共享（Decision Tree Based State Tying）策略等。本节的最后还介绍了 MAP 自适应方法。

2.2.1　声学模型的训练流程

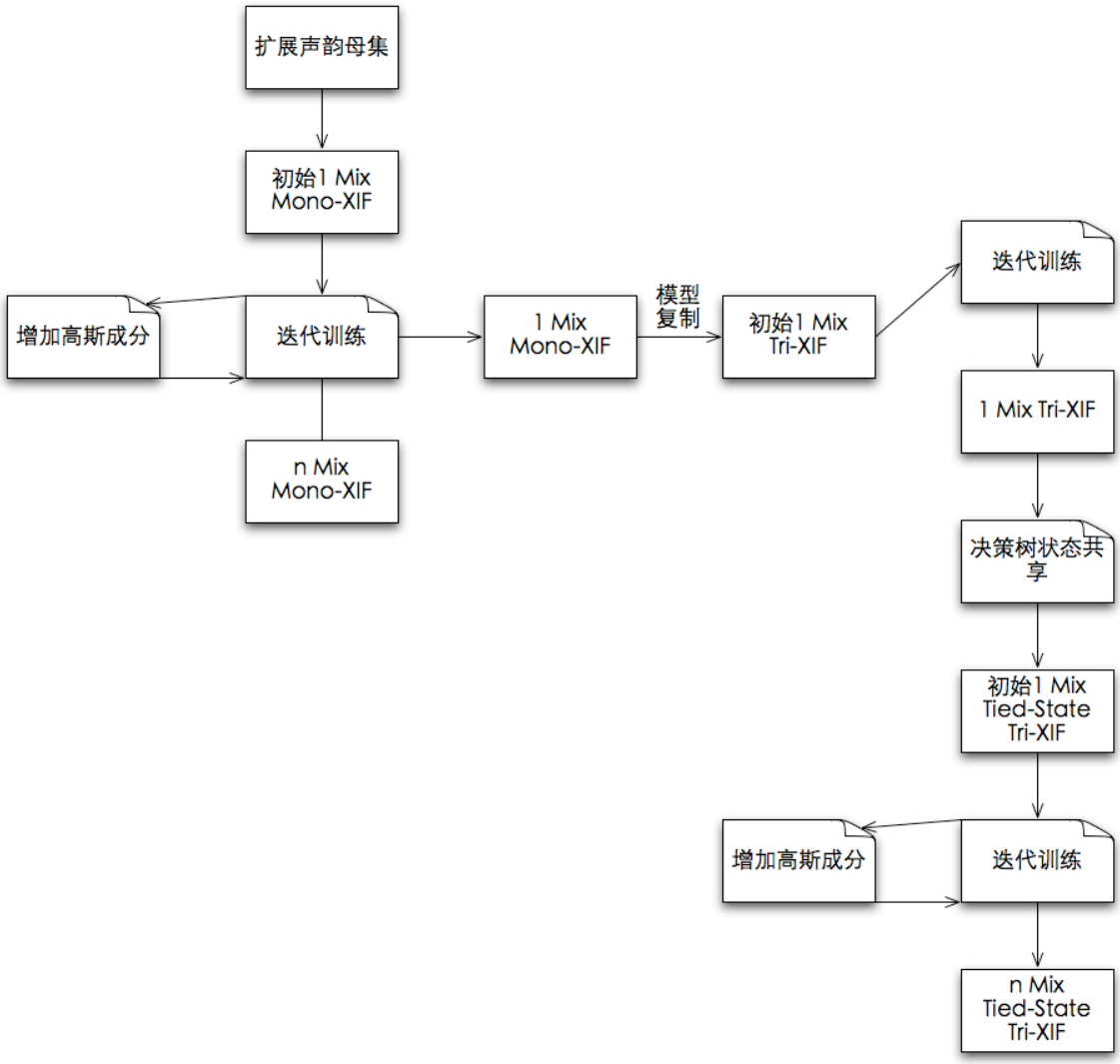


图 2-1　Mono-XIF 和 Tri-XIF 模型训练流程

3 下一章中的 Monophone HMMs、Triphone HMMs、Mono-AF HMMs、Tri-AF HMMs 也使用本节所述的训练方式得到。

对于 Mono-XIF 声学模型，其中每个 HMM 对应于一个 XIF 单元。初始时每个 HMM 状态的观测密度中都只含有一个高斯成分[4]，对 HMM 进行迭代训练直到参数收敛，迭代训练的详细算法见附录 1 公式(1)-(9)；然后等数目增加观测密度中的高斯成分并进行迭代训练；重复以上过程直到高斯成分的数目达到指定阈值，就完成了 Mono-XIF 模型的训练，如图 2-1 所示。增加高斯成分的时，首先在观测密度中选择权值最大的高斯成分并对其参数进行复制，然后将它的权值均分给新增的高斯成分，再对这两个高斯成分的均值和方差进行逆向的微调[29]。

对于 Tri-XIF 模型，首先获得在训练数据中的所有 Tri-XIF 单元，复制中心单元对应的 Mono-XIF 模型作为 Tri-XIF 的初始模型，再使用 Tri-XIF 标注迭代训练就可以得到未进行状态共享的 Tri-XIF 模型。由于一些 Tri-XIF 单元在语料中可能很少出现，未进行状态共享的 Tri-XIF 模型中某些 HMM 可能并没有得到充分训练。通常使用基于决策树的状态共享算法对中心单元相同的单高斯成分 Tri-XIF 模型中对应状态进行聚类，每类只保留一个被所有模型共享的状态，得到的模型称为状态共享（Tied-State）的 Tri-XIF 模型。再用类似 Mono-XIF 模型的高斯成分增加和迭代训练策略，就得到具有指定数目高斯成分的 Tri-XIF 模型[5]，如图 2-1 所示。

2.2.2　基于决策树的状态共享策略

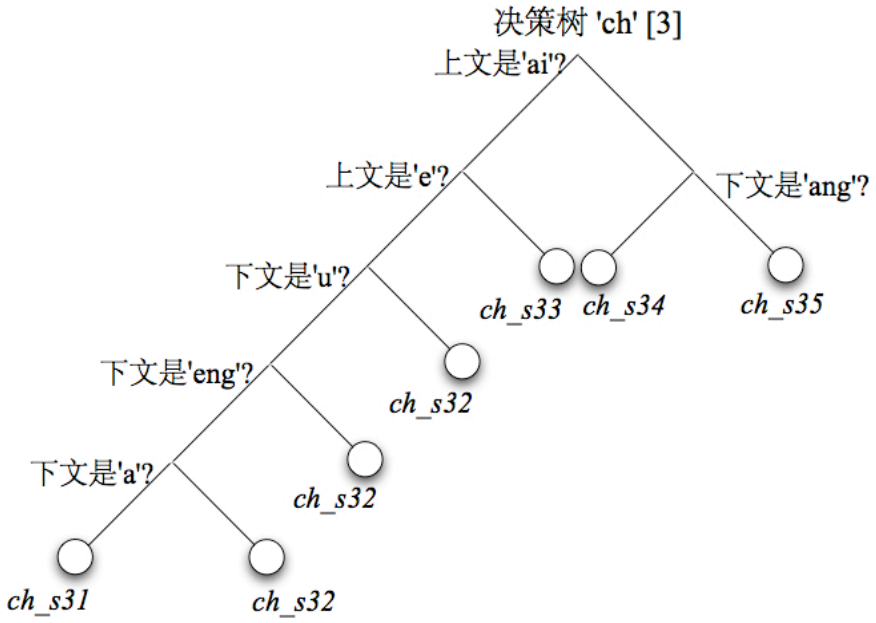


图 2-2　基于决策树进行状态共享的例子

4 对于每个状态的观测密度中包含 n 个高斯成分的 HMM，简记为 *n* Mix HMM，下同。
5 后文默认用 Tri-XIF 模型指代状态共享的 Tri-XIF 模型，类似的还有 Tri-AF 和 Triphone 模型。

决策树（Decision Tree）是一种用于聚类（Clustering）的二叉树（Binary Tree）。决策树的每个节点均为待聚类元素的集合[31]。根节点包含所有的元素，根据一定的标准将它们分为两类，作为根节点的两个子节点。类似这样自顶向下对每个节点中的元素进行二分类，直到满足停止条件。这时二叉树中每个叶节点为一类。

基于决策树的状态共享算法通过引入语音学知识对 HMM 状态进行启发式的无指导（Unsupervised）聚类[29][120]。为每个 XIF 单元的每个状态分别构建一棵决策树，如图 2-2。图中'ch'[2]表示决策树对声母'ch'的所有上下文相关模型的第 2 个状态聚类，记为这些状态为集合 S。基于语音学知识构造问题集（Question Set），问题的定义为状态对应的 Tri-XIF 单元的上文、下文或中心 XIF 是否为某个 XIF 或是否具有某个发音特征。于是根据是否满足答案可以将状态分为 2 类。例如将上下文的 XIF 是否圆唇记为问题 q，它将 S 分为了 S_Y 和 S_N 两个子集合。所有的发音特征及 XIF 与发音特征的对应关系列在附录 2 表 1、表 2、表 3 中。对决策树的某个叶节点，设聚类后该类的共享状态为 s_0，转移概率可以忽略，状态共享后帧和状态的对应关系保持不变，d 为特征维数。训练集中有 R 条语音，第 r 个观测序列为 $O^r = o_1 o_2 \ldots o_{T_r}$，它的似然度为 $P(O^r \mid \Lambda) = P_r$，则近似计算 S 中状态生成所有训练样本的对数似然度 $L(S)$ 为

$$L(S) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{s \in S} \ln(P(o_t^r; \mu_{s_0} \Sigma_{s_0})) P(s \mid o_t^r). \tag{2-1}$$

对于单高斯成分的观测密度，有

$$L(S) = -\frac{1}{2}(\ln((2\pi)^d \mid \Sigma_{s_0} \mid) + d) \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{s \in S} P(s \mid o_t^r). \tag{2-2}$$

$P(s \mid o_t^r)$ 利用前向后向算法在前一次迭代训练中获得。这时公式(2-2)中所有参数均保持不变，从而简化了计算。这也是 2.2.1 节中使用单高斯成分的 Tri-XIF 模型进行状态共享的主要原因。使用问题 q 对 S 进行分类后的得到的对数似然度增益为

$$\Delta L^q(S) = L(S_Y) + L(S_N) - L(S). \tag{2-3}$$

分类后多使用了一个状态，可以更准确的为训练数据建模，故(2-12)的对数似然度总是增加的。使用 $\Delta L^q(S)$ 可以度量多基于问题 q 增多的 1 个状态给识别带来的收益。于是在决策树的每层，从问题集中贪心（Greedy）地选择带来最大对数似然

度增益的问题进行分类，每个叶节点为可以共享相同参数的状态即得到该类的共享状态。

决策树的节点停止分类的准则包括 3 种：(i)节点中所有状态的样本数少于指定阈值；(ii)节点分类所产生的最大对数似然度增益小于指定阈值；(iii)决策树中的一条路径已经使用了所有问题。当决策树完全停止分类后，对不同父节点的任何两个叶节点，若合并它们产生的对数似然度损失小于指定阈值，则合并这两个节点，如图 2-2 中状态 *ch_s32* 和 *ch_s34*。

基于状态共享的 Tri-XIF 的声学模型结构如图 2-3。

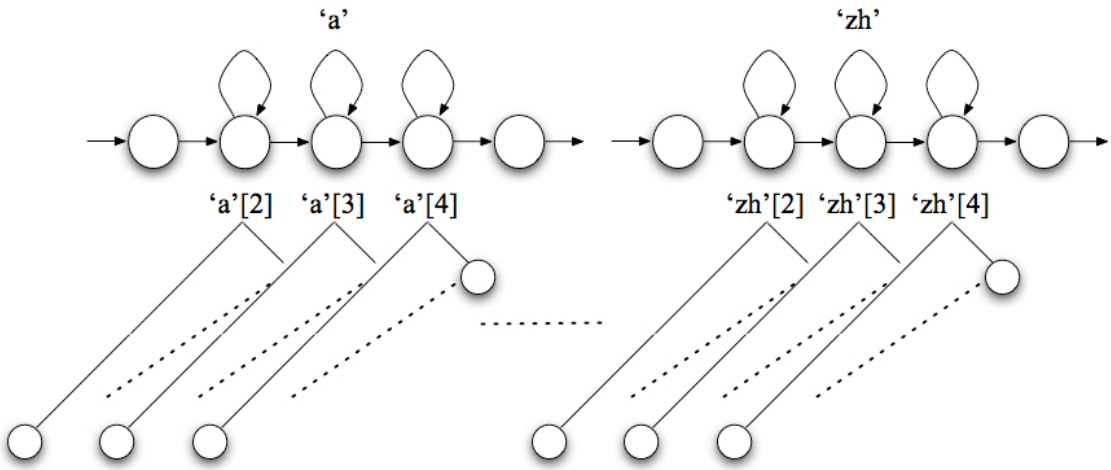


图 2-3　基于决策树进行状态共享的 Tri-XIF 声学模型

基于决策树的状态共享技术结合了语音学知识和数据驱动的优点[120]，并且容易控制聚类的程度和模型的规模，因而得到了广泛的应用[29]。

2.2.3　基于最大后验概率的说话人自适应

本章实验中的基线系统中使用了 MAP 自适应[121]，故本节简介这种技术。

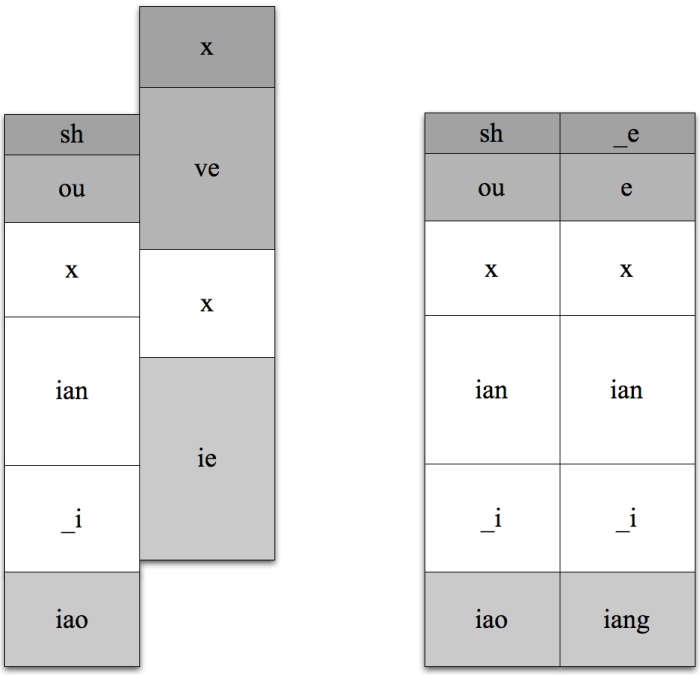
通过最大化似然度 $P(O|\Lambda)$ 求解声学模型的参数，相当于在求解后验概率 $P(\Lambda|O)$ 时假设了先验概率 $P(\Lambda)$ 是常数，即声学模型的所有参数都服从均匀分布。而 MAP 自适应技术基于人为选定的模型参数的先验概率 $P_0(\Lambda)$，再最大化后验概率 $P(O|\Lambda)P_0(\Lambda)$，从而利用先验知识降低模型估计所需的数据，使用少量数据提高了特定说话人与声学模型的匹配性。

在使用 GMM 作为观测密度的 CDHMM 模型中，通常使用指定参数 τ 对 HMM 的参数进行加权来实现最大化后验概率的目的。自适应后 GMM 的均值为，

$$\hat{\mu}'^{(h)}_{jm}=\frac{\tau\hat{\mu}^{(h)}_{jm}+\sum_{r=1}^{R'}\sum_{t=1}^{T_r}L^{(h)r}_{jm}(t)o^r_t}{\tau+\sum_{r=1}^{R'}\sum_{t=1}^{T_r}L^{(h)r}_{jm}(t)}. \tag{2-4}$$

用于自适应的数据中包含 R' 个观测序列，其它参数同附录公式(7)。MAP 自适应只对自适应数据中有对应样本的 HMM 的参数进行更新。

2.3　基于时间对准的音子识别和可靠口音相关单元的生成



(a) 传统口音相关单元　　(b) 可靠口音相关单元

图 2-4　帧错位导致 ASU 实例不准确相比于可靠 RASU 实例的真实例子。(a)和(b)是对应于相同的语句，每个矩形代表一个 XIF 单元，矩形的高与 XIF 的时常正相关。每一部分中左边为基准发音序列，右边为表象发音序列。

使用数据驱动方法从带口音数据集中自动生成 ASU 时，通常将人工标注的参考标注序列作为基准发音序列，将使用自由语法 XIF 识别的识别结果作为表象发音序列。利用 DP 算法对这两个序列进行对齐，每个替换错误都作为一个 ASU 候选。除替换错误外自由语法的 XIF 识别通常还会产生大量的插入和删除错误，但汉语口音变异中却很少有与之对应的 XIF 插音现象。这些插入、删除错误可能导

致 ASU 候选的基准发音和表象发音对应于完全不同的语音片段，产生帧错位现象。如图 2-4 (a)中’sh’→‘x’、’ou’→‘ve’、’iao’→‘ie’的基准发音和表象发音都对应不同的语音片段，所以不是真实的口音变异，导致 ASU 不准确。

我们提出使用 TAPR 算法消除识别结果中的插入错误和删除错误，从而消除帧错位现象。在 TAPR 中，ASR 首先通过 FA 得到每个基准发音的起止时间，然后进行上下文相关的自由语法的 XIF 识别，但加入 2 条额外的识别结果选择规则：(i) 识别结果和基准发音中 XIF 的数目必须的相等；(ii) 识别结果中每个 XIF 必须与基准发音中对应位置 XIF 的起止时间相同。也就是说，TAPR 使用上下文相关的 XIF 分类来获得表象发音序列。

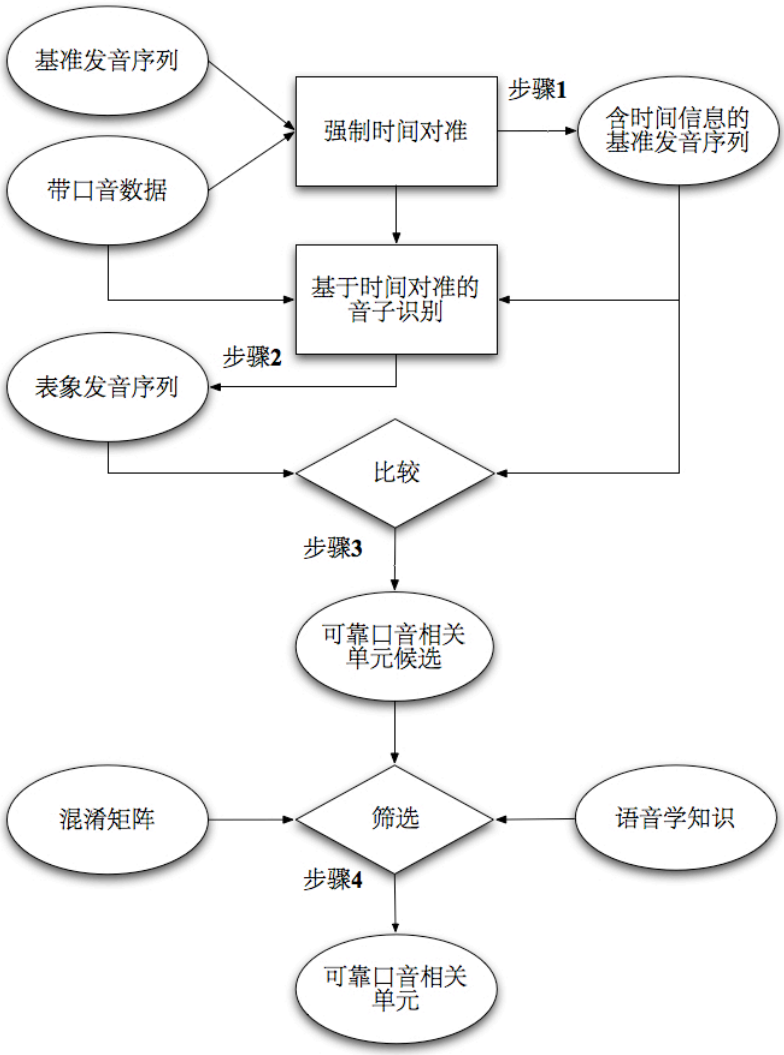


图 2-5 生成 RASU 的流程

图 2-5 中给出了利用 TAPR 生成 RASU 的流程，对各个步骤的详细解释如下：

(1) 获得基准发音的时间信息：利用预先训练的 PTH 声学模型对参考序列进行 FA，得到每个基准发音 XIF 的时间。

(2) 获得表象发音序列：使用 TAPR 算法生成相应的表象发音序列，基准发音的时间信息来自步骤(1)。

(3) 生成 RASU 候选：比对基准发音和表象发音序列对应位置的 XIF，将所有替换错误作为 RASU 的候选。

(4) 筛选 RASU：如 1.3.4 节所述，步骤(3)得到的候选同时包含有发音层变异、声学层变异，以及数据和算法原因产生的错误。所以我们结合语音学知识以及 XIF 的混淆矩阵进行筛选，选出的单元为 RASU；RASU 的每个出现都是它的一个实例。

(5) 生成 RASU 标注序列：在参考序列中，对每个 RASU 实例，用 RASU 替换基准发音，得到 RASU 标注序列。

步骤(4)中的筛选策略如下：

(1) 移除 PTH 固有混淆：例如’i2’→‘i1’是在 PTH 中很容易产生的混淆，故相应的识别错误可能不是由口音变异造成的。将’i2’→‘i1’视为 PTH 固有混淆直接从 RASU 候选中移除。

(2) 对可疑的 RASU 候选，将其表象发音用它在 PTH 中的固有混淆替代后再进行判断：例如’an’→‘ai’不是典型的川语口音变异，但川语使用者通常倾向于把’an’发为/ae/。而/ae/不属于 PTH-IF，使用 PTH 中与/ae/最接近的韵母’ai’替换/ae/，这样得到的’an’→‘ai’是 RASU。

(3) 移除因数据库和识别器导致的错误：例如川、粤、吴方言中都没有声母’zh’，于是’z’→‘zh’不符合语音学知识；同时在数据中’z’→‘zh’比’zh’→‘z’的实例少得多，由口音变异的单向性，’z’→‘zh’不是 RASU。

图 2-4 (b)中给出了 TAPR 通过消除帧错位来获得准确的口音变异的例子：韵母’ian’和’_i’都被正确识别而不再是图 2-4 (a)中的删除错误；而生成的 RASU 候选为’sh’→‘_e’、’ou’→‘e’和’iao’→‘iang’而不是图 2-4 (a)中的’sh’→‘x’、’ou’→‘ve’和’iao’→‘ie’。特别的，在本章实验中将 TAPR 与基于编辑距离（Edit Distance）和 DP 算法[63][76]生成 ASU 候选的方法进行对比：(i) TAPR 可以得到更多的口音变异，如川口音中的’ch’→‘s’。(ii) TPAR 生成的 RASU 实例更可靠，如使用 TAPR 得到的川口音 RASU ‘uen’→‘en’有 38 个不符合 1.2.2 节中的语音学规则的实例，而使用传统方法会得到 61 个这样的实例。ASU 实例作为 ASU 模型的训练样本，其可靠性非常重要。(iii) 由于 TAPR 通常比自有语法 XIF 产生的识别错误少，故 RASU

候选更容易进行筛选。综上，使用 RASU 及其实例可以更准确的为发音变异建立模型，这对静态声学模型重构方法非常关键。

对川语口音数据得到的一些代表性的 RASU 包括：

‘an’→‘ai’	‘ch’→‘c’	‘eng’→‘en’	‘ing’→‘in’
‘n_l’→‘j’	‘sh’→‘s’	‘uen’→‘en’	‘zh’→‘z’

对粤语口音数据得到的一些代表性的 RASU 包括：

‘ch’→‘c’	‘er’→‘e’	‘ou’→‘ao’	‘sh’→‘s’
‘zh’→‘c’	‘zh’→‘j’	‘zh’→‘s’	‘zh’→‘z’

对吴语口音数据得到的一些代表性的 RASU 包括：

‘an’→‘ang’	‘ch’→‘c’	‘d’→‘n’	‘en’→‘ing’
‘r’→‘l’	‘r’→‘n’	‘sh’→‘s’	‘zh’→‘z’

可以发现上述 RASU 可以很好的对应 1.2.2 节语音学家给出的口音变异规则。

2.4　声学模型的静态重构

口音变异导致的识别错误相当于带口音样本出现在预料外的声学特征子空间，即样本落入了划分给它的表象发音而非基准发音的子空间中。SLPM 方法通过训练 RASU 的模型来获得的位于表象发音子空间的高斯成分，然后将这些高斯成分通过决策树融合算法加入到 PTH Tri-XIF 模型的共享状态中，从而通过扩充基准发音模型的观测密度调整了声学特征空间的划分。SLPM 算法将存在识别错误的表象发音子空间重新划分给基准发音以覆盖相应的口音变异，使得声学模型发生了永久、稳定的改变，称为静态声学模型。

由于重构后声学模型同时具有 PTH 高斯成分和带口音高斯成分，故能在有效解决口音变异的同时保持对 PTH 语音的识别率。算法还可以用于多方言口音的问题：为每种口音分别提取 RASU 并构造模型，用所有 RASU 模型重构 PTH 声学模型，以提高 1 套 ASR 对多口音发音变异的鲁棒性。需要注意的是，不同口音中的相同 RASU，如川、粤、吴口音的’zh’→‘z’，不能简单合并，而应该独立处理。这

是因为它们的语速等特性不同，具有不同的声学特性。最后，SLPM 方法相当于高斯成分层次重新划分声学空间，HMM、ASR 等层次解决多口音问题更灵活，甚至能够解决由混合口音产生的发音变异。

2.4.1 静态声学模型重构的原理

对 $O=o_1o_2\ldots o_T$，将每帧的基准发音构成的序列记为 $B=b_1b_2\ldots b_T$，则公式(1-4)中 $P(O|B)$ 为使用 PTH 语音训练的声学模型。记表象发音帧序列为 $S=s_1s_2\ldots s_T$，假设有口音语音及标注，则 $P(O|S)$ 为只使用口音数据得到的声学模型。由于真实语音中可能存在 PTH 或带口音语音，$P(O|B)$ 和 $P(O|S)$ 在真实语音中都不是最优的。为求得最优声学模型，需要同时考虑 PTH 和带口音语音，由加法公式[25]可将(1-4)改写为

$$\hat{B}=\underset{B}{\operatorname{argmax}}\left[P(B)\sum_{S}P(O|B,S)P(S|B)\right]. \tag{2-5}$$

公式(2-5)中 $P(O|B,S)$ 为考虑了表象发音后混合口音数据和 PTH 数据得到的声学模型，$P(B|S)$ 称为发音模型。

对 HMM，由(1-12)式有

$$P(O|B,S)=\prod_{i=1}^{T}P(o_i|b_i,s_i). \tag{2-6}$$

o_i 可看作由基准发音模型 b_i 和表象发音模型 s_i 共同生成的，本章中 $P(o_i|b_i,s_i)$ 使用声学模型的静态重构方法得到[122]。

设 $B=b^{(1)}b^{(2)}\ldots b^{(n)}$ 为基准发音模型序列，表象发音模型序列为 $S=s^{(1)}s^{(2)}\ldots s^{(m)}$，则发音模型 $P(B|S)$ 可以写为

$$P(S|B)=P(s_1|B)P(s_2|B,s_1)\ldots P(s_m|B,s_1s_2\ldots s_{m-1}). \tag{2-7}$$

2.4.2 决策树融合算法

为得到能够准确描述口音变异的高斯成分，为每个 RASU 训练上下文相关模型。将 RASU 加入 XIF 集合作为扩展发音基元，使用带口音数据及相应的 RASU 标注序列训练 Tri-RASU 模型。决策树的状态共享的问题集来自对 2.2.2 节中问题集按照如下规则进行扩展：对每个问题，若包含某个 RASU 的表象发音的作为上

下文条件，则将该 RASU 的相同上下文条件加入该问题。例如 RASU ‘zh’→‘z’，包含’z’作为上文、下文或中心单元的 Tri-XIF 单元在问题“是否唇齿音”中，所以将包含’zh’→‘z’的所有 Tri-RASU 单元加入该问题中。

对于 RASU 单元，其基准发音和表象发音都为上下文无关 XIF，而 Tri-RASU 模型和 Tri-XIF 模型都具有图 2-3 中的结构，以下将 PTH Tri-XIF 模型中的决策树称为标准决策树（Conventional Decision Tree），将 Tri-RASU 模型中的决策树称为辅助决策树（Auxiliary Decision Tree）。将 Tri-RASU 模型中 $b \rightarrow s$ 的每个状态的辅助决策树分别融合到 PTH Tri-XIF 中基准发音 b 的对应状态的标准决策树中，就实现了利用 RASU 模型扩充基准发音模型的目的。由于 1 个基准发音可能产生多种发音变异，即多个 RASU 可能具有相同的基准发音。于是决策树融合时的可能出现将不同 RASU 的辅助决策树融合到 1 个标准决策树的情况，如图 2-6。

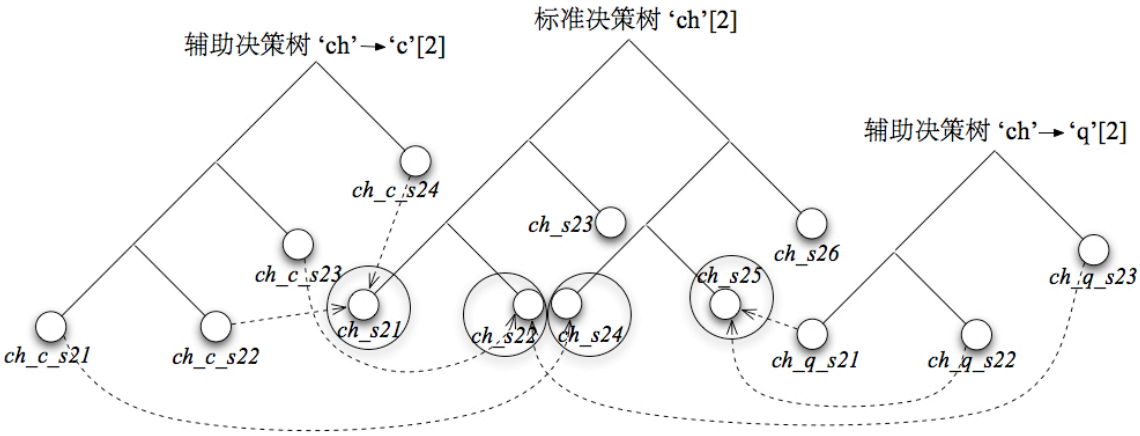


图 2-6 辅助决策树’ch’→‘c’[2]和’ch’→‘q’[2]融合到标准决策树’ch’[2]中

同时，决策树的每个叶节点可以代表 1 个共享状态，将辅助决策树融合进标准决策树可以看做将该辅助决策树的所有叶节点分别融合到标准决策树的某些叶节点中，如图 2-6 所示。为保持融合后标准决策树叶节点的变化尽可能小，以减少对 PTH 性能的影响，将辅助决策树叶节点加入与它距离最近的标准决策树中叶节点中。对于辅助决策树叶节点的观测密度 $b_A(o)$ 为单高斯 $N(o|\mu^A,\Sigma^A)$ 的情况，叶节点间的距离定义为 μ_A 到标准决策树叶节点的观测密度 $b_C(o)=\sum_{m=1}^{M} w_m N(o|\mu'_m,\Sigma'_m)$ 的马氏距离（Mahalanobis Distance）[31]，

$$d(b_A,b_C)=\sum_{m=1}^{M} w_m(\mu^A-\mu_m)^T\Sigma_m'^{-1}(\mu^A-\mu_m). \tag{2-8}$$

对$b_A(o)$包含多高斯成分的情况，使用高斯成分合并公式将其缩减为单高斯成分的情况进行计算。将权值分别为w_1^A和w_2^A的高斯成分$N(o|\mu_1^A,\Sigma_1^A)$和$N(o|\mu_2^A,\Sigma_2^A)$进行合并的公式定义为

$$\begin{aligned} w_{1+2}^A &= w_1^A + w_2^A \\ \mu_{1+2}^A &= \frac{w_1^A\mu_1^A + w_2^A\mu_2^A}{w_{1+2}^A} \\ \Sigma_{1+2}^A &= \frac{w_1^A(\Sigma_1^T(\mu_1^A-\mu_{1+2}^A)(\mu_1^A-\mu_{1+2}^A)^T)}{w_{1+2}^A} + \frac{w_2^A(\Sigma_2^T(\mu_1^A-\mu_{1+2}^A)(\mu_1^A-\mu_{1+2}^A)^T)}{w_{1+2}^A} \end{aligned} \quad . \tag{2-9}$$

得到权值为w_{1+2}^A的高斯成分$N(o|\mu_{1+2}^A,\Sigma_{1+2}^A)$。

将一个辅助决策树叶节点融合到某个标准决策树叶节点是指把辅助决策树叶节点的观测概率中所有高斯成分都扩充到标准决策树叶节点的观测概率中，如图 2-6 所示。为保证扩充后的观测概率中的所有高斯成分的权值之和仍为 1，需要对融合后高斯成分的权值进行统一调整。设共有V个辅助决策树叶节点$P(x|s_i),1\le i\le V$融合到标准决策树叶节点$P(x|b)$中，则融合后的模型$P(x|b,s)$为

$$P(x|b,s) = \lambda P(x|b) + \frac{(1-\lambda)}{\sum_{j=1}^{V} P(v_j|b)} \sum_{i=1}^{V} P(x|s_i)P(s_i|b). \tag{2-10}$$

其中λ为b被正确识别的概率，$P(s_i|b)$为b被误识为s_i的概率，它们都可以从 PTH Tri-XIF 模型生成 RASU 时产生的混淆矩阵中获得。

这样就通过决策树融合算法利用 RASU 模型中带口音数据训练的高斯成分完成了对 PTH Tri-XIF 声学模型的静态重构。

2.5　基于动态高斯混合选择算法的声学模型动态重构

静态声学模型重构能够让 1 套 ASR 在保持对 PTH 识别率的前提下同时提高对多种口音的鲁棒性，而代价仅为有针对的增加一定数量的高斯成分。但 SLPM 算法仍存在以下问题：

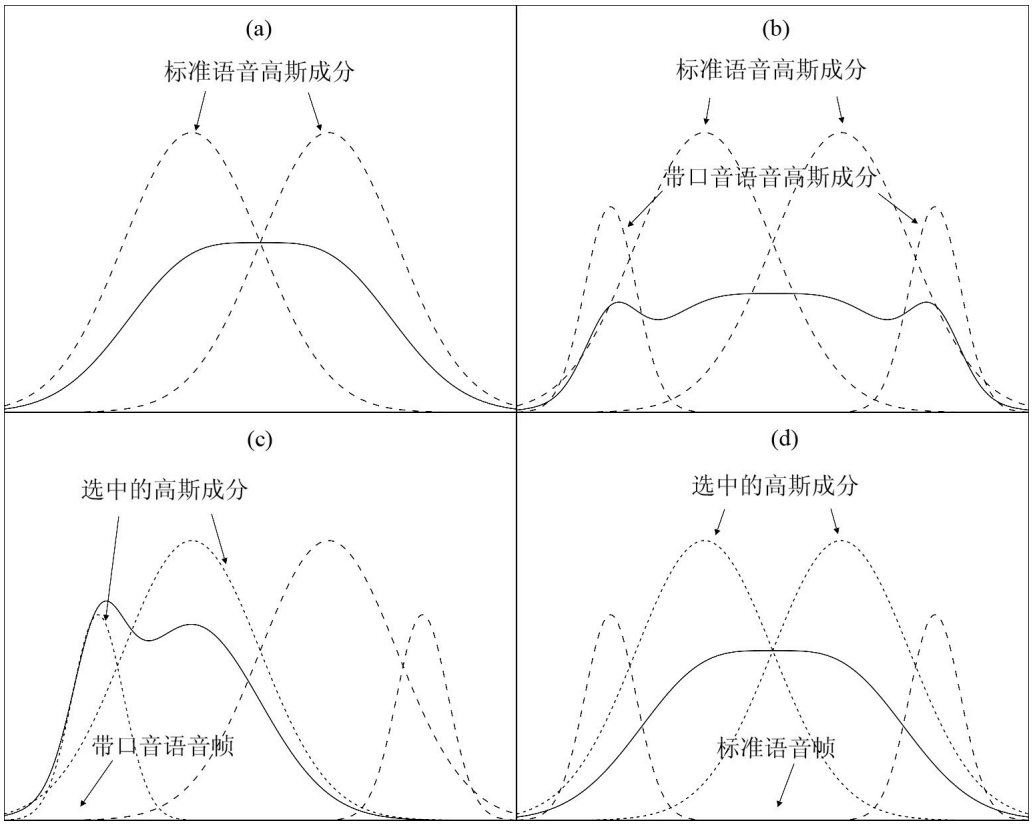


图 2-7 静态声学模型重构和动态高斯混合选择算法产生的输出分布

(1) 决策树融合算法对应该从辅助决策树叶节点中引入多少高斯成分、这些高斯成分应该引入到哪个标准决策树叶节点中缺乏可靠的数学依据，很难保证增加最少的高斯成分且对声学空间的划分最优。尤其在多方言口音问题中，这样经常会导致在声学模型中新增了很多冗余的高斯成分。

(2) 静态重构后不同状态的观测密度通常具有不同数目的高斯成分，某些状态中的高斯成分可能增多的更加明显，比如在本章的实验中，共增加了 6,620 个高斯成分到 546 个共享状态中，重构后观测密度变化最明显的状态新增了 120 个高斯成分。同时，新增的高斯成分可能被放置到声学空间的不同位置，某些高斯成分的距离甚至可能比较远，使得重构后的状态生成它对应的基准发音的某些样本时似然度有所降低，而该样本在竞争者模型上的似然度可能保持不变，这样就降低了重构后声学模型的精度，如图 2-7 (a)和 2-7 (b)所示。

模型中冗余的高斯成分不但会增大模型的大小，降低 ASR 的识别速度，还会造成模型精度的下降。模型精度下降可能限制模型性能的提高，并导致进行剪枝束搜索时正确结果更容易被提前剪掉，从而严重降低剪枝搜索的准确率。由于剪枝搜索是几乎所有实用 ASR 系统的必备配置，故提高静态重构后声学模型的精度

可以提高 ASR 的实用性。

本节提出 DGMS 算法。在搜索解码中，对每个输入观测，DGMS 在每个状态的观测密度中按照 k 最近邻（k Nearest Neighbor）原则分别选择恰当的高斯成分构成动态观测密度，即利用与当前观测距离最近的 k 个高斯成分定制一个新的观测密度以计算观测概率。由于每个高斯成分具有不同的方差，故使用马氏距离度量高斯成分 $N(o\,|\,\mu,\Sigma)$ 到观测 o 的距离，

$$d(N(o\,|\,\mu,\Sigma),o)=(o-\mu)^T\Sigma^{-1}(o-\mu). \tag{2-11}$$

于是，DGMS 算法进行更形式化的表述如下。

记 $b_r(o)=\sum_{m=1}^{M}w_m N(o\,|\,\mu_m,\Sigma_m)$ 是静态重构后声学模型中共享状态 r 的观测密度函数，并简记 $N_m=N(o\,|\,\mu_m,\Sigma_m)$。设 $N'_1,N'_2,\ldots,N'_k$ 为 $b_r(o)$ 的 M 个高斯成分中与 o 最近的 k 个高斯成分，则状态 r 的动态观测密度 $\tilde{b}_r(o)$ 为

$$\begin{cases}\tilde{b}_r(o)=\sum_{m=1}^{k}w''_m N(o\,|\,\mu'_m,\Sigma'_m)\\ w''_m=\dfrac{w'_m}{\sum_{m=1}^{k}w'_m}\end{cases} \tag{2-12}$$

对位于静态重构后输出分布边界的带口音观测，k 最近邻原则选择 k 个与该观测距离最近的高斯成分，这些高斯成分位于分布边界附近并最能够代表观测对应的口音变异。这样得到的动态观测密度对生成该观测的输出分布具有更强的描述能力，可以提高模型精度以减少剪枝束搜索造成的性能损失，如 2-7 (c)所示。对于位于分布中央区域的 PTH 观测，它的动态观测密度可能与模型重构前的观测密度相同，如 2-7 (d)所示。于是动态重构的声学模型可以保持对 PTH 的识别率。

为了使动态重构的声学模型有足够灵活以适应口音变异多样化的声学特性，每个静态重构的 HMM 共享状态具有独立的参数 k。这些预先选定的参数 k 为利用 DGMS 优化静态重构的声学模型提供了足够的自由度。以下将所有参数 k 视为一个参数向量。优化 DGMS 参数向量的主要难度包括：(i) 如何评估不同的参数向量；(ii) 如何得到最优的参数向量。

本节最后从机器学习的视角重新审视 DGMS 算法。统计模型的决策边界是统计机器学习中的核心问题之一[24]，它决定特征空间中每个子空间如何进行模式分类，训练统计模型的目的就是获得决策边界。ASR 中应用的传统机器学习模型的

特征空间划分是固定的；而 DGMS 算法对不同的观测可能使用不同的高斯成分，从而导致不同的特征空间划分。很容易看出，DGMS 算法最多可以产生指数数量级的不同决策边界，从而极大的增加了模型的灵活性。

2.6　使用最小化分类错误准则和遗传算法进行离散区分性训练

为了能够准确的获得 DGMS 的最优参数向量，本节选用 MCE 准则以直接优化语句层级的错误率[83]。由于 DGMS 算法中每个共享状态的高斯成分选择数 k 是 MCE 损失函数的离散变量，不存在导数，因此很难用对均值、方差等连续变量进行区分性训练的传统优化方法优化 DGMS 参数向量。事实上 DGMS 参数向量的优化问题属于整数规划问题，本节使用 GA 对它进行启发式的高效求解。这一过程相当于对观测密度的高斯成分数目这组离散变量进行区分性的参数优化，称为离散区分性训练。

2.6.1　基于最小化错误率的区分性准则评参数向量

常用的统计模型优化准则包括生成性准则和区分性准则两类，使用它们得到的模型分别称为生成性模型和区分性模型。训练 HMM 模型相当于使用公式(1-13)最大化模型生成观测序列的似然度 $P(O|B)$，B 是基准发音序列。最大化似然度（Maximum Likelihood）准则是生成性的，得到了生成性的 HMM。与生成性准则相比，区分性准则通常是优化后验概率、错误率等与模型性能更直接相关的因素，故得到的区分性模型通常具有更好的性能但也需要耗费更多的训练时间，常见的区分性模型包括 CRF 等。对于 HMM，通常要在 ML 训练后叠加 1 次额外的区分性训练才能得到区分性的 HMM。ASR 中常用于 HMM 的区分性准则包括最大化互信息（Maximum Mutual Information Estimation, MMIE）[132]、最小化音子错误（Minimum Phone Error, MPE）[84]、MCE[83]等。其中 MMIE 和 MPE 准则优化的目标是(2-13)中后验概率的某种变形，

$$P(B|O)=\frac{P(O|B)P(B)}{\sum_{S}P(O|S)P(S)}, \tag{2-13}$$

S 为假设序列。从公式(2-13)的分母可以看出 MMIE 和 MPE 都需要利用点阵等数据结构获得所有的假设序列，运算量的复杂度很高。相比之下，MCE 准则最小化识别结果中包含的分类错误，只需 1 最优假设，运算复杂度较小。同时 MCE 准则直接优化训练数据集上的分类错误，比 MMIE 的优化目标更直接、得

到的模型性能更好[133]，与 MPE 得到的模型性能接近[85]。故本文选择 MCE 准则作为区分性 DGMS 参数向量的优化标准。

假设O为训练集中的观测序列，定义区分度函数$g_j(O,\Lambda,c)$为 DGMS 使用参数向量c在声学模型Λ上得到的假设序列S_j的对数似然度。定义分类错误度函数$d_i(O,\Lambda,c)$以评估参考标注序列S_i和识别结果间的对数似然度差异，

$$d_i(O,\Lambda,c) = -g_i(O,\Lambda,c) + \max_j g_j(O,\Lambda,c). \tag{2-14}$$

在此基础上定义 MCE 的损失函数为，

$$l(d_i(O,\Lambda,c)) = \frac{1}{1+e^{\alpha \cdot d_i(O,\Lambda,c)}}. \tag{2-15}$$

我们使用 MCE 损失函数作为对 DGMS 进行参数优化的目标函数。MCE 取得最小值时 DGMS 通过最小化训练集中的因口音等原因造成的分类错误实现了最小化经验风险（Empirical Risk）[31]的目的，增加了模型对口音变异的覆盖能力。

在不使用 DGMS 的情况下，分别用 FA 和自由语法 XIF 识别得到参考标注序列和 n 最优识别结果中每个状态的起止时间。实验发现若参数向量c改变，在Λ上得到的识别结果会发生显著的变化。但所有c分别通过解码搜索求其识别结果的复杂度过高，故利用包含有限假设序列的 n 最优识别结果近似包含所有假设序列的点阵，选择在c和Λ上具有最大对数似然度的假设序列近似这时的识别结果。因此，根据公式(2-14)和(2-15)，正确识别时$d_i(O,\Lambda,c) \le 0$（存在小于 0 的情况是因为 n 最优识别结果中可能未包含正确的识别结果，所以每个假设序列的对数似然度都比正确结果小[6]）；否则有$d_i(O,\Lambda,c) > 0$。使公式(2-15)最小化的c就是 DGMS 的最优参数向量。

值得说明的是，由于 DGMS 的参数向量优化属于离散变量优化问题，优化的目标函数不需要保证具有导数。本节使用公式(2-15)而非帧层次的分类错误数作为目标函数并不是为了利用公式(2-15)的连续性保证可导[83]，而是出于以下原因：

(1) MCE 损失函数中包含对度量错误分类的函数，错误程度不同的帧对

6 事实上只要当参考标注序列比固定 n 最优识别结果中的每个假设序列的对数似然度都高时都会有$d_i(O,\Lambda,c) < 0$，也就是说这时也可能存在某个不属于固定 n 最优识别结果中的不正确的假设序列比参考标注序列对数似然度更高，这是使用固定 n 最优识别结果近似点阵的代价。

目标函数产生不同的影响，且这些影响会在语句层级折中，从而进行语句层次而非语音帧层次识别率的最优化。

(2) 通过赋予α不同的值可以给训练集中分类错误度不同的语句以不同的权值[123]。当α取 0.01 等较小的值时，许多训练语句都被映射到了图 2-8 中的线性区域，增加了正确和错误假设序列间的区分度，使得参数有较好的推广性；当α增长为 2.0 等足够大的值时，MCE 损失函数相当于对所有识别错误的语句进行计数，如图 2-8。

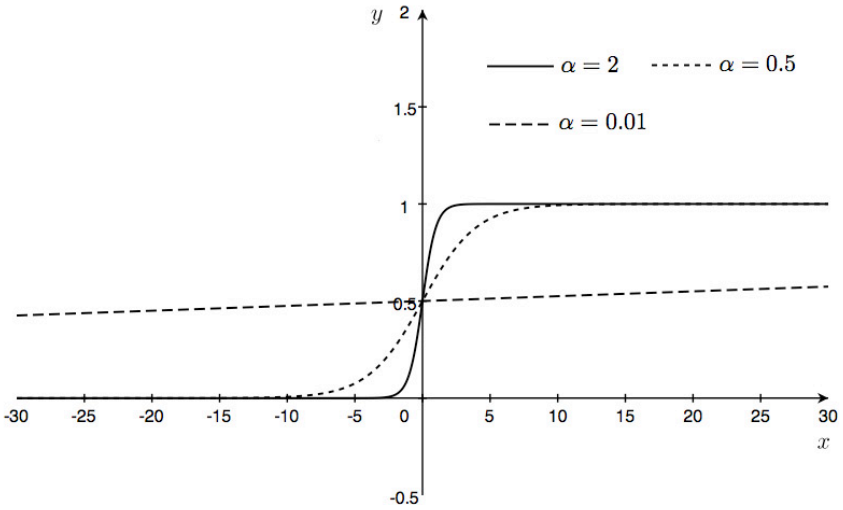


图 2-8　α与不同误分类度的关系

2.6.2　利用遗传算法加速参数向量的优化

设扩充的声学模型中包含G个被静态重构的共享状态，第g个状态的观测密度中包含M_g个高斯成分。DGMS 算法中参数向量的所有可能候选为$\prod_{g=1}^{G} M_g$，它随着R的增长进行指数级增长。由于扩充的声学模型中通常包含数百个被静态重构的状态，每个状态可能包含上百个高斯成分，通过遍历参数向量的所有取值得到最优参数在计算上是不可行的（Computationally Infeasible），属于整数规划问题。

GA 是一种用搜索求解问题的解搜索（Search-for-Solution）算法。它通过模拟自然界物种“物竞天择适者生存（Survival of the Fittest）”的进化[124]的过程，可以在仅遍历小部分搜索空间的情况下获得问题的全局最优解[125]。将每个参数向量视作进化过程中一个独立的染色体（Chromosome），则寻找最优参数向量的优化问题相当于寻找整个进化过程中最适合生存的染色体，于是可以利用 GA 高效的

优化 DGMS 的参数向量。令染色体c由G个正整数顺序组成，其中第g正整数对应第g个静态重构状态的动态高斯成分选择数目k_g，$1 \le k_g \le M_g$。

定义 GA 中用于评估每个染色体适合生存程度的适应性函数（Fittest-Function）为 MCE 损失函数，记为$f(c)$，有

$$f(c) = \frac{1}{R}\sum_{r=1}^{R} l(d_i(O^r, \Lambda, c)). \tag{2-16}$$

GA 按照下面步骤求解 DGMS 的最优参数向量：

(1) 随机生成N个染色体作为初始种群（Population）。

(2) 对种群中每个染色体c计算其适应性函数值$f(c)$。

(3) 从当前种群中随机选择C个染色体，C是偶数。染色体c的选择概率可以如下计算得到，

$$P(c) = \frac{1}{f(c)} \Big/ \sum_{c'} \frac{1}{f(c')}. \tag{2-17}$$

c'用来指代当前种群中每个染色体。使用轮盘赌采样（Roulette-Wheel Sampling）算法[125]对染色体进行随机选择：首先生成随机数$rand1$，$0 \le rand1 \le 1$，将当前种群中每个染色体的选择概率$P(c)$顺序相加，若当前染色体为c^*时得到的和超过随机数$rand1$的大小，选择c^*。

(4) 对随机选择的C个染色体进行交配（Reproduce）。首先这些染色体顺序划分为$C/2$个染色体对，对每对染色体使用单点交换（One-Point Crossover）方法[125]生成 2 个子代染色体。单点交换方法首先在染色体中随机选择某个位置，再将两个父代染色体该位置以下的片段（整数子序列）互换，生成子代染色体。这 1 步共生成C个子代染色体。

(5) 将新得到的C个子代染色体与步骤(3)中未被选中的所有染色体合并作为新种群。对新种群中每个染色体中的每个位置g，生成随机数$rand2$，$0 \le rand2 \le 1$。若$rand2$大于给定的突变概率（Mutation Probability）ξ，则对位置g的整数k_g以等概率随机增加或减小 1。注意需要保证增减后k_g仍然满足$1 \le k_g \le M_g$。

(6) 使用步骤(5)中新得到的种群作为新的当前种群。

(7) 如果当前种群中不存在染色体c使得$f(c) \approx 0$，且算法的迭代次数未达到预先指定的算法最大迭代次数限制I，重复步骤(2)-(6)。

算法终止后，所有曾出现的染色体中，具有最小适应函数值的染色体c_0即为区分性 DGMS 的最优参数向量。

2.7　实验结果

2.7.1　实验数据集

本章实验使用的川、粤、吴 3 种带口音数据均选自 RASC863 数据库。为严正本章提出的算法对多方言口音问题的有效性，分别建立了开发集（DevC、DevY、DevW）和测试集（TestC、TestY、TtestW），来自根据数据库记录信息选择的具有最重口音的说话人的语音。PTH 的训练集和测试集都来自 SONY PTH 数据库。所有语音均为 16 kHz 采样率 16b 采样精度，上述各数据集的详细情况见表 2.1。

表 2.1 本章所有数据集详情

名称	时长（小时）	音节总数	说话人总数	语句总数	口音类型
DevC	6.5	51,907	20	3,205	川语口音
TestC	4.3	33,847	20	2,000	
DevY	6.1	51,341	20	3,091	粤语口音
TestY	3.5	31,191	20	2,000	
DevW	6.6	52,584	20	3,471	吴语口音
TestW	3.8	29,888	20	2,000	
TrainP	51.5	340,556	100	25,920	PTH
TestP	3.9	23,158	10	2,000	

2.7.2　识别系统

本章实验共对 5 套 ASR 系统进行测试，每套系统的详情如下。

系统 1：基线系统。声学特征采由 CMN 处理过的 12 维倒谱系数和 1 维归一化的能量构成的 13 维 MFCC，13 维一阶差分 MFCC 和 13 维二阶差分 MFCC。采用附录 2 表 1 和表 2 给出的包含 22 个声母、36 个韵母、6 个零声母的 XIF[14]作为构造 HMM。每个 HMM 的拓扑结构都如图 1-5 所示，包含 5 个共享状态，按照从左到右的顺序排列且没有状态可能被跳过。声学模型使用 TrainP 数据集按 2.2.1 节的训练流程得到。模型使用基于决策树的状态共享，包含 3,000 个共享状态，每个状态有 12 个高斯成分。

系统 2：基于 ASU 进行静态重构的系统。ASU 使用基于动态规划和耗散对齐的灵活对齐工具（Flexible Alignment Tool）[63]进行对齐并筛选得到。从 DevC、DevY、DevW 中分别生成了 160、187、166 个 ASU，并基于这些 ASU 分别构造了 480、561、498 棵辅助决策树，共分别包含 531、582、569 个共享状态。将这些辅助决策树融合到系统 1 的标准决策树中，得到静态重构后的声学模型中包含 42,728 个高斯成分。构造系统 2 是为了体现 ASU 的有效性。

系统 3：基于 RASU 进行静态重构的系统。从 DevC、DevY、DevW 中分别生成了 165、191、166 个 RASU。基于这些 RASU 为川语口音构建了包含 517 个共享状态的 495 棵辅助决策树，为粤语口音构建了包含 605 个共享状态的 573 棵辅助决策树，为吴语口音构建了含有 533 个共享状态的 498 棵辅助决策树。将这些上下文相关的 RASU 模型利用决策树融合算法融合到系统 1 中，得到的静态重构的声学模型包含 42,620 个高斯成分，平均每个状态有 14.2 个高斯成分。

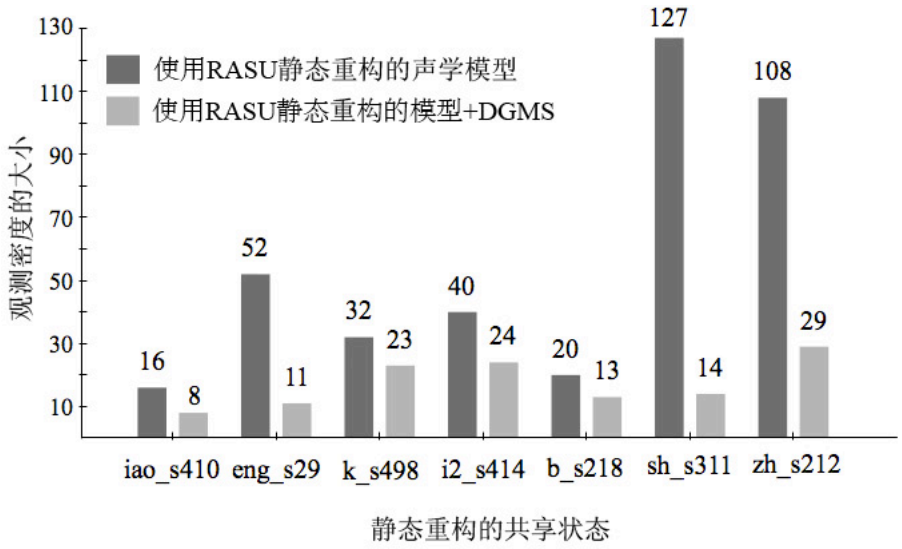


图 2-9 使用/不使用区分性 DGMS 得到的一些代表性状态的观测密度的大小

系统 4：系统 3 + 区分性 DGMS。系统 3 中声学模型共有 546 个被静态重构的共享状态，因此 DGMS 的参数向量中有 546 个整数参数。为优化这些离散变量，选定 MCE 准则的参数α为 0.01；GA 的种群规模 N、交配规模 C 和突变概率 ξ 分别设定为 400、200 和 0.3。使用系统 3 生成的 5 最优结果在 MCE 准则中近似点阵来获得不同参数向量下的 1 最优结果。混合 DevC、DevY、DevW 的所有数据进行离散区分性训练来优化 DGMS 的参数向量。GA 的最大迭代次数为 1,200 次，迭代终止后目标函数从 0.86 收敛到 0.62，得到区分性 DGMS

的参数向量用于本章所有相关实验。图 2-9 给出了一些代表性的静态重构状态中观测密度的大小及其高斯成分选择数，可以看出动态观测密度的大小显著小于静态重构的观测密度的大小。

系统 5：MAP 自适应的系统。对系统 1 中的声学模型混合 DevC、DevY、DevW 数据进行 MAP 自适应以使它符合川语口音、粤语口音、吴语口音的声学特性。这 1 系统作为对照组对本章提出的算法和口音自适应方法进行比较。

2.7.3　实验结果和讨论

下文所有实验都使用自由语法的音节识别，5 个系统都使用插入错误和删除错误数目相近情况下的音节准确率进行评估，详细的实验结果在表 2.2 中给出。与基线系统（系统 1）相比，基于 ASU 进行静态重构的系统（系统 2）在所有口音的测试集上的准确率都得到了显著提高。这是因为 SLPM 算法通过静态重构声学模型的观测密度使更多的高斯成分被放置到了决策边界附近，从而覆盖了造成识别错误的口音变异 [13][79]。与系统 2 相比，系统 3 的准确率在测试集 TestC、TestY 和 TestW 上分别相对提升了 4.36%、1.63%和 0.96%。这些结果说明了 RASU 对口音变异具有比 ASU 更好的覆盖能力。

增加了区分性的 DGMS 单元后，系统 4 在系统 3 的基础上，TestC、TestY 和 TestW 上的准确率相对提升了 3.23%、4.77%和 3.42%。离散区分性训练通过在训练集上最小化分类错误以增强模型对多种口音发音变异的覆盖能力。DGMS 算法限制了动态观测密度中最少高斯成分数为 6，以使 PTH 观测的动态观测密度接近原始 PTH 模型的观测密度。联合使用 RASU 和区分性 DGMS（系统 4）使系统在川语口音、粤语口音、吴语口音上的准确率比基于 ASU 进行声学模型静态重构的方法（系统 2）分别相对提高了 7.73%、5.26%和 4.41%。系统 4 在所有口音上都取得了最高的识别率。

从表 2.2 还可以看出，在 3 种口音上系统 4 比系统 5 的准确率分别相对提高了 7.57%、3.69%和 4.19%，这表明：(i) 显式且准确地为每种口音变异分别建模优于对声学模型进行自适应以使其匹配多种口音变异的方法；(ii) 由于无法求解 MAP 准则中模型参数的先验分布 [121]，区分性的 MCE 准则优于生成性的 MAP 准则。同时，系统 5 对 PTH 的准确率下降到了 69.39%，而系统 4 对 PTH 的准确率几乎保持不变。在本章的实验设置中，MAP 自适应通过调整声学模型的参数使它更加匹配多口音的声学特点，所以不再适合 PTH 的声学特点。

表 2.2 使用 RASU 和 DGMS 可以获得比使用 ASU 或 MAP 自适应更高的准确率

编号	系统名称	音节准确率%			
		TestP	TestC	TestY	TestW
1	基线系统	77.52	43.43	41.69	44.54
2	基于 ASU 进行静态重构的系统	77.79 (+0.27)	50.68 (+7.25)	50.97 (+9.28)	53.08 (+8.54)
3	基于 RASU 进行静态重构的系统	77.89 (+0.37)	52.89 (+9.46)	51.70 (+10.01)	53.59 (+9.05)
4	基于 RASU 进行静态重构的系统 + DGMS	77.82 (+0.30)	54.60 (+11.17)	53.65 (+11.96)	55.42 (+10.88)
5	混合 DevC、DevY、DevW 进行 MAP 自适应的系统	69.39 (-8.13)	50.76 (+7.33)	51.74 (+10.05)	54.15 (+8.61)

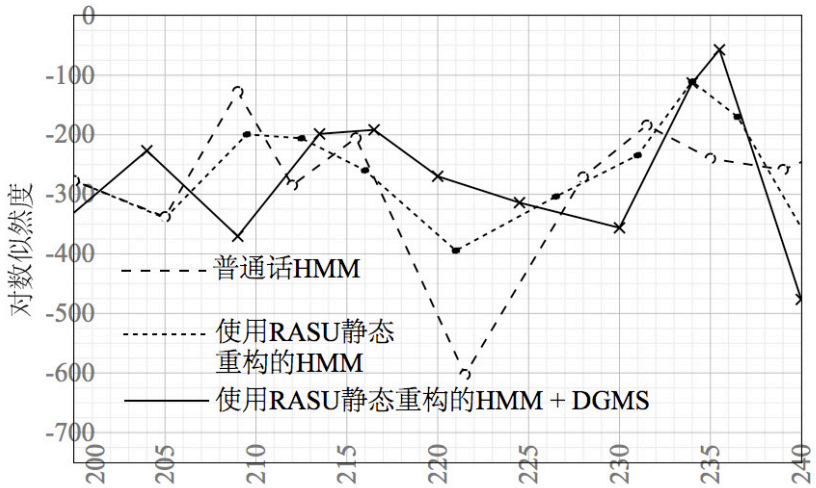


图 2-10 一个使用区分性 DGMS 修正粤语口音造成的局部模型失配的例子

图 2-10 中给出了 1 个通过使用区分性 DGMS 算法修正粤语口音造成的局部模型失配的例子。当使用基线系统和系统 3 时，粤语口音变异'zh'→'z'造成音节[zhi]的声母'zh'被错识别为'z'。被错误识别的声母'z'的 3 个状态的观测概率位于图 2-10 中第 215 到第 230 帧。基线系统在第 222 帧附近的声学似然度出现了急剧下降。对于使用 RASU 进行静态声学模型重构的系统，RASU 'zh'→'z'的高斯成分在一定程度上增加了 222 帧附近的对数似然度，但还不足以修正模型的局部失配。而使用区分性 DGMS 后，动态观测密度进一步提升了模型对口音变异的鲁棒性，如图

中实线所示。故系统 4 成功的解决了模型的局部失配，并生成了正确的识别结果。

本节最后给出使用 DGMS 提升静态重构后声学模型精度的量化分析。使用公式(2-18)对模型精度进行度量，

$$D_i(\Lambda)=\sum_O d_i(O,\Lambda,c)\big/ g_i(O,\Lambda,c). \tag{2-18}$$

$g_i(O,\Lambda,c)$ 和 $d_i(O,\Lambda,c)$ 类似公式(2-14)，定义为基准状态 i 上的对数似然度和基准状态与表象状态间的错误分类度。当状态 i 的模型精度增加时 $D_i(\Lambda)$ 也随之增加。令 $D_i(\Lambda_0)$ 指代静态重构后 HMM 的精度，Λ' 是另一套声学模型，它相对于声学模型 Λ 产生的模型精度的提升可以使用公式(2-19)计算，即统计错误分类程度的变化代表模型区分度的变化，

$$\mathrm{sgn}(D_i(\Lambda'))\cdot|D_i(\Lambda')\big/D_i(\Lambda_{0.})|. \tag{2-19}$$

图 2-11 给出了 利用公式(2-19)得到的一些代表状态的相对精度提升。可以发现 DGMS 显著提高了模型的相对精度，即增大了基准状态和表象状态的间距（Separation）。模型间距越大，精度越高。

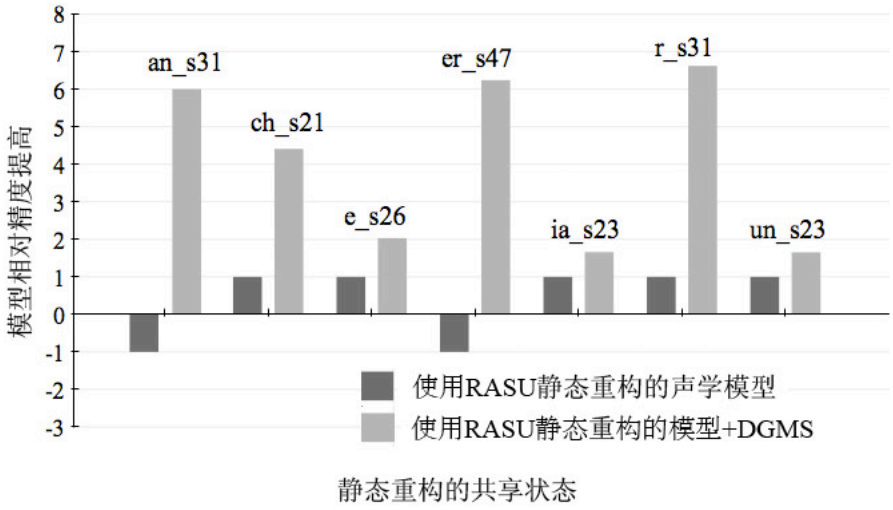


图 2-11 使用粤语口音数据评估使用/不使用区分性 DGMS 时一些代表性共享状态的模型精度变化

模型精度提升带来的效果同样可以反映在剪枝束搜索中，剪枝参数的定义如 1.3.4 节。当声学模型的精度更高时，DGMS 所保留的搜索路径比不使用 DGMS 时

的准确度更高，从而降低了由声学模型静态重构导致的剪枝束搜索性能下降，如图 2-12。使用 DGMS 时系统准确率总是高于不使用 DGMS 的情况。进一步可以发现当剪枝参数 t 越小时使用 DGMS 获得的准确率提升就越大。当 t 减小时图 2-12 中两条折线的间距变窄，表明区分性 DGMS 同时增强了系统的识别率和鲁棒性。

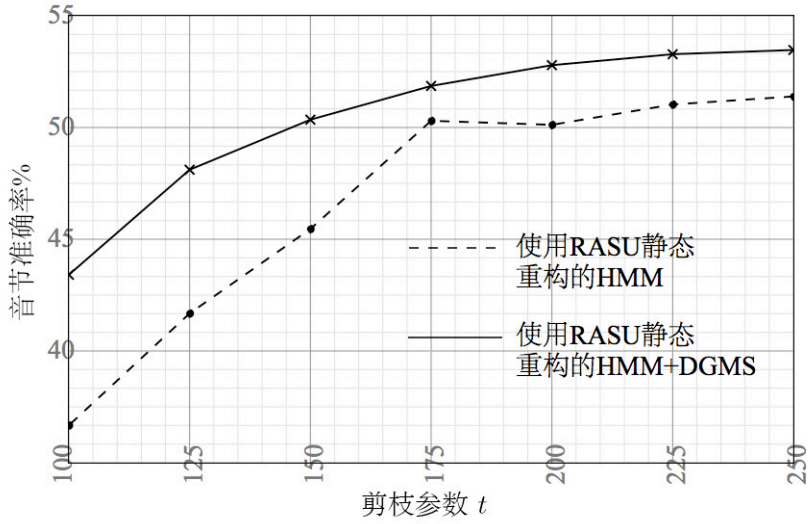


图 2-12 基于 RASU 静态重构的声学模型中使用/不使用区分性 DGMS 得到当剪枝参数为 100.0 到 250.0 时系统在粤语口音测试数据上的音节准确率

2.8 本章小结

在 ASR 的使用中通常不断遇到训练模型时未预料的口音变异。本章使用基于单一语音属性的 ASR，提出了使用 TAPR 高效生成 RASU 的流程，并证实了 RASU 可以准确、显式的获得发音层变异和声学层变异。本章还针对多种口音发音变异提出了 DGMS 算法，它对预先训练的 GMM 模型针对每个输入观测的声学特性进行动态重构以解决口音变异。本章提出了利用 MCE 准则和 GA 算法对 DGMS 的参数向量进行离散变量的区分性训练，实验证明可以提高声学模型的精度，从而在增强模型对口音变异鲁棒性的同时减小剪枝束搜索的性能损失。实验结果表明，联合使用本章提出的所有算法得到的系统与使用 ASU 对声学模型进行静态重构的系统相比，使 ASR 对川语、粤语、吴语口音的音节准确率分别相对提升了 7.73%、5.26%、4.41%；与标准 MAP 自适应方法得到的系统相比，在 3 种口音上分别获得了 7.57%、3.69%和 4.19%的相对音节准确率提升，并保持系统对 PTH 的识别率。

第 3 章　基于多语音属性的探测式 ASR

3.1　本章引论

不同于传统基于单一语音属性的 ASR，探测式 ASR 通过整合探测得到的声学和语音学知识实现语音识别，从而在 ASR 中导入更多的先验知识。ASAT 是一种代表性的探测式 ASR 框架，它通过模拟 HSR 将一系列语音属性探测器的探测结果自底向上地逐层融合以产生识别结果，期望能够利用更丰富的信息实现 ASR 识别率和鲁棒性的突破[104]。ASAT 主要包含 3 个组成部分[88]：(i) 从连续语音中寻找特定声学线索的探测器组；(ii) 将探测到的语音属性逐步整合为音子等更高层信息的事件融合器；(iii) 对可能的识别结果进行验证和选择的线索验证器。

以往的研究表明，ASAT 能够以特征或规则的形式利用语音学信息来获得鲁棒性更好的模型[93]。从第一、二章的研究和分析可以发现，现有基于单一语音属性的 ASR 很难完全解决多样的发音变异，导致它在带口音语音上的性能比标准语音仍有显著差距；同时相关方法往往或者太过复杂，或者只适用于有限的场合。这表明基于 HMM 的 ASR 的线性结构并不完全适合为语音建模[99]。本章我们应用 ASAT 方法进行带口音语音识别研究，期望 ASAT 整合多样语音属性的能力、利用丰富的先验知识的能力和灵活的系统结构能够更匹配汉语多方言口音中复杂的发音变异。同时，语音学研究表明发音层、声学层等不同层次的口音变异与发音特征的变异和替换有重要关联，我们在 ASAT 的层次化结构中引入发音特征以匹配口音变异的这些特点，方便在系统中进一步导入更多先验知识，从而利用语音学研究指导系统的设计。

本章提出使用基于发音特征的 ASAT 技术进行官话、粤语、吴语等 3 种口音的语音识别。使用上下文相关的 HMM 作为发音特征探测器可以更好的结合语音学知识，并获得带口音语音中因协同发音等产生的上下文相关的发音变异。由于语音学中的发音变异规则通常为发音特征值的组合，将发音特征值存在与否作为 2 元特征可以更容易地组合特征以匹配先验知识。使用 CRF 作为将输入发音特征序列整合为音子序列的帧层次融合器，从而可以更恰当、更灵活的利用语音学规则。通过在 CRF 中仅使用状态特征函数（State Feature Function），并输出带概率的音子点阵，可以解决 CRF 常遇到的音子欠生成问题，在提高系统识别率的同时为音节等更高层的处理提供了方便。我们使用基于 HMM 和 FSN 的线索验证器来实现

插入和删除错误的平衡。系统设计符合汉语非线性音系学的研究成果，比基于单一语音属性的 ASR 更符合语音的本质[11][21][99]。实验结果表明面向单一口音的探测式 ASR 在达到与传统 Triphone HMMs 相近的识别率的情况下具有更快的识别速度。另外，本章还提出了语音属性区分单元，它可以在无需修改系统中任何模型的情况下改善系统对口音变异的覆盖能力，提升了系统的灵活性。本章的最后还实现了并行探测式 ASR，它对多口音产生的复杂发音变异具有好的鲁棒性。

本章剩余部分按照如下方式组织：3.2 节简介了 ASAT 及其研究现状；3.3 节介绍了本章使用的发音特征、非线性音系学等语音学知识；3.4 节从图概率模型等角度介绍了 CRF 的主要理论；3.5 节按照各个组成部分介绍了面向汉语带口音语音识别的探测式 ASR；3.6 节给出了实验结果，并提出了语音属性区分单元和面向多口音问题的并行探测式 ASR；3.7 节总结了基于多语音属性的探测式 ASR 相对于传统单一语音属性 ASR 的主要优势。最后是本章小结。

3.2　自动语音属性转译技术及其研究现状

ASAT 方法规范了基于多语音属性 ASR 的设计思想和系统结构，如图 3-1 (a)所示。主要包括语音属性探测器（Speech Attribute Detectors）、事件融合器（Event Merger）和线索验证器（Evidence Verifier）3 个主要部分[1]。对于输入语音，首先利用一系列探测器获得语音属性在语音中的存在状况，每种语音属性的一个被检测到的实例都称为一个事件（Event），事件的长度不受限制，这样就将一段语音转换成了多个非时齐的时序事件序列。融合器接受探测器输出的事件序列，模拟 HSR 的机制自底向上的对低层语音事件进行融合，逐步得到音子、音节、词语等高层事件，如图 3-1 (b)。以点阵等形式保留高层事件出现的所有可能性，并将其视为得到最终识别结果的线索。接着将融合器输出的高层事件信息输入验证器，依据线索对所有可能的假设序列进行加工，排除可能性较低的假设。验证器输出的剩余假设序列仍保留在点阵中，可以将它们反馈回融合器和验证器做进一步加工选择，直到确定最终的识别结果。每个模块输出的事件、线索都可以使用概率来代表它们存在的可能性，如图 3-1 (b)中，$\vec{s}$ 表示输入的语音序列，$\vec{F}$ 为从 $\vec{s}$ 可以获得的所有参数，$P(\vec{A}\,|\,\vec{F})$、$P(\vec{P}\,|\,\vec{F})$、$P(\vec{S}\,|\,\vec{F})$ 和 $P(\vec{W}\,|\,\vec{F})$ 分别表示语音属性探测器以及语音属性到音子、音子到音节和音节到词语的各层融合器输出的条件概率。这样继承使用已有的各种统计模型实现各个模块。另外，每个模块中都可以结合

1 后文简称为探测器、融合器和验证器。

任何需要的知识对识别结果进行优化，如图 3-1 (b)。

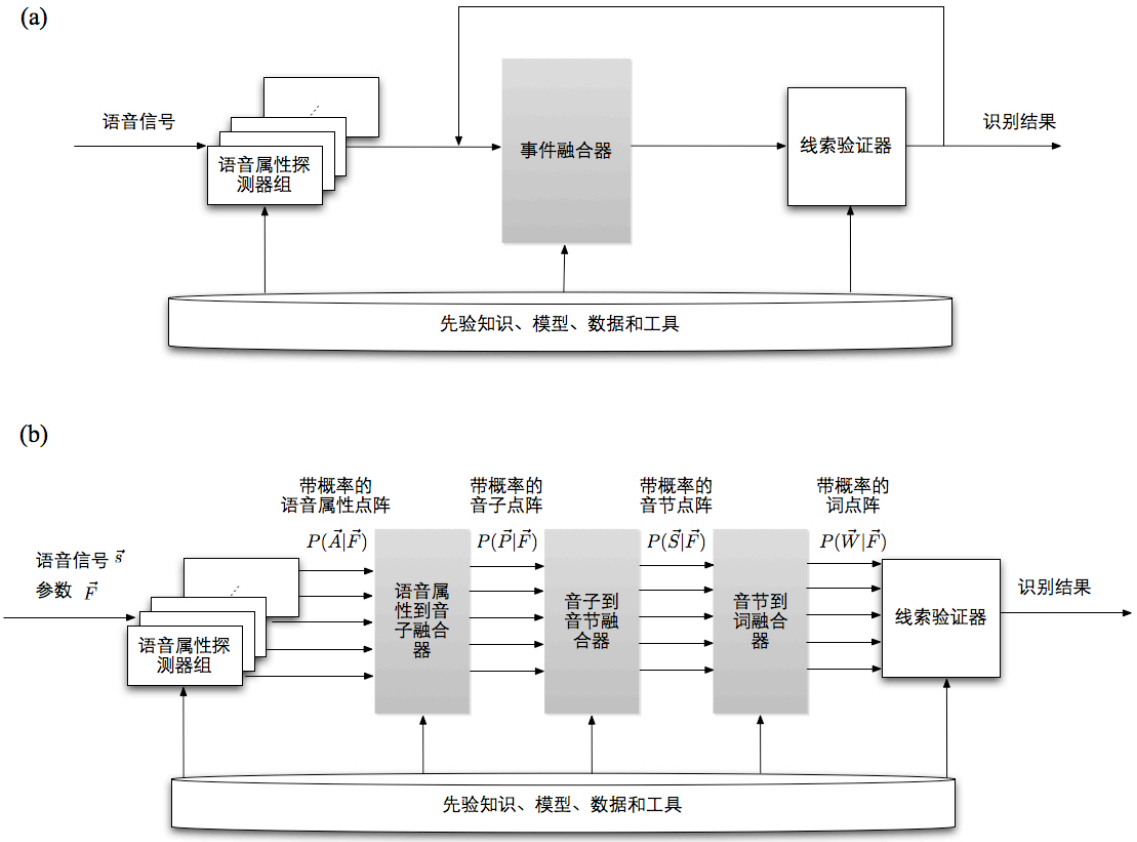


图 3-1 ASAT 的结构图及其工作流程

不同于统计 ASR 使用的自顶向下结构，ASAT 使用的自底向上的融合过程可以更方便地在各层利用不同特点的语音属性和知识提高系统性能。同时，模块化和分而治之（Divide and Conquer）的设计方便对系统的每个部分进行相对的独立修改和调整，体现了即插即用（Plug ‘n’ Play）的思想[88][93]。另外，通过逆向分析 ASAT 的识别过程还可以增进我们对语音的认识[88][93]，从而可以将 ASAT 作为语音学家的实验平台。

目前 ASAT 框架已有许多成功的实现[42][43][101][102][103][104][110]，它们可以按照融合器的输入语音属性类型分为帧级别和段级别两种系统[110]。帧级别的系统是指融合器的输入以语音帧为单位，如使用 ANN 探测帧级别发音特征属性，并将 ANN 输出的后验概率分别输入 HMM[102]、CRF[103]、ANN[104]进行融合的方法；使用 DNN 替代 ANN 作为发音特征探测器和融合器的方法[104]；使用 HMM 在段级别探测发音特征属性并将探测到的结果展开为帧级别特征输入 CRF 融合器的方法等[101]。这

种系统的优点是通过将非时齐事件序列转化为语音帧层次的时齐序列从而降低了融合器的复杂度。相对的，段级别系统的融合器接受不定长度的输入事件，容易利用更符合语音段本质且更具有区分性的段层次特征。例如将事件在基于图的特征空间中利用维特比搜索算法获得识别结果的方法[110]；利用基于 HMM、ANN、DNN 等探测段长、音素、音节、词语等语音属性，并使用 SCRF 作为段级别的融合器的方法等[42][43]。

可以注意到的是，ASAT 对多种语音属性进行探测的思想接近于传统的声学-语音学方法，而其自底向上的系统结构以及在识别器中大量使用知识的思想则与专家系统的思路相似。但不同于声学-语音学方法 ASAT 的探测器大量利用了统计语音识别的技术，同时对语音属性的利用也从通过恰当选取声学、语音特征进行识别发展为综合各种冗余的语音属性改善识别结果。另外，ASAT 的融合器也通常以统计机器学习方法为主，先验知识用来指导系统设计或修正识别结果，而非专家系统纯粹基于规则进行决策的方式。可以看出，历史总是循环往复的出现，但每次循环都不是简单的重复，而是在继承中不断向前发展[112]。另外值得一提的是近年兴起的基于稀疏表示定理（Sparse Representation Theory）[113]的 ASR，它在标本库中通过求解最小化 1 阶范数（1-Norm）来进行语音识别[114][115]，可以看做模板匹配方法的一种发展。

3.3　汉语语音学中的口音和口音变异

由于本章的方法涉及了许多语音学知识，为了方便讨论，本节将从现代语音学主流的生成音系学的角度重新分析汉语语音及口音变异。

3.3.1　语音的产生

从生理学（Physiology）角度，语音是人类调节呼吸器官（Respiratory Organs）产生的气流通过发音器官（Vocal Organs）时所产生的声音，气流通过的部位、方式不同就导致形成的语音不同[8]。这里所指的发音器官其实是呼吸器官和消化器官（Digestive Organs）的一部分[8]。具体来说，由肺部（Lungs）呼出的气流经过气管（Trachea）到达声带（Vocal Cords），位于声带中间的声门（Glottis）关闭使气流积蓄，增大的气压冲开声带并使之不断颤动而形成声波（Wave）；声波继续传播，依次经过由喉腔（Laryngeal Cavity）、咽（Pharynx）、口腔（Oral Cavity）和唇腔（Labial Cavity）、鼻腔（Nasal Cavity）等所组成的声道（Vocal Tract），并不断的共振（Resonance），最终才形成语音[2][9][17]，如图 3-2 所示[2][9][17][18]。由

于声道的形状非常灵活多变，可以形成不同的共振，故而造就了千变万化的语音[2][9]。下面将仔细介绍人类声道的各种变化及其产生的语音的特点。

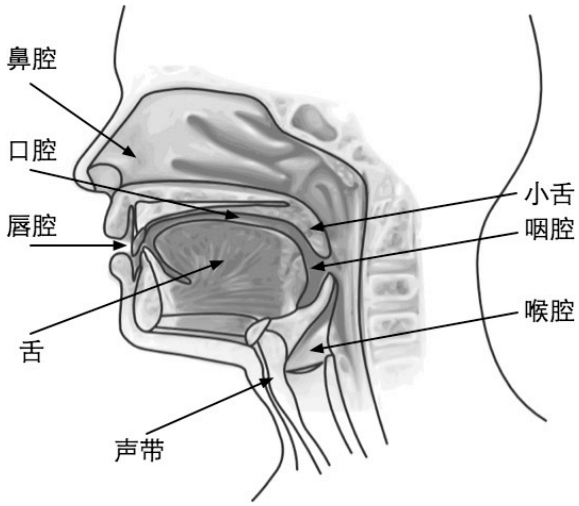


图 3-2 发音器官的组成和结构

喉腔和咽位于声带和小舌（Uvula）之间，它们的形状和大小会随着舌（Tongue）和喉的动作发生变化，从而改变共振效果，改变形成的语音[9]。

口腔是人类最重要的发音部位，其中集中了包括双唇（Bilabial）、舌、软腭（Velum）等主要可活动的发音器官，可以通过改变口腔的形状、容积、气流通路等，产生发音过程中的一切复杂变化[2][9]。其中舌是口腔中最灵活的发音器官，在发音过程中，舌可以上下升降、前后移动，并且舌的组成部分如舌尖、舌面、舌根等也都可以独立运动，形成不同共振效果。双唇是声道的主要出口，它们和牙齿之间形成唇腔；双唇可以拢圆以延长唇腔、可以形成狭缝使气流摩擦通过、也可以完全闭塞。软腭则主要用以改变与鼻腔的气流通路，当它上升挡住到鼻腔的通路时，声音主要在口腔形成共振；而当软腭下降时，鼻腔通路开放，如果此时口腔的通路也开放，声音同时在口腔和鼻腔共振，则形成所谓鼻化音（Nasal Twang），否则仅在鼻腔共振，形成鼻音（Nasal）[9]。

口腔中多变的共振效果会产生多样的语音声波。当声道通畅时，声带的周期性颤动顺利传递，产生周期性的声波叫做浊音（Voiced Sound）。如果声道的某部分部分或暂时完全闭合形成的非周期性声波称为清音（Unvoiced Sound），常见的清音包括擦音（Fricative）、塞音（Stop）和塞擦音（Affricative）。于是语音学家根据发音时声道的开放情况对音素分类：声道完全开放的音素称为元音（Vowel, V），否则称为辅音（Consonant, C）[19]。元音都属于浊音，如’a’。辅音可能是清

音或浊音：而当声道部分开放、先完全封闭再开放、或介于两者之间时，辅音可能分别为擦音（如’f’)、塞音（如’d’)、或塞擦音（如’j’）等清音；如果声带保持震动，则辅音还可能属于浊音，如吴语中对应 PTH-IF 中’d’的浊音声母。

声波的性质由音质、音高（Pitch）、音强、音长（Length）组成。音质是语音最重要的性质，从感知角度，一段语音总可以看做由音质连续变化的（听觉上能够分辨的）最短音组成，这种最短音称为音素；由一个或若干个音素组成的（听觉上）最容易分辨的音段称为音节[2][8]。音高、音强、音长通常称为超音质（Super Tone Quality）成分，它们通常依附于音节或更大的音段上[9]。汉语声调是音节的音高，本文 1.2.2 节在 IF 层次分析口音变异相当于分析口音产生的音质变异。

3.3.2　汉语声母的形成

很自然的，我们可以将从生理角度得到的定性描述发音状态的要素，如圆唇、鼻音等用于描述音素的形成，称为发音特征[17]。本节我们使用发音特征来刻画 PTH 的 22 个声母。

PTH 中每个声母都是 1 个辅音，语音学家使用舌位（Place）、发音方法（Manner）、送气情况（Aspirated/Unaspirated）、清浊情况（Voiced/Unvoiced）4 种不同的发音特征角度来唯一确定 1 个辅音，而能够唯一确定 1 个音素的发音特征特征也称为区别特征（Distinctive Feature）[11][17][19]。下面我们简介各类发音特征。

(1) 根据辅音定义，发辅音时声道中要形成阻碍。舌位主要刻画口腔中成阻时舌头的位置，下面简介舌位角度的发音特征:

双唇音（Bilabial）：是由上下双唇接触成阻，如’b’。

唇齿音（Labiodental）：是由上齿和下唇接触成阻，如’f’。

舌尖前音（Dental）：舌尖与齿背成阻，如’z’。

舌尖中音（Alveolar）：由舌尖和上齿成阻，如’d’。

舌尖后音（Palatal Alveolar）：舌尖与硬腭前成阻，如’zh’。

舌面前音（Palatal）：舌面前部与硬腭前部接触成阻，如’j’。

舌面后音（Velar）：舌面后部与软腭成阻，如’g’。

(2) 发音方法描述气流经过声道形成辅音的过程，可以简单分为形成阻碍的成阻、肌肉保持紧张使阻碍持续的持阻、以及肌肉放松出去阻碍的除阻阶段，下面按照这 3 个阶段简介发音方法角度的发音特征[10]：

塞音（Stop）：在成阻和持阻时声道完全阻塞，出现声音的间歇；到除阻阶段

阻碍突然消失，出现声音爆破的感觉，如’b’。

塞擦音（发音特征 fricative）：在成阻时声道完全阻塞；持阻时略微开放使气流挤出而产生摩擦直到阻碍消失，如’zh’。

擦音（Fricative）：成阻和持阻时声道并不完全阻塞，气流从缝隙摩擦通过，产生紊音直到阻碍消除，如’f’。

鼻音（Nasal）：在成阻时口腔完全封闭但软腭下降时鼻腔通路开放；持阻时气流从鼻腔释放，直到阻碍消除，如’n’。

边音（Lateral）：舌尖成阻，但舌尖一侧或两侧有空隙可以释放气流，如’l’。

(3) 辅音除阻之后有送气流跟随称为送气情况[9]，下面属于这一角度的发音特征：

送气音（Aspirated）：有气流跟随，如’b’。

不送气音（Unaspirated）：无气流跟随，如‘s’。

均可（Not Available）：如’f’。

(4) 从清浊情况的角度：

浊音（Voiced）：如’b’。

清音（Unvoiced）：如’r’。

综合以上各种发音特征，我们可以将 PTH 辅音归类到附录 2 表 1 中[9][11]。表中每项都是 1 个 PTH 声母，可以由记做<舌位, 发音方式, 送气情况, 清浊情况>形式的 1 个 4 元组唯一确定。当 4 元组中某 1 项的即发音特征发生变异，声母就发生了变化，即发音变异。例如<舌尖中音, 塞音, 不送气, 清音>在表中对应声母’d’，当发音特征“清音”变为“浊音”时，即组合变为<舌尖中音, 塞音, 不送气, 浊音>，’d’就变异为吴语中对应的浊辅音。类似的还有川、粤、吴语共有的’zh’→‘z’、’ch’→‘c’、’sh’→‘s’等口音变异，从附录 2 表 1 中可以清楚的看到它们均对应发音特征“舌尖后音”到“舌尖前音”的变异。可见发音特征比 IF 更容易描述辅音的口音变异。

3.3.3　汉语韵母的形成

PTH 中共有 36 个韵母，每个韵母由可能由 1~3 个音素顺序组成，依据其结构可分为单元音（Vowel）、二合元音（Diphthong）、鼻韵母（Nasalized Vowel）、三合元音（Triphthong）共 4 种类型。下面分类简介这些韵母。

(1) 单元音：不与 1 其它元音结合就能在音节中单独存在的元音叫做单元音[9]，每个单元音都由 1 个元音构成，如’a’。PTH-IF 中共有 10 个单元音韵母。

(2) 二合元音：复韵母是由 VV[2]形式构成的，PTH 中包括 8 个二合元音。在拼读二合元音时说话人的舌头是滑动运动的，故而产生的音质是连续变化的，一般比较难在两个元音间找到准确界限；同时各元音的音长、音强等超音质成分不同，一般只有一个元音听起来最响亮，如 PTH-IF 韵母'ai'中'a'比'i'更响，前一个元音较响的韵母称为前响复韵母；后一个元音较响则成为后响复韵母[11]。

(3) 鼻韵母：在 PTH 中是指由单元音或二合元音接辅音'n'或/ng/结尾所构成的韵母，其组成可以写为 VC 或 VVC 的形式。PTH 中有 16 个鼻韵母。

(4) 三合元音：PTH 共有 4 个三合元音，它们都由 VVV 的形式构成，分别是'iao'、'iou'、'uai'和'uei'。其韵母连续变化的特点与二合元音类似；但往往是中间一个元音发音较响，所以也称为中响复韵母[11]。

从上面的介绍可以看出，PTH 的韵母的结构可以统一写为 V(V)(V/C) [3]的形式，语音学家将可能出现的 3 个元音或辅音依次称为韵首、韵腹和韵尾[11]。PTH 中所有韵母的组成方式按其分类总结在附录 2 表 2[4]中。还可以发现，元音是组成韵母的核心音素，本文将从发音时的舌位、舌位前后（Front-End）、舌位高低（Height）、以及唇形（Rounded/Unrounded）4 个发音特征的角度介绍 PTH 中的各种元音。

(1) 描述元音的舌位与辅音的名词相似但含义不同，形成元音时声道通畅无阻碍，此时舌位角度发音特征描述的不再是形成阻碍的位置而是发音时舌部肌肉用力的位置和舌的形状。下面是元音舌位角度的发音特征：

舌尖前音：发音时主要靠舌尖部分的肌肉用力，但舌中线所构成的马鞍形曲线在舌尖的高点比较靠前同时舌面后部的高点比较靠后[8]，如图 3-3 (a)所示[20]。PTH 中的声母'i'是舌尖前音。

舌尖后音：发音时也依靠舌尖用力，但舌尖的高点靠中间且舌面后部的高点较靠前，如图 3-3 (b)所示[20]，如 PTH-IF 中的'u'。

舌面前音：发音时舌头各部分的肌肉用力比较均衡，同时舌中线两个高点的情况与舌尖前音一致，如图 3-3 (c)所示[20]。例如 PTH-IF 中的'e'。

卷舌音（Retroflection）：发音舌面前音时舌尖向硬腭卷起，称为卷舌音，如图 3-3 (d)所示[20]。卷舌音最典型的代表是 PTH 中的儿化音'er'。

(2) 发音时舌在纵向的位置被分为高（High）、半高（Mid-High）、半低（Mid-Low）、低（Low）4 种发音特征，合称为舌位高低角度。

2 V 指代一个元音，VV 表示两个元音依次相接；类似的，C 指代一个辅音，下同。

3 括号表示可能出现；C/V 则表示该音素可能是元音或辅音。

4 附录 2 表 2 中'a'对应的发音包含了 3 种音位变体的发音/A/，/a/和/ɑ/，因区分与否它们对意义没有影响，故简单起见，本文将它们合并处理。

(3) 舌位前后角度描述发音时舌在横向所处的位置，包括前（Front）、央（Central）、后（End）3 种不同的发音特征。受限于舌部肌肉的物理性质，发音时舌在前后可活动的范围受舌位高低的影响，语音学中使用图 3-5 中的梯形来描述这种约束关系：舌位越高在梯形中对应的位置越靠上；梯形的上底长于下底，表示舌头在更高的位置时有更大的前后活动范围。

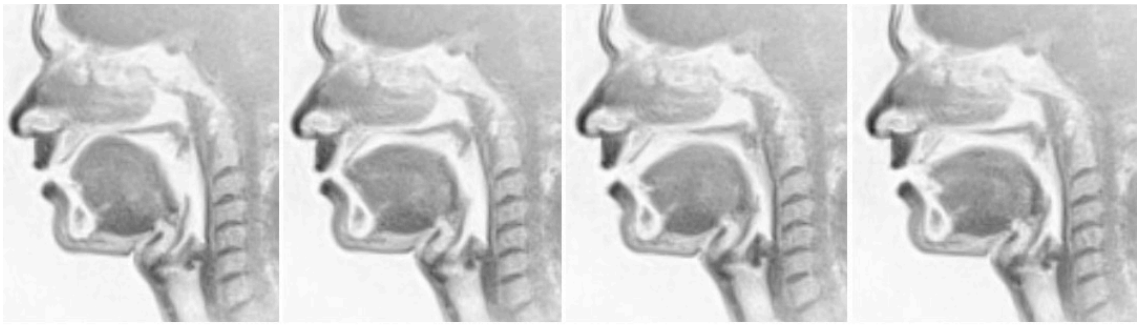


图 3-3 使用核磁共振技术得到的元音’i’、’u’、’e’、’er’对应的舌位特性

(4) 发元音时唇形也是决定元音音质的因素之一。唇形角度包括圆唇与（Rounded）与非圆唇（Unrounded）两种发音特征，一般定义唇形与舌头的状态相互独立，但事实上随着舌位的降低，圆唇的程度也降低。唇形也可表示在图 3-4 中，圆唇的音在竖线右侧，非圆唇音列在竖线左侧。

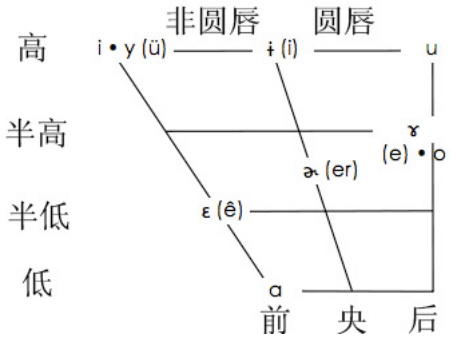


图 3-4 标准汉语元音的发音特征

于是，利用图 3-4 可以发音方式清楚的描述所有元音。图中已表示了 PTH-IF 中的所有单元音，如左上角右侧的点对应的发音特征元组是<舌面前音, 高, 后, 圆唇>，它代表 PTH-IF 的’u’，而它下方右侧的元音/ø/则对应了 W-IF 中“南”字读音的韵母。这样，当某个发音特征值变异时就发生了元音的口音变异。对于上一

节所提到 PTH 中后响复韵母’ia’、’ua’、’ie’、’uo’、’ue’的韵首在粤语中通常不发音的情况，利用发音特征可以简单的总结为：当二合元音的前一元音比后一元音的舌位高时，前一元音的舌位高低发音特征丢失，从而导致插音现象。这样发音特征也可以简洁的描述汉语韵母的口音变异。清楚起见，我们将 PTH 中单元音对应的发音特征组合列在附录 2 表 3 中。

3.3.4　连续语音中的协同发音现象和非线性音系学

人类用语音自由交流的时候，总是连续不断的对各个音素进行发音的，如图 1.1 所示，这种语音称为连续语音，包括自然语音以及连续发音的朗读语音等。由于这些紧密连续出现的音素对应的发音方式和部位各不相同，在连续语音中它们难免因相互影响而造成说出的音子带有明显的发音变异[9]。这些变异既包括前文提到的上下文相关的口音变异，也包括在口音和标准语音中都有的有规律或随意的发音变异，后一种称之为协同发音现象。

协同发音可能受到说话人的情绪、态度、个人习惯等很多因素的影响，变化非常复杂。对协同发音造成最大影响的因素是语速，比如词语“不知道”[bu zhi dao]在语速很快的时候可能出现声母脱落的现象，变成[bu i2 dao][9]。除增减发音的情况外，协同发音现象还包括把不同的发音变得相同的同化作用（Assimilation），以及把相同发音变得不同的异化作用（Dissimilation）等[19]。这些由协同发音造成的发音变异在音质层面主要依靠结合发音上下文的方法进行研究。另外一种代表性的受规则约束的协同发音现象是儿话音现象，儿化音是指由作为语音结尾的’er’与它前面的音节合成 1 个特殊发音所导致的发音变异现象[9] [11]，如 PTH 中的“哪儿”[na er]常连读为[nar]。传统的生成音系学研究方法主要集中在音素层面，即研究前 1 个和后 1 个因素对当前音素读音的影响，但研究表明，这种方法无法准确解释许多协同发音现象[11]。

近些年来，语音学研究的主流转向了非线性音系学（Non-Linear Phonology）[21]，这种理论可以更好的解释语音中包括口音变异、协同发音等在内的各种发音变异现象，并已成功应用于汉语研究[11]。传统语音学认为音素在由它们自身按照时序构成的单一线性结构上组成语音[9]，而非线性音系学则认为不同特征之间有着互相排斥或共同出现等特点，同时每种特征自身也像音素一样分别构成一条时序的线，从而组成多线性结构[11]。非线性音系学的理论有着明确的生理学意义：对于不同的发声器官，如喉、软腭、舌体、舌根、双唇等，分别被不同的肌肉群所控制，可以相对独立的进行自主运动；而出于发音的目的，器官自主运动时又会存在相

互的影响和制约，从而导致描述器官运动状态的发音特征存在影响和制约，如临近的器官总是一起运动、一个人的舌头不可能既在前又在后等；而连续语音中某种器官的时序运动则会使器官的当前状态受到其前一状态和下一状态的影响，即该器官对应的发音特征间会存在协同发音的情况[11]，例如前后音素都是非圆唇的情况下中间圆唇音素的唇形很可能受到影响而接近于非圆唇 。

于是，非线性音系学将描述同一音素的发音特征按照相互间的影响和制约关系进行分层并组织为树形结构，如图 3-5 中音素对应的树形横截面所示；而不同音素的同类发音特征则在时间轴上进行连线，如图 3-5 不同树形截面间对应节点的连线。

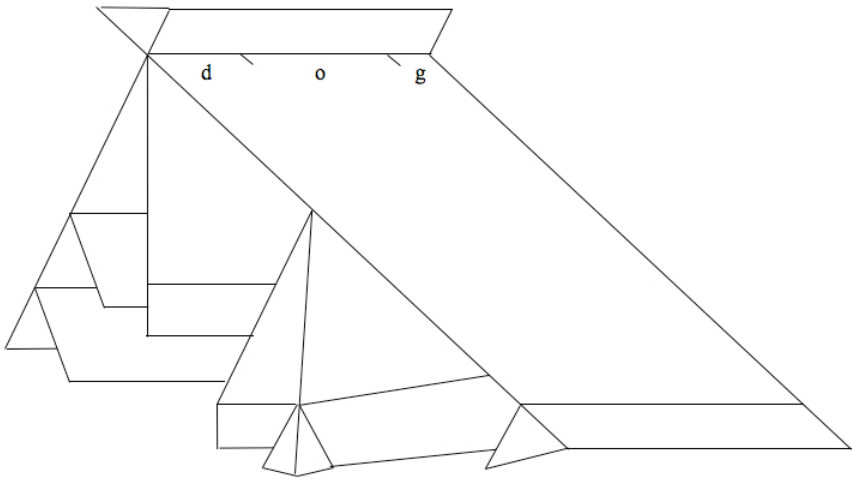


图 3-5 连续语音中英文单词“dog”在非线性音系学中的结构

研究表明利用图 3-5 所示结构，配合节点的扩展、删除、连改等规则，可以很好的解释因协同发音或口音所造成的音质和超音质方面的发音变异。

3.4 条件随机场

CRF 是一种无向图模型（Undirected Graph Model），如图 3-6 (a)所示。CRF 能够将冗余的输入序列 $\vec{x} = x_1 x_2 \ldots x_T$ 整合为在条件概率 $P(\vec{y} \mid \vec{x})$ 意义下可能性最高的输出序列 $\vec{y} = y_1 y_2 \ldots y_T$。

根据贝叶斯公式和加法公式，CRF 的优化目标可以写为

$$P(\vec{y} \mid \vec{x}) = \frac{P(\vec{x}, \vec{y})}{\sum_{\vec{y}'} P(\vec{x}, \vec{y}')}. \tag{3-1}$$

CRF 对应的无向图中，每个节点表示一个随机变量，节点间的边表示条件约束关系。定义图模型中任何两个节点间都有边连接的节点集合为团；若团中再加入任何节点就不再是团，称之位为最大团[31][36]，记为C。设C中随机变量为$(\vec{x}_C,\vec{y}_C)$，$\psi_C(\vec{x}_C,\vec{y}_C)$是实值势函数，最大团间是条件独立的，由乘法公式[25]，随机变量的联合分布为

$$P(\vec{x},\vec{y})=\frac{1}{Z}\prod_C \psi_C(\vec{x}_C,\vec{y}_C), \tag{3-2}$$

Z为归一化因子，以保证(3-2)满足概率意义，

$$Z=\sum_{\vec{y}}\prod_C \psi_C(\vec{x}_C,\vec{y}_C). \tag{3-3}$$

为保证势函数大于 0 且方便计算，通常取势函数为指数形式，即

$$\psi_C(\vec{x}_C,\vec{y}_C)=\exp\{-E_C(\vec{x}_C,\vec{y}_C)\}. \tag{3-4}$$

其中$E_C(\vec{x}_C,\vec{y}_C)$为团的能量函数，$\psi_C(\vec{x}_C,\vec{y}_C)$为玻尔兹曼分布（Boltzmann Distribution），则$P(\vec{x},\vec{y})$可视为每个最大团上的能量之和。可以证明能够以公式(3-2)的概率形式表示的模型集合与从无向图表示中得到的模型集合相同[127]。这是因为可以将(3-2)视为对$\vec{x}$的对应节点序列构成的图χ进行染色的问题，记$\Omega=\prod_{x\in\chi}\Omega_x$是节点$x$的所有可能着色，$\Gamma$是$\Omega$的幂集，$P$是$(\Gamma,\Omega)$上的概率测度，$P(\vec{x},\vec{y})$是概率密度，即图上的概率定义与通常的概率定义一致[31]。概率空间(Γ,Ω,P)为图χ上的随机场（Random Fields）。图中每个最大团内的节点间满足 1 阶马尔科夫性，将上述无向图模型称为马尔科夫随机场（Markov Random Fields）[31]，CRF 属于马尔科夫随机场[126]。

当对随机变量分布仅知道它所属的类别时，通常选择最大熵（Maximum Entropy）分布的能量函数。这是因为信息相当于负熵，最大熵就相当于最小化分布所固有的先验知识，从而保证了对估计的无偏性；同时许多物理系统随着时间的推移也都倾向于变为最大熵的形式。于是马尔科夫随机场的最大熵分布为

$$P(\vec{x},\vec{y})=\frac{1}{Z}\prod_C \exp\{\sum_{k=1}^{K}\lambda_k f_k(\vec{x}_C,\vec{y}_C)\}. \tag{3-5}$$

其中 $f(\vec{x},\vec{y})$ 为特征函数，λ 为特征的对应权值，$f(\vec{x},\vec{y})$ 和 λ 的取值都可以为任意实数。于是，对于图 3-6 (a)中的线性结构，联合分布的最大熵可以写为

$$P(\vec{x},\vec{y})=\frac{1}{Z}\prod_{C}\psi_C(\vec{x}_C,\vec{y}_C)=\frac{1}{Z}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\vec{x}_t,)\}. \tag{3-6}$$

相应的最大熵条件分布为

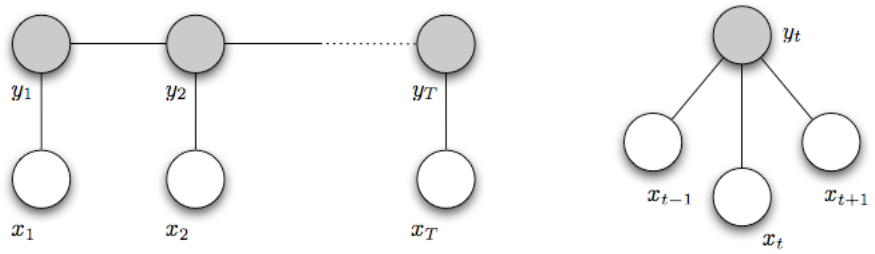
$$P(\vec{y}\mid\vec{x})=\frac{P(\vec{x},\vec{y})}{\sum_{\vec{y}'}P(\vec{x},\vec{y}')}=\frac{\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\vec{x}_t,)\}}{\sum_{\vec{y}'}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\lambda_k f_k(y'_t,y'_{t-1},\vec{x}_t,)\}}. \tag{3-7}$$

即 CRF 可以写为

$$P(\vec{y}\mid\vec{x})=\frac{1}{Z(\vec{x})}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\vec{x}_t)\right\}, \tag{3-8}$$

$Z(\vec{x})$ 为归一化常数，通过将所有可能假设 $\vec{y}'$ 加和计算得到，

$$Z(\vec{x})=\sum_{\vec{y}'}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(y'_t,y'_{t-1},\vec{x}_t)\right\}. \tag{3-9}$$



(a)　　(a)

图 3-6 CRF 和区分性逻辑回归的图表示

于是训练CRF相当于从 R 个训练数据中为所有 K 个特征函数分别学习适合的权值以使条件概率最大化，如图 3-7 所示。条件概率的对数似然度为，

$$l(\{\lambda_k\}_{k=1}^{K})=\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{k=1}^{K=1}\lambda_k f_k(y_t'^{(r)},y_{t-1}'^{(r)},\vec{x}_t^{(r)})-\sum_{r=1}^{R}\ln(Z(\vec{x}^{(r)}))-\sum_{k=1}^{K}\frac{\lambda_k^2}{2\delta^2}. \tag{3-10}$$

其中 $\sum_{k=1}^{K}\lambda_k^2/2\delta^2$ 为 L2 正则化惩罚（L2-Regularization Penalty），即通过假设

λ是均值为 0、方差为δ的参数来避免(3-10)出现参数过拟合。对(3-10)中每个权值求导，有

$$\frac{\partial l}{\partial \lambda_k}=\sum_{r=1}^{R}\sum_{t=1}^{T_r} f_k(y_t^{(r)},y_{t-1}^{(r)},\vec{x}_t^{(r)})-\sum_{r=1}^{R}\sum_{\vec{y}} P(\vec{y}',\vec{x}^{(r)})\sum_{t=1}^{T_r} f_k(y_t'^{(r)},y_{t-1}'^{(r)},\vec{x}_t^{(r)})-\frac{\lambda_k}{\delta^2}. \quad (3\text{-}11)$$

对于图 3-6 (a)中线性结构的随机变量，$P(\vec{y}'|\vec{x})$类似于 HMM 中的情况可以使用前向后向算法得到，

$$\begin{aligned} P(y_t,y_{t-1}|\vec{x}) &\propto \alpha_{t-1}(y_{t-1})\exp\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\vec{x}_t)\}\beta_t(y_t) \\ \alpha_t(y_t) &= \sum_{y_{t-1}}\alpha_{t-1}(y_{t-1})\exp\{\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},\vec{x}_{t-1})\}. \\ \beta_t(y_t) &= \sum_{y_{t+1}}\beta_{t-1}(y_{t-1})\exp\{\sum_{k=1}^{K}\lambda_k f_k(y_{t+1},y_t,\vec{x}_{t+1})\} \end{aligned} \quad (3\text{-}12)$$

则$\partial l/\partial\lambda_k=0$可以使用 L-BFGS 等梯度计算方法得到，从而完成了 CRF 的训练。其中计算量最大的部分为归一化常数$Z(\vec{x})$的计算。

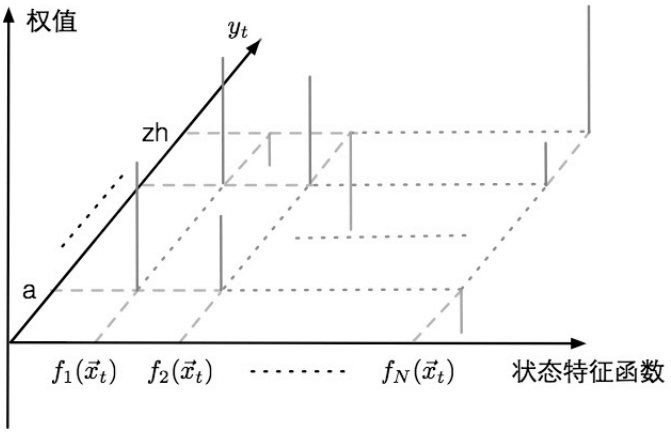


图 3-7 对特征进行线性加权组合的条件随机场模型

在使用 CRF 进行解码时，求解

$$\vec{y}^{*}=\arg\max_{\vec{y}} P(\vec{y}|\vec{x})=\arg\max_{\vec{y}} P(\vec{x},\vec{y}). \quad (3\text{-}13)$$

仍然类似于 HMM 可以使用维特比算法实现，

$$\delta_t(y_t) = \max_{y_t} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, \vec{x}_t)\} \delta_{t-1}(y_{t-1}). \tag{3-14}$$

这时无需计算$Z(\vec{x})$。

CRF 可以看做 HMM 在区分性模型中的直接扩展。首先标准的 HMM 是生成性模型，而 CRF 是区分性模型。HMM 限制了观测间是相互独立的，而 CRF 中可以利用任意实值特征函数，特征函数对团内观测间的关系没有限制。HMM 中状态转移与观测无关，而 CRF 中的转移状态函数可以使用团内所有观测。可以将 HMM 的观测概率和转移概率视作特征函数，于是(1-13)中 HMM 的表达式可以写为(3-6)的指数形式，HMM 相当于依据 ML 准则联合训练了两种特征函数，并使用人工设定的特征权值构成马尔可夫随机场。

3.5 面向带口音语音的探测式ASR

图 3-8 为本章提出的面向带口音语音的探测式 ASR 的系统结构，包含：(i) 基于发音特征属性和音子属性探测器组，(ii) 一个语音属性到音子的融合器，和(iii)基于 HMM 和 FSN 的验证器。与以往研究中通常使用的 ANN 探测器不同[102][103][104]，我们使用 Tri-AF HMMs 对发音特征进行高性能探测。我们使用输入为 2 元特征的 CRF 作为融合器，以便利用声学、语音学规则生成音子。我们使用的 CRF 不包含转移特征函数，并输出带概率的音子点阵而非 1 最优假设，我们利用这样的系统配置解决 CRF 的音子欠生成问题。

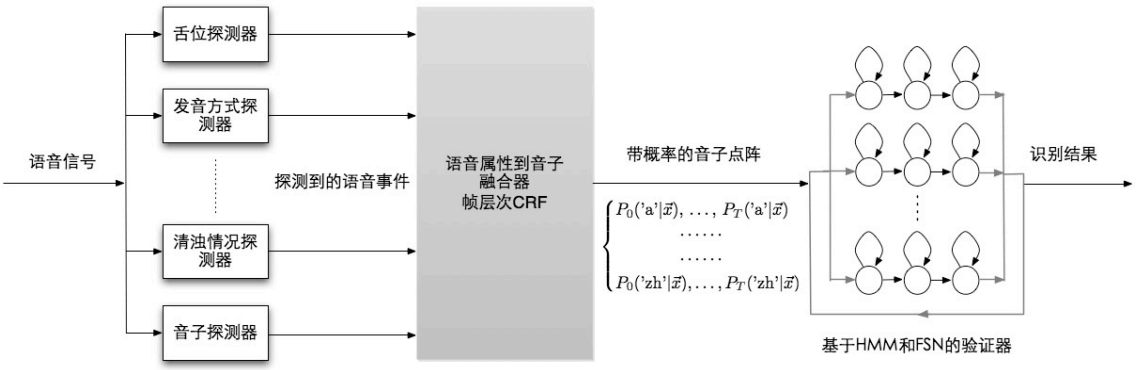


图 3-8 本章使用的探测式 ASR 的结构示意图

3.5.1　基于发音特征属性的探测器设计

图 3-9 中[18][137]给出了语音产生和接收的过程：发音器官通过运动产生长度不同的连续语音段，语音在物理介质中传播直到接收者，接收者的人耳对语音进行处理和感知。由 1.3.2 节可知 MFCC 考虑了声波在介质中传播时所受介质声学特性及人耳的听觉特性的影响，取得了巨大的成功。但 MFCC 也面临着 1.4.2 节给出的诸多问题。由图 3-9，解决这些问题的一个自然角度是引入直接描述语音生成的特征，即发音特征。与 MFCC 不同，发音特征可以直接反映语音段的不定长等特性。

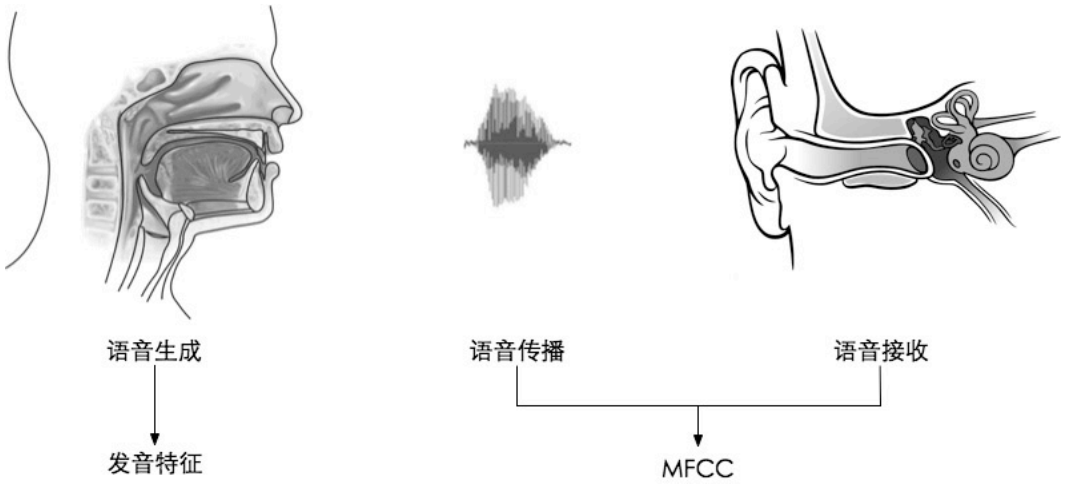


图 3-9 语音的生成、传播和接收过程

本章系统中同时使用发音特征和音子作为语音属性，我们根据附录 2 表 1 和表 3 将音子和发音特征间的对应关系总结在表 3.4 中。为表 3.1 中每一类发音特征特征分别构造 1 套上下文相关的 Tri-AF HMMs 探测器，这样可以更好的与非线性音系学知识结合：(i) 每个探测器中的不同发音特征值属于同一器官的不同状态，不可能同时出现，故将它们作为同一探测器中的不同分类以利用这一特点；(ii) 为每类发音特征分别进行上下文相关建模，从而更好的描述每种发音器官各自连续运动产生协同发音发音变异的过程；(iii) 可以从发音特征角度更好的捕捉上下文相关的口音变异规则。 我们为音子构建了 Monophone HMMs 探测器，从而在系统中引入了冗余信息修正识别结果，如图 3-6 所示。训练音子模型使用的音子标注来自于对原有的 XIF 标注基于附录表 2 分解得到，零声母不对应任何音子；训练发音特征模型的标注来自对音子标注按照附录 2 表 1 和表 3 中的规则转换得到，即 1 个声母对应于 1 个由舌位、发音方法、送气情况、清浊情况中任何可能语音属性的组合构成的 4 元组；1 个韵母对应于 1 个由舌位、舌位高低、舌位前后、唇

形构成的 4 元组。4 元组中不对应于任何 PTH 音子的语音属性组合（如附录 2 表 1、表 3 的空白表项所示）以及额外的音子特征都作为冗余特征，用于修正识别结果。

需要注意的是，如 3.3 节所述，元音和辅音的舌位特征分别定义在不同的语音学空间中。但由于它们都属于对舌头位置的描述，会产生协同发音，所以将元音、辅音的舌位特征合并成 1 类。

表 3.1 PTH 音素与发音特征的对应关系

发音特征分类	发音特征值	音子
舌位	舌尖中音	d l n t
	双唇音	b m p
	舌尖前音	c s z i1
	唇齿音	f
	舌面前音	j q x a o e ei i u v
	舌尖后音	ch r sh zh i2
	卷舌音	er
	舌面后音	g h k ng
发音方法	塞擦音	c ch j q z zh
	擦音	f h r s sh x
	边音	l
	鼻音	m n ng
	塞音	b d g k p t
送气情况	送气音	c ch k p q t
	不送气音	b d g j z zh
	均可	f h l m n r s sh x ng
清浊情况	浊音	l m n r ng
	清音	b c ch d f g h j k p q s sh t x z zh
舌位高低	高	i i1 i2 u v
	低	a
	半高	o e
	半低	ei er
舌位前后	前	o e u
	央	a er i2
	后	ei i v i1
唇形	圆唇	o u v
	非圆唇	a e ei er i i1 i2

本文 2.3 节基于 IF 为每种口音生成上百个基于 XIF 的 ASU 以表述相应的口音变异[4][13][76]，为这些 ASU 建模是一项琐碎且繁杂的工作。然而基于发音特征，我

们可以根据 3.3.3 节和 3.3.4 节中的分析将这些 XIF 的口音变异转化少得多的发音特征变异。本节剩下的部分我们以粤语口音为例实证 XIF 口音变异和发音特征变异的关系。表 3.2 中给出使用 5 小时官话口音数据训练的探测器分别在官话口音和粤语口音测试数据上的识别结果。音段正确率使用表 1.3.4 节中的正确率获得。

表 3.2 所有官话发音特征探测器的正确率

发音特征分类	音段正确率%		帧正确率%		
	官话口音	粤语口音	T = 0	T = 1	T = 2
舌位	79.54	**68.09**	58.63	62.31	65.22
发音方法	84.61	83.30	73.31	77.13	79.23
送气情况	82.57	82.59	76.52	79.18	80.82
清浊情况	88.26	80.14	78.19	82.38	84.84
舌位高低	85.31	83.44	74.92	79.23	81.70
舌位前后	85.59	85.12	77.93	82.14	84.55
唇形	88.26	88.20	75.54	79.98	83.06

众所周知的，语音段的边界并不稳定，即分割不同语音段的时间可能并不精确，通常会有 1~2 帧的浮动。于是在评估语音帧的分类正确率时，对处在音段边界的帧，若它与指定浮动范围内的某帧相同，则分类正确。表 3.2 中“T=0, 1, 2”分别表示浮动范围为 0、1、2 帧。我们对不同浮动范围进行比较，以评估是否可以通过在帧层次融合器中引入附近的帧来降低音段边界的敏感度。

表 3.2 的结果显示了粤语口音变异对官话探测器性能的影响。我们以发音位置探测器为例，通过分析发音特征探测器的混淆矩阵得到它的各个发音特征值的正确率，列在表 3.3 中。观察结果可以很容易发现，双唇音、唇齿音以及舌面前音在两种口音中表现出良好的鲁棒性，而舌尖前音、舌尖后音和卷舌音在粤语口音中出现了显著的性能下降。进一步分析两种口音中的混淆矩阵我们发现：(i)舌尖后音的性能下降主要被误识别成了唇齿音，这符合粤语口音说话人容易把’zh’、’ch’、’sh’拼读成’z’、’c’、’s’ 的规律；(ii) 粤语口音中卷舌音的大多数识别错误是由它被错识别为舌面前音或被删除造成的，这与粤语中因为没有卷舌音元音’er’做为音节结尾，从而导致粤语口音说话人或者倾向于将其拼读成发音最接近的元音’e’，或者倾向于或者放弃拼读儿化音的规律相符；(iii) 粤语口音测试集中存在更多对舌尖前音的删除错误，而相应的，我们利用官话 Monophone 模型在同一测试机上也发现了对’z’、’c’、’s’的更多删除错误。于是，在 ASR 中可以利用发音特征变异对口音变异进行更简明更本质的表述。

表 3.3 官话口音舌位探测器中所有发音特征的正确率

舌位发音特征值	音段正确率%	
	官话口音	粤语口音
双唇音	81.8	79.0
唇齿音	92.2	90.9
舌尖前音	82.3	**76.7**
舌尖中音	75.4	72.3
舌尖后音	79.3	**71.0**
舌面前音	80.7	80.2
舌面后音	80.9	77.3
卷舌音	90.8	**66.1**

3.5.2　基于条件随机场的语音属性融合器

语音属性到音子融合器的主要作用是把探测器探测得到的各个事件流赋予不同的权值整合为带概率的音子点阵，并将其传递给线索验证器。融合器应该能够依据语音学规则从数据中自动挖掘深层的口音变异规律，并根据它们在数据中出现的频率和上下文为学习到相应的权值。CRF 是实现这一目标的一种强大的工具[126]。$\vec{x}$ 是探测器输出的帧层次的语音事件流，$\vec{y}$ 是 1 最优的音子假设序列，CRF 对每帧输入 x 自动赋予 1 个标注 y。

CRF 应用中的一个核心问题是特征函数的选择。特征函数 $f(y_{t-1}, y_t, \vec{x}_t)$ 通常可以分为两类：状态特征函数和转移特征函数。一个状态特征函数用于描述一个帧的一个特别状态。例如，用于描述韵母'er'被错发音成舌面前音'e'的状态函数定义为，

$$s(y_t, \vec{x}_t) = \begin{cases} 1, & \text{if } y_t = \text{'er' and palatal}(x_t) = true \\ 0, & \text{otherwise} \end{cases} . \qquad (3\text{-}15)$$

转移特征函数的使用相似的方法定义，它用来对在某个特殊情况下发生的转移进行计数，例如，

$$t(y_t, y_{t-1}, \vec{x}_t) = \begin{cases} 1, & \text{if } y_t = \text{'er'}, y_{t-1} = \text{'a' and retroflection } (x_t) = true \\ 0, & \text{otherwise} \end{cases} , \qquad (3\text{-}16)$$

表示对韵母'a'发生儿化的情况进行计数。

以往使用 CRF 作为语音属性融合器的研究中通常同时使用状态特征函数和转移特征函数。ASR 中因为输入特征比较复杂，故状态特征函数的变化很频繁，导致状态特征函数总表示标注 y 应该改变；但语音中具有相同标注的音段通常都比较长，所以模型中代表 y 不变的转移特征函数的权值很大。这样使得识别时 y 总倾向于稳定不变，导致音子的删除错误是插入错误的若干倍，产生音子欠生成问题[39][101][128]。由于 ASR 的准确率通常在插入错误和删除错误接近的时候达到最高，故音子欠生成问题会伤害 ASR 的性能。同时底层融合器中过多的音子被删除会给音节等更高层融合器造成大量的删除错误，给在音节结构等层次分析口音变异带来困难[99]。为解决音子欠生成问题，我们在 CRF 中仅使用状态特征函数而不使用转移特征函数；同时令 CRF 生成带概率的音子点阵而非 1 最优假设序列，输出包含所有可能假设音子序列信息，以便在后续处理中引入状态转移信息。

我们结合使用规则和数据驱动的方法生成状态特征函数。基于语音学知识定义状态特征函数的模式，然后从实际数据中寻找所有匹配这些模式的实例作为特征函数。这样既可以引入先验知识来约束特征函数的形式，又可以使规则完全来源于实际数据从而更加有效。我们使用的特征函数的模式如下：

(1) 存在特征函数：为每个音子标注，为所有输入语音属性分别构造特征函数。

(2) 区别特征函数：依据 3.3.2 和 3.3.3 节中区别性特征的定义，用分别属于<舌位, 发音方式, 送气情况, 清浊情况>的 4 个发音特征的所有组合构成辅音的区别特征，用分别属于<舌位, 舌位高低, 舌位前后, 唇形>的 4 个发音特征的所有组合作为元音的区别特征。并为每个音子分别构造这样 2 个区别特征函数。区别特征函数可以有效的反应发音器官间协同出现或互斥的关系，从而在 CRF 中引入了非线性音系学的思想。

(3) 窗特征函数：使用当前观测的前 1 帧、前 2 帧、后 1 帧、后 2 帧的每个存在特征函数和区别特征函数为当前标注构造特征函数。使用窗特征函数是因为表 3.2 中显示了引入前后 2 帧的信息可以有效提高探测器的正确率。

值得说明的是，本节使用的不包含转移特征函数的 CRF 相当于通过区分性训练得到的逻辑回归（Logistic Regression）模型，如图 3-6 (b)所示[36]。

我们在系统的探测器和融合器中整合了非线性音系学的核心知识，但不同于专家系统等方法利用先验知识的思路，我们只使用语音学中的定性研究的成果指导系统设计，而没有利用先验知识定量确定系统参数。系统中的所有参数都使用统计方法获得，更加准确有效。

3.5.3　线索验证器

图 3-10 中为线索验证器的工作原理。图中上半部分为带概率的音子点阵，包括了每一时刻输出为每个音子的概率，可以使用多种常见的技术搜索得到识别结果，即所有可能其中所有假设序列中使得后验概率最大化的音子序列。我们使用基于 HMM 和 FSN 的搜索技术，如图 3-10 下半部分所示。我们为每个音子分别构造 1 个类似图 1-5 中拓扑结构的上下文无关 HMM，HMM 在某时刻的观测概率为点阵中对应音子在该时刻的概率，转移概率来自 Monophone HMM 探测器的转移概率。我们使用类似图 1-6 中自由语法规则构造 FSN 对音子识别进行约束，每个 HMM 的初始概率均相等。使用维特比算法在网络中进行搜索以获得最终的识别结果。利用 1.3.4 节中的词惩罚分技术，当路径中每生成一个音子就扣除一定的对数似然分，从而可以实现插入错误和删除错误的平衡。

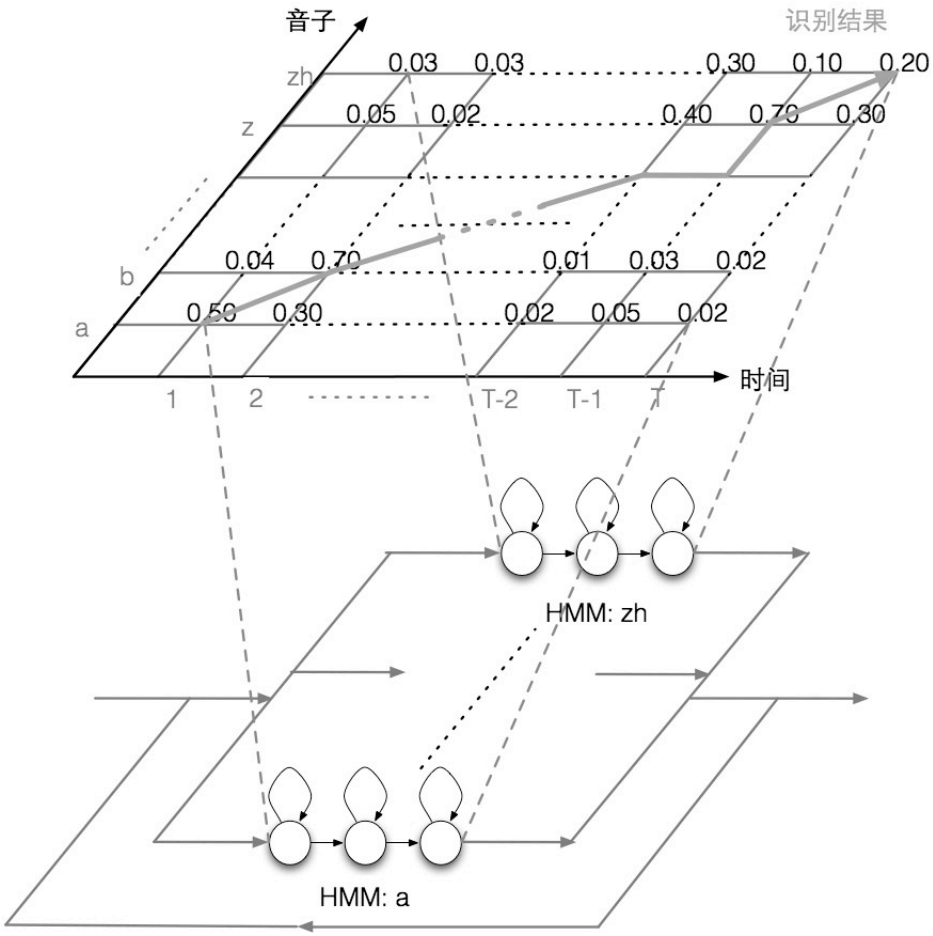


图 3-10 在带概率的音子点阵中根据自由语法获得音子层次的识别结果

3.6 实验结果

3.6.1 数据和基线系统

本章实验使用的粤、吴 2 种带口音数据均选自 RASC863 数据库，并根据数据库的记录信息选口音最明显的说话人；官话口音数据选自 RASC863-G2 数据库属于北方官话方言区的数据，选择其中口音最轻的说话人以使官话口音尽可能接近于 PTH。所有语音均为 16 kHz 采样率 16b 采样精度，各数据集的详细情况见表 3.4。TrainG、TrainY、TrainW 均为训练数据集，其中每个音子平均有 3,193 个训练样本，足够训练可靠的 HMM 模型。TestG、TestY、TestW 均为测试数据集。

表 3.4 本章所有数据集详情

数据库标记	**TrainG**	**TestG**	**TrainY**	**TestY**	**TrainW**	**TestW**
时长 (小时)	4.2	2.9	5.1	2.8	5.1	2.9
音子总数	102,646	70,397	105,167	70,297	108,315	68,173
说话人总数	60	20	60	20	60	20
语句总数	2,144	1,274	2,317	1,309	2,611	1,283
口音类型	官话		粤语		吴语	

本章实验的所有声学特征均采用经过 CMN 处理的 12 维倒谱系数和 1 维归一化的能量构成的 13 维 MFCC，13 维一阶差分 MFCC 和 13 维二阶差分 MFCC。所有探测器和基线系统中的 HMM 均使用图 1-5 的拓扑结构，包含 5 个状态，按照从左到右无跳转的顺序进行跳转。利用基于决策树的状态共享技术训练 Tri-AF HMMs 探测器、Monophone HMMs 系统和 Triphone HMMs 系统。决策树的问题集来自对 2.2.3 节中的 XIF 问题集使用附录 2 表 1、表 2、表 3 的映射关系转换得到。

为准确评估本章提出的算法对口音变异建模所产生的绝对系统性能提高，本章所有实验均使用自由语法的音子识别，识别结果均使用 1.6 节中的音子准确率和音子正确率进行度量。

3.6.2 面向单一口音的探测式ASR

对每种口音分别构造 8 个探测器：为 7 类发音特征分别构造基于决策树进行状态共享的 Tri-AF HMMs 探测器，每个状态包含 6 个高斯成分；为音子属性构造 Monophone HMM 探测器，每状态包含 16 个高斯成分。各组探测器的详细情况如表 3.5 所示。

表 3.5 所有 Tri-AF HMMs 探测器详情

训练集	**TrainG**	**TrainY**	**TrainW**
发音特征分类	**共享状态数**	**共享状态数**	**共享状态数**
送气情况	48	46	47
舌位前后	183	185	182
舌位高低	179	176	176
发音方式	147	140	148
舌位	230	233	228
唇形	104	109	106
清浊情况	68	68	69

我们为每种口音分别构造 1 个 CRF 语音属性-音子融合器。官话口音 CRF、粤语口音 CRF 和吴语口音 CRF 分别包含 260,730、272,245 和 272,745 个特征函数。表 3.6 给出了探测式 ASR 与 Triphone HMMs 基线系统间的性能比较。官话口音、粤语口音、吴语口音的 Triphone HMMs 声学模型分别包含 514 个、572 个和 574 个共享状态。表 3.6 中每个系统均使用与它训练口音相同的测试集测试。

表 3.6 使用一种口音训练和测试得到的系统性能比较

系统	**口音类型**	**音子正确率%**	**音子准确率%**	**耗时 (秒)**
探测式 ASR	官话	73.25	64.48	2,744
	粤语	71.66	63.06	3,304
	吴语	72.41	63.59	2,608
Monophone HMMs	官话	68.77	59.44	408
	粤语	67.54	58.38	424
	吴语	67.23	57.53	384
Triphone HMMs	官话	74.34	66.17	5,128
	粤语	72.57	64.79	5,920
	吴语	73.38	64.71	4,912

在 ASR 中应用 CRF 时通常会产生大量的删除错误，在导致系统性能下降的同时会给音节、语句等更高层处理带来不必要的困难。由表 3.6，探测式 ASR 的识别结果中插入和删除错误相平衡，这是因为：(i) 由于取消了转移特征函数，CRF 解码时不再受到具有极大权值的同状态转移函数的影响；(ii) 验证器中 HMM 的 5 状态无跳转的结构对音子长度进行了约束；(iii) 验证器利用词惩罚分提高了插入

错误和删除错误相近的搜索路径被选出的可能性。

在单一口音问题中，探测式ASR系统的性能显著好于Monophone HMMs系统，这表明使用 2 元特征的 CRF 能够根据语音学规则的指导从数据中挖掘以发音特征形式存在的深层规律，从而对发音相近的音子进行有效的区分。同时虽然我们系统的绝对音子准确率比 Triphone HMMs 系统低 1.51%，但平均识别速度却加快了 5.71 倍。探测式 ASR 中还可以整合更多的语音、语言学知识来进一步提高性能。

3.6.3　语音属性区分单元的集成

比较表 3.6 和 3.7 容易发现，粤语口音和吴语口音严重降低了官话 ASR 的识别率。我们提出在系统中集成口音相关的“语音属性区分单元（Speech Attribute Discrimination Module, SADM）”，从而可以再不重新训练系统中任何模型的情况下提升对系统口音变异的覆盖能力。SADM 通过覆盖发音特征变异达到解决音子、XIF 等口音变异的目的。

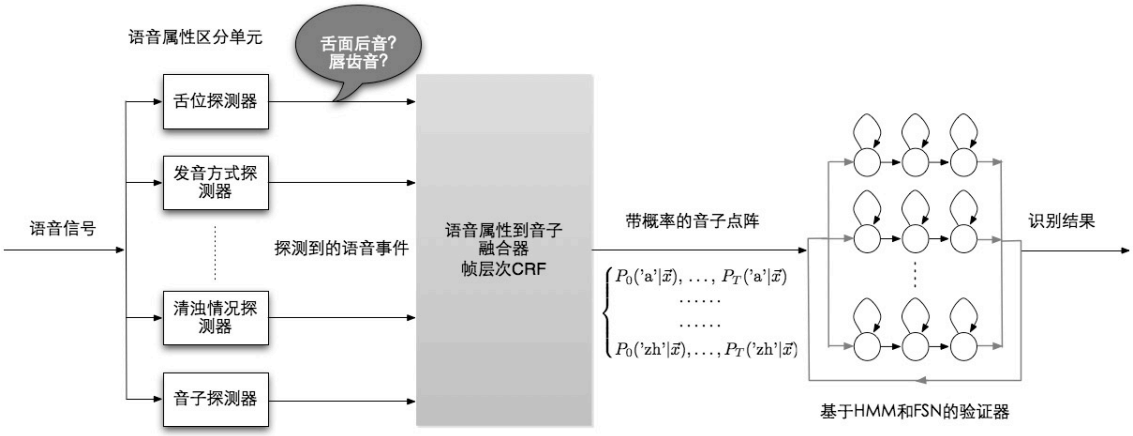


图 3-11　在官话口音系统中使用粤语口音“舌面后音”/“唇齿音”ADM

对特定的发音特征变异，使用粤口音数据中对应发音特征的样本训练一个“支持向量机（Support Vector Machine, SVM）”。SVM 是一种在最小化结构风险（Minimum A Structured Risk）意义下理想的 2 元分类器，并已经被成功应用到 ASR 研究中[129][130][131]，对发音特征分类问题它比 GMM 模型具有更高的准确率。当测试数据为粤语口音时，对于某个发音特征变异，使用 SVM 对探测器探测到的属于相应发音特征的样本进行重分类，并用重分类的结果取代探测结果输入 CRF，如图 3-11。我们使用基于径向基核函数（Radius Based Function）的 SVM 为“舌面后音/唇齿音”发音特征变异训练 1 套 SADM，SVM 的输入观测为按照 3.5.1 节

设置得到的 MFCC。表 3.7 给出集成 SADM 后探测式 ASR 的识别结果。

表 3.7 官话 ASR 在粤语口音和吴语口音上的识别率

系统	**官话口音系统**		**官话口音系统 + 粤语口音 SADM**
测试集	TestY	TestW	TestY
音子正确率%	66.22	63.92	68.16
音子准确率%	52.15	51.18	59.21

使用 SADM 后，舌面后音和唇齿音的准确率分别从 71.0%和 76.7%上升为 83.0%和 92.1%，同时官话口音探测式 ASR 的绝对音子准确率提高了 7.06%，而训练 SADM 仅使用了 6.49%的 TrainY 数据（1,830,778 个样本中的 118,751 个样本）。可以以通过即插即用的方式在探测式 ASR 中集成针对任何口音和任何发音特征变异的 SADM，而无需重新训练系统中任何已有模型。使用 SADM 可以增加探测式 ASR 的灵活性，同时便于进行口音变异的相关研究。

通过分析音子混淆矩阵，我们发现与舌面后音变异为唇齿音相关的一系列口音变异的识别错误都得到了降低，如’zh’→‘z’、’ch’→‘c’、’sh’→‘s’等。同时，一些与舌面后音和唇齿音无关的元音正确率也通过它们删除错误的减少得到了提高。这一结果表明 CRF 利用了语音学规则整合探测器输出，但并不限于仅使用这些规则。事实上 CRF 用启发式的方法从数据中挖掘出了可以区分音子的深层规律。图 3-9 中给出了一些代表性音子识别率的提高。

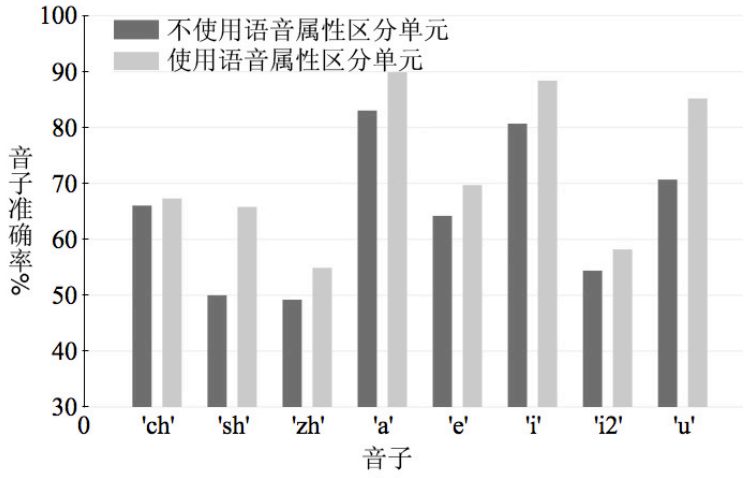


图 3-12 官话系统使用/不使用 ADM 的时在粤语口音上的识别率比较

3.6.4　面向多种口音的并行探测式ASR

最后我们提出一种对多口音具有鲁棒性的探测式 ASR 的构造方法。从前文的讨论可以看出，探测器的探测结果受很多符合一定规律的口音变异的影响。直观的，使用不同口音数据训练的探测器在某种口音数据上会表现出不同的规律，这就为从不同口音的探测器组中提取与多方言口音发音变异有关的模式提供了可能性。于是我们在一组探测式 ASR 中同时使用官话探测器、粤语口音探测器、吴语口音探测器并行提取发音特征，再使用 CRF 对探测结果进行整合。我们在 3.5.2 节中给出的特征生成模式的基础上新增以下模式：

(4) 交互特征函数：对不同口音中的同一类语音属性的联合出现，在每个音子分别作为当前帧、前后 1 帧、前后 2 帧的输出标注的情况下各构造 1 个特征函数。

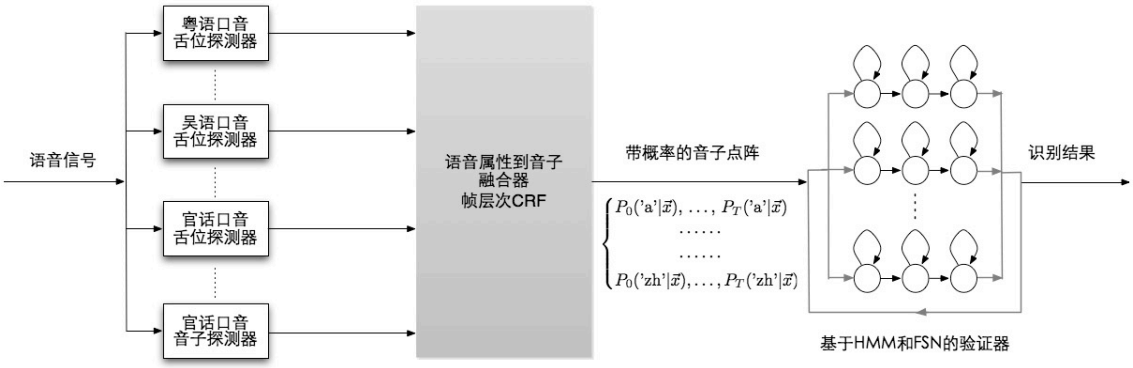


图 3-13　对多口音发音变异鲁棒的并行探测式 ASR

于是，我们在系统中联合使用表 3.5 中的所有探测器和每种口音的上下文无关的 HMM 音子探测器，如图 3-13 所示。CRF 融合器中共包含 788,488 个特征函数。表 3.8 对并行探测式 ASR 及使用 TrainG、TrainY 和 TrainW 数据混合训练的 Triphone HMMs 系统进行比较。Triphone HMMs 使用基于决策树的状态共享，共包含 1,203 个共享状态，每状态包含 6 个高斯成分。

表 3.8　系统对多口音鲁棒性的比较

系统	口音	音子正确率%	音子准确率%	耗时 (秒)
并行探测式 ASR	官话	73.83	65.39	13,128
	粤语	73.75	66.49	13,384
	吴语	73.13	67.09	12,740
Triphone HMMs	官话	74.49	65.40	45,208
	粤语	74.79	66.34	46,740
	吴语	73.70	65.87	43,008

从表 3.8 可以看出，并行探测式 ASR 在所有口音上都得到了与 Triphone HMMs 相近的识别率，但识别速度快 3.44 倍，同时无需重新训练探测器。

3.7　使用基于多语音属性的ASR进行带口音语音识别的优势

由 3.5 节和 3.6 节，本章提出的基于多语音属性的探测式 ASR 进行带口音语音识别相比于传统基于单一语音属性的 ASR 方法具有以下优势：

(1) 容易利用具有复杂关系的发音特征作为多语音属性。发音特征与发音过程直接相关，可以更好的反应语音的本质特性，可以简洁高效的刻画并解决口音变异。同时，语音学中已有大量应用发音特征进行方言口音发音变异相关研究的先验知识，应用这些知识可以提高系统的鲁棒性。

(2) 探测式 ASR 利用层次化的系统结构可以方便的利用非线性音系学的框架，将因口音或协同发音产生的上下文相关的发音变异从音子层次解构到发音特征层次进行上下文相关建模，从而极大的提高了系统速度。另外，研究表明这种结构还非常适合为汉语声调等超音质成分的建模，并容易口音产生的音节结构等更高层次发音变异的建模处理。

(3) 探测式 ASR 系统各模块功能明确，系统结构灵活，可以很容易的为系统添加、减少新的功能和模块而无需修改整个系统。方便在发音特征、音子、声韵母、音节等各个层次发音变异的研究。同时，通过在系统中引入不同的语音属性也有助于增进对语音本质的了解。

3.8　本章小结

本章我们提出了使用基于 ASAT 技术和发音特征的探测式 ASR 解决汉语带口音语音识别问题，系统按照非线性音系学知识设计。我们首先介绍了 ASAT 技术及其研究现状，然后详细介绍了汉语语音的发音特征及非线性音系学。我们将 7 类发音特征和音子视为语音属性并分别为之构造探测器。将探测到的多个语音事件流作为 2 元帧层次特征，基于语音学知识启发式的利用 CRF 对所有特征灵活地进行融合。我们通过在 CRF 中仅使用状态特征函数，并使用线索验证器取代转移特征函数来解决 CRF 的音子欠生成问题。我们还提出了口音相关的语音属性区分单元，它可以简洁的解决多样的口音变异而无需重新训练系统中的任何模型，增强了系统的鲁棒性和灵活性。实验结果表明，在官话、粤语口音、吴语口音上，

我们提出的系统与传统的 Monophone HMMs 系统相比，绝对音子准确率分别提高了 5.04%、4.68%、6.06%；与传统的 Triphone HMMs 系统相比，我们的系统具有相近的识别率，但识别速度快 5.71 倍。我们还提出了并行探测式 ASR，它可以在单一系统中灵活地解决多方言口音产生的发音变异。

第 4 章　总结和展望

汉语多样的方言口音通常会导致语音的声学特性发生改变，从而与使用标准语音训练的 ASR 的声学特性失配，导致 ASR 识别率下降。由于普通话使用者中带有一定程度方言口音的说话人比例很高，故使 ASR 对由多种方言口音产生的发音变异具有好的鲁棒性是汉语 ASR 实用化的关键之一。但受限于语音学、认知科学等学科的发展，目前缺乏对语音本质的足够认识，很难使用单一一种理想的特征实现这一目标。

研究表明，语音中很多特征对于语音的区分有重要作用，称它们为语音属性。于是，突破基于 HMM 的经典 ASR 中只使用单一语音属性的限制，合理利用更多的语音属性来提高 ASR 的性能和鲁棒性是推动 ASR 进步的重要方向之一。本文从语音属性这一 ASR 研究的核心角度出发，分别使用基于单一语音属性的传统 ASR 和前沿的基于多语音属性的探测式 ASR 两种方法进行带多种方言口音的普通话语音识别的研究。

在基于单一语音属性的 ASR 方法中，我们首先提出了基于时间对准的音子识别算法以更准确地从数据中获得口音产生的发音变异。我们利用声学模型静态重构的方法提高了系统对 3 种方言口音鲁棒性。但静态重构会导致导致声学模型的精度下降，使系统在剪枝束搜索中性能受损。我们提出了动态高斯混合选择算法以提高模型精度，在搜索解码中为每个语音帧动态选择指定数目的高斯成分来定制观测密度。对由每个状态的高斯成分选择数构成的参数向量，我们提出使用基于最小化分类错误准则和遗传算法的离散区分性训练进行优化。由于动态高斯混合选择算法相当于在统计机器学习模型中引入了动态变化的决策边界，故算法对其它机器学习问题也具有应用价值。实验表明，区分性动态高斯混合选择算法可以提高系统在剪枝和非剪枝情况下的性能及对多种方言口音的鲁棒性。联合使用上述算法使 ASR 在川、粤、吴 3 种口音数据上比 MAP 自适应算法的音节准确率相对提高了 7.57%、3.69%和 4.19%，同时不影响系统对标准普通话的识别率。

在基于多语音属性的探测式 ASR 方法中，我们使用音子和发音特征作为语音属性。发音特征可以依据对应的发音器官分为 7 类，为每类分别构建上下文相关探测器，在发音器官层次解决语音中因协同发音和口音产生的上下文相关的发音变异。使用 CRF 整合探测器的输出，整合过程依赖于根据非线性音系学知识从数

据中获得的深层规律，这些深层规律反映了发音特征之间协同出现与互斥的关系。我们取消了 CRF 的转移特征，并结合使用了基于 HMM 和有限状态网络的线索验证器，解决了在 ASR 中应用 CRF 时经常遇到的音子欠生成问题，在提高系统性能的同时为更高层处理提供了方便。针对单一口音的探测式 ASR 在官话、粤、吴口音上音子准确率分别比 Monophone HMMs 系统相对提高了 5.04%、4.68%和 6.06%。通过分析和实验，我们证明了在 ASR 中可以利用发音特征描述并解决口音变异。我们还提出了语音属性区分单元和并行探测式 ASR。语音属性区分单元可以在语音属性层次解决发音变异，并且不需要修改系统中已有的任何模型。并行探测式 ASR 可以解决多种方言口音产生的发音变异，它的识别率与经典的 Triphone HMMs 系统相近，但识别速度快 3.44 倍。

本文虽然提出了一些新算法并取得了比较好的效果但仍然存在着很多不完善的地方。对于动态高斯混合选择算法，目前构造动态观测密度的策略是选择与当前语音帧距离最近的一些高斯成分。事实上对任何一个语音帧，在有 N 个高斯成分的观测密度中，共有 $\sum_{i=1}^{N} C_N^i = 2^N - 1$ 种可能的高斯成分选择方法；对所有的 M 个状态，在每一时刻，都存在 $\prod_{m=1}^{M} \sum_{n=1}^{N_m} C_{N_m}^n = \prod_{m=1}^{M} (2^{N_m} - 1)$ 种可能的动态观测密度组合，可能产生同样数目的动态决策边界。最近邻策略只能产生这些动态决策边界中的一小部分，算法仍有继续提升系统性能的很大可能性。继续提升系统性能需要更复杂的高斯成分选择策略，比如上下文相关的选择方式等。

对基于多语音属性的探测式 ASR，首先可以考虑换用更强大的探测器，如 HCRF 模型，它在具有比 HMM 更高准确率的同时还可以使用任意特征，方便引入额外的时域、频域特征以提高探测器性能；每类发音特征只含有很少的值，容易使用上下文相关的 HCRF。其次，可以音子融合器中加入神经网分数及更多的发音变异规则。第三，增加音节等更高层次的融合器，可以使用 SCRF 等段结构的模型以在音子层次处理音节结构等高层口音变异，容易利用音段特征并具有较低复杂度。第四，根据非线性音系学可以在发音器官层次为 ASR 引入声调信息。最后，可以将探测式 ASR 视为一种深层模型，对系统的各个部分进行联合优化。

参考文献

[1] Pinker S. The language instinct: how the mind creates language. New York: Perennial, 1995.

[2] Rabiner L, Juang B H. Fundamentals of speech recognition. 2nd ed. Englewood Cliffs: Prince Hall, 1999.

[3] 中国语言文字使用情况调查资料领导小组办公室. 中国语言文字使用情况调查资料. 北京: 语文出版社, 2006.

[4] Fung P, Liu Y. Effects and modeling of phonetic and acoustic confusions in accented speech. Journal of Acoustical Society of America, 2005, 118(4): 1-15.

[5] McKean E. The new Oxford American dictionary. 2nd ed. Oxford: Oxford University Press, 2005.

[6] DeFrancis J. The Chinese language: fact and fantasy. Hawaii: University of Hawaii Press, 1984.

[7] 黄景湖. 汉语方言学. 厦门: 厦门大学出版社, 1987.

[8] Kane D. The Chinese Language: its history and current usage. North Clarendon: Tuttle Publishing, 2006.

[9] 林焘, 王理嘉. 语音学教程. 北京: 北京大学出版社, 1992.

[10] 袁家骅. 汉语方言概要. 第 2 版. 北京: 语文出版社, 2001.

[11] 王洪君. 汉语非线性音系学. 北京大学出版社. 增订版. 北京: 北京大学出版社, 2008.

[12] Tsai M Y, Lee L S. Pronunciation variation analysis based on acoustic and phonemic distance measures with application examples on Mandarin Chinese. Proceedings of the 8th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Virgin Island, USA, 117-122. 2003.

[13] Liu Y, Fung P. Multi-accent Chinese speech recognition. Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pennsylvania, USA, 1-4, 2006.

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

[15] Ng T, Siu M, Ostendorf M. A quantitative assessment of the importance of tone in Mandarin speech recognition. IEEE Signal Processing Letters, 2005, 12(12): 867-870.

[16] Zhou J L, Tian Y, Shi Y, Huang T. Tone articulation modeling for Mandarin spontaneous speech recognition. Proceedings of the 30th International Conference on Audio, Speech, and Signal Processing (ICASSP). Quebec, Canada, 17-21, 2004.

[17] Jurafsky D, Martin J H. Speech and language processing: an introduction to natural language

processing, speech recognition, and computational linguistics. 2nd ed. New Jersey: Prentice Hall, 2009.

[18] Semhur. English: Pharynx Overview [EB/OL]. [2011-11-09]. http://commons.wikimedia.org/wiki/File: Pharynx_diagram-1-el.svg

[19] 胡壮麟. 语言学教程. 第 3 版. 北京: 北京大学出版社, 2008.

[20] Real-Time MRI

[21] Sagey E C. The representation of features and relations in non-linear phonology [Ph.D. Thesis]. Cambridge, MA: Department of Linguistics and Philosophy, MIT, 1986.

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

[23] Rabiner L. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 1989: 77(2): 257-286.

[24] Duda R O, Hart P E, Stork D G. Pattern classification. 2nd ed. New York: Wiley-Interscience, 2000.

[25] Ross S M. Introduction to probability models. 10th ed. Singapore: Elseveier, 2009.

[26] Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980: 28(4): 357-366.

[27] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. Journal of Acoustical Society of America, 1990, 87(4): 1738-1752.

[28] Oppenheim A V, Schafer R W. Discrete-Time Signal Processing. 3rd ed. New Jersey: Prentice Hall, 2009.

[29] Young S, Evermann G, Gales M, et al. The HTK book. 3.4 ed. Cambridge: Cambridge Research Laboratory, 2009.

[30] Lee C H, Juang B H, Soong F K, et al. Word recognition using whole word and subword models. Proceedings of the 14th International Conference on Audio, Speech, and Signal Processing (ICASSP). Glasgow, Scotland, 683-686, 1989.

[31] Bishop C M. Pattern recognition and machine learning. New York: Springer, 2007.

[32] Sorenson H W, Alspach D L. Recursive Bayesian estimation using Gaussian sums. Automatica, 1971: 7: 465-479.

[33] Balakrishnan V, Sivaram G S V S, Khudanpur Sanjeev. Dirichlet mixture models of neural net posteriors for HMM-based speech recognition. Proceedings of the 36th International Conference on Audio, Speech, and Signal Processing (ICASSP). Prague, Czech, 5028-5031, 2011.

[34] Juang B H, Levison S E, Sondhi M M. Maximum likelihood estimation for multivariate mixture observations of Markov Chains. IEEE Transactions on Information Theory, 1986: 32(2): 307-309.

[35] Frankel J, King S. A hybrid ANN/DBN approach to articulatory feature recognition.

Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Lisbon, Portugal, 3045-3048, 2005.

[36] Sutton C, McCallum A. An introduction to conditional random fields for relational learning. Cambridge: MIT Press, 2006.

[37] Gunawardana A, Mahajan M, Acero A, et al. Hidden conditional random fields for phone classification. Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Lisbon, Portugal, 1117-1120, 2005.

[38] Kuo H K J. Maximum entropy direct models for speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2006: 14(3): 873-881.

[39] Morris J, Fosler-Lussier E. Conditional random fields for integrating local discriminative classifiers. IEEE Transactions on Audio, Speech, and Language Processing, 2008: 16(3): 617-628.

[40] Oura K, Zen H, Nankaku Y, et al. Hidden semi-Markov model based speech recognition system using weighted finite-state transducer. Proceedings of the 31th International Conference on Audio, Speech, and Signal Processing (ICASSP). Toulouse, France, 33-36, 2006.

[41] Sarawagi S, Cohen W W. Semi-Markov conditional random fields for information extraction. Advances in Neural Information Processing Systems 17 (NIPS). Vancouver, Canada, 1185-1192, 2004.

[42] Zweig G, Nguyen P. A segmental CRF approach to large vocabulary continuous speech recognition. Proceedings of the 11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Merano, Italy, 152-157. 2009.

[43] Nguyen P, Compernolle V D, Demuynck K, et al. Speech recognition with segmental conditional random fields: a summary of the JHU CLSP 2010 Summer Workshop. Proceedings of the 36th International Conference on Audio, Speech, and Signal Processing (ICASSP). Prague, Czech, 5044-5047, 2011.

[44] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural network. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989: 37(3): 328-339.

[45] Siniscalchi S M, Svendsen T, Lee C H. Toward a detection-based universal phone recognizer. Proceedings of the 33th International Conference on Audio, Speech, and Signal Processing (ICASSP). Las Vegas, USA, 4261-4264, 2008.

[46] Hinton G E, Osindero, S, The Y. A fast learning algorithm for deep belief nets. Neural Computation, 2006: 18(7): 1527-1554.

[47] Dong Y, Deng L. Deep learning and its applications to signal and information processing. IEEE Signal Processing Magazine, 2011: January: 145-154.

[48] Mohamed A R, Dahl G E, Hinton G E. Deep belief networks for phone recognition. Proceedings of NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, Vancouver, Canada, 1-9, 2009.

[49] Seide F, Li G, Dong Y. Conversational speech transcription using context-dependent deep neural networks. Proceedings of the 12th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Florence, Italy, 437-440, 2011.

[50] Jelinek F. Up from trigrams! The struggle for improved language models. Proceedings of the 2nd European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Genoa, Italy, 1037-1040, 1991.

[51] Xu P Y, Khudanpur S, Gunawardana A. Randomized maximum entropy language models. Proceedings of the 12th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA, 226-230. 2011.

[52] Roark B, Saraclar M, Collins M, et al. Discriminative language modeling with conditional random fields and perceptron algorithm. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). Barcelona, Spain, 47-54, 2004.

[53] Miklov T, Deoras A, Povey D, et al. Strategies for training large scale neural network language models. Proceedings of the 12th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA, 196-201. 2011.

[54] Lowerre B T. The HARPY speech recognition system [Ph.D. Thesis]. Pittsburgh, PA: Department of Computer Science, CMU, 1976.

[55] Siniscalchi S M, Lee C H. A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Communication, 2009: 51: 1139-1153.

[56] Tsai M Y, Chou F C, Lee L S. Pronunciation variation analysis with respect to various linguistic levels for Mandarin Chinese. Proceedings of the 7th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Aalborg, Denmark, 1445-1448, 2001.

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

[58] Sloboda T, Waibel A. Dictionary learning for spontaneous speech recognition. Proceedings of the 4th International Conference on Spoken Language Processing (Interspeech-ICSLP). Philadelphia, USA, 2328-2331, 1996.

[59] Huang C, Chang E, Zhou J L, et al. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. Proceedings of the 6th International Conference on Spoken Language Processing (Interspeech-ICSLP). Beijing, China, 818-821, 2000.

[60] Tajchman G, Jurafsky D, Fosler-Lussier E. Learning phonological rule probabilities from speech corpora with exploratory computational phonology. Proceedings of the 33rd

Annual Meeting of the Association for Computational Linguistics (ACL). Cambridge, USA, 9-15, 1995.

[61] Li A, Zheng T F, Byrne W, et al. CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. Proceedings of the 6th International Conference on Spoken Language Processing (Interspeech-ICSLP). Beijing, China, 485-488, 2000.

[62] Liu Y, Fung P. Rule-based word pronunciation networks generation for Mandarin speech recognition. Proceedings of the 2nd International Symposium on Chinese Spoken Language Processing (ISCSLP). Beijing, China, 35-38, 2000.

[63] Fung P, Byrne W, Zheng T F, et al. Pronunciation modeling of Mandarin casual speech. Technical Report at John Hopkins University Summer Research Workshop. Baltimore, USA, 1-45, 2000.

[64] Chen Y J, Wu C H, Chiu Y H, et al. Generation of robust phonetic set and decision tree for Mandarin using chi-square testing. Speech Communication. 2002: 38(3): 349-364.

[65] Ding G H. Phonetic confusion analysis and robust phone set generation for Shanghai-accented Mandarin speech recognition. Proceedings of the 10th International Conference on Spoken Language Processing (Interspeech-ICSLP). Brisbane, Australia, 1129-1132, 2008.

[66] Bacchiani M, Ostendorf M. Joint lexicon, acoustic unit inventory and model design. Speech Communication. 1999: 29: 99-114.

[67] Hain T, Woodland P C. Dynamic HMM selection for continuous speech recognition. Proceedings of the 6th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Budapest, Hungary, 1327-1330, 1999.

[68] Fosler-Lussier E. A tutorial on pronunciation modeling for large vocabulary speech recognition. // Renals S, Grefenstette G, eds. Text and Speech Triggered Information Access. Berlin, Germany: Springer Verlag, 2003: 38-77.

[69] Woodland P C. Speaker adaptation for continuous density HMMs: a review. Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition. Sophia Antipolis, France, 11-19, 2001.

[70] Humphries J J, Woodland P C, Pearce D. Using accent-specific pronunciation modeling for robust speech recognition. Proceedings of the 4th International Conference on Spoken Language Processing (Interspeech-ICSLP). Philadelphia, USA, 2324-2327, 1996.

[71] Oh Y R, Kim H K. MLLR/MAP adaptation using pronunciation variation for non-native speech recognition. Proceedings of the 11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Merano, Italy, 216-221. 2009.

[72] Brown P F, Lee C H, Spohrer J C. Bayesian adaptation in speech recognition. Proceedings of the 8th International Conference on Audio, Speech, and Signal Processing (ICASSP). Boston, USA, 761-764, 1983.

[73] Kuhn R, Junqua J C, Nugyen P. Rapid speaker adaptation in eigenvoice space. IEEE Transactions on Speech and Audio Processing, 2000: 8(6): 695-707.

[74] Gunawardana A, Byrne W. Discriminative speaker adaptation with conditional maximum likelihood linear regression. Proceedings of the 7th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Aalborg, Denmark, 1203-1206, 2001.

[75] Saraclar M, Nock H, Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language. 2000: 13(4): 137-160.

[76] Byrne W, Venkataramani V, Kamm T, et al. Automatic generation of pronunciation lexicons for Mandarin spontaneous speech. Proceedings of the 26th International Conference on Audio, Speech, and Signal Processing (ICASSP). Salt Lake City, USA, 569-572, 2001.

[77] Riley M, Ljolje. Automatic generation of detailed pronunciation lexicons. // Lee C H, Soong F K, Paliwal K K, eds. Automatic Speech and Speaker Recognition: Advanced Topics. London: Kluwer Academic Publishers, 1-17, 1996.

[78] Liu L Q, Zheng T F, Akabane M, et al. Using a small development set to build a robust dialectal Chinese speech recognizer. Proceedings of the 10th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Antwerp, Belgium, 1729-1732, 2007.

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

[80] Fung P, Liu Y. Triphone model reconstruction for Mandarin variations. Proceedings of the 27th International Conference on Audio, Speech, and Signal Processing (ICASSP). Hong Kong, China, 760-763, 2003.

[81] Nakamura A. Restructuring Gaussian mixture density functions in speaker-independent acoustic models. Speech Communication. 2002: 36(3): 277-289.

[82] Li Y, Fung P, Xu P. Asymmetric acoustic modeling of mixed language speech. Proceedings of the 36th International Conference on Audio, Speech, and Signal Processing (ICASSP). Prague, Czech, 5004-5007, 2011.

[83] Juang B H, Wu C, Lee C H. Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Audio Processing. 1997: 5(3): 257-265.

[84] Povey D, Woodland P C. Minimum phone error and I-smoothing for improved discriminative training. Proceedings of the 26th International Conference on Audio, Speech, and Signal Processing (ICASSP). Orlando, USA, 105-108, 2002.

[85] Fu Q, Zhao Y, Juang B H. Automatic speech recognition based on non-uniform error criteria. IEEE Transactions on Audio, Speech, and Language Processing. 2012: 20(3): 780-793.

[86] Zheng Y L, Sproat R, Gu L, et al. Accent detection and speech recognition for Shanghai-accented Mandarin. Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Lisbon, Portugal, 217-220, 2005.

[87] Chen T, Huang C, Chang E, et al. Automatic accent identification using Gaussian mixture models. Proceedings of the 7th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Madonna di Campiglio, Italy, 343-346. 2001.

[88] Lee C H, Clements M A, Dusan S, et al. An overview on automatic speech attribute transcription (ASAT). Proceedings of the 10th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Antwerp, Belgium, 1825-1828, 2007.

[89] Levner I, Bulitko V, Lin G. Existing feature extraction and classification methods. // Guyon I, Gunn S, Nikravesh M, et al., eds. Feature Extraction, Foundations and Applications, New York: Springer, 2006: 265–296.

[90] Zhu Q F, Chen B Y, Grezl F, et al. Improved MLP structures for data-driven feature extraction. Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Lisbon, Portugal, 2129-2132, 2005.

[91] Hirschberg J. "Everytime I fire a linguist, my performance goes up," and other myths of the statistical natural language processing revolution. Invited talk on the 15th National Conference on Artificial Intelligence (AAAI), Madison, USA, 1998.

[92] Meyer B T, Brand T, Kollmeier B. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. Journal of Acoustical Society of America, 2011, 129(1): 388-403.

[93] Lee C H. From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition. Invited talk on the 8th International Conference on Spoken Language Processing (Interspeech-ICSLP). Jeju Island, Korea, 2004.

[94] Ohde R N, German S R. Formant onsets and formant transitions as developmental cues to vowel perception. Journal of Acoustical Society of America, 2011, 130(3): 1628-1642.

[95] Ma J K Y, Ciocca V, Whitehill T L. The perception of intonation questions and statements in Cantonese. Journal of Acoustical Society of America, 2011, 129(2): 1012-1023.

[96] Fogerty D. Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure. Journal of Acoustical Society of America, 2011, 129(2): 977-988.

[97] Shannon R V, Zeng F G, Kamath V, et al. Speech recognition with primarily temporal cues. Science. 1995: 270(5234), 303-304.

[98] Lin C Y, Wang H C. Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection. Journal of Acoustical Society of America, 2011, 130(1): 514-525.

[99] Ostendorf M. Moving beyond the "beads-on-a-string" model of speech," Proceedings of the 6th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Keystone, USA, 79-84. 1999.

[100] Stuuker S, Metze F, Schultz T, et al. Integrating multilingual articulatory features into speech recognition. Proceedings of the 8th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Geneva, Switzerland, 1033-1036, 2003.

[101] Lin C Y, Wang H C. Attribute-based Mandarin speech recognition using conditional random

fields. Proceedings of the 10th European Conference on Speech Communication and Technology (Interspeech-Eurospeech). Antwerp, Belgium, 1833-1836, 2007.

[102] Hermansky H, Ellis D P W, Sharma S. Tandem connectionist feature extraction for conventional HMM systems. Proceedings of the 25th International Conference on Audio, Speech, and Signal Processing (ICASSP). Istanbul, Turkey, 1635-1638, 2000.

[103] Morris J, Fosler-Lussier E. Conditional random fields for integrating local discriminative classifiers. IEEE Transactions on Audio, Speech, and Language Processing. 2008: 16(3): 617-628.

[104] Yu D, Siniscalchi S, Lee C H. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. Proceedings of the 37th International Conference on Audio, Speech, and Signal Processing (ICASSP). Kyoto, Japan, 4169-4172, 2012.

[105] Wand M, Schultz T. Analysis of phone confusion in EMG-based speech recognition. Proceedings of the 36th International Conference on Audio, Speech, and Signal Processing (ICASSP). Prague, Czech, 757-760, 2011.

[106] Kirchhoff K. Combining articulatory and acoustic information for speech recognition in reverberant environments. Proceedings of the 5th International Conference on Spoken Language Processing (Interspeech-ICSLP). Sydney, Australia, 1-4, 1998.

[107] Chomsky N. Syntactic structures. 2nd ed. Berlin: Walter de Gruyter, 2002.

[108] Shen W, Chen N, Reynolds D A. Dialect recognition using adapted phonetic models. Proceedings of the 10th International Conference on Spoken Language Processing (Interspeech-ICSLP). Brisbane, Australia, 763-766, 2008.

[109] Zhang C, Liu Y, Lee C H. Detection-based accented speech recognition using articulatory features. Proceedings of the 12th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA, 500-505, 2011.

[110] Glass J R. A probabilistic framework for segment-based speech recognition. Computer Speech and Language. 2003: 17: 137-152.

[111] Chelba C, Jelinek F. Structured language modeling. Computer Speech and Language. 2000: 14: 283-332.

[112] 维柯. 新科学. 朱光潜, 译. 合肥：安徽教育出版社. 2006.

[113] Candes E J, Romberg, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory. 2006: 52(2): 489-509.

[114] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009: 31(2): 210-227.

[115] Gemmeke J F, Van hamme H. An hierarchical exemplar-based sparse model of speech, with and application to ASR. Proceedings of the 12th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA, 101-106, 2011.

[116] ChineseLDC.org. RASC863 – 四大方言普通话语音语料库[EB/OL]. [2009-12-03].

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

[117] ChineseLDC.org. RASC863-G2 – 863 地方普通话语音语料库（第 2 批，6 地语音库）[EB/OL]. [2010-10-09]. http://www.chineseldc.org/doc/CLDC-SPC-2007-002/intro.htm.

[118] Zhang Q Q, Pan J L, Chan S D, et al. Nonnative speech recognition based on bilingual model modification at state level. Advances in Intelligent and Soft Computing, 2009: 56: 299-309.

[119] Zhang C, Liu Y, Xia Y Q, et al. Discriminative dynamic Gaussian mixture selection with enhanced robustness and performance for multi-accent speech recognition. Proceedings of the 37th International Conference on Audio, Speech, and Signal Processing (ICASSP). Kyoto, Japan, 4749-4752, 2012.

[120] Young S J, Odell J J, Woodland P C. Tree-based state tying for high accuracy acoustic modeling. Proceedings of the ARPA Workshop on Human Language Technology. Plainsboro, USA, 307-312, 1994.

[121] Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chain. IEEE Transactions on Speech and Audio Processing. 1994: 2: 291-298.

[122] Liu Y, Fung P. Pronunciation modeling for spontaneous Mandarin speech recognition. International Journal of Speech Technology. 2004: 7: 155-172.

[123] McDermott E, Hazen T J. Minimum classification error training of landmark models for real-time continuous speech recognition. Proceedings of the 28th International Conference on Audio, Speech, and Signal Processing (ICASSP). Montreal, Canada, 937-940, 2004.

[124] 达尔文. 物种起源. 舒德干, 译. 北京：北京大学出版社. 2005.

[125] Mitchell M. An introduction to genetic algorithm. Cambridge: MIT Press. 1999.

[126] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (ICML). Williamstown, USA, 282-289, 2001.

[127] Clifford P. Markov random fields in statistics. // Grimmett G R, Welsh D J A, eds. Disorder in Physical System. A Volume in Honour of John M. Hammersley. Oxford，UK: Oxford University Press. 1990: 19-32.

[128] Sung Y H, Jurafsky D. Hidden conditional random fields for phone recognition. Proceedings of the 11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Kyoto, Japan, 107-112, 2009.

[129] Vapnik V N. Statistical learning theory. New York: Wiley-Interscience, 1998.

[130] Clarkson P, Moreno P J. On the use of support vector machines for phonetic classification. Proceedings of the 24th International Conference on Audio, Speech, and Signal Processing (ICASSP). Phoenix, USA, 585-588, 1999.

[131] Zhang S X, Ragni A, Gales M J F. Structured log linear models for noise robust speech recognition. IEEE Signal Processing Letters. 2010: 17(11): 945-948.

[132] Woodland P C, Povey D. Large scale discriminative training of hidden Markov models for

speech recognition. Computer Speech and Language. 2002: 16: 25-47.

[133] Schluuter R, Macherey W, Muuller B, et al. Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication. 2001: 34: 287-310.

[134] Young S. A review on large-vocabulary continuous –speech recognition. IEEE Signal Processing Magazine. 1996: 13(5): 45-57.

[135] 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.

[136] Wu Y, Gohuler. Map of Sinitic Dialect (Simplified Chinese) [EB/OL]. [2011-05-10]. http://commons.wikimedia.org/wiki/File:Map_of_sinitic_dialect(Simplified_Chinese).svg.

[137] Chittka L, Brockmann. Anatomy of the Human Ear [EB/OL]. [2009-02-15]. http://en.wikipedia.org/wiki/File:Anatomy_of_the_Human_Ear.svg.

致 谢

感谢导师刘轶副教授对我的悉心指导，刘老师平易近人，工作忘我，学术水平高，就读硕士的三年以来刘老师的言传身教使我受益匪浅。感谢郑方教授、夏云庆副教授、徐明星副教授给予我的指点和帮助，各位老师在科研和生活中多次为我指点迷津，我从中受惠良多。

感谢美国佐治亚理工学院的李锦辉教授，李老师孜孜不倦的追求真理，对待学术非常严谨，待人宽厚慈祥，对工作高度负责。和李老师一起修改论文的那个下午将永远是我学术生涯中最难忘的经历之一。

感谢杨士强老师，王家歆老师，刘卫东老师，王诚老师在我本科期间对我关心和照顾。在我学业压力最大的时候是几位老师和计 55 班一些同学的帮助才使我走出阴霾，对此我没齿难忘。

感谢黄石磊博士教我语音识别的基础，感谢刘建博士在许多个夜晚陪我熬夜进行学术讨论，感谢王璇同学传授我语音学知识，这些都对我的学业有很大帮助。

感谢我的女友王琳琳，没有她的耐心、善良和支持我很难完成这篇论文。

感谢我的父母和家人，你们给了我一切。

在即将告别生活了 7 年的清华园之际，再次衷心感谢在这个园子里关心和帮助过我的各位，衷心感谢美丽可爱钟灵毓秀的清华园，感谢我的青葱岁月。

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1 HMM的训练算法

对于 1 个 HMM，需要通过训练数据获得其各状态间的转移概率、除入口和出口状态外每个状态的观测密度中所有高斯成分的权值、均值向量和协方差矩阵。需要说明的是由于倒谱系数的各维相互独立，故协方差都为对角阵[25][29]。

假设输入观测序列 $O=o_1o_2\ldots o_T$ 由 H 个 HMM 构成的序列生成，带括号的上标 h 表示 HMM 在序列中序号，N_h 为序列中第 h 个 HMM 的状态数。一个语句可以视为相关 HMM 的所有状态构成的序列，从而可以类似 1 个 HMM 时使用前向后向算法进行训练，α 和 β 分别称为前向因子和后向因子。

对于前向因子，$t=1$ 时 α 的初始值为

$$\alpha_1^{(h)}(1)=\begin{cases}1 & h=1\\ \alpha_1^{(h-1)}(1)a_{1,N_{h-1}}^{(h-1)} & \text{otherwise}\end{cases}$$

$$\alpha_j^{(h)}(1)=a_{1,j}^{(h)}b_j^{(h)}(o_1). \tag{1}$$

$$\alpha_{N_h}^{(h)}(1)=\sum_{i=2}^{N_h-1}\alpha_i^{(h)}(1)a_{i,N_h}^{(h)}$$

当 $t>1$ 时，有

$$\alpha_1^{(h)}(t)=\begin{cases}0 & h=1\\ \alpha_{N_{h-1}}^{(h-1)}(t-1)+\alpha_1^{(h-1)}(t)a_{1,N_{h-1}}^{(h-1)} & \text{otherwise}\end{cases}$$

$$\alpha_j^{(h)}(t)=\left[\alpha_1^{(h)}(t)a_{1,j}^{(h)}+\sum_{i=2}^{N_h-1}\alpha_i^{(h)}(t-1)a_{i,j}^{(h)}\right]b_j^{(h)}(o_t) \quad . \tag{2}$$

$$\alpha_{N_h}^{(h)}(t)=\sum_{i=2}^{N_h-1}\alpha_i^{(h)}(t)a_{i,N_h}^{(h)}$$

对于前向因子，$t=T$ 时 β 的初始值为

$$\beta_{N_h}^{(h)}(T)=\begin{cases}1 & h=H\\ \beta_{N_{h+1}}^{(h+1)}(T)(1)a_{1,N_{h+1}}^{(h+1)} & \text{otherwise}\end{cases}$$

$$\beta_i^{(h)}(1)=a_{1,N_h}^{(h)}\beta_{N_h}^{(h)}(T). \tag{3}$$

$$\beta_1^{(h)}(T)=\sum_{j=2}^{N_h-1}a_{1,j}^{(h)}b_j^{(h)}(o_T)\beta_j^{(h)}(T)$$

当$t<T$时，有

$$\beta_{N_h}^{(h)}(t)=\begin{cases}0 & h=H\\ \beta_1^{(h+1)}(t+1)+\beta_{N_{h+1}}^{(h+1)}(t)a_{1,N_{h+1}}^{(h+1)} & \text{otherwise}\end{cases}$$

$$\beta_i^{(h)}(t)=a_{i,N_h}^{(h)}\beta_{N_h}^{(h)}(t)+\sum_{j=2}^{N_h-1}a_{i,j}^{(h)}b_j^{(h)}(o_{t+1})\beta_j^{(h)}(t+1)\,. \tag{4}$$

$$\beta_1^{(h)}(t)=\sum_{j=2}^{N_h-1}a_{1,j}^{(h)}b_j^{(h)}(o_t)\beta_j^{(h)}(t)$$

通过前向因子或后向因子的计算都可以得到生成观测序列的似然度

$$P(O\mid M)=\alpha_{N_H}(T)=\beta_1(1). \tag{5}$$

设共有R条训练语音，第r个观测序列为$O^r=o_1o_2\ldots o_{T_r}$，记它的似然度为$P(O^r\mid M)=P_r$，则从公式(2-1)~(2-5)可以解出状态间的转移概率为

$$\hat{a}_{i,j}^{(h)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(h)r}(t)a_{i,j}^{(h)}b_j^h(o_{t+1}^r)\beta_j^{(h)r}(t+1)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(h)r}(t)\beta_i^{(h)r}(t)}$$

$$\hat{a}_{i,j}^{(h)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_1^{(h)r}(t)a_{1,j}^{(h)}b_j^h(o_t^r)\beta_j^{(h)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_1^{(h)r}(t)\beta_1^{(h)r}(t)+\alpha_1^{(h)r}(t)a_{1,N_h}^{(h)}\beta_1^{(h+1)r}(t)}\,. \tag{6}$$

$$\hat{a}_{i,N_h}^{(h)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(h)r}(t)a_{i,N_h}^{(h)}\beta_{N_h}^{(h)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(h)r}(t)\beta_i^{(h)r}(t)}$$

作为状态观测密度的 GMM 模型需要使用 EM 算法求解。EM 算法的核心思想是通过对每个变量分别求导并让导数为 0 来求得期望的最大值。类似于其它优化

方法，EM 算法也使用迭代策略。第 j 个状态的观测密度中第 m 个高斯成分的参数的计算公式为

$$\hat{\mu}_{jm}^{(h)}=\frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{(h)r}(t)o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{(h)r}(t)}$$

$$\hat{\Sigma}_{jm}^{(h)}=\frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{(h)r}(t)(o_t^r-\hat{\mu}_{jm}^{(h)})(o_t^r-\hat{\mu}_{jm}^{(h)})^{\mathrm{T}}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{(h)r}(t)}. \tag{7}$$

$$c_{jm}^{(h)}=\frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{(h)r}(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{m=1}^{M}L_{jm}^{(h)r}(t)}$$

其中 $L_{jm}^{r}(t)$ 为高斯成分的占有率（Occupancy），有

$$L_{jm}^{(h)r}(t)=\frac{1}{P_r}U_j^{(h)r}(t)c_{jm}b_{jm}(o_t^r)\beta_j^{(h)r}(t). \tag{8}$$

$$U_j^{(h)r}(t)=\begin{cases}\alpha_1^{(h)r}(t)a_{1j}^{(h)} & t=1\\ \alpha_1^{(h)r}(t)a_{1j}^{(h)}+\sum_{i=2}^{N_{h-1}}\alpha_i^{(h)r}(t-1)a_{ij}^{(h)} & \text{otherwise}\end{cases}.$$

对于迭代优化的算法，一般需要给变量赋予初值以保证算法快速收敛到较好的解。初始时 HMM 中每个状态的观测密度中都只有 1 个高斯成分，记 $N=\sum_{r=1}^{R}T_r$ 为训练集的总帧数，所有均值和方差的初值都相同，为

$$\hat{\mu}=\frac{1}{T}\sum_{t=1}^{T}o_n \tag{9}$$

$$\hat{\Sigma}=\frac{1}{N}\sum_{n=1}^{N}(o_n-\mu)(o_n-\mu)^T.$$

附录 2 汉语声韵母与发音特征对应关系表

表 1 汉语辅音与发音特征的对应关系

			双唇音	唇齿音	舌尖前音	舌尖中音	舌尖后音	舌面前音	舌面后音
塞音	不送气	清音	‘b’			‘d’			‘g’
	送气		‘p’			‘t’			‘k’
塞擦音	不送气				‘z’		‘zh’	‘j’	
	送气				‘c’		‘ch’	‘q’	
擦音	均可			‘f’	‘s’		‘sh’	‘x’	‘h’
		浊音					‘r’		
鼻音			‘m’			‘n’			/ng/
边音						‘l’			

表 2 汉语韵母的分类

类型	韵母	韵首	韵腹	韵尾
单元音	‘a’	‘a’		
	‘o’	‘o’		
	‘e’	‘e’		
	‘i’	‘i’		
	‘u’	‘u’		
	‘v’	‘v’		
	‘i1’	‘i1’		
	‘i2’	‘i2’		
	‘er’	‘er’		
二合元音	‘ai’	‘a’		
	‘ao’	‘a’		

	‘ia’	‘i’	‘a’	
	‘ie’	‘i’	‘e’	
	‘ou’	‘o’	‘u’	
	‘ua’	‘u’	‘a’	
	‘uo’	‘u’	‘o’	
	‘ve’	‘v’	‘e’	
鼻韵母	‘an’	‘a’		‘n’
	‘ang’	‘a’		/ng/
	‘en’	‘e’		‘n’
	‘eng’	‘e’		/ng/
	‘ian’	‘i’	‘a’	‘n’
	‘iang’	‘i’	‘a’	/ng/
	‘in’	‘i’		‘n’
	‘ing’	‘i’		/ng/
	‘iong’	‘i’	‘o’	/ng/
	‘ong’	‘o’		/ng/
	‘uan’	‘u’	‘a’	‘n’
	‘uang’	‘u’	‘a’	/ng/
	‘un’	‘u’		‘n’
	‘van’	‘v’	‘a’	‘n’
	‘vn’	‘v’		‘n’
三合元音	‘iao’	‘i’	‘a’	‘o’
	‘iou’	‘i’	‘o’	‘u’
	‘uai’	‘u’	‘a’	‘i’
	‘uei’	‘u’	‘e’	‘i’

表 3 汉语元音与发音特征的对应关系

单元音	舌位	舌位高低	舌位前后	唇形
‘a’	舌面前音	低	央	非圆唇
‘o’	舌面前音	半高	后	圆唇
‘e’	舌面前音	半高	后	非圆唇
‘ei’	舌面前音	半低	前	非圆唇
‘i’	舌面前音	高	前	非圆唇
‘u’	舌面前音	高	后	圆唇
‘v’	舌面前音	高	前	圆唇
‘i1’	舌尖前音	高	前	非圆唇
‘i2’	舌尖后音	高	央	非圆唇
‘er’	卷舌音	半高	央	非圆唇

个人简历、在学期间发表的学术论文与研究成果

个人简历

1986 年 7 月 24 日出生于内蒙古呼和浩特市。

2005 年 8 月考入清华大学大学计算机科学与技术系计算机科学与技术专业，2009 年 7 月本科毕业并获得工学学士学位。

2009 年 9 月免试进入清华大学计算机科学与技术系攻读计算机语音技术专业学硕士至今。

发表的学术论文

[1] 张超, 刘轶, 郑方. 面向多口音语音识别的声学模型重构. 清华大学学报(自然科学版), 2011, 51: 1161-1166. (EI 收录, 检索号:20114514493895)

[2] Zhang C, Liu Y, Xia Y Q, Lee C H. Discriminative dynamic Gaussian mixture selection with enhanced robustness and performance for multi-accent speech recognition. Proceedings of the 37th International Conference on Audio, Speech, and Signal Processing (ICASSP). Kyoto, Japan, 4749-4752, 2012. (EI 待检索)

[3] Zhang C, Liu Y, Lee C H. Detection-based accented speech recognition using articulatory features. Proceedings of the 12th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA, 500-505, 2011. (EI 收录, 检索号:20121314909218)

[4] Zhang C, Liu Y, Xia Y Q, Zheng F, et al. Reliable accent specific unit generation with dynamic Gaussian mixture selection for multi-accent speech recognition. Proceedings of the 18th IEEE Conference on Multimedia and Expo (ICME). Barcelona, Spain, 1-6, 2011. (EI 收录, 检索号:20114514487600)

[5] Zhang C, Liu Y, Zheng F. Asymmetric acoustic model for accented speech recognition. Proceedings of the 3rd APSIPA Annual Summit and Conference (ASC). Xian, China, 1-5, 2011. (EI 待检索)

[6] Hou J, Liu Y, Zhang C, Huang S L. An in-car Chinese noise database for speech recognition. Proceedings of the 4th International Conference on Asian Language Processing (IALP). Penang, Malaysia, 1-4, 2011. (EI 收录, 检索号:20120414712101)