# Research on Pronunciation Modeling for Spontaneous Chinese Speech Recognition

Dissertation submitted to

**Tsinghua University**

in partial fulfillment of the requirements

for the degree of

**Doctor of Engineering**

by

Zhanjiang SONG

(Computer Application)

Dissertation Supervisors:

Prof. Wenhu WU and Associate Prof. Fang ZHENG

April, 2001

# 摘　要

相对于朗读式发音而言，自然、随意发音的语音在声学层面上具有发音变化频繁、语速不均匀、协同发音严重等特点，这大大增加了对其进行正确识别的难度。“发音建模”的目标，就是要在传统的连续语音识别技术的基础上，进一步研究更能体现自然语音发音变化特性的建模方法。对于英语语言，这方面的研究已经取得了一些可喜的成果。

本文针对汉语，研究焦点为自然语音在纯声学层面上的发音建模问题，从相互作用且密切相关的识别基元集、发音词典和声学模型等角度出发，提出了如下的方法和策略。

**第一，汉语声韵集的推广及其精细建模。**对标准的汉语声韵基元集合进行推广，以更加全面地覆盖汉语中可能出现的音变现象，并借鉴模型自适应技术的思想，提出了基于广义声韵集的精细建模方法。与基准方法相比，该方法可更好地反映自然语音中的细微差异，并为在训练数据稀疏情况下进行发音建模提供了新思路。

**第二，基于上下文的发音词典加权策略。**从减少发音词典固有的音节间混淆度的角度出发，定义了对其进行定量评价的指标，并在此基础上提出基于上下文的发音词典加权策略。该方法在保证有效降低音节间混淆度的同时，提高了对发音词典加权值进行估计的精度，并因而改善了整体的声学识别性能。

**第三，基于声韵集和广义声韵集的参数共享方法。**该方法在采用具有最低音节混淆度的单发音词典的同时，通过对 HMM (Hidden Markov Model) 状态内的高斯混合参数进行共享或组合，保证了在单一基元内部最大限度地包容自然语音中最关键的发音变化现象，使整体声学识别性能得到较大提高。

本文在自然语音发音建模的若干方面进行的研究，虽然是基于汉语的，但是这些的思路与方法可以很容易地推广到对其它语言进行发音建模的任务中。

**关键词**：发音建模、广义声韵集、精细建模、词典加权策略、参数共享

# Abstract

Compared with read speech, spontaneous or casual speech has some prominent acoustical characteristics, such as rich pronunciation variants, instable uttering speeds, and notable co-articulations. Apparently, this will greatly increase the difficulties to recognize speech correctly. The main target of *Pronunciation Modeling* is just about to better model pronunciation variants based on traditional continuous speech recognition technologies. For casual English speech, encouraging progress has been made for its pronunciation modeling.

Regarding Chinese speech, this dissertation focuses on pronunciation modeling methods for spontaneous speech at the pure acoustic level and following methods or strategies are proposed on the basis of tightly related speech recognition units, pronunciation lexicons, and acoustic models.

**One, a Chinese Generalized Initial/Final (GIF) set with its refined modeling.** The standard Initial/Final (IF) set is generalized to cover more important pronunciation variants in spontaneous Chinese, while a refined GIF modeling method is proposed based on model adaptation techniques. Compared with the baseline methods, GIF with refined modeling can better reflect elaborate variations for spontaneous speech and weaken the passive effects of training data sparseness.

**Two, a Context-Dependent Weighting (CDW) strategy for pronunciation lexicons.** Originated from intrinsic syllable confusion of pronunciation lexicons, a quantitative measure is defined and a context-dependent lexicon weighting strategy is proposed. This strategy can decrease the syllable confusion degree for a lexicon and improve the precision for the lexicon weighting estimation, thus the overall performance of the recognizer is improved.

**Three, the IF and GIF sets based parameter sharing.** While the standard lexicon with lowest intrinsic syllable confusion being adopted, the inner-state Gaussians of hidden Markov models (HMM) are shared or combined together. This guarantees that the important pronunciation variants of spontaneous speech can be included into model parameters to contribute to the satisfying overall recognition performance despite no utterance variation is given in the lexicon.

Although the approaches proposed in this dissertation are based on Chinese, it is easy to apply these methods or strategies to the pronunciation modeling of other languages.



**Key words**: Pronunciation Modeling, Generalized Initial/Final, Refined Modeling, Lexicon Weighting Strategy, Parameter Sharing

# 目　　录

# 插图索引

# 表格索引

# 主要符号和缩略语表

| 符号或缩写 | 含义及其解释说明 |
|---|---|
| HMM | Hidden Markov Model. 隐马尔可夫模型 |
| CASS | Chinese Annotated Spontaneous Speech database.<br>为进行自然语音发音建模研究而构建的一个汉语口语语音库 |
| IPA | International Phonetic Alphabets. 国际音标(符号集合) |
| SAMPA-C | Speech Assessment Method Phonetic Alphabets (for Chinese).<br>一种由 IPA 导出的、用于汉语口语标注的音标(机内码)集合 |
| IF | Initial/Final. (标准的)声韵母(或其集合) |
| GIF | Generalized Initial/Final. 广义声韵母(或其集合) |
| IF-GIF | 基于 GIF 的精细(Refined)模型 (对 IF 和 GIF 进行细化的基元) |
| S | Syllable. (标准的)音节(或其集合)，包含若干 IF |
| GS | Generalized Syllable. 广义音节(或其集合)，包含若干 GIF |
| CI | Context-Independent. 上下文无关的 (如 CI-IF, CI-GIF 等) |
| CD | Context-Dependent. 上下文相关的 (如 CD-IF, CD-GIF 等) |
| PLIC | Pronunciation Lexicon Intrinsic Confusion.<br>发音词典固有的(音节间的)混淆度 |
| DOP | Direct Output Probability. (基元发音序列的)直接输出概率 |
| CDW | Context-Dependent Weight(ing). 上下文相关权重(加权) |
| $H_b$，$H_s$ | HMM 参数集合 (分别对应 Base-form 基元和 Surface-form 基元) |
| $\hat{S}_b$，$\hat{S}_s$ | 最优的状态序列 (分别对应 Base-form 模型和 Surface-form 模型) |
| $Q_{ij}$ | 两个 HMM 状态 $b_i$ 和 $s_j$ 之间的相似度 (混叠概率) |
| SCR, UCR，<br>SER, SER↓ | 用来衡量声学识别性能的指标。分别为：音节识别正确率、基元识别正确率、音节识别错误率、音节误识率相对下降百分比 |

# 第一章 绪 论

## 1.1 引 言

语音技术，包括语音识别、语音合成、关键词检出、说话人识别与确认、口语对话系统等，是现代人机交互的重要方式之一，具有广泛的应用前景。其中语音识别技术，尤其是连续语音识别技术，是最基础、最重要的部分，而且已经逐步走向成熟与实用。

到目前为止，基于朗读式发音的连续语音识别技术已经可以达到比较好的识别性能，然而对于自然、随意发音的连续语音（例如口语和自由对话）来说，由于其本身发音过程的随意性，使得识别性能与实用化目标之间，仍有非常大的差距。

本文以汉语为例，针对自然随意的连续语音，从分析其发音变化规律入手，尝试从纯声学的角度对其建模方法进行研究，并提出若干新的思路与方法，以期提高对自然连续语音的整体识别性能。

## 1.2 语音识别：人机交互的重要方式

### 1.2.1 语音识别的重要意义

语言是人类进行交流的最为直接、最为自然的方式，而语音则和文字一样，是语言的重要载体。近几十年来计算机技术的飞速发展，使得人类社会发生了有史以来最空前的变化，计算机也早已成为我们日常工作和生活中必不可少的组成部分。

但是，人与计算机之间的交互水平却严重地滞后于计算机其他技术的发展，远远不能满足人类的需求。因此，如何提高人机交互的友好程度，是一个有着重要意义的研究课题。

语音处理技术的重点是要解决计算机的“听”和“说”的能力，即：如何使计算机理解人类给出的语音信息，以及如何让计算机产生人类满意的语音信息。这包括了语音合成、语音识别、关键词检出、说话人识别与确认、口语对话系统等方面的研究内容。

自动语音识别 (ASR，Automatic Speech Recognition) 技术的目的，正是要让计算机能够“听懂”人的说话。它是发展人机语音通信和新一代智能计算机的主要组成部分，也是目前业界普遍认为很有前途的一项技术。随着计算机处理能力和存储能力的不断增强，一些非常复杂的语音识别算法能够得以实现，语音识别器的性能也得以不断提高，这就为更加自然方便的人机交互水平以及更加广阔的语音技术应用前景提供了可能。

### 1.2.2 语音识别的历史发展

对自动语音识别的研究，可以追溯到本世纪 50 年代。1952 年美国的 Davis 等人研究成功了世界上第一个识别 10 个英文数字发音的实验系统[Davis52]。我国在 50 年代后期，也曾研制出一套“自动语音识别器”，用来识别汉语的十个元音。1960 年，Denes 等人研究成功了第一个计算机语音识别系统[Denes60]，从此开始了计算机语音识别的阶段。进入 70 年代之后，语音识别，尤其是小词汇量、特定人、孤立词的识别方面，取得了许多实质性的进展，象线性预测分析技术、动态时间规整算法、矢量量化技术等。

七十年代后期，语音识别开始沿着三个不同的方向来扩展其研究领域：特定人向非特定人扩展；孤立词向连接词扩展；小词汇量向大词汇量扩展。在具体系统中，采用了更加复杂的聚类算法，同时也产生了新的基于动态规划的匹配算法等。

八十年代中期以来，新技术的不断出现使语音识别有了实质性的进展。特别是隐马尔可夫模型 (HMM) 的广泛应用和研究，推动了语音识别的迅速发展，使得语音识别的任务得以由连接词向连续语音扩展，并陆续出现了许多基于 HMM 模型的语音识别系统。其中美国 CMU 的 *SPHINX* 系统被认为是 80 年代末 90 年代初的典型代表[Lee89a, Huang93]，该系统在英语的大词汇量、非特定人连续语音识别方面能够达到 97％的识别率。诸如此类的实际系统还有 DRAGON

公司的 *Dragon Dictate* 系统[Dragon]等。但这一阶段的连续语音研究，还主要集中于匀速、受限的、基于标准口音的朗读式发音。

进入九十年代，语音识别技术更是展现出其蓬勃生机和广阔的发展前景。首先，基于关键词识别及朗读式连续语音识别的语音应用系统已经逐步走向市场，并进入实用或半实用化阶段，例如 IBM 的 *ViaVoice*[IBM]以及 L&H[L&H]、Philips[Philips]、Dragon[Dragon]等公司的听写机和其它语音产品；其次，语音识别、语音合成、语音编码等技术在手机、PDA 等嵌入设备中，以及在进行 Internet 网页浏览控制和阅读等方面，也已得到了广泛应用。随着其应用领域的进一步扩大，尤其是基于自然语音或电话语音的信息查询、预约、订票、导游等新需求的不断提出，语音识别研究的内容也开始由朗读式的连续发音向非受限的、自然的口语或对话发音方面扩展。

## 1.3 自然语音：语音识别的未来与难点

### 1.3.1 自然语音识别的重要性

这里的自然语音是相对于朗读式语音而言的，它是指连续非朗读式的、自然的口语或对话发音。基于标准发音、标准口音、朗读式的关键词识别系统和连续语音识别系统在过去的几十年里，已经从理论研究、技术实现和推广应用方面取得了巨大的成功，但上述限制条件也制约了其应用的进一步发展。研究重点从朗读式语音向自然语音的转移，是由语音技术的应用背景和需求决定的。首先，人们期待更为自然和方便的人机交互方式，提出了基于自然语音或电话语音的听写录入、信息查询、预约、订票、导游等新的需求，并在这些方面进行大量的研究和开发工作，有些工作已经进入实用阶段；其次，在任何一个国家，要求每个人都使用标准发音和速度发音是不可能的事情，所以对自然随意的语音识别技术进行研究，就为进一步扩大用户群体和应用面提供了可能。可以说，在未来的语音应用系统中，以自然语音为处理对象是一个必然的趋势。随着近年来语音技术的飞速发展，在这方面的研究已经积累了一定的经验和技术基础，目前人们在进行自然语音的识别方面不断提出一些新的思路，其识别

器的性能也在不断提高。

口语对话系统是语音技术走向实用化的一个重要的研究方向，其研究内容也是目前语音领域内最具挑战性的方向之一。它的技术除了涉及语音识别与合成外，还包括语言理解、对话管理、口语现象等，属于语音处理和自然语言处理两个领域的交叉学科，从声学层面上看，大部分口语对话系统的语音识别部分都是基于连续自然发音的。在国外，有专门的研究计划开展这项研究，像美国的 TRAINS 计划[Allen94]和 DARPA-Communicator 计划[Goldschen99]、欧洲的 ARISE 计划[Os99]、REWARD 计划[Failenschmid98]、VERBMOBIL 计划[Wahlster93]等，有许多著名的学府和研究机构从事这项研究，比如 MIT 的 SLS 实验室[Zue97]、CMU 大学[Rudnicky99]、Lucent-Bell 实验室[Lee98, Potamianos99] 、OGI 的 CSLU 中心[Sutton98]、法国的 LIMSI[Lamel98, Lamel99]、德国的 Erlangen-Nuremberg 大学[Gallwitz98]、日本的 ATR 实验室[Takezawa98]和 Philips 公司[Aust98]等，在国内，清华大学[Huang00]、中国科学院[Huang99]、中国科学技术大学、香港大学、香港中文大学[Meng00]、香港科技大学、台湾大学[Lin00]等也都投入了相当大的精力开展这项研究。可以预计，在刚刚到来的二十一世纪，语音技术将得到前所未有的发展。

### 1.3.2 自然语音的特点以及汉语的特殊性

已有实验表明，对于采用标准发音的、仔细朗读的语音，传统的连续语音识别器一般都能够达到很高的词或音节识别率。然而，对于无准备的、自然随意的发音，这些识别器的性能却会急剧下降[Lussier99]，这在很大程度上是由于自然语音本身的特殊性决定的。

在声学层面，自然语音往往包含了多变的语速、语气、韵律和真实的情绪，以及严重的协同发音，这就会造成大量的音素级的插入、删除和替换现象。此外，不同的人具有不同的口音背景和发音习惯，因此即使说话人努力去按照标准的读音去发音，实际的音素序列也不会完全相同[Decker99, Greenberg99]。

汉语相对于西方语言来说，又具有其特殊性。汉语的发音是基于标准音节的声韵结构（包括零声母现象），这种结构非常短，因此在自然发音中很容易受到口语上下文的影响而发生畸变[Lin92]。

同时，汉语中多音字的广泛存在以及某些非标准读音的使用，复杂的方言和口音背景[Li57, Yuan83, Chen98]，以及儿化音和变调现象的普遍存在，使得音节或音素间的混淆程度更加严重。从另一个角度来看，汉语的标准声韵集合以及它所覆盖的音素集合非常小，涵盖不了自然语音中大量的似是而非的发音，这也是影响声学建模精度的一个重要因素。

表 1-1 给出了若干个汉语自然口语发音中的音变现象的例子，从中可以初步体会到自然语音的发音变化及其识别任务的复杂性。

表 1-1. 汉语自然语音中的音变现象举例

| 序号 | 汉语文本 | 标准的与实际的发音 | 发音变化现象的分析[1] |
|---|---|---|---|
| 1 | 这个<br>走了 | zhe (zhei) ge<br>zou le (la) | 口语化读音： zhe => zhei<br>非标准读音： le => la |
| 2 | 当官的 | dang guan (-er) de | 儿化音： guan => guan-er |
| 3 | 什么时候<br>呢 | shen me shi (h) ou<br>ne (na, ni, nia, nie) | 声母的删除： hou => ou<br>略带口音： ne => na, ni, nia, nie |
| 4 | 还得有<br>知识啊 | hai de (dei) you<br>zhi shi (r-) a | 口语化读音： de => dei<br>插入声母： a => ra |

表格中所列举的现象仅仅是示意性的。在现实的发音中，有许多发音由于上下文的影响而变得似是而非，例如读音的弱化、清化或浊化，或者产生根本不在标准音节表里的新音节（例如上表中的 *ra* 和 *nia*）等现象，此时若仅仅使用标准的音节和标准的声韵集合或声韵对应的标准音素序列来刻画这些发音就显得力不从心了。

[1] 在现代汉语中，“*zhe*”和“*zhei*”都是多音字“这”的合理读音，“*de*”和“*dei*”也都是多音字“得”的合理读音。通常在朗读式（如播音员）的发音中，“这”和“得”更多地倾向于被读作“*zhe*”和“*de*”，而在随意自然的发音中，它们各自的两种发音则会同时存在。

### 1.3.3 连续语音识别中的基本技术

一般地，对于给定的声学输出观测矢量 $\mathbf{X} = X_1 X_2 \cdots X_n$，连续语音识别目的是给出它对应的词序列 $\hat{\mathbf{W}} = W_1 W_2 \cdots W_m$，这是通过最大化后验概率 $P(\mathbf{W} \mid \mathbf{X})$ 来实现的[Yang95, Huang93, Zheng99c]，即：

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}) = \arg\max_{\mathbf{W}} \frac{P(\mathbf{X} \mid \mathbf{W})P(\mathbf{W})}{P(\mathbf{X})} = \arg\max_{\mathbf{W}} P(\mathbf{X} \mid \mathbf{W})P(\mathbf{W}) \quad (1.1)$$

其中，$P(\mathbf{X} \mid \mathbf{W})$ 代表声学模型得分，$P(\mathbf{W})$ 代表语言模型得分。与此对应，目前在连续语音识别任务中广泛采用的一些成熟的建模技术，就是即包括声学层面的，也包括语言学层面的。

图 1-1 以清华大学计算机系语音技术中心开发的汉语连续语音识别的原型系统 *EasyTalk*[Zheng99a]的功能结构[Song00a, Song00b]为例，说明了连续语音识别系统中的各个基本组成部分。其中最主要的是声学建模及搜索功能模块和语言建模及处理功能模块（也有些识别系统将这两部分功能合并在一个模块中），此外，还应包括数据采集和预处理、特征提取、说话人自适应等功能模块。

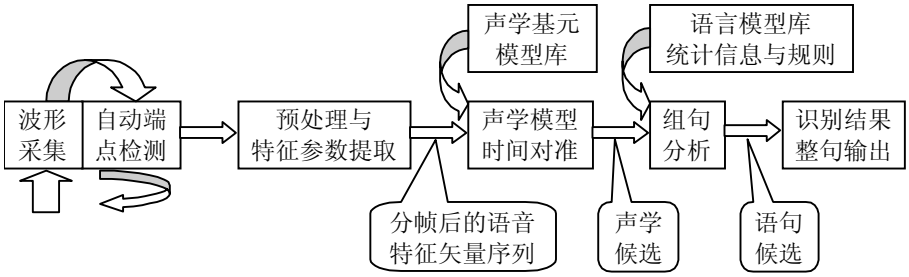


图 1-1. 汉语连续语音识别系统框架举例

迄今为止，一些成熟的基本建模技术在基于朗读式发音的识别系统中已经取得了巨大的成功，在目前基于自然发音的识别系统中也得到了广泛的应用。

这些连续语音识别中的基本建模技术包括：

1) 利用基于子词（Sub-Word）的识别基元（如声韵、半音节、音素等）集

合来进行建模，既可以保证声学模型有较小的规模、较大的灵活性和较快的识别搜索速度，又可以保证在训练数据有限时，仍可以对模型参数进行充分的估计[Zheng96]；

2) 采用 HMM（Hidden Markov Model，隐马尔可夫模型）。HMM 是基于一个隐含的状态空间 Markov 链以及与其相关联的输出概率随机函数集。它既反映了语音的短时平稳特性，又具备了描述隐含的发音状态改变的能力[Rabiner89a, Rabiner89b, Kriouile90]，它是声学建模的最核心技术之一；

3) 在 HMM 状态内部采用多个高斯混合的特征空间分布描述，可以更加准确而精细地拟合各个发音状态的声学输出特征分布[Bahl88, Huang90a]；

4) 采用上下文相关的建模方法，并结合基于决策树或数据驱动方式的 HMM 状态共享技术，使得在声学模型中既包容了协同发音现象，又保证了每个上下文相关模型都能得到充分的训练[Rabiner93, Liu99]；

5) 采用宽度优先（Breadth-First）的搜索算法（例如 Frame-Synchronous 算法[Lee89b]、Viterbi 束搜索[Viterbi67]、Token-Passing 算法[Odell95]等），深度优先（Depth-First）算法（例如 A*算法[Nilsson71, Paul92]等），或两者的结合算法（例如前向—后向搜索算法[Li96]、Quasi-Stack 算法[Renals96]）及其各种改进算法来进行搜索解码识别；

6) 为特定用户进行声学模型的自适应[Lee91, Lee93, Gauvain94, Leggetter95, Gales96b]，使得与说话人无关的模型更接近于当前用户的声学特性，从而获得更为理想的识别性能；

7) 采用基于 *N*-gram 的统计语言模型[Witten91, Katz87, Jelinek80, Ronald00]，在声学识别的基础上进行组句分析，从而获得最终的文本识别结果，等等。

### 1.3.4 自然语音声学建模中的关键技术

上节中介绍的目前在朗读式连续语音识别任务中广泛采用的一些成熟的建模技术，对于自然语音识别的任务来说，同样也是非常重要的。但对于自然语音的识别任务来说，仅有这些技术还是远远不够的。例如，对于自然语音中复杂的发音变化，如果仍采用仅具有标准发音序列的词典，就无法准确反映大量

的音素级的插入、删除和替换现象。但是，当在词典中为每个词或音节引入多个可能的发音变化序列后，又会增加音节间的混淆度[Decker99, Zheng00]。因此，对自然语音的声学建模问题，需要在原来的主流技术基础上进行更加深入的研究。

类似地，对于自然语音的识别和理解的研究，也主要集中在如下两个层面上：

1) 声学层面：从声学的角度对自然语音进行精细的建模，以充分反映在声学层面上可能的发音变化；
2) 语言层面：对复杂的语言现象进行建模、理解或语义处理，包括口语中的重复、省略、修正等。

本论文的研究焦点定位在自然语音的声学建模技术，而不涉及任何语言层面的处理（即与统计语言模型或自然语言处理相关的研究内容）。

目前，国际上对于自然语音的声学建模问题，在如下五个方面进行了卓有成效的探索，其中前两部分的内容是最主要的研究焦点。

### 1. 多发音词典的构造

构造一个包含了必要的、可能的发音变化的发音词典，是对自然语音进行识别的基础。通常，发音词典的构造方法可以分为两类：

一类是基于专家的知识，这可以通过语音学家对口语数据库进行详细的声学标注，然后对其中包含的发音变化进行统计并构造发音词典[Byrne01]；也可以由语言学家利用语言学和语音学的各种先验知识与规则生成各种可能的发音变化，并将其反映到发音词典中[Finke97]。

而另一类则是基于数据驱动的方法，其共性是首先训练一个初始的声学识别器，然后通过如下方法来构造发音词典：第一种方法是进行基元的自由识别，确定一个大的基元识别混淆矩阵[Liu00c]，从中直接生成发音词典或将识别基元序列与标准基元序列进行动态规划的时间对准，然后统计得出发音词典；第二种方法是基于特定语法的生成规则[Cremelie97, Cremelie99, Liu00a]，例如定义重写规则 $LBR \rightarrow S$（其中 $B$ 为标准发音形式，$S$ 为实际观测到的发音形式，$L$ 和 $R$ 为上下文），通过对标准的和识别的基元序列进行时间对准，生成符合上述语法的一

系列规则，然后进行迭代和精简，最后将这些规则反映到发音词典中；第三种方法是在给定标准基元序列的情况下，利用神经网络来预测词典中各种可能的发音变化及其概率[Fukada97]；第四种方法是在给定标准基元序列的前提下，利用一个决策树（可以由规则生成、由数据库中统计得到、或者由识别器训练出来）来预测词典中各种可能的发音变化及其概率[Byrne98, Riley96]。

实际上，基于专家知识和基于数据驱动这二者并不是完全可分的，在某些情况下，应该有机地结合。例如，对于利用决策树来预测音变的方法[Byrne98, Riley99]，可能就需要根据对人工标注的数据库进行统计得到初始的决策树，之后才能利用决策树和一些音韵学规则来对发音变化进行预测和平滑。同时，当手工标注的训练数据不是很充分时，数据驱动得到的概率不是很准确，可扩展性差，反之，利用规则生成的发音变化又存在过饱和（Over-Generation）的问题。文献[Ma98]提出了对二者进行结合的思路：首先，将某些高频的，经常连接在一起使用且容易发生词间弱读或变音的词对（二元词对或三元词组等）定义为复合词（Multi- Word，例如 *gonna*，它是 *going to* 的一个很常见的口语化读音），对其单独建模并加入发音词典中，然后利用数据驱动的方法对高频词和复合词直接统计概率，同时利用音韵学规则来预测低频词的可能音变，并为不同的说话人对这些发音变化的权重进行适应，取得了较好的效果。

从另一个角度来看，当词典中包含了复杂的发音变化后，相应的词或音节的混淆程度也就相应地增加，这就需要根据一定原则对所有的发音变化进行取舍。一种最直观的方法是根据各个音节每种发音的出现频度对词典项进行加权；另一种方法则是根据最大似然的准则，为每个词典项确定一个信任度阈值，然后进行取舍和加权，以减小词或音节的混淆度。

### 2. 自然语音的声学建模

这方面的研究主要集中在针对自然语音在声学上多变的特性，对 HMM 模型的训练和构造方法进行改进或提出新思路。

第一方面是对传统模型的优化[Liu00c]，包括利用上述的多发音词典，对原始的标准标注进行强制时间对准（Force-Alignment）的声学解码，并更新训练数据的标注信息，以数据驱动的方式反映发音变化，然后根据新的统计信息不断

调整发音词典和 HMM 模型参数，进行反复迭代求精，以得到一个更加贴近于实际变化发音的声学模型集合。

另一方面是数据的借用和模型插值[Jelinek98, Huang96, Kim97]。这些方法针对的问题是，由于口语数据库的基元分布不均衡和训练数据稀疏所造成的局部模型训练不充分。为了增加训练数据量，可以利用上面提到的决策树对发音变化的预测能力，将现有数据库包容的发音变化现象向其它的口语数据库进行推广，以数据驱动的方式产生合理的新标注并引入新的训练数据，或者利用删除插值（Deleted-Interpolation）技术来结合未能充分训练、但包含了丰富音变现象的自然语音声学模型和其它的充分训练过的、但包含较少音变现象的传统声学模型，将二者进行加权组合并重估参数[Liu00b]，使得在不损失音变信息的前提下，自然语音的声学模型也得到了较充分的训练。

在自然语音的声学建模方面的工作，还包括如下方法：利用 HMM 状态间的混淆程度来共享这些状态内部包含的高斯混合参数[Saraclar99]，以达到更精细地在声学模型内部描述发音变化的目的；利用混淆矩阵并结合可能的发音变化规律来为词内和词间发音变化进行建模的方法[Liu00c]；利用全数据驱动的、最大似然的方法对发音变化建模[Holter99]的方法等。

### 3. 识别基元的选择和生成

包括两方面的内容。首先是合适的识别基元集合的选择问题，对于不同的自然语音识别任务，确定识别基元的合适的长度和集合大小，对最终的识别性能有很大的影响。目前常用的识别基元有音素（Phoneme）、子音素（Sub-phoneme）、音位变体（Allophone）、词（如英文中的单词）或复合词（Multi-Word）等，对于汉语，还可以有音节、半音节（Semi-Syllable）、声韵（Initial/Final）等基元。

其次是识别基元的自动生成问题。上述的各种基元集合往往是由语言或语音学家根据先验知识定义，对于特定领域的自然语音来说，它们的分布是极其不均衡的，因此文献[Bacchiani99]提出了纯数据驱动的、根据最大似然的原则来自动产生合理的识别基元集合，并构造相应的多发音词典的方法。其思想是将每个词初始分割为相同的段数，根据最大似然和动态规划的准则调整标注分点，然

后利用 *K*-均值（*K*-Means）方法[Selim84]对这些分段进行聚类，并抛弃样本数不足的类，然后迭代重新训练声学模型，不断调整自动的基元集合和生成的多发音词典，这样就保证了生成的基元集合与发音词典从最大似然的角度来看也是最优的。

**4. 定制搜索解码算法**

在自然语音的识别中，当引入多发音词典后，传统的声学搜索解码网络规模急剧膨胀，尤其是采用上下文相关建模以后。因此如何在不减小识别精度的前提下，尽可能地提高识别器的搜索速度，也是一个很值得研究的课题。例如，文献[Finke99]中介绍了一个改进的时间同步搜索算法，通过在路径扩展时引入中介的共享节点，使得路径扩展的规模得到了很大的压缩，从而提高了声学搜索的速度。文献[Holter99]中介绍了另一种树状网格（基于 A*的搜索算法）搜索算法，通过修正 A*目标函数，使得在路径中可以同时支持对多种发音变化的打分。

**5. 修正统计语言模型**

当采用 *N*-gram 统计语言模型时，若直接利用公式 (1.1)，显然是不合适的，这是因为对于任何一个 **W**，由于发音变化的存在，$P(\mathbf{X}\mid\mathbf{W})$ 并不能准确地反映出各种发音变化形式的声学输出概率，所以需要对这个目标公式进行修正[Strik99]。

一种可能的修正方法是直接采用实际的发音序列 **V** (包含了插入、删除和替换等音变现象)来统计 *N*-gram，此时识别的目标是找到最优的发音序列 $\hat{\mathbf{V}}$，即：

$$\hat{\mathbf{V}} = \arg\max_{\mathbf{V}} P(\mathbf{X}\mid\mathbf{V})P(\mathbf{V}) \tag{1.2}$$

其优点在于对 $P(\mathbf{V})$ 进行统计时已经隐含了发音的上下文信息，但由于集合 **V** 远远大于集合 **W**，因此语言模型的训练语料将显得更加稀疏。

另一种修正方法是通过引入中间的信息层 $P(\mathbf{V}\mid\mathbf{W})$ 使得音变信息得以在语言模型中反映出来，即：

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{X}\mid\mathbf{V})P(\mathbf{V}\mid\mathbf{W})P(\mathbf{W}) \tag{1.3}$$

与公式 (1.2) 相比，其优点在于最终的目标序列直接就是词序列，而不是公式(1.2)中的最优发音序列 $\hat{\mathbf{V}}$（例如最优的音素序列）从而需要再次进行组句分析。它的缺点在于 $P(\mathbf{V}|\mathbf{W})$ 并没有反映出发音变化形式 $\mathbf{V}$ 的上下文相关性，而实际上，由上下文的协同发音而造成的音变对声学的输出 $P(\mathbf{X}|\mathbf{V})$ 却是至关重要的。总的来说，对于一个实际的语音识别系统，在各种可能方案的优缺点之间进行一些折衷是必要的。

很明显，上面介绍的这些关键技术是面向自然随意发音的、对其中大量的发音变化现象进行建模的一些“特殊技术”，这有别于传统的、面向朗读式语音的建模技术。从国际上看，这方面的研究内容通常被称为“发音建模”(Pronunciation Modeling)，并且绝大多数情况下是指在声学层面上为发音变化进行建模的一些技术。这也正是本文的研究重点。

## 1.4 总体研究思路与创新点

### 1.4.1 研究目标与思路框架

本论文将研究焦点定位在汉语自然语音的识别问题，尝试从纯声学建模的角度解决由于发音的随意性而给识别任务带来的巨大的困难，并提出一些行之有效的新思路和方法。

论文的研究目标为：

1) 结合语言学家为新的自然语音数据库提供的详尽标注，对标准的汉语声韵基元集合进行推广，以求更加全面地覆盖汉语中主要的、可能的音变现象；
2) 改进多发音的词典内不同发音变化序列的加权策略，以达到进一步降低自然语音中各个音节间混淆度和提高识别率的目的；
3) 通过更加精细的声学建模技术，来减弱训练数据的稀疏问题对模型精度的影响，并更加准确地反映可能的发音变化，以及进一步提高识别器的

性能。

图 1-2 从总体上反映了论文研究的思路和出发点，以及各部分主要工作之间的联系。

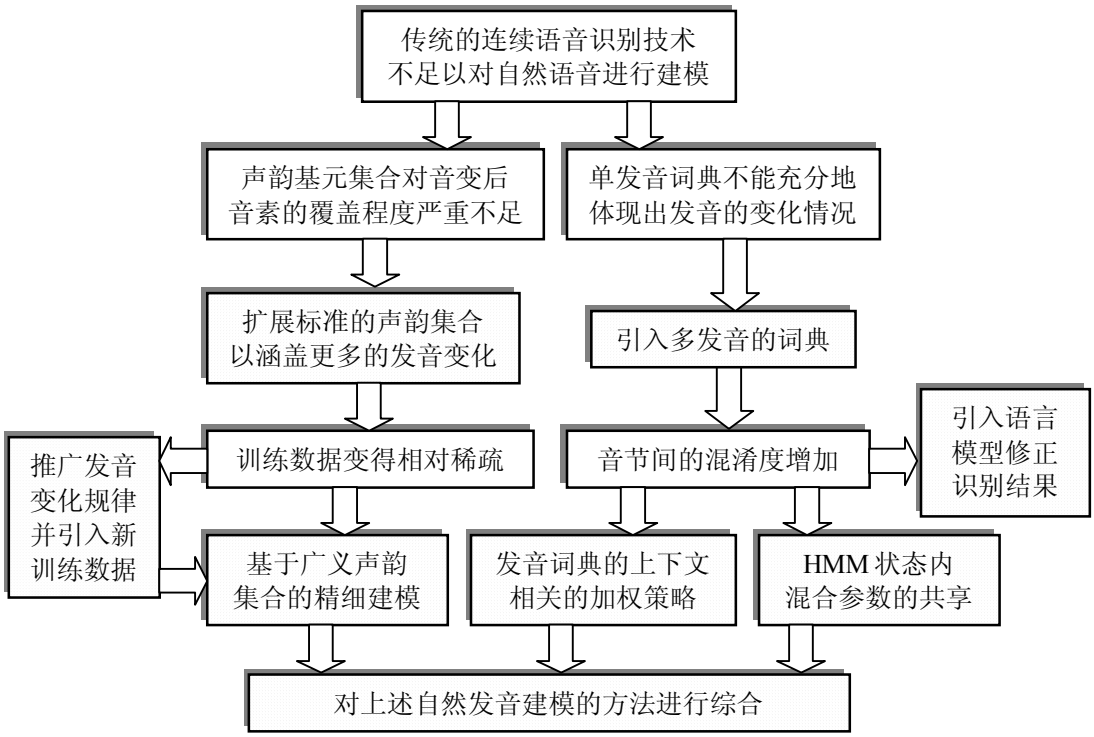


图 1-2. 论文总的研究思路以及各部分间的关系

其中，语言模型的作用是显而易见的。但正如前面所提到的那样，在本论文的研究工作中，将不涉及任何“语言”模型（包括汉语词的 *N*-gram 和汉语音节的 N-gram 等）的作用，而是将研究和比较的重点集中在纯声学的层面上，从而更清晰地体现这些“声学”建模方法对总体识别性能的改善能力。

### 1.4.2 该研究的创新点概述

本论文在基于汉语的自然语音识别的声学建模方面进行了初步的探索，提

出了若干新思路并通过实验证明了其有效性，同时也为这方面后续的深入研究积累了一定的经验。其创新点可以概括为如下三点：

1) 提出广义声韵集的概念，以及对它进行的精细建模方法，可以刻画更为丰富与精细的发音变化，并为解决训练数据的稀疏问题提供了新思路；

2) 提出减小音节之间的混淆度的新方法，即根据半音节上下文对多发音词典进行加权的策略，该方法与传统的基于发音序列出现概率的方法相比，更加稳定和有效；

3) 提出通过模型状态之间共享参数的思想为发音变化进行建模，在保证 HMM 模型具备足够的描述发音变化的能力的同时，进一步降低音节间的混淆度和提高识别率。

## 1.5 论文的组织结构

论文的内容和结构是按照上述三个方面的主要工作来组织的。

在第二章中，介绍了为进行汉语自然语音识别研究而构建的语音库及其标注的特点，在此基础上定义了广义的声韵集，并对它和标准的声韵集合进行基准的声学建模。并在第三章中，根据模型自适应技术的思路，提出在基准声学模型的基础上，对广义声韵集进行精细建模的方法。

在第四章中，从发音词典固有的音节间混淆度的角度出发，定义了在理想情况下对其进行评价的指标，并提出利用基元上下文来对多发音词典进行加权处理并减小音节间混淆度的新方法。

在第五章中，提出两种通过模型参数共享的方法进行发音建模的方法。通过采用单发音词典以降低音节间的混淆度，并在声学模型进行在声学模型内部尽可能地包容发音变化，从而达到提高系统性能的目的。

在第六章中，对各章的思路之间的联系以及语言模型的作用等进行归纳和阐述。最后，在第七章中对论文的研究进行总结和展望。

在附录中，简单介绍了在博士就读期间完成的其它两项主要研究工作。

# 第二章　广义声韵集与基准模型

## 2.1　从标准声韵集到广义声韵集

### 2.1.1　识别基元的选择原则

识别基元的确定是任何一个语音识别系统进行声学建模的基础和首要任务。汉语语音识别基元的确定可以有多种选择，例如音节、声韵、半音节、音素等[Yang95]，不同的方案具有不同的优缺点，不同的研究机构对它们的定义和表示方法也会略有差别，需要综合考虑确定。

例如，汉语的发音是基于音节的结构，直接将音节作为识别基元是一个显而易见的选择[Zheng99b]，但由于其个数较多（汉语音节的全集为 418 个无调音节，或大约 1200 多个有调音节），需要有很大的数据库才能避免训练数据的稀疏；音素通常可以认为是最小的发音单位，在连续语音中，很容易受上下文的影响，因此协同发音现象严重，因此，若采用音素作为识别基元，必须要进行上下文相关的建模，才能保证有比较满意的识别率。

声韵和半音节具有相似的定义和集合大小，它们都隐含了音节内部的部分音素间上下文相关信息，差别在于半音节中的声母是根据其后面可能连接的韵母进行了细化，因此额外包含了声母尾部与韵母首部音素的相关性，且半音节集合比声韵集合略大。当二者都采用上下文相关建模时，其声学描述能力差别不大。

由于汉语音节是基于标准的声韵结构，而且声韵集合的大小适中，它们的定义及其与音节的对应关系也是统一的，因此本论文的研究选择标准声韵集合作为基准的识别基元。

与此相关的、汉语音节 S 到其标准读音 IF 序列的映射表包含了一系列的三元组 $(S,I,F)$，其中 $S$、$I$、$F$ 分别为音节、声母、韵母，如果 $S$ 是一个零声母的音节，则 $I$ 为空白。例如：

| | | | |
|---|---|---|---|
| a | | a | （“啊”的标准发音） |
| ba | b | a | （“八”的标准发音） |
| hou | h | ou | （“候”的标准发音） |

显然，这是一个单发音的词典，在大多数基于声韵基元的、朗读式的连续语音识别系统中被广泛采用。但当应用于自然语音识别中时，它并不能准确地反映出频繁的声母或韵母的插入、删除和替换等音变现象，因此对于自然语音的识别任务，首要的问题就是要确定一个合适的多发音词典。

### 2.1.2 汉语自然语音库的构建

目前绝大多数基于汉语的连续语音识别系统所采用的识别基元都是上面介绍的标准集合，例如声韵、半音节、音素等等。一些尝试在识别器中引入发音变化的研究，也主要是采用汉语声韵或音素集合内部的识别混淆矩阵，并由此构造多发音词典的方法，来反映可能的发音变化[Liu00c]。但正如前面所分析的那样，对于一个真实的自然语音识别任务来说，标准的汉语声韵或音素集合对各种可能的发音变化的覆盖程度是非常不充分的。因此有必要将标准的声韵集合进行扩展，并在此基础上构造合适的多发音词典。

由于目前国内并没有完全适用于上述需求的、用于自然语音识别的数据库，为此，清华大学计算机系语音技术中心联合中国社会科学院语言所、香港科技大学，以及美国约翰斯·霍普金斯大学的语言和语音处理中心，共同构建了一个专用于自然语音识别研究的汉语语音库，命名为 CASS (Chinese Annotated Spontaneous Speech)[Li00a, Li00b, Li00c, Chen00]数据库[2]。本论文的研究也是全部基于这个新构建的 CASS 数据库的。

CASS 数据库的设计目标包括：

1) 说话人以极其自然的、口语化的方式发音，其口音基本符合汉语的普通话发音；

2) 期望数据库覆盖尽可能多的发音变化实例，并由语言和语音学家（或专业工作者）给出详尽的、准确的标注，并用它们来先验地统计音变规律

[2] 作者本人参与了其中部分工作，包括组织数据采集，编写数据处理和标注校验的程序等

和概率；

3) 同时，将这个数据库作为进行自然语音识别声学建模研究的训练和测试数据。

CASS 数据库区别于其它汉语语音库的一个最显著的特征和优势是，它首次在汉语数据库的标注中采用了由 IPA (International Phonetic Alphabets，国际音标)[IPA]导出的 SAMPA (Speech Assessment Method Phonetic Alphabets) 符号集[SAMPA, Kessens99]，并针对汉语进行了定制，命名为 SAMPA-C (SAMPA for Chinese)[Chen00]。利用 SAMPA-C 进行的标注与传统的汉语音节、声韵或音素标注符号集相比，包含了更多的对发音变化的描述能力，这是因为它对于汉语自然发音中多种可能的音位变体(Allophone)分别采用了不同的符号表示，并通过一些特殊定义的后缀，可以明确地标记出辅音的浊化、元音的清化等细节信息。

表 2-1 给出了以 SAMPA-C 形式定义的基本汉语辅音集合，表 2-2 给出了以 SAMPA-C 形式定义的基本汉语元音集合，表 2-3 给出了以后缀形式表示的更加精细的发音变化表示，表 2-4 给出了以 SAMPA-C 形式表示的汉语标准声韵母和儿化音的若干实例。

在 CASS 的实际标注中，声韵的发音变化既包括了其内部的细微变化（通过音位变体以及 SAMPA-C 后缀来与其标准形式区分。例如元音 *a* 可以变为 *a*, *a_"*, *a_"_v* 等），又包括了声韵间的发音变化（即由一个声韵基元完全音变到了另一个声韵基元，例如辅音 *sh* 变为 *r*），以及大量的声韵基元本身（及其内部个别音素，如 *an* 或 *iang* 变为 *ang*）的插入（如零声母拼音 *a* 插入声母 *r* 变为 *ra*）和删除（如拼音 *hou* 的声母被弱读或删除）等发音变化现象。

表 2-1. 汉语辅音的 SAMPA-C 表示

| Phoneme | | Allophone | SAMPA-C | Phoneme | | Allophone | SAMPA-C |
|---|---|---|---|---|---|---|---|
| Pinyin | IPA | | | Pinyin | IPA | | |
| b | p | p | p | z | ts | ts | ts |
| p | pH | pH | p_h | c | tsH | tsH | ts_h |
| m | m | m | m | s | s | s | s |
| f | f | f | f | zh | t© | t© | ts` |
| d | t | t | t | ch | t©H | t©H | ts`_h |
| t | tH | tH | t_h | sh | © | © | s` |
| n | n | n | n | r | ¸ | ¸ | z` |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (a)n | n | n | _n | j | t» | t» | ts\ |
| l | l | l | l | q | t»ʰ | t»ʰ | ts_h |
| g | k | k | k | x | » | » | s\ |
| k | kʰ | kʰ | k_h | | | | |
| h | x | x | x | | | | |
| ng | ŋ | ŋ | N | | | | |

表 2-2. 汉语元音的 SAMPA-C 表示

| Phoneme | | Allophone | SAMPA-C | Phoneme | | Allophone | SAMPA-C |
|---|---|---|---|---|---|---|---|
| Pinyin | IPA | | | Pinyin | IPA | | |
| a | A | a | a | i | i | i | i |
| | | A | a_" | | | I | I |
| | | ɐ | A | | | j | j |
| | | Q | E | | ï | ɣ̈ | i` |
| | | P | { | | | ɨ | ɨ\ |
| o | o | o | o | u | u | U | U |
| | | ʊ | U | | | u | u |
| | | u | u | | | w | w |
| | | È | @ | | | ʊ | v\ |
| e | ° | ° | 7 | ü | ʏ | y | y |
| | | e | e | | | á | H |
| | | E | E_r | er | Ä | Ä | @` |
| | | È | @ | | | | |

表 2-3. 以后缀形式表示的发音变化

| 音变分类 | SAMPA-C | 举例 | | |
|---|---|---|---|---|
| | | IPA | SAMPA-C | Explanation |
| 鼻　化　Nasalized | ~ | a | §~ | '§' is *nasalized*. |
| 央　化　Centeralized | _" | e | e_" | 'e' is *centeralized*. |
| 清　化　Voiceless | _u | n | n_u | 'n' is *voiceless*. |
| 浊　化　Voiced | _v | d | d_v | 'd' is *voiced*. |
| 圆唇化　More Rounded | _O | $f_w$ | f_O | 'f' is more *rounded*. |
| 成音节　Syllabic | = | $m_l$ | M= | 'M' is *syllabic*. |
| 喉　化　Pharyngealized | ɑ̲?\ | t/ | A_?\ | 'A' is *pharyngealized*. |
| 增　音　Inserted | (+) | | (N+) | 'N' is *inserted*. |
| 减　音　Deleted | (-) | | (i-) | 'i' is *deleted*. |

表 2-4. 以 SAMPA-C 表示的声韵母及儿化音举例

| 标准声母的 SAMPA-C 表示举例 | | | |
|---|---|---|---|
| 声母 | IPA | SAMPA-C | 拼音举例 |
| b | p | p | ba, bi, bu … |
| d | t | t | de, da, di … |
| g | k | k | ge, gu, ga … |
| p | pʰ | p_h | po, pa, pu … |
| k | kʰ | k_h | ke, ku, kan … |
| z | ts | ts | zi, zu, zou … |
| zh | tʂ | ts` | zhi, zha … |
| r | ʐ | z` | ri, ru, ruan … |
| x | ɕ | s\ | xi, xia … |
| h | x | x | ha, he, hun… |
| 儿化韵母的 SAMPA-C 表示举例 | | | |
| 儿化 | IPA | SAMPA-C | 拼音举例 |
| ar | ar | a` | Par |
| anr | ar | a` | ganr |
| angr | ar | a~` | gangr |
| ier | iɚ | iE_r` | jier |
| ir | iɚ | i@` | jir |
| uan | uar | ua` | tuanr |
| üanr | yar | ya` | yuanr |

| 标准韵母的 SAMPA-C 表示举例 | | | |
|---|---|---|---|
| 韵母 | IPA | SAMPA-C | 拼音举例 |
| a | A | a_" | ba, a … |
| ai | aI | aI | ai, sai, shai,… |
| an | an | a_n | an, fan, san … |
| eng | ɛɤ | @N | beng, feng … |
| i1(zi) | ɿ | i\ | zi,ci,si … |
| i2(zhi) | ʅ | i` | zhi,chi,shi … |
| i | i | i | yi, ji, bi, ji … |
| iao | i§U | iAU | yao, xiao … |
| ian | iɑn | iE_n | yan, qian … |
| ua | uA | ua_" | wa, kua … |
| uai | uaI | uaI | wai, kuai … |
| uan | uan | ua_n | wan, duan … |
| uen | uɛn | u@_n | wen, cen … |
| ui | uei | uei | wei, gui,… |
| ou | ɛu | @U | ou, hou, fou … |
| ong | uɤ | UN | hong, gong … |
| ü | y | y | yu, xv, … |
| üe | yE | yE_r | yue, que … |
| üan | yPn | y{_n | yuan, xuan … |
| ün | Yn | y_n | Yun … |

CASS 数据库的详细标注信息包括了五个层次，即：汉字（标准的书面语）、有调音节（隐含着标准的声韵序列信息）、实际发音的声韵（在声韵层面体现出插入、删除和声韵替换等信息）、SAMPA-C 序列（参考上述声韵的 SAMPA-C 表示以及后缀形式表示的音变符号）、杂类信息（如噪音、咳嗽、咂嘴、笑声等）标注等。其中第三层标注，即实际发音的声韵标注是按照一定的规则，由程序自动生成的，这将在后面章节介绍。

表 2-5 给出了 CASS 标注信息的一个实际例子（这里省略了各层标注的时间边界信息）。

表 2-5. CASS 数据库的标注举例

| 标注层次 | 标注信息举例 |
|---|---|
| 汉字标注层 | 多　认　识　点　人　啊 |
| 拼音标注层 | duo1　ren4　shi2　dian3　ren2　- a0 |
| 实际声韵层 | d uo2　r en4　r　i4　d ianr0　- er2　r　a4 |
| SAMPA-C 层 | t_v uo　z` @_n　z` ts`　t　ia`　- @`　z` a_" |
| 杂类信息层 | smack<　smack>　noise<　noise> |

从这个例子中可以看出一些复杂的发音变化，例如：

1) 部分的音变：拼音 *duo* 的声母 *d* 清化位 *t_v*；拼音 *dian* 的韵母 *ian* 儿化为 *ianr*；
2) 完全的音变：拼音 *shi* 的声母 *sh* 变为 *r*；拼音 *ren* 的韵母 *en* 变为 *er*；
3) 声韵的插入：拼音 *a* 被插入声母 *r*，从而变为（非标准的）拼音 *ra*；
4) 声韵的删除：拼音 *ren* 的声母被删除，其韵母也同时发生音变，从而变为拼音 *er*。

CASS 数据库共包括约 6 个小时、17 人（5 个女声＋12 个男声）的自然发音，以 16 KHz、16 位精度采样。本论文的研究采用 CASS 的全部男声数据，从其中 5 个男声的数据中随机抽取约 30 分钟的发音(共计 500 句)作为测试集，其余的全部男声发音（约 5 个小时的数据，共计 5500 句）作为训练集。本论文的发音建模研究中均不考虑标注中的音调信息。

### 2.1.3 广义声韵集的定义及相关概率

如前所述，在自然随意的语音中，从音素的层次来看，标准的识别基元（通常称为基准形式，即 Base-form）与实际观测到的识别基元（通常称为表象形式，即 Surface-form）之间存在着如下两类变化：

1) 第一类是声音变化（Sound Change），即一个音素产生略微的发音变化，

例如发生了鼻音化、圆唇化、央化、清化、浊化等。此时虽然从听觉上或语谱图上可以觉察或观察到一些略微的畸变（相对于朗读式发音），但是仍可判断出所发的音属于期望的声韵母。这往往是由较快的发音速度以及前后音素间的协同发音造成的；

2) 第二类是音素变化（Phoneme Change），即一个音素完全变为另一个音素，或者发生音素的插入或删除现象。此时从听觉上或语谱图上都可以觉察或观察到较大的变化（相对于朗读式发音）。这种情况往往是由较快的发音速度、不同的口音背景或非标准读音造成的；

此外，汉语中大量的儿化音现象也是不容忽视的。在汉语的自然发音中，某些词汇的读音基本上都会发生儿化，而严格遵循其标准读音的情况反倒非常少见。例如在标准的书面语中，“（完成）指标”和“（骑）自行车”一般都不会被写为“指标儿”和“自行车儿”，但是人们几乎都会将它们读作“*zhi biao-er* (*biaor*)”和“*zi xing che-er* (*cher*)”（括号中的拼音是 SAMPA-C 标注中对儿化拼音的写法），而不是其标准的发音“*zhi biao*”和“*zi xing che*”。对于韵母儿化这类音变，视其相对于非儿化读音的畸变程度，既可以归类为声音变化，也可以归类为音素变化。

概括地讲，相对于声韵的标准读音来说，上述两类发音变化也可以笼统地分为声韵层次上的替换、插入和删除等三种类别，同时，这些发音变化现象，在 SAMPA-C 的符号设计以及 CASS 数据库的标注中都得到了体现。

为了对汉语自然语音进行更加有效的声学建模，这里将标准声韵的各种可能出现的表象形式定义为 GIF（Generalized Initial/Final，广义声韵）集合，以体现这些有意义的发音变化现象。与此对应，原来的标准声韵集合则简称为 IF（Initial/Final，声韵）集合。

表 2-6 以 CASS 数据库的部分标注为例，给出了 IF 和 GIF 之间的对应关系。

表 2-6. 标准声韵(IF)及其发音变化形式(GIF)举例

| IF (Pinyin) | GIF (SAMPA-C) | 说　明 |
|---|---|---|
| *z* | /ts/ | 缺省发音 |
| *z* | /ts_v/ | 被浊化 |
| *z* | /ts`/ | 变为 *zh* |

| | | |
|---|---|---|
| *z* | /ts`_v/ | 变为浊化的 *zh* |
| *e* | /7/ | 缺省发音 |
| *e* | /7`/ | 发生“儿化”音变 |
| *e* | /@/ | 变为某个GIF: @ |

显然，IF 集合是 GIF 集合的一个子集，它采用新华词典中定义的标准汉语声韵集合。而 GIF 集合通过对 CASS 标注进行统计得到——从 CASS 数据库中挑选出所有出现频度高于某个阈值（例如 10 次）的 SAMPA-C 序列，作为初始的 GIF 集合，并将其它不属于 GIF 集合的、但在 CASS 中实际出现的 SAMPA-C 序列映射到 GIF 集合中某个发音最相近的基元。

由于汉语自然语音中的发音变化非常复杂，而且 CASS 数据库的容量也是有限的，因此这样得到的 GIF 集合显然不能涵盖任何可能的汉语发音变化。但实际上，这些发音变化形式的出现频度相差是很大的。对于某些非常少见的发音变化形式，完全可以将其归结为某个与其发音最相近的 GIF 基元（如上面的方法），而不会对总的识别器性能造成明显影响。

表 2-7 中列出了 IF 和 GIF 集合所包含的识别基元数目。

表 2-7. 声韵(IF)和广义声韵(GIF)集合的基元数目

| 集合类别 | 声母数目 | 韵母数目 | 总数目 |
|---|---|---|---|
| IF 集合 | 21 | 38 | 59 |
| GIF 集合 | 37 | 43 | 80 |

确定 GIF 集合之后，就可以从 CASS 数据库的 SAMPA-C 标注中导出以 GIF 表示的新标注，并可以对它进行统计以得到 GIF 集合相关的概率：

1) GIF 的观测概率（observation probability）

上下文无关：$P(GIF|IF)$，　$P(<Del>|IF)$，　$P(GIF|<Nul>)$　　(2.1)

上下文相关：$P(GIF|IF, C)$，　$P(<Del>|IF, C)$，　$P(GIF|<Nul>, C)$　(2.2)

其中，*<Nul>*代表一个插入位置，这多发生在零声母的情况下（例如零声母的音节 *a* 被插入一个声母 *r* 变为 *ra*）；*<Del>*代表声母或韵母的一个被删除的位置；*C* 代表发音变化发生时的上下文，例如当前 *IF* 基元左右相邻的 *IF* 基元。

2) GIF 的转移概率（transition probability）

$$P(GIF),\quad P(GIF_2|GIF_1),\quad P(GIF_3|GIF_1, GIF_2) \tag{2.3}$$

上述转移概率的定义类似于统计语言模型中的 unigram, bigram 和 trigram，统计方法也与其相同，在此不再赘述。同理可得 IF 的转移概率。这些概率值在后面的研究中将会用到。

一个值得注意的现象是，从声韵 IF 集合到广义声韵 GIF 集合之间，存在多对多的映射关系，即一个标准的 IF 可以音变为不同的 GIF，而同一个 GIF，也可以由不同的标准 IF 发生音变得到，因此在音节层面会造成大量的发音混淆是显而易见的。

### 2.1.4 广义音节与多发音词典

与 GIF 的定义类似，将标准音节 S（或称为规范音节）的各种可能的实际发音基元序列称为广义音节（GS，Generalized Syllable），显然，S 集合是 GS 集合的一个子集。将 CASS 数据库的音节层面标注与 GIF 层面标注利用动态规划的方法进行时间对准，即可得到标准音节与其各个实际发音的 GIF 序列的对应关系，并可由此确定 GS 集合的内容以及每个 GS 的观测概率：

$$P(GS \mid S) = P(GIF_1, GIF_2, \cdots \mid S) \tag{2.4}$$

其中，*GS* 中的 GIF 序列 $GIF_i$（$i = 1, 2, \cdots$）是这个 *S* 的实际发音序列，在大多数情况下，*GS* 至多包含一个具有声母特性的 GIF 和一个具有韵母特性的 GIF，此时，将它们分别称为广义声母和广义韵母。表 2-8 给出了一个音节 *chang* 在 CASS 数据库中出现的广义声母、广义韵母及其输出概率。

表 2-8. 标准音节及对应的广义音节举例

| 音节 | 广义声母 | 广义韵母 | 输出概率 |
|---|---|---|---|
| *chang* | /ts`_h/ | /AN/ | 0.7850 |
| *chang* | /ts`_h_v/ | /AN/ | 0.1217 |
| *chang* | /ts`_v/ | /AN/ | 0.0280 |
| *chang* | <Del> | /AN/ | 0.0187 |
| *chang* | /z`/ | /AN/ | 0.0187 |

| *chang* | <Del> | /iAN/ | 0.0093 |
|---|---|---|---|
| *chang* | /ts_h/ | /AN/ | 0.0093 |
| *chang* | /ts`_h/ | /UN/ | 0.0093 |

显然，根据 GIF 及其到 GS 集合的对照关系，就可以初步构造用于自然语音识别的多发音词典。另一方面，从表中也可以看出，对于同一个标准音节来说，它的各个 GS 所具有的观测概率差异非常大，对于那些出现频度非常低的 GS，完全没有必要放入词典中，否则只会增加音节间的混淆程度，从而抵消掉由于引入多重发音而对识别性能的改善能力。因此，在这个词典中，只保留了那些累计概率高于某个阈值（例如 90%）的高频的 GS。

与基于标准 IF 的单发音词典相对照，汉语音节 S 到 GIF 序列（即 GS）的映射表包含了一系列的四元组 $(S,G_i,G_f,P)$，其中 $S$、$G_i$、$G_f$、$P$ 分别为音节、广义声母、广义韵母、$(G_i,G_f)$ 的观测概率（即词典中发音变化项的加权值）。其中 $(G_i,G_f)$ 从概念上相当于一个 GS；当 $P$ 为空白时，表示对词典进行等概率加权；在不同的发音上下文中，$G_i$ 或 $G_f$ 都有可能是空白。例如：

```
a          a_″    0.87   (“啊”的实际发音 1：等同于标准读音)
a          a_″_u  0.07   (“啊”的实际发音 2：发生韵母清化)
a    z`    a_″    0.06   (“啊”的实际发音 3：插入额外声母)
ba   p_v   a_″    0.71   (“八”的实际发音 1：发生声母浊化)
ba   p     a_″    0.29   (“八”的实际发音 2：等同于标准读音)
hou  x     @U     0.54   (“候”的实际发音 1：等同于标准读音)
hou  x_v   @U     0.26   (“候”的实际发音 2：发生声母浊化)
hou        @U     0.20   (“候”的实际发音 3：声母被丢失)
```

## 2.2　基准的声学建模

### 2.2.1　声学建模的理论依据

根据公式(1.1)，连续语音识别的目标就是，对于给定的声学输出观测矢量 $\mathbf{X}=X_1X_2\cdots X_n$，通过最大化后验概率 $\mathbf{X}=X_1X_2\cdots X_n$ 来求解最优的词序列 $\hat{\mathbf{W}}=W_1W_2\cdots W_m$。当不考虑任何语言模型 $P(\mathbf{W})$ 时，这个目标函数可以写为：

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}) \cong \arg\max_{\mathbf{W}} P(\mathbf{X} \mid \mathbf{W}) \tag{2.5}$$

对于汉语连续语音识别任务来说，由于不考虑语言模型，因此在音节序列的层面而不是汉语语句的层面来衡量语音识别器的声学识别性能是比较适宜的。

假设任意可能的标准音节序列为 $\mathbf{B} = b_1 b_2 \cdots b_N$，并且假设各个音节之间的声学输出矢量是相互独立的，因此，

$$P(\mathbf{X} \mid \mathbf{W}) = P(\mathbf{X} \mid \mathbf{B}) \approx \prod_{n=1}^{N} P(a_n \mid b_n) \tag{2.6}$$

其中，$a_n$ 是与 $b_n$ 相对应的声学输出矢量。对于自然语音识别任务来说，上式并没有体现出广泛存在的发音变化情形，因此需要进行修正。这是通过引入发音变化项实现的：

$$P(a \mid b) = \sum_{s} P(a \mid b, s) P(s \mid b) \tag{2.7}$$

其中 $s$ 是任意可能由 $b$（属于标准音节 S 的集合）发生音变后而观测到的一个广义音节（属于广义音节 GS 的集合），$a$ 为实际的声学输出矢量。于是这个声学模型就可以分为两部分：$P(a \mid b, s)$ 和 $P(s \mid b)$。后者是 GS 的观测概率，即多发音词典中的概率项 $P(GS \mid S)$，而前者可以称为 GIF 的“细化”或“精细”（Refined）声学模型，这是因为这个概率的条件项中的 $s$ 是与原始的 $b$ 有关的（例如 $IF_0$-$GIF_1$, $IF_0$-$GIF_2$, $IF_1$-$GIF_1$ 等）。它的直观物理含义是，由相同 *IF* 变音得到的不同 *GIF* 的声学输出概率是不同的，反之，由不同 *IF* 变音得到的相同 *GIF* 的声学输出概率也是不同的。

细化 GIF 与上下文相关的 IF（或上下文相关的 GIF）是不同的概念。其中，细化 GIF 是根据可能变为这个 GIF 的不同的原始 IF 而进行的细化，而上下文相关的 IF（或 GIF）则是根据训练数据中某个 IF（或 GIF）左右（或者仅仅考虑左边或右边）可能出现的相邻的 IF（或 GIF）而进行的细化。

显然，对于自然语音识别来说，上述两个概率公式部分是研究的焦点，这正与 1.2.3 小节中介绍的国际上的主流研究趋势相吻合。本论文的研究也正是基于这两个方面而展开的。其中第四章侧重于研究 $P(s \mid b)$ 的改进策略，第三章及

第五章则侧重于研究 $P(a \mid b,s)$ 的建模方法。

### 2.2.2 构建基准的声学模型

对于细化 GIF，若直接采用通常的建模方法是不行的，这是因为细化后的 GIF 集合将远远大于原来的 IF 集合或 GIF 集合，从而造成训练数据的严重稀疏。根据 CASS 数据的统计，IF 集合大小为 59，GIF 集合的大小为 80，而细化 GIF 集合（也可形象地表示为 IF-GIF 集合）的大小则为 237。若采用如此多的识别基元，将使得其中很多基元得不到充分训练，从而不能保证有比较满意的整体识别性能，因此必须考虑采用其它方法，这将在后续章节中研究。

为了避免训练数据的稀疏问题，可以对公式 2.7 进行简化，从而在数据量与模型精度之间进行一些折衷。对于细化 GIF 模型 $P(a \mid b,s)$，可以分别考虑利用 $P(a \mid b)$ 和 $P(a \mid s)$ 来近似它，从而得到以 IF 和 GIF 作为识别基元的声学模型，这是本文后面研究内容的基准（Baseline）声学模型。

#### 1. 上下文无关的 IF 和 GIF 建模

进行声学建模的训练数据和测试数据集合在 2.1.2 节中进行了说明，后面不再重复叙述。声学特征参数采用 12 维 MFCC 参数＋对数帧能量，并加上它的一阶与二阶差分，共 39 维；识别基元全部采用三个状态、自左向右无跳跃的 HMM 模型拓扑结构；各个基元状态的声学输出概率采用 8 个混合的高斯分布描述，见图 2-1 的表示。其中，最前和最后的两个是虚状态，不产生输出，仅仅用来在搜索解码时连接相邻的 HMM。

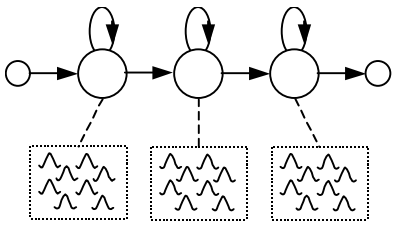

图 2-1. 上下文无关的 IF 和 GIF 的 HMM 模型结构

首先进行上下文无关（CI，Context-Independent）的 IF 和 GIF 建模，其模型分别简称为 CI-IF 和 CI-GIF。识别基元除了 59 个 IF 和 80 个 GIF 之外，还增加了一个静音模型 sil 和三个垃圾模型——NOISE、BURST、NOCH，来分别吸收静音、平稳噪音（如呼吸、背景白噪声等）、突发噪音（如开门、翻书、咳嗽、笑声等），以及非中文或含混不清的发音等，以使得目标模型更加鲁棒。

训练方法为，首先根据训练集中的时间标注，对每个基元以孤立词方式进行训练，以求出其初始的单高斯混合模型参数，之后利用 Baum-Welch（也称为 Forward-Backward）算法[Rabiner89a, Yang95]，对模型进行若干次迭代，直到模型集的参数收敛，即训练集整体的平均对数概率值（似然值）趋于稳定；然后逐步增加状态内的高斯混合数（提高模型的复杂度），并重复利用 Baum-Welch 算法进行迭代训练。如此反复，直到达到期望的高斯混合数（这里为 8），并且使模型的参数收敛。

### 2. 上下文相关的 IF 和 GIF 模型

在连续语音识别中，协同发音现象非常普遍，为此，当训练数据比较充分时，人们一般都会采用上下文相关（CD，Context-Dependent）的方法来为协同发音建模，并取得了良好的识别效果。对于自然随意的连续语音来说，由于发音速度相对较快且多变，以及可能存在的非标准读音或轻微口音背景，使得上下文对发音的影响更为严重，因此，在这里为了进一步提高识别性能，在训练好的 CI-IF 和 CI-GIF 模型基础上，也进行了上下文相关的建模。类似于英文中的 triphone，这里同时采用了左右上下文相关，可以将这些基元形象地称为 tri-IF 和 tri-GIF，将目标模型分别简称为 CD-IF 和 CD-GIF。

对 CD-IF 的训练方法如下：

1) 训练上下文无关的单混合的声韵模型；
2) 对原始数据中标注的上下文无关的声韵母进行上下文扩展，从而得到上下文相关的声韵标注信息。通过对上下文无关模型的简单复制得到上下文相关的声韵模型。然后通过 Baum-Welch 算法的迭代训练来重估参数。
3) 使用基于决策树的状态共享方法进行状态的合并。这样，一个新的基于

状态共享的上下文相关的声韵模型便可生成，训练集中没有出现的上下文相关基元也通过决策树来估计参数。然后通过迭代训练来重估参数。

4) 通过分裂的方法来增加状态中的混合数目，从而细化模型。

对 CD-GIF 的训练过程与此完全相同，只需将基元从 IF 换为 GIF 即可。

其中，基于决策树（Decision Tree）[Breiman84]的状态共享（State Tying）[Young94]方法是建立上下文相关模型的关键。这个方法已被广泛地应用于大词表连续语音识别系统中[Bahl91, Hwang93, Reichl98, Young92]，其目的是通过状态共享解决训练数据相对稀疏的问题，并在模型中合成训练集中从未出现过的上下文相关基元。

决策树的基础是根据先验知识定义的一个问题集合。问题集的选取对于基于决策树的声学建模是很重要的。一些研究者的工作就集中在怎样选取更好的问题集[Willet99, Beulen98]。在本文的实验中，问题集是从音韵学角度选取的[Zhu86, Wu89, Cao90, Luo00, Wu97]。本文的决策树是针对声韵母基元提出的，还有一些研究使用音素作为识别基元，并提出了相应的问题集[Ma00]。一般根据需要，为每个基本的识别基元（例如所有可能的 triphone 中心基元）的每个状态构造一棵决策树。决策树从结构上讲是完全的二分搜索树，每个结点上携带一个合适的问题，根据数据驱动的原则对它们进行确定，即每个结点所选择的问题应从整体上使得将来训练集数据划分出的类别间具有尽可能大的声学差异。

决策树方法对模型间的状态共享一般是通过离线（Offline）的方式对模型进行修正得到的。对每个训练集中出现和未出现（但可能在将来的测试集中出现）的 triphone（这里是 tri-IF 和 tri-GIF）中心基元的各个状态，从它所对应的决策树根结点出发，将这个 triphone 的上下文相关部分的基元与当前结点携带的问题进行匹配，根据匹配与否，分别沿着结点的左分枝或右分枝向下推进，直至到达决策树的叶结点。最后，对所有到达相同叶结点的 triphone 状态内部的参数（例如观测概率分布和状态转移概率等）进行完全的共享。

问题集的内容应既包含一般问题，又包含特殊问题，它们所覆盖的范围可以由宽泛到狭窄，彼此之间亦可重叠或相互包含；同时，这些问题既应该对 triphone 的左上下文进行询问，也应该对其右上下文进行询问。

表 2-9 是本文对 CD-IF 和 CD-GIF 进行建模时所使用的决策树中的问题集（均包含 61 个问题，每个问题都可同时作用于 tri-IF 的左上下文和右上下文）

的一个例子。对 IF 和 GIF 进行上下文相关建模时，它们用来训练决策树并进行相似 HMM 状态间的参数共享。

表 2-9. 用于 tri-IF 和 tri-GIF 决策树的问题集举例

| 模型类别 | 采用的问题集举例 | |
|---|---|---|
| tri-GIF | Initial | b p m f d t n l g k h j q x z c s zh ch sh r |
| | I_Sonorant | m n l |
| | I_Affricate | z zh j c ch q |
| | I_Stop | b d g p t k |
| | I_AspStopAff | p t k c ch q k |
| | I_UnAspStop | b d g |
| | I_AspStop | p t k |
| | I_UnAspAff | z zh j |
| | I_AspAff | c ch q |
| | I_Nasal | m n |
| | F_Open | a o e ai ei er ao ou an en ang eng i1 i2 ng |
| | F_Stretched | i ia ie iao iou ian in iang ing iong |
| | F_Round | u ua uo uai uei uan uen uang ong |
| | F_Protruded | v ve van vn |
| | F_OpenV | a o e ai ei er ao ou i1 i2 |
| | F_OpenN | an en ang eng ng |
| | F_Open_n | an en |
| | F_Open_ng | ang eng ng |
| | F_StretchedV | i ia ie iao iou |
| | F_StretchedN | ian in iang ing iong |
| | ... | ... |
| tri-GIF | Initial | p p_v p_h p_h_v m f f_v t t_v t_h t_h_v n l k k_v k_h k_h_v x x_v ts\ ts_v ts_h ts_h_v s\ s_v ts ts_v ts_h ts_h_v s s_v ts` ts`_v ts`_h ts`_h_v s` s`_v z` |
| | I_Sonorant | m n l |
| | I_Affricate | ts ts_v ts` ts`_v ts\ ts_v ts_h ts_h_v ts`_h ts`_h_v ts_h ts_h_v |
| | I_Stop | p p_v t t_v k k_v p_h p_h_v t_h t_h_v k_h k_h_v |
| | I_AspStopAff | p_h p_h_v t_h t_h_v k_h k_h_v ts_h ts_h_v ts`_h ts`_h_v ts_h ts_h_v |
| | I_UnAspStop | p p_v t t_v k k_v |
| | I_AspStop | p_h p_h_v t_h t_h_v k_h k_h_v |
| | I_UnAspAff | ts ts_v ts` ts`_v ts\ ts_v |
| | I_AspAff | ts_h ts_h_v ts`_h ts`_h_v ts_h ts_h_v |
| | I_Nasal | m n |
| | F_Open | a_″ a_″_h o @ 7 aI ei @` AU @U a_n @_n AN @N I\ i` N |
| | F_Stretched | i ia_″ ia` iE_r iAU i@U iE_n i_n iAN iN iUN |
| | F_Round | u ua_″ uo uaI ua` uei ua_n u@_n uAN UN |
| | F_Protruded | y yE_r y{_n y_n |
| | F_OpenV | a_″ a_″_h o @ 7 aI ei @` AU @U i\ i` |

| | |
|---|---|
| | F_OpenN　　a_n @_n AN @N N<br>F_Open_n　　a_n @_n<br>F_Open_ng　　AN @N N<br>F_StretchedV　　i ia_" ia` iE_r iAU i@U<br>F_StretchedN　　iE_n ia` i_n iAN iN iUN<br>...　　... |

图 2-2 给出了在对 IF 进行上下文相关建模时，由数据驱动方法得到的一棵实际的决策树，用于对上下文相关的声母 h 的中间状态（第 2 个状态）进行共享参数的归类。

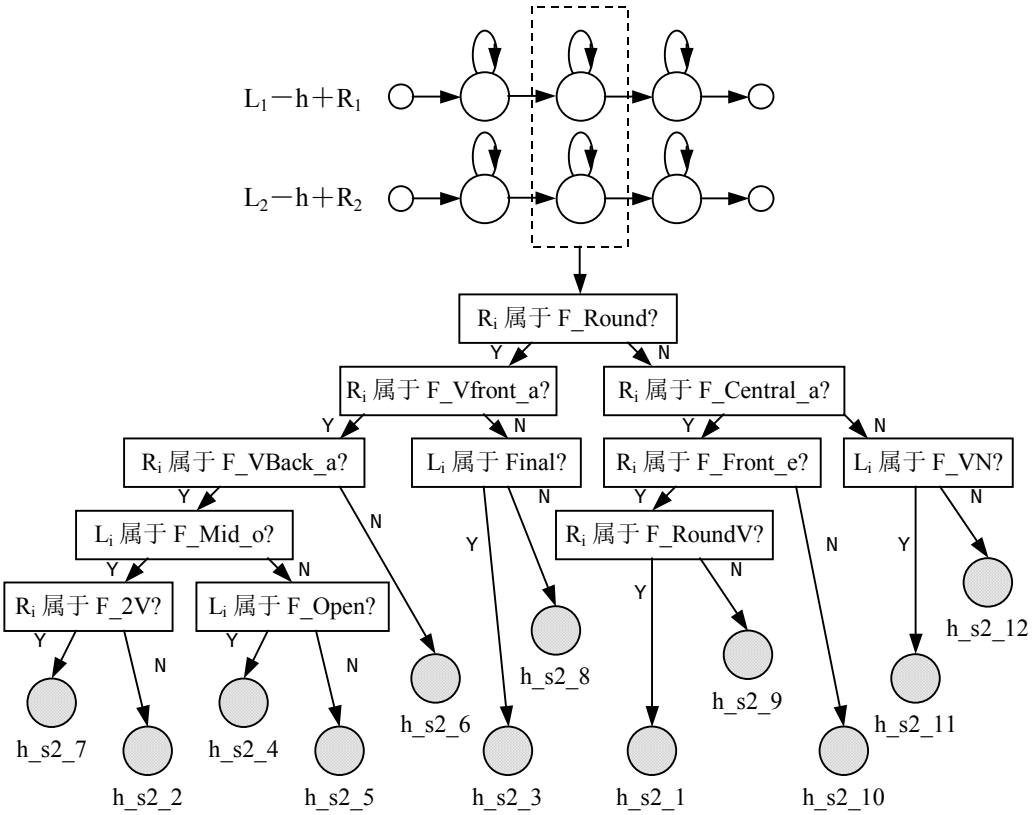


图 2-2. 上下文相关的基元 h 的中间状态的决策树结构

这里是以声母 h 的两个 tri-IF（$L_1$-h+$R_1$ 和 $L_1$-h+$R_1$，其中 h 为中心基元，“－”和“＋”分别表示左、右相关的基元上下文）作为例子的。其中，叶子结点表

示可共享的状态参数（例如高斯混合分布等）集合。当这些状态沿着决策树到达叶子结点时，就可共享叶子结点所表示的参数集合。

通过决策树方法完成模型状态参数共享后，与 CI 模型的训练方法相似，利用 Baum-Welch 算法，对模型进行若干次迭代，直到模型集的参数收敛，然后逐步增加状态内的高斯混合数（提高模型的复杂度），并重复利用 Baum-Welch 算法进行迭代训练。如此反复，直到达到期望的高斯混合数且模型参数收敛。图 2-3 给出了进行状态共享后 tri-IF 和 tri-GIF 的 HMM 模型结构示意图。

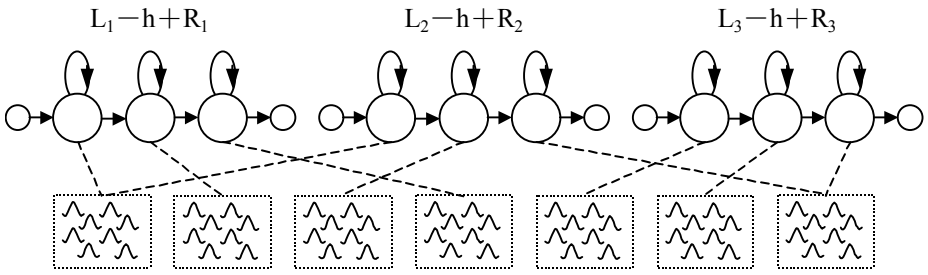


图 2-3. 上下文相关且状态共享的 HMM 模型结构

基于决策树的状态参数共享方法具有如下两个突出的优点：首先，声学上相近状态的参数共享起来，可以起到一种平滑作用，使得每个模型都可得到比较充分的训练，从而减小数据稀疏的影响；其次，对于训练集中没有出现、但将来可能在测试集或其它场合出现的 triphone，也可以通过决策树，让它们的每个状态与某个声学特性相近的、已经充分训练好的状态进行参数共享，从而保证识别性能。

### 2.2.3 对基准模型的实验和分析

在这部分，对基准的声学模型进行如下三方面的比较：

#### 1. 对模型与数据库的匹配程度的比较

首先对 IF 和 GIF 模型本身与训练数据的声学匹配程度进行了比较。这是采

用训练集整体的平均对数概率值（平均似然值）$L(\mathbf{O};\mathbf{M})$ 作为评价指标的。其中，$\mathbf{O}$ 表示训练集的所有训练样本，$\mathbf{M}$ 表示训练得到的 HMM 模型参数集，它既可以是 CI（上下文无关）的，也可以是 CD（上下文相关）的。

声学模型 $\mathbf{M}$ 对于数据集 $\mathbf{O}$ 的平均似然值 $L(\mathbf{O};\mathbf{M})$ 的定义如下：

$$L(\mathbf{O};\mathbf{M}) = \ln P(\mathbf{O} \mid \mathbf{M}) \approx \frac{1}{N}\sum_{i=1}^{N} \ln P(\mathbf{O}_i \mid \mathbf{M}) \tag{2.8}$$

$$P(\mathbf{O}_i \mid \mathbf{M}) = \sum_{S} \left\{ a_{s_{0,1}} \prod_{t=1}^{T_i} b_{s_t}(o_t^i) \cdot a_{s_{t,t+1}} \right\} \tag{2.9}$$

其中，$\mathbf{O}_i = o_1 o_2 \cdots o_T$ 是训练集中的第 $i$ 个训练样本$(1 \le i \le N)$；$\mathbf{S} = s_1 s_2 \cdots s_T$ 是与 $\mathbf{O}_i$ 相关的任意可能的 HMM 状态序列；矩阵 $\mathbf{A} = [a_{ij}]$ 和矢量 $\mathbf{B} = [b_i(\cdot)]$ 分别是与 $\mathbf{S}$ 对应的各个 HMM 模型的转移概率矩阵和观测概率矢量。

表 2-10 给出了在不同模型、不同高斯混和数条件下，模型迭代至收敛时，训练集的整体的似然值。其中，“#GM（Number of Gaussian Mixtures）”表示模型的“高斯混合数”。

表 2-10. 不同的 HMM 相对于训练集的平均似然值

| 模型 \ #GM | 1 | 2 | 4 | 8 |
|---|---|---|---|---|
| CI-IF | -58.75 | -57.78 | -57.01 | **-56.25** |
| CI-GIF | -58.56 | -57.62 | -56.81 | **-56.01** |
| CD-IF | -57.42 | -56.30 | -55.28 | **-54.13** |
| CD-GIF | -57.35 | -56.27 | -55.20 | **-54.08** |

从上表可以看出如下趋势：

1) 随着 HMM 状态内高斯混和数的增加，同一模型使得训练集的平均似然值不断增大。这是因为更多的高斯混合可以更加精确的拟合输出概率，从而增大似然值；

2) 相同高斯混合数时，同类 HMM 的 CD 模型的似然值总是高于 CI 的。这是因为 CD 模型根据不同的上下文对模型进行了细化（同时利用决策

树进行状态参数共享，在一定程度上解决了数据稀疏问题），从而增加了模型的描述精度；

3) 无论是 CI 的，还是 CD 的，对 GIF 建模使得训练集的平均似然值始终高于对 IF 的建模，这说明 GIF 模型可以更好地与训练数据相匹配。

其中，第三点趋势是非常重要的。这说明，由于 GIF 集合在确定时，充分考虑到了自然语音中音素的频繁插入、删除和替换现象，并为此而采用了包含多种可能发音变化形式的多发音词典，因而使得 GIF 模型更细致地体现出了对发音变化的描述能力。这也说明，从 IF 基元集合扩展到 GIF 基元集合，是非常现实的，也是很有必要的。

对测试集进行 Viterbi 搜索识别解码，即将公式(2.9)改为如下的(2.10)，得到了同样趋势的曲线和规律，见表 2-11。其中 CI 模型和 CD 模型的每个 HMM 状态内部均采用 8 个混合的高斯分布描述。

$$P(\mathbf{O}_i \mid \mathbf{M}) = \max_{s}\left\{ a_{s_{0,1}} \prod_{t=1}^{T_i} b_{s_t}(o_t^i) \cdot a_{s_{t,t+1}} \right\} \tag{2.10}$$

表 2-11. 对测试集进行识别解码得到的平均似然值

| CI-IF | CI-GIF | CD-IF | CD-GIF |
|---|---|---|---|
| -57.85 | -57.58 | -57.04 | -56.89 |

显然，由于识别解码时存在着对 IF 或 GIF 的误识，因此这些模型对测试集整体的平均似然值要比对训练集的平均似然值略低一些，但仍然可以得出结论，无论是 CI 的，还是 CD 的，GIF 模型加上基于 GIF 的多发音词典，从总体上对自然语音的匹配程度都要优于采用标准 IF 模型及其单发音词典的方案。

### 2. 对音节识别率的比较

然后对训练完成的基准声学模型（CI-IF，CI-GIF，CD-IF，CD-GIF）进行了识别测试。进行识别的搜索解码时，CI-IF 和 CD-IF 采用基于 IF 的单发音词典，CI-GIF 和 CD-GIF 采用基于 GIF 的多发音词典，这些词典用来构造音节或

识别基元的搜索网络。

在统计识别结果时，将采用如下标记来衡量纯声学的识别性能：

- 音节识别正确率 SCR（Syllable Correct Rate）
- 音节识别错误率 SER（Syllable Error Rate）
- 音节误识相对下降率 SER↓（Syllable Error Rate Reduction）
- 基元识别正确率 UCR（Unit Correct Rate）

其中，SER = 100%－SCR，而 SER↓是 SER 的相对下降百分比，即当 $SER_0$ 是基准系统的 SER 声学识别结果时，SER↓ = ($SER_0$-SER) / $SER_0$。此外，UCR 表示进行自由识别（识别解码网络不受发音词典的限制）时的基元正确率。

表 2-12 给出了上述基准模型的音节和基元的识别率。

表 2-12. 基准模型的音节与基元识别结果(%)

| 模型 / 识别率 | CI-IF | CI-GIF | CD-IF | CD-GIF |
|---|---|---|---|---|
| SCR | 32.81 | 33.06 | 44.98 | 45.93 |
| UCR | 46.03 | 43.66 | 54.72 | 52.19 |

从上表中可以看到如下趋势：

1) 无论是 CI 还是 CD 的模型，GIF 的音节识别率都比 IF 的要略高，而且从 CI 模型到 CD 模型，无论是 IF 还是 GIF，其音节识别率都有一个较大的跳跃（>12%）；

2) 但无论是 CI 的还是 CD 的模型，GIF 的基元识别率都要略低于 IF 的基元识别率。

与前面的训练集和测试集的似然值变化趋势类似，从第 1 点中的声学识别率对比来看，基于 GIF 的建模也是要优于基于 IF 的建模的。而第 2 点的现象也并不奇怪，这是因为 GIF 集合是 IF 集合的超集，并且在统计识别率时，某些从音韵学角度来看，差别非常细小的 GIF 基元之间并没有被看作是等同的（例如 *GIF* $a\_'' \neq a\_''\_v$），于是从表象上看来 GIF 的 UCR 就小于 IF 的 UCR，但由于使用了适当构造的 GIF 多发音词典，并包容了必要的发音变化，从而使得 GIF 总

体的 SCR 均要高于 IF 的 SCR。

从总体上看，IF 或 GIF 的识别率都不算很高，这是因为 CASS 本身是一个极其口语化的数据库，其中包含了大量有意义的发音变化现象，但同时这些发音的过分“随意性”也造成了自动识别的困难，因为其中的部分语句即使由人来听，也很难准确地判断出其规范的发音应该是什么。另外，由于 CASS 数据库是在教室、礼堂、会议室等非专业环境里进行的录音，因此不可避免地含有大量的背景噪声，但噪声的问题并不是这里所关注的内容，因此建模时没有进行任何去噪或相关的鲁棒性处理，从而在一定程度上影响了建模与识别的精度。

### 3. 与 IF 识别混淆矩阵方法的比较

从上述的结果分析中，很容易产生一个疑问，即若不考虑 GIF 对 IF 的扩充能力，仅仅考虑其中多发音词典的作用，是否利用 IF 的识别混淆矩阵构造出一个基于 IF 的多发音词典，也可以达到优于基于单发音词典的传统 IF 建模甚至接近或优于 GIF 建模方法呢？

为此，尝试由已经训练好的 IF 的 CD 模型出发，利用类似于识别混淆矩阵方法，对训练数据进行基元层面的识别解码，然后从得到的 IF 层面的新标注中构造出包含了一定发音变化现象的、基于 IF 的多发音词典，并进行相关的识别测试，实验结果见表 2-13。

表 2-13. 为 IF 构造多发音词典后的识别结果(%)

| 模型 / 识别率 | CI-IF | CD-IF |
|---|---|---|
| SCR | 32.28 | 43.36 |
| UCR | 45.63 | 54.31 |

与表 2-12 相比可以看出，基于混淆矩阵方法构造发音词典后，识别率均要比 IF 与 GIF 的基准系统低一些。这是因为：

1) 与 GIF 基准相比，多发音词典的 IF 虽然考虑到插入、删除，甚至一部分完全音变，但没有考虑到大量的细微音变（如 SAMPA-C 标注中的后

缀表示），造成了模型与实际语音间的匹配有相对较大的失真；

2) 与 IF 基准相比，采用多发音词典后，虽然引入了发音变化，增强了对自然语音中音变的刻画能力，但同时也增大了音节间的混淆度，这个因素若不加以适当的处理，有可能会抵消引入发音变化带来的优点，甚至会降低系统的识别率。这一点将在第四章进行详细讨论；

3) 在构造GIF的多发音词典时，是根据专业工作者的手工标注统计得到的，因而在基元层面上看，这些发音变化项的误差较小。但在基于混淆矩阵的发音词典构造方法中，由于 IF 层面具有不可避免的识别错误，就会造成词典中可能会包含一些错误的发音变化项，并因而影响多发音词典的精度，使得整体识别误差被放大。

综上所述，基于 GIF 集合及其相关的多发音词典相对于 IF 集合及其传统的单发音词典的若干优势，后面的研究将主要集中在提高 GIF 的建模能力并完善多发音词典等方面。

## 2.3 对基准系统的修正

对于 CASS 这类非常随意自然发音的口语数据库，某些似是而非的发音在实际的标注中，不可避免地会有一定的偏差，例如在文献[Li00c]中，四位语言和语音学领域的专业工作者同时独立地对 CASS 的一个子集（15 分钟的发音数据）进行了仔细的标注，结果他们给出的 SAMPA-C 标注间的一致性仅有 80－90％!

虽然客观上存在着 GIF 标注的不一致性问题，但经过观察发现，这些不一致之处是很细微的，例如大多数是具有不同“后缀”的“部分音变”的差别，例如 *a_"*与 *a_"_v* 等，而“完全音变”或插入、删除的差别相对较少。

为了消除这类标注的不一致性，并进一步提高系统的识别性能，可以利用已经训练好的模型，对整个基准系统进行一定程度上的修正。这个过程是完全数据驱动的。包括如下三个方面的修正：

1) 修正 CASS 标注。根据音节层面的标注，利用 HMM 模型（这里采用精度较高的 CD-GIF）和多发音词典，对训练集进行强制的时间对准，并输出 GIF 层面的新标注；

2) 修正多发音词典。通过对修正后的 GIF 层面的 CASS 新标注，重新统计多发音词典中各个音节的发音变化项，以及相应的权重；

3) 修正基准模型。利用修正后的标注，重新训练基于 GIF 的模型（CI-GIF 和 CD-GIF）。

然后，利用修正后的基准模型和多发音词典，重新对测试集进行识别，结果见表 2-14。

表 2-14. 修正后的基准模型识别结果(%)

| 模型 / 识别率 | CI-GIF | CD-GIF |
|---|---|---|
| SCR | 33.18 | 46.07 |
| UCR | 43.64 | 52.17 |

与表 2-12 相比可以看出，对基准系统进行修正后，音节识别率略有提高，而 GIF 基元的识别率却略有下降，但升高和下降的幅度均非常小。后续的实验将采用修正后的 CASS 标注和多发音词典，并以表 2-12 的 IF 识别结果与表 2-14 的 GIF 识别结果作为比较的基准。

本论文的研究目标，是希望通过对总体识别率的“相对”提高，来验证所提出的新方法的有效性及其对声学识别的“整体”改善能力。因此在后面的研究中，将不再考虑基元层面的识别结果 UCR，而只是对音节层面的识别结果进行比较，作为对总体声学识别性能的评价指标。即后续实验对识别性能的统计将仅仅包括 SCR、SER 和 SER↓。

为便于查阅，将上述基准模型的识别结果统一列在表 2-15 中，同时，在表 2-16 中给出了各个模型识别结果之间的 SER↓值。

表 2-15. 基准模型的识别结果汇总(%)

| 模型 / 识别率 | CI-IF | CI-GIF | CD-IF | CD-GIF |
|---|---|---|---|---|
| SCR | 32.81 | 33.18 | 44.98 | 46.07 |
| SER | 67.19 | 66.82 | 55.02 | 53.93 |

表 2-16. 基准模型的 SER↓值对照表(%)

| 起始模型 | 目标模型 | SER↓ |
|---|---|---|
| CI-IF | CI-GIF | 0.55 |
| CD-IF | CD-GIF | 1.98 |
| CI-IF | CD-IF | 18.11 |
| CI-GIF | CD-GIF | 19.29 |

# 第三章　对广义声韵集的精细建模

## 3.1　问题的提出

在上一章中对基准的声学模型进行了讨论。考虑变量 $b$、$s$ 和 $a$ 分别表示标准音节（或标准声韵 IF）、广义音节（或广义声韵 GIF），以及实际的声学输出矢量。第 2.2.2 节指出，当把 $(b,s)$ 各种可能出现的组合作为识别基元时，其集合容量将大大增加，现有训练数据也因此显得极其不充分。无论是基于 IF 的建模，还是基于 GIF 的建模，都是对发音模型 $P(a\,|\,b,s)$ 的近似，也是对训练数据稀疏问题的一个折衷。但从另一个角度来看，如下假设是合理的（见 2.1.3 节：IF 集合与 GIF 集合间的多对多的映射关系），即对于不同的上下文，

1) 由同一个 IF 发生音变得到的不同 GIF 之间的声学输出特性是不同的；
2) 同一个 GIF，根据发生音变而得到它的不同的 IF，其声学输出特性也是有一定差异的。

因此，将 $(b,s)$ 作为识别基元，可以更好地体现出发音的变化。这里将 $(b,s)$ 集合形象地记为 IF-GIF 集合，并把基于 IF-GIF 的建模称为对 GIF 的精细（Refined）建模。显然，对 GIF 进行精细建模与对 IF 或 GIF 的基准建模相比，将是一个更好的选择，但其前提是要有比较充分的训练数据，然而，对于 CASS 来说，这个条件并不具备。

虽然 IF-GIF 的训练数据对于 IF-GIF 基元来说很稀疏，但是对于 IF 或 GIF 基元来说，却相对地比较充分。为了在声学模型中更好地体现出发音的变化，并弥补训练数据量不足的矛盾，本章尝试在现有 IF 和 GIF 基准模型的基础上，利用模型自适应技术来解决数据稀疏问题，从而提出了对 GIF 集合进行精细建模的方法。

按照这个思路，对 GIF 进行精细建模将涉及到如下三个主要问题：

1) 如何产生与 IF-GIF 基元集合相对应的训练数据标注；
2) 如何对 IF 和 GIF 基准模型进行自适应，以得到较理想的精细模型；

3) 如何构造音节到 IF-GIF 序列的多发音词典，供识别搜索时使用。

在 3.2 节中首先讨论对 GIF 进行精细建模的理论依据，然后在 3.3 节中对上述三个主要问题的解决过程进行介绍。

## 3.2 精细建模的理论依据

### 3.2.1 建模的总体思路

根据汉语的发音的特性，标准音节都是由声母和韵母组成的（零声母情况将在后面讨论）。假设 $b=(i,f)$，$s=(g_i,g_f)$，其中 $b$ 和 $s$ 分别代表音节 S 和广义音节 GS 的声韵结构，$i$ 和 $f$ 分别代表 S 中的声母和韵母，而 $g_i$ 和 $g_f$ 分别代表 GS 中其对应的 GIF（广义声母和广义韵母）。为了简化问题，假设声母和韵母之间的声学输出是彼此独立的，上述精细发音模型可以表示为：

$$P(a\mid b,s)\approx P(a\mid i,g_i)\cdot P(a\mid f,g_f) \tag{3.1}$$

其中 $a$ 为声母或韵母部分的实际声学输出矢量。于是对 GIF 进行精细建模的问题就转化为对现有的 IF 和 GIF 基准模型进行修正的问题了。

根据 3.1 节的讨论，一个标准的 IF 可以音变为不同的 GIF，而同一个 GIF，也可以是由不同的标准 IF 发生音变得到。对于有限的训练数据来说，虽然对 IF-GIF 基元建模将显得数据非常稀疏，但是对于 IF 或 GIF 基元集合来说却并不稀疏。因此，可以从已经得到比较充分训练的 IF 模型或 GIF 模型出发，借鉴模型自适应技术的思想，将它们的模型参数向与之相关的 IF-GIF 模型集偏移，从而得到更加精细的模型。

### 3.2.2 模型自适应技术简介

自适应技术的目的是利用较少的注册（Enrollment）数据，来修正初始模型或者特征参数，以减小训练集和测试集之间的差异，提高系统在不同应用环境

下的鲁棒性。通常，自适应技术可以应用在特征参数和声学模型两个层次，分别用来减小当前应用环境与训练条件在特征参数空间和声学模型空间的偏差；这里的环境差异包括不同说话人、背景噪声和传输信道等等。此外，还可以利用语言模型的自适应技术，减少不同领域文本语料的概率分布特性的统计差异。

本文正是借鉴了声学模型自适应技术的思想来解决训练数据的稀疏问题并对 GIF 进行精细建模。其中，比较成熟的声学模型自适应技术包括 MLLR（最大似然线性回归，Maximum Likelihood Linear Regression）和 MAP（最大化后验概率，Maximum a Posteriori）方法以及将二者结合在一起的方法等。

对于采用多高斯混合的连续 HMM 模型集来说，最重要的待自适应的参数就是输出概率分布的高斯混合参数，即均值矢量与协方差矩阵（通常情况下取对角矩阵形式）。MLLR 方法通过一系列线性变换矩阵，将初始模型参数映射到一个新的参数空间，在变换矩阵的估计中采用了最大化自适应数据的似然值的方式。

MLLR 对均值矢量的自适应变化矩阵的估计方法为：

$$\hat{\mu} = \mathbf{W}\xi, \qquad \xi = [w, \mu_1, \mu_2, \cdots, \mu_n]^T$$

其中，$n$ 为特征矢量的维数，$w$ 是偏移分量标识，$\mathbf{W}$ 是 $n \cdot (n+1)$ 维的变换矩阵，可以表示为 $\mathbf{W} = [b, \mathbf{A}]$，其中 $\mathbf{A}$ 是待估的 $n \cdot n$ 维变换矩阵，当 $w$ 为 1 时，$b$ 代表偏移矢量。

MLLR 对协方差的自适应变化矩阵的估计方法为：

$$\hat{\mathbf{\Sigma}}_m = \mathbf{B}_m^T \mathbf{H}_m \mathbf{B}_m, \qquad \mathbf{\Sigma}_m^{-1} = \mathbf{C}_m \mathbf{C}_m^T, \qquad \mathbf{B}_m = \mathbf{C}_m^{-1}$$

其中，$\mathbf{H}_m$ 是待估计的线性（对角形式）变换矩阵，$\mathbf{C}_m$ 是 $\mathbf{\Sigma}_m$ 逆矩阵的 Choleski 因子，$\mathbf{B}_m$ 是 $\mathbf{C}_m$ 的逆矩阵。

MLLR 的变换矩阵 $\mathbf{W}$ 和 $\mathbf{H}$ 都是通过期望值最大化（EM，Expectation Maximisation）[Gales96b, Dempster77]方法，对一个标准辅助函数进行最大化来获得的。MLLR 还根据可用的自适应数据量的多少来为 HMM 模型集合构造一个二分回归树（binary regression class tree）[Gales96b]，依据各个状态或高斯混合间的相似性对其进行分类，为每个类统一进行上述的线形变换，从而保证了自适应过程

的鲁棒性和灵活性。对 MLLR 的详细叙述见文献[Gales96a, Leggetter95]。

MAP 方法又被称为贝叶斯（Bayesian）自适应。它需要根据先验知识，对模型参数的分布形式给出假设，并将初始模型信息作为模型参数的先验估计纳入估计式，与有限的自适应数据中蕴含的与当前应用相关的信息结合，尤其适用于训练数据稀疏的条件。

对基于混合高斯分布的 HMM，对于状态 $j$ 内的第 $m$ 个高斯混合，其 MAP 重估公式为

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\overline{\boldsymbol{\mu}}_{jm} + \frac{\tau}{N_{jm}+\tau}\boldsymbol{\mu}_{jm}$$

其中，$\tau$ 是根据对自适应数据的先验知识确定的一个权重，该权重决定了自适应的调整幅度。N 是自适应数据的出现概率（或似然值）。对 MAP 的详细叙述见文献[Lee93, Lee91, Gauvain94]。

由于基于 MAP 的自适应方法是直接针对各个高斯混合参数进行的，因此与 MLLR 相比，它需要更多的自适应数据。在某些情况下，结合 MLLR 与 MAP 共同进行自适应，会取得比单用 MLLR 或单用 MAP 更好一些的识别性能。

## 3.3　精细建模的实现方法

### 3.3.1　自动标注的生成

为了对广义声韵 GIF 的基元集合进行精细建模，首先要产生相应的 IF-GIF（声韵—广义声韵）标注。正如 2.1.2 节指出的，CASS 数据库包含了实际发音中的音节标注和 GIF 标注，并且根据音节标注与标准的单发音词典可以导出理想情况下的 IF 标注，而实际发音的声韵标注（这里记为 IF-a）是自动产生的，其方法将在这里介绍。

在自然语音中，由于存在着大量的发音变化，使得标注 IF 序列与 GIF 序列并不是一一对应的。例如，某些 IF 直接发生音变而生成了对应的 GIF；某些 IF

在发音时被删除（吸收），反映在 GIF 序列中是一个空缺的位置，而在某些情况下，由于协同发音现象以及发音器官的惯性，使得某些 GIF 被插入，反映在 IF 序列中也是一个空缺的位置。为了生成 IF-GIF 的标注，就需要知道与 GIF 序列对应的原本的发音序列，即 IF-a 标注。

表 3-1 以举例的形式说明了 IF-a 标注和 IF-GIF 标注的生成规则：

表 3-1. IF-a 与 IF-GIF 标注的生成

| 基元类型 | 标注信息 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *IF* | $i_1$ | $f_1$ | $i_2$ | $f_2$ | $i_3$ | | $F_3$ | ... |
| *GIF* | $gi_1$ | $gf_1$ | $gi_2$ | | $gi_3$ | $gif_4$ | $gf_3$ | ... |
| *IF-a* | $i_1$ | $f_1$ | $i_2$ | | $i_3$ | $if_4$ | $F_3$ | ... |
| *IF-GIF* | $i_1$-$gi_1$ | $f_1$-$gf_1$ | $i_2$-$gi_2$ | | $i_3$-$gi_3$ | $if_4$-$gif_4$ | $f_3$-$gf_3$ | ... |

首先需要根据标准的 IF 序列和实际的 GIF 序列以动态规划的方式进行对齐，然后即可获得 IF-a 标注，其要点为，根据 GIF 序列中每个存在 GIF 的位置，将原始的 IF 标注进行拷贝，其中如果 GIF 是属于发音的插入现象，则将它映射到一个缺省的、最相近的 IF 基元；最后，将 IF-a 序列与 GIF 序列对应位置一一拼接，即可得到所需的 IF-GIF 标注。

### 3.3.2 精细模型的生成

正如前面所述，可以通过在现有的 IF 和 GIF 模型基础上，通过模型自适应技术来对 IF-GIF 进行精细的建模。这个思路是基于一个客观存在的事实，即 IF 集合与 GIF 集合在发音变化的环境中是多对多的关系，以及两种可能的、比较合理的假设，即：

1) 由相同 IF 发生音变得到的不同 GIF 之间的声学输出特性差异不大，或

2) 由不同 IF 发生音变得到的相同 GIF 之间的声学输出特性差异不大。

因此，利用模型自适应技术来偏移原有的 IF 或 GIF 参数，就可以在 IF-GIF 模型中体现出上述微小的差异，从而达到精细建模的目的。

**1. 构造初始的精细模型**

根据上述两个假设，可以得到两种相应的构造初始精细模型的方法，分别命名为 B-GIF（Baseform IF to IF-GIF）方法和 S-GIF（Surface-form IF to IF-GIF）方法。

用 B-GIF 方法构造初始模型的出发点为：同一个 IF 在发生音变时，可以变为不同的 GIF，如图 3-1 所示。

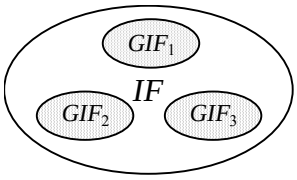


图 3-1. 用 B-GIF 方法构造初始的精细模型

其具体方法可以分为两步。首先根据自动生成的 IF-a 标注，重新训练新的 $N_{IF}$ 个 IF 模型的集合 $\{B_i : 1 \le i \le N_{IF}\}$，注意，它与根据标准 IF 标注生成的 IF 模型集略有不同。然后，对每个 IF 模型 $B_i$，根据它所有可能由于发音变化而变为它的 $N_{Bi}$ 个 GIF 的集合 $\{S_{ij} : 1 \le j \le N_{Bi}\}$ 进行一一拷贝，从而生成初始的 IF-GIF 模型集合 $\{B_i - S_{ij} : 1 \le i \le N_{IF}, 1 \le j \le N_{Bi}\}$。

而用 S-GIF 方法构造初始模型的出发点为：同一个 GIF，可以由不同的 IF 发生音变后得到，如图 3-2 所示。

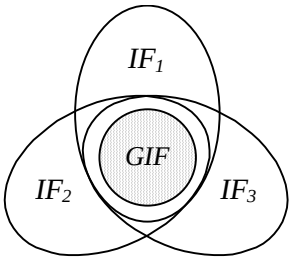


图 3-2. 用 S-GIF 方法构造初始的精细模型

其具体方法也可以分为两步。首先可以直接根据 CASS 的 GIF 标注训练出包含了 $N_{GIF}$ 个 GIF 模型的集合 $\{S_j : 1 \le j \le N_{GIF}\}$，然后对于每个 GIF 模型 $S_j$，根据所有可能由于发音变化而变为它的 $N_{Sj}$ 个 IF 的集合 $\{B_{ij} : 1 \le i \le N_{Sj}\}$ 进行一一拷贝，从而生成初始的 IF-GIF 模型集合 $\{B_{ij} - S_j : 1 \le j \le N_{GIF}, 1 \le i \le N_{Sj}\}$。

**2. 利用自适应方法修正精细模型**

根据 B-GIF 方法或 S-GIF 方法生成的初始 IF-GIF 模型集合，实际上与原来的 IF 或 GIF 是等价的。为了进一步体现出它们之间的精细变化，这里采用上面介绍的模型自适应技术来对上述初始模型进行修正。

具体方法为，采用自动生成的 IF-GIF 标注，将训练集中所有样本都作为自适应数据，利用 MLLR 和 MAP 相结合的方法进行有监督（Supervised）方式的自适应：先根据对 IF-GIF 基元在训练集中所占帧数的统计及其彼此的声学相似性，构造一个适当规模的二分回归树来对 HMM 的状态进行分类，并为每一类进行 MLLR 自适应，以获得相应的参数线性变换，之后再进行 MAP 自适应，使得不同状态的高斯混合可以分别获得更加精细的模型参数偏移。

### 3.3.3 多发音词典的构造

为了对 IF-GIF 模型进行声学识别测试，需要生成相应的多发音词典，即标准音节到 IF-GIF 序列的影射表。其生成方法与 3.3.1 中自动标注的生成方法完全相同，做法是：利用现有的基于 GIF 的多发音词典，根据其中每行的音节符号，从基于标准 IF 的单发音词典中抽取出对应的 IF 序列，以相同的方法将其变换为 IF-a 序列，然后再与同一行的 GIF 序列组成 IF-GIF 序列。

表 3-2 举例说明了基于 IF-GIF 序列的多发音词典的生成过程。

表 3-2. 基于 IF-GIF 的多发音词典的生成

| 基于 IF 的单发音词典 | 基于 GIF 的多发音词典项 | 基于 IF-GIF 的多发音词典项 |
|---|---|---|
| hou h ou | hou 0.4410 x @U<br>hou 0.2889 @U<br>hou 0.2701 x_v @U | hou 0.4410 h-x ou-@U<br>hou 0.2889 ou-@U<br>hou 0.2701 h-x_v ou-@U |
| a a | a 0.7487 a_"<br>a 0.2302 a_"_h<br>a 0.0211 z` a_" | a 0.7487 a-a_"<br>a 0.2302 a-a_"_h<br>a 0.0211 r-z` a-a_" |

考虑到对初始的 IF-GIF 模型进行自适应之后，虽然模型的精度有了一定程度的提高，但是从 IF 和 GIF 标注自动生成的 IF-GIF 标注中的时间边界信息的准确程度降低了。因此在完成自适应后，需要利用新的 IF-GIF 模型以及这个新的 IF-GIF 多发音词典，对训练集的所有数据进行一次强制性的时间对准，以产生新的 IF-GIF 标注和更为精确的时间边界信息。

然后，利用标准的 Baum-Welch 算法，对整体的 IF-GIF 模型参数进行重估。考虑到IF-GIF基元的训练数据稀疏性，在重估时仅仅更新各个高斯混合的均值，而不改变其它的协方差或转移概率矩阵等参数，以获得较鲁棒的模型参数。

## 3.4　实验结论与分析

表 3-3 给出精细模型刚刚完成自适应以及进行均值重估后（每个 HMM 状态内均保持 8 个高斯混合分布），训练集的平均似然值的变化情况。

表 3-3. 精细模型的训练集平均似然值的变化

| 过程 / 模型 | 仅进行自适应 | 再进行均值重估 |
|---|---|---|
| B-GIF 模型 | -56.15 | **-56.05** |
| S-GIF 模型 | -55.97 | **-55.93** |

与表 2-10 相比可以看出，进行模型自适应以及均值重估后，无论是 B-GIF 还是 S-GIF 模型，都使得训练集的整体概率值较其初始模型有所提高。由于

B-GIF 是由 IF 模型初始拷贝而来，因此这里 B-GIF 与 CI-IF 具有可比性；同理，由于 S-GIF 是由 GIF 模型初始拷贝而来，因此这里 S-GIF 与 CI-GIF 具有可比性。这说明，通过这种对精细模型进行构造的过程，使得其模型参数与训练数据的匹配程度有了进一步的提高。

利用精细的 IF-GIF 模型，以及相应的基于 IF-GIF 基元的多发音词典，进行了与基准的 GIF 模型同等条件下的识别测试。结果见表 3-4。

表 3-4. 对 IF-GIF 精细模型的识别测试(%)

| 识别率 / 模型 | | SCR | SER | SER ↓ |
|---|---|---|---|---|
| 基准的 GIF 模型（CI-GIF） | | 33.18 | 66.82 | —— |
| 仅进行自适应的 IF-GIF | B-GIF | 33.60 | 66.40 | 0.63 |
| | S-GIF | 33.37 | 66.63 | 0.28 |
| IF-GIF＋自适应＋均值重估 | B-GIF | **34.44** | **65.56** | **1.89** |
| | S-GIF | 33.66 | 66.34 | 0.72 |

从表 3-4 中可以看出，对 GIF 进行精细建模得到的 IF-GIF 模型要比基准 GIF 模型的识别性能好一些，这是因为它既考虑了 GIF 的不同来源（不同上下文中的 IF）之间在发生音变时的细微差异，使得声学模型更加精确，又利用了模型自适应技术对相对训练得比较充分的 IF 和 GIF 模型参数进行偏移得到 GIF 的精细模型，从而避免了训练数据的稀疏问题。

另外，以 B-GIF 方式初始化的 IF-GIF 模型要比以 S-GIF 方式初始化的 IF-GIF 模型的识别性能略好。这个现象并不是很容易给出定性的解释，这是因为在实际的自然发音中，IF 与 GIF 之间的映射关系是很复杂的，从直观上，相当于把图 3-1 和图 3-2 进行结合或重叠。而且，自适应后识别性能的好坏，与自适应前模型内部各个高斯混合的内聚性或分散程度之间，并没有必然的正向或反向趋势，而是依赖于具体的自适应数据分布情况。因此，实验结果中 B-GIF 与 S-之间的差异仅仅是数据驱动意义上的，本文不再进行深入的定性分析，但总的趋势，即对 GIF 进行精细建模的有效性，通过实验得到了验证。

对 GIF 进行精细建模的思路，并没有扩展到上下文相关（即对 tri-GIF 进行

精细建模）的情况，这是因为 tri-GIF 本身的模型规模就已经很大，若再进行精细建模，将使其规模过分增大、结构更加趋于复杂。

此外，对 GIF 进行精细建模时，仅仅从一个 IF 可以变为多个 GIF 以及同一个 GIF 可以由多个 IF 发生音变而得到的角度进行考虑，但具体在什么情况下会产生这些精细的发音变化，即自然语音的上下文信息如何影响音变的发生，却尚未顾及。有关这些方面的问题将在第四章和第五章中讨论。

# 第四章　基于上下文的词典加权策略

## 4.1　问题的提出

### 4.1.1　发音词典对识别性能的影响

在本章中，重点研究发音词典的规模及其构成方式等对发音建模的识别性能的影响。

在针对自然随意语音的发音建模研究中，通过在数据库标注和发音词典中引入适当的发音变化，有助于提高声学模型对子词基元的分辨能力并提高整体的识别性能。但反过来，在词典中引入发音变化，同时又会增加词或音节间的混淆程度（这是因为词典中会不可避免地存在某些相同的发音变化项），反而有可能会降低识别率。

为了从直观上说明这个问题，采用与基准系统相同的测试集，以基准的 CD-GIF 作为声学模型，针对基于 GIF 的多发音词典进行了如下不同条件下的识别测试。

1) 为发音词典中每个音节保留最可能的、累积概率约为 90%的发音变化。这与第二章中基准模型采用的多发音词典相同；
2) 为每个音节在词典中保留所有可能的发音变化，及其对应的出现概率；
3) 为发音词典中的每个词仅仅保留最高频的，即最可能出现的一种发音形式（近似于单发音词典）；
4) 为每个音节在发音词典中保留累积概率为 90%的发音变化，但去除每个发音序列的出现概率（视作等概率情况）。

在识别前，对多发音词典中各个发音序列的出现概率均进行了归一化处理。表 4-1 给出了识别结果。

表 4-1. 不同发音词典对识别性能的影响(%)

| 识别率 / 发音词典 | SCR | SER | SER ↓ |
|---|---|---|---|
| 基准，含 90% 发音变化 | 46.07 | 53.93 | —— |
| 全部（100%）发音变化 | 45.93 | 54.07 | **- 0.26** |
| 仅含最可能的（单）发音 | 44.82 | 55.18 | **- 2.32** |
| 90%发音变化，等同概率 | 44.90 | 55.10 | **- 2.17** |

从表 4-1 中可以看出，当采用上述各种发音词典的变化形式后，其识别率均要低于基准的、包含了 90%发音变化的、含概率的多发音词典。下面对这些现象进行分析。

将多发音词典简化为单发音词典时，自然语音中的一些很常见的、很重要的发音变化（如替换、删除和插入等）就不能在识别搜索的网络中体现出来，从而降低了识别器对自然语音的匹配能力以及整体的识别率。

但当在发音词典中包含了全部可能的（100%）发音变化时，会有大量相同的发音序列存在，而这些相同的发音序列又分别对应着不同的音节。于是在不借助于任何语言模型的前提下，当这些实际的发音序列在自然语音中出现时，在音节的层面必然发生混淆，从而降低了整体的识别性能。

另一个很值得注意的现象是，在采用同等规模（这里采用具有 90%累积概率）的发音词典时，当去除其中的概率（即将所有的发音变化看作是同等概率出现的）时，识别率会有很大的下降。这是因为不同发音变化的出现频度是非常不均衡的。将它们不加区别地看待，只会造成音节间的混淆度增大，使识别性能下降。

由此可以看出，在自然语音识别的任务中，发音词典的规模（即它所包含的发音变化序列的多少）以及对发音序列出现概率的取舍和使用方式，对识别器的整体性能的影响是很大的。

### 4.1.2 基准的多发音词典的不足

在第二章的基准系统中采用了基于 GIF 的多发音词典，它为每个音节包含了累积概率约为 90%的发音变化项，且这些初始概率是由 CASS 数据库的标注直接统计得到的，即为 $P(s|b)$，其中 $b$ 为（音节层标注的）标准音节 S，$s$ 为 $b$ 的某一种可能的变化形式，即在 GIF 层实际出现的广义音节 GS。

为便于讨论，在此将这个概率定义为直接输出概率（DOP，Direct Output Probability）。为了使得 DOP 更加精确，根据数据驱动的方法对这个作为基准的多发音词典进行了一定程度上的修正，具体过程见 2.3 节，并根据实验确定了一个相对较优的发音词典规模（这里是取发音变化项的累积概率约为 90%）。

根据上节的分析，发音词典中各个发音变化形式的概率加权值对于降低音节间的混淆程度和提高识别率都是非常重要的。但一个不容忽视的问题是，即使进行了上述修正，仍不能保证 DOP 估计的准确性。这是因为标准音节的集合相对较大，于是在训练数据中每个音节出现的频度较小，而且很不均衡，至于每个音节 $b$ 的可能变化形式 $s$ 的出现频度，则更加小而且更加不均衡。因此这里 DOP 统计的不精确性，必定会对整体的识别性能产生一定的负面影响。

### 4.1.3 改进发音词典的思路

上节对 DOP 不足的原因的分析可以归结两点。首先，是训练数据在音节层面的稀疏问题导致 DOP 统计的不精确；其次，对 DOP 的统计完全没有考虑到上下文信息，然而在自然语音中，各种发音变化的出现频度，恰恰是与不同的上下文环境密不可分的，这也是影响发音词典精度的一个重要因素。

但是，音节层面的数据稀疏并不意味着半音节（IF/GIF）或音素层面的数据稀疏。于是，一个很有吸引力的思路就是尝试从相对比较充分的 IF 和 GIF 基元来估计发音词典内部各个发音变化项的条件输出概率。为此，本文提出利用上下文相关的发音词典加权策略。

在采用多发音词典的自然语音识别任务中，音节间的混淆程度是影响整体识别率的一个重要因素，减少音节间的混淆程度也正是这里对发音词典进行改

进的目标。因此在介绍新的词典加权策略前，首先提出一种对音节间混淆程度进行度量的指标。

## 4.2　音节间混淆度的度量方法

### 4.2.1 音节识别混淆的产生

首先举例说明音节间的混淆现象是如何造成的。假如一个多发音词典含有如下内容：

| | | | |
|---|---|---|---|
| $B_1$ | $P(S_{11} \mid B_1)$ | $S_{11}$ | $=S_0$ |
| $B_1$ | $P(S_{12} \mid B_1)$ | $S_{12}$ | |
| $B_2$ | $P(S_{21} \mid B_2)$ | $S_{21}$ | |
| $B_2$ | $P(S_{22} \mid B_2)$ | $S_{22}$ | $=S_0$ |

其中 $B_i$ 是某个标准的音节，$S_{ij}$ 是 $B_i$ 的第 $j$ 种可能的发音变化（识别基元序列），而 $P(S_{ij} \mid B_i)$ 则是这个发音变化项对应的条件输出概率或加权值。其中音节 $B_1$ 的发音序列 $S_{11}$ 与音节 $B_2$ 的发音序列 $S_{22}$ 完全相同，均记为 $S_0$。

假设某个自然语音样本中含有如下的发音片断（识别基元序列）：

… … … $S_0$ … … $S_0$ … … …

当不考虑使用任何语言模型时，若 $P(S_0 \mid B_1) > P(S_0 \mid B_2)$，则识别结果（音节序列）必定为

… … … $B_1$ … … $B_1$ … … …

在这种情况下，永远不可能从基元序列中正确地识别出具有发音序列 $S_0$ 的 $B_2$ 来。反之亦然，即若 $P(S_0 \mid B_1) < P(S_0 \mid B_2)$，则永远不可能从基元序列中正确地识别出具有发音序列 $S_0$ 的 $B_1$ 来（当 $P(S_0 \mid B_1) = P(S_0 \mid B_2)$ 时，识别结果或者为前者，或者为后者，这依赖于具体搜索算法的实现细节）。也就是说，这个发音词典造成了音节 $B_1$ 和音节 $B_2$ 之间固有的、一定程度的混淆现象。

### 4.2.2 发音词典固有混淆度的定义

为了可以定量地分析不同的发音词典可能造成的音节间混淆度的大小，本文提出一种针对特定多发音词典的、对音节间混淆程度进行度量的指标，命名为发音词典的固有混淆度（PLIC，Pronunciation Lexicon Intrinsic Confusion）。

PLIC 定义为在满足如下三个条件时，自然语音的音节误识率的下限（即 1－音节识别正确率的上限）。

1) 使用某个“理想”的声学模型，即对于任何测试集，基元层面上的识别率为 100%；
2) 识别过程中不借助任何语言模型，例如（汉语的）词或者音节层面的统计语言模型等；
3) 使用待评估的多发音词典，其中为每个音节包含了若干种可能的发音变化项及其相应的概率或权重。

从中可以看出，PLIC 将焦点完全集中在多发音词典本身，目的是度量不同的发音词典可能造成的音节间混淆程度的大小。因此，它反映了一个自然语音识别系统内在的、固有的识别错误率。

将待评估的多发音词典记为 $\boldsymbol{L}$，它包含的所有音节的集合为 $\boldsymbol{B}$，包含的所有音节的各个发音变化项的集合为 $\boldsymbol{S}$（相同的发音序列已被合并），那么当使用这个多发音词典 $\boldsymbol{L}$ 时，音节误识率的下限，即 PLIC 值为

$$\begin{aligned} PLIC(\mathbf{L}) &= \frac{1}{N}\sum_{s\in S}\left\{N\cdot P(s)\cdot(1-\max_{b\in B}P(b\mid s))\right\} \\ &= \sum_{s\in S}\left\{P(s)\cdot(1-\max_{b\in B}P(b\mid s))\right\} \end{aligned} \tag{4.1}$$

其中 $N$ 为测试集中被理想模型识别出来的基元序列中所含发音变化项的个数。公式 4.1 的物理意义比较明显：$N\cdot P(s)$ 为某个发音变化项 $s$ 的出现次数，$P(b\mid s)$ 为 $s$ 属于某个音节 $b$ 的后验概率，所有的 $s$ 都被识别为具有最大 $P(b\mid s)$ 概率的那个 $b$，因此对 $s$ 识别出错的概率为 $1-\max P(b\mid s)$。于是公式 4.1 就是理想情况下的音节误识率的最小值。

$N$ 随后被消掉，这说明 PLIC 并不依赖于具体的测试数据，而只与发音词典

***L*** 本身有关（是一个固有的指标）。

这里没有更精细地考虑一种特例，即当相邻两个发音序列的邻接部分的某个基元的子串能够组成另一个发音序列时，上述公式就会有一定的误差了，但统计结果表明，从总体上来看，这种现象出现频度的比例非常小，因此可以忽略不计。

继续对公式 4.1 进行推导。利用贝叶斯公式，可以得到

$$\begin{aligned} PLIC(L) &= \sum_{s\in S}\left\{P(s)\cdot(1-\max_{b\in B}\frac{P(s\mid b)\cdot P(b)}{P(s)})\right\} \\ &= \sum_{s\in S}\left\{P(s)-\max_{b\in B}P(s\mid b)\cdot P(b))\right\} \\ &= \sum_{s\in S}\left\{\sum_{b\in B}P(s\mid b)\cdot P(b)-\max_{b\in B}P(s\mid b)\cdot P(b))\right\} \end{aligned} \tag{4.2}$$

其中 $P(s\mid b)$ 已经被包含在多发音词典中，$P(b)$ 是音节 $b$ 的出现概率。当然若彻底不考虑语言模型的影响的话，也可以将它改为等概率的方式，即 P($b$)=1 / |**B**|（|**B**|表示集合 **B** 的大小），但这样会降低 PLIC 值的精度和可信度。因此在本文中，对 $P(b)$采用了前者的定义。

### 4.2.3 对基准发音词典的评估

根据 PLIC 的定义，就可以定量地对采用直接输出概率 DOP 的、不同规模的多发音词典所造成的固有音节混淆度进行比较，见表 4-2。

与 4.1.1 相对应，这里比较了两种情况，即对不同规模的多发音词典中的各个发音变化项，分别采用 DOP 和等概率的情况，来分析 PLIC 值的变化趋势。

表 4-2. 利用 PLIC 对基准的多发音词典进行评估

| 词典规模：累积概率％ / 加权方案 | | 100 | 95 | 90 | 85 | 80 | 75 | 70 | 65 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|
| PLIC 值 | 等概率加权 | 12.29 | 3.34 | 1.52 | 0.66 | 0.60 | 0.42 | 0.42 | 0.30 | 0.24 |
| | DOP 加权 | **1.73** | **1.16** | **0.83** | **0.54** | **0.51** | **0.41** | **0.33** | **0.33** | **0.24** |

从中可以看到，PLIC 值的变化趋势与 4.1.1 中的分析结论是吻合的。即随着词典规模的下降，发音变化项间的重复程度也不断下降，从而音节混淆度也不断下降，当各个音节的发音序列累计概率值降为 60%或更低时，已经基本上退化为单发音词典，此时音节混淆度和 PLIC 值均很小；此外，若对词典采取等概率方案（即不对发音变化项加权），则 PLIC 值要大许多，这是因为某些低频的、但大量重复的发音变化项的存在，加重了音节间的混淆程度。

当然将表 4-2 和 4.4.1 节将要给出的表 4-3 与相应的识别结果进行对照就可以看出，音节间的混淆度（PLIC 值）与音节的误识率之间并不是严格的正比关系或相同的增减趋势，它们还依赖于词典的规模、发音模型的精度、识别基元的频度分布情况等。只有在一定的条件（例如相同的发音词典规模）下，这些指标才具有可比性和参考价值。另外由于在实际系统中，上述理想声学模型的假设很难成立，因此实际的音节误识率要远远高于对应的 PLIC 值。

总的来说，PLIC 可以在理想情况（即独立于任何具体的声学模型）下，定量而直接地评价一个发音词典构成方式的优劣，并为评估实际系统的整体识别性能提供了一定的预测能力和参考价值。

## 4.3　上下文相关的词典加权策略

### 4.3.1　基本的推导过程

根据 4.1.3 节中提出的思路，在本节介绍上下文相关的词典加权策略。

在发音词典中，对各个发音变化项的概率加权值为 $P(s|b)$，这是在给定标准音节 $b$ 的前提下，发音变化项 $s$ 的条件输出概率。假设 $b=(i,f)$，$s=(g_i,g_f)$，其中 $i$ 和 $f$ 属于 IF 基元集合，$g_i$ 和 $g_f$ 属于 GIF 基元集合（可在与 IF 的删除现象对应的位置放入一个特殊的<*Nul*>或<*Del*>符号参与概率统计）。为简化问题起见，假设声母和韵母的输出之间彼此独立，于是

$$P(s|b)\approx P(g_i|i)\cdot P(g_f|f) \tag{4.3}$$

显然，公式 4.3 右边的乘式前后两部分均可归结为 P (*GIF* | *IF*) 的形式，这里 *IF* 为变量，表示 *i* 或 *f*，*GIF* 也为变量，表示 $g_i$ 或 $g_f$。即在给定某个 IF 基元的情况下，在自然语音流中实际观测到某个 GIF 基元的条件输出概率，这就将词典的概率加权值求解问题从训练数据比较稀疏的声韵层面转换到了训练数据相对比较充分的 IF 和 GIF 层面了。

但上式的一个明显不足是，它没有考虑到上下文的影响，这是因为不同的发音上下文是导致不同发音变化现象的一个很重要的因素，尤其对于自然随意的发音来说更是如此。为此，在 $P(GIF \mid IF)$ 中同时引入发音上下文的相关性，根据全概率公式得到

$$P(GIF \mid IF) = \sum_{C_{IF}} P(GIF \mid IF, C_{IF}) P(C_{IF} \mid IF) \tag{4.4}$$

其中 $C_{IF}$ 是与 *IF* 相关的任何发音上下文，例如它的左相关或右相关（*IF* 的 bigram），或者是左右相关（*IF* 的 trigram），或者是更宽的上下文序列等。在这里仅仅取 *IF* 左边紧紧相邻的 *IF* 作为其（左）上下文，表示为 $C_{IF}=L_{IF}$，于是公式 4.4 可以改写为

$$P(GIF \mid IF) = \sum_{L_{IF}} P(GIF \mid (L_{IF}, IF)) P(L_{IF} \mid IF) \tag{4.5}$$

公式 4.5 右边和式中的两个乘积项，分别为 *IF* 上下文相关的 *GIF* 条件输出概率和 *IF* 之间的转移概率，见 2.1.3 节。它们可以由数据库的原始标注或 2.3 节中修正后的标注中直接统计出来。

由于在公式 4.3 中对概率 $P(s \mid b)$ 进行了近似，这样得到的值一般不会满足 $\Sigma_s P(s \mid b) = 1$，因此它不再是一个严格意义上的“概率”，而是一种“权重”。根据其推导过程，本文将上述方法称为对发音词典的上下文相关加权策略，所计算出的权重称为上下文相关权重（CDW，Context- Dependent Weight）。显然它的值与直接输出概率 DOP 是不同的。

### 4.3.2 上下文相关权重的确定

一个很有意义的现象是，上下文相关权重 CDW 将直接输出概率 DOP 推广到了“跨词”相关的情形。这是因为当公式 4.5 中的 *IF* 是某个音节内部的声母

时，它对应的 $L_{IF}$ 必定为上一个音节结尾的韵母，这样就体现出了“音节间”的相关性。与此对照，DOP 考虑的仅仅是音节变化项的出现频度（即 Unigram），不包含任何上下文的相关性，因此就显得粗糙一些。

直观上看来，公式 4.5 中的 *IF* 既可以施加于声母，也可以施加于韵母，它从形式上是对所有可能的 *IF* 上下文求和。在此基础上，还可以考虑对 *P*(*GIF*|*IF*) 取如下两种变化：

1) 从所有可能的 *IF* 上下文中，仅仅取其中概率最大的一个 *IF* 上下文；
2) 仅仅取与其声母相关的一个 *IF* 上下文。这对于韵母部分是有意义的，因为在一个特定的音节内部，韵母的左上下文是固定的（即音节的声母部分）。

现在对公式 4.5 及上述两种变化形式进行重写。首先定义

$$\widetilde{M}_L(GIF \mid IF) = P(GIF \mid (L, IF))P(L \mid IF) \tag{4.6}$$

其中 $L$ 为与 *IF* 相关的上下文符号。于是有

$$\widetilde{P}(GIF \mid IF) = \sum_{L_{IF}} \widetilde{M}_{L_{IF}}(GIF \mid IF) \tag{4.7}$$

$$\widetilde{Q}(GIF \mid IF) = \max_{L_{IF}} \widetilde{M}_{L_{IF}}(GIF \mid IF) \tag{4.8}$$

公式 4.7 与公式 4.5 是等价的。将公式 (4.6)、(4.7)和(4.8) 分别作为声母或韵母代入公式 4.3，得到如下三种比较合理的组合（即 CDW 的三种可能的定义）：

CDW-SUM: $$P(s \mid b) \approx \widetilde{P}(g_i \mid i) \cdot \widetilde{P}(g_f \mid f) \tag{4.9}$$

CDW-MAX: $$P(s \mid b) \approx \widetilde{Q}(g_i \mid i) \cdot \widetilde{Q}(g_f \mid f) \tag{4.10}$$

CDW-MATCH: $$P(s \mid b) \approx \widetilde{P}(g_i \mid i) \cdot \widetilde{M}_i(g_f \mid f) \tag{4.11}$$

上式中 $b$、$s$、$i$、$f$、$g_i$、$g_f$ 的定义与 4.3.1 节中相同。其中，CDW-SUM 的声母和韵母部分均对各种可能的左上下文求和，CDW-MAX 的声母和韵母部分均取可使其概率达到最大的左上下文，而 CDW-MATCH 的声母部分对各种可能的左上下文求和，但其韵母部分仅仅取音节 $b$ 的声母作为其可能的左上下文。

## 4.4　实验结论与分析

### 4.4.1 在基准系统上的实验与分析

首先利用 PLIC 值来分析 CDW 对音节混淆程度的影响。与 4.2.3 节类似，对不同规模的多发音词典中的各个发音变化项，分别采用不同的 CDW 加权值来分析 PLIC 值的变化趋势。

表 4-3. 利用 PLIC 对采用 CDW 的多发音词典进行评估

| 加权方案 \ 词典规模：累积概率% | | 100 | 95 | 90 | 85 | 80 | 75 | 70 | 65 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|
| PLIC 值 | CDW-SUM | 1.11 | 0.68 | 0.45 | 0.31 | 0.28 | 0.26 | 0.26 | 0.24 | 0.24 |
| | CDW-MAX | 1.19 | 0.60 | 0.40 | 0.30 | 0.30 | 0.26 | 0.26 | 0.24 | 0.24 |
| | CDW-MATCH | **0.86** | **0.52** | **0.36** | **0.28** | **0.27** | **0.26** | **0.26** | **0.24** | **0.24** |

与表 4-2 相比，各个 CDW 的 PLIC 值均要低于同等词典规模下采用 DOP 时的 PLIC 值，这说明采用 CDW 方案后，从概率意义上看，由于发音词典内所含重复的发音变化项而造成的音节混淆程度有所降低。

在同等词典规模情况下，CDW-SUM 和 CDW-MAX 的 PLIC 值相比，有时前者低些，有时后者低些。而总的来看，CDW-MATCH 的 PLIC 值要略低于 CDW-SUM 和 CDW-MAX 的 PLIC 值。由于计算精度的影响，当词典规模低于发音变化的累积概率 75%以后，三种 CDW 的值变得相等。

图 4-1 直观地给出了在采用等概率、DOP、CDW 等三种多发音词典的加权方案下的 PLIC 曲线趋势对比图，其中 CDW 的曲线为采用 CDW-MATCH 的加权方案。

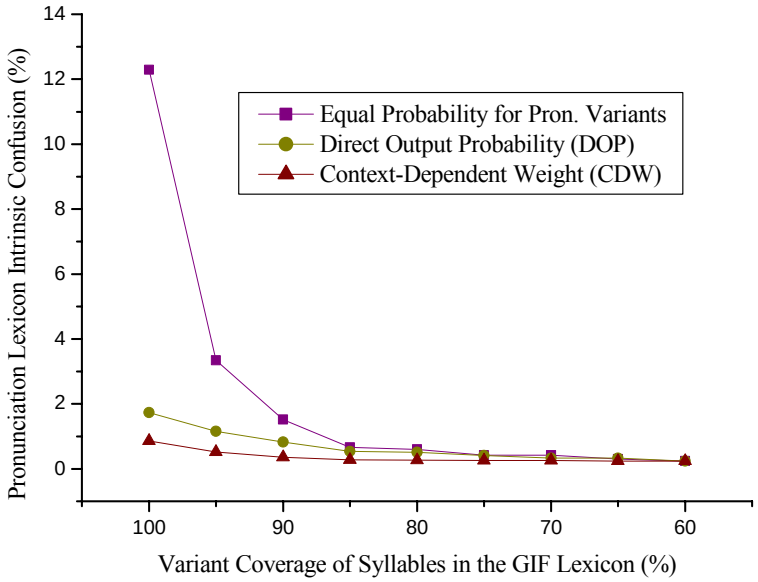


图 4-1. 多发音词典采用不同加权方案时的 PLIC 曲线

下面进行识别测试和比较。采用与基准系统同等规模的发音词典（90%的发音变化累积概率）以及 DOP 和不同的 CDW 加权方案，声学模型同时采用 CI-GIF 和 CD-GIF。

表 4-4. 采用 CDW 加权方案的识别结果(%)

| 模型 / 加权方案 | | CI-GIF | | | CD-GIF | | |
|---|---|---|---|---|---|---|---|
| | | SCR | SER | SER↓ | SCR | SER | SER↓ |
| 基准 | DOP | 33.18 | 66.82 | —— | 46.07 | 53.93 | —— |
| CDW | CDW-SUM | 33.30 | 66.70 | 0.18 | 46.66 | 53.34 | 1.10 |
| | CDW-MAX | 32.95 | 67.05 | - 0.34 | 46.42 | 53.58 | 0.65 |
| | CDW-MATCH | **34.28** | **65.72** | **1.65** | **46.92** | **53.08** | **1.58** |

从中可以看出，除了 CDW-MAX 在 CI-GIF 模型上的音节误识率略有上升外，其它情况下的音节误识率均下降，其中 CDW-MATCH 方案的下降幅度最大，

对于 CI-GIF 和 CD-GIF，其音节误识率分别下降 1.65%和 1.58%。这说明 CDW 相对于 DOP 来说，不仅对于降低音节的混淆程度有效，而且对多发音词典加权后，对总体识别性能的改善也是有效的。

对 CDW 的有效性分析如下。与 DOP 相比，CDW 绕开了训练数据比较稀疏的音节，而是从相对比较充分的半音节（IF 和 GIF）入手对发音词典的权重进行估计，使得其数值的可信度有所提高；同时，它又考虑了半音节的上下文相关性，从而在权重中体现出了上下文的因素，而这一点对刻画自然语音中发音变化的产生及其分布规律是非常重要的。

其中尤为值得注意的是 CDW-MATCH 加权方案。从它的推导公式可以看出，它在声母中对所有左边邻接的上下文进行求和，相当于引入了“跨词”（或者说是“音节间”）的相关性，同时对韵母又仅仅取与其声母相匹配的上下文，这更符合音节本身的结构特性，因此它与其它两种 CDW 加权方案相比，具有更大的优越性。实验结果也证明了这一点。

### 4.4.2 在 GIF 精细模型上的实验与分析

为进一步验证 CDW 的有效性，在第三章中的 GIF 精细模型（IF-GIF 模型）上也进行了类似的实验。基准模型采用 3.4 节中识别性能最好的“IF-GIF＋自适应＋均值重估”模型，利用 3.3.3 中的方法将采用各个 CDW 方案的 GIF 多发音词典转换为 IF-GIF 的多发音词典格式。实验结果见表 4-5。

表 4-5. 在 GIF 精细模型上采用 CDW 的识别结果(%)

| 模型 / 加权方案 | | B-GIF | | | S-GIF | | |
|---|---|---|---|---|---|---|---|
| | | SCR | SER | SER↓ | SCR | SER | SER↓ |
| 基准 | DOP | 34.44 | 65.56 | —— | 33.66 | 66.34 | —— |
| CDW | CDW-SUM | 34.25 | 65.75 | - 0.29 | 33.89 | 66.11 | 0.35 |
| | CDW-MAX | 34.39 | 65.61 | - 0.08 | 33.54 | 66.46 | - 0.18 |
| | CDW-MATCH | **35.85** | **64.15** | **2.15** | **34.86** | **65.14** | **1.81** |

从结果中可以看出，CDW-SUM 和 CDW-MAX 对于 B-GIF 或 S-GIF 方法得到的 IF-GIF 模型，音节误识率大部分略有上升，仅有一种情况略有下降，但无论上升或下降，其幅度均非常小。而 CDW-MATCH 方案在这两种情况下，其误识率均有大幅下降。因此进一步验证了 CDW（主要是 CDW-MATCH 方法）对发音词典加权方案的有效性。

### 4.4.3 对 CDW 的深入讨论

在本文提出的 CDW 方法中，利用了半音节（IF 和 GIF）的上下文相关性来得到所需的词典加权值。实际上，这个方法并不仅仅局限于半音节，对于基于音素或其它子词基元的自然语音识别系统来说，很容易将类似的方法进行推广使用。

此外，这里的 CDW 仅仅考虑了左上下文，同理可以推广到使用右上下文的情况。但若同时使用左右上下文，则不能肯定地预测其有效性。这是因为，当同时使用了左右上下文时，从半音节或音素的层面来看，其训练数据也将显得过分稀疏，这就有可能会抹煞 CDW 相对于 DOP 的优势，因此需要根据具体的训练数据量和实验结果来确定其可行性。

# 第五章　发音建模的参数共享策略

## 5.1　问题的提出

从前面的研究中可以看出，在自然语音的发音建模中，存在着两个互相矛盾的需求：

1) 为了准确地反映出自然语音中重要的发音变化情况（例如音素或声韵的替换、插入和删除等），需要在发音词典中为每个音节加入尽可能多的发音基元序列，并为这些序列中的各个基元进行声学建模（例如采用GIF 和与之相关的多发音词典，并进行基于 GIF 的精细建模）；

2) 但随着词典中发音序列的增加，音节间的混淆度随之增加（这从 PLIC 曲线中可以看出来，并且对 CDW 的研究也是为了改善发音词典和减小音节混淆度）。只有采用单发音词典时，这种混淆度才会达到最低（例如基于 IF 模型及其标准的单发音词典），但这又与上述需求矛盾（因为此时无法体现出上述重要的发音变化现象）。

在前面的章节中，都分别是孤立地、集中地研究上述的某一个方面。那么是否可以有一种比较有效的方法将二者的优点综合起来，从而达到更好的识别性能呢？这正是本章要研究的目标。

基本的思路是，采用单发音的标准词典，使得音节间的混淆程度最小（甚至为零）；同时改进声学模型的构造方法，使得在基元的 HMM 模型内部就可以直接体现出原来需要由多发音词典来包含的发音变化现象，这正是下面的研究重点。

为此本文提出在现有的基准 IF 与 GIF 模型基础上进行模型参数共享的发音建模方法。按照其实现细节，可以分为两类：

1) 基于状态相似度的高斯混合共享方法

2) 基于状态相似度的模型参数组合方法

其中，高斯混合的共享方法是在 GIF 模型内部进行的，而模型参数的组合

方法是在 IF 模型与 GIF 模型之间进行的。

本章的所有研究都是基于上下文相关的模型。

## 5.2　模型参数共享的原理和实现

### 5.2.1 基本的实现思路

考虑第二章中的采用 GIF 基元和多发音词典的基准系统。假设一个音节 $S$ 的标准发音是由 $(a,b,c)$ 三个音素（对于声韵的表示与此类似）组成的，在某种自然发音环境下，音素 $b$ 会变化为音素 $s$（即在实际的发音中被观测到）。于是在发音词典中就应当同时为这个音节 $S$ 包含 $(a,b,c)$ 和 $(a,s,c)$ 两种可能的发音序列，并在识别搜索时的解码网络中体现出来，见图 5-1。其中图 5-1($a$)是采用 CI（上下文无关）模型时的情形，而图 5-1($b$)是采用采用 CD（上下文相关）模型时的情形。

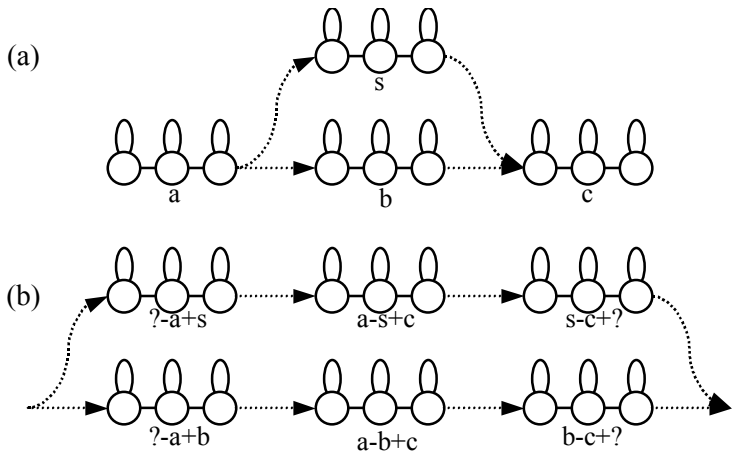


图 5-1. 多发音词典与相应的解码网络举例

根据 5.1 节的思路，需要采用单发音词典以尽量减少音节间的混淆程度，并且在其对应的声学模型中包含必要的发音变化。以图 5-1 为例，首先，需要

在发音词典中抛掉这个音节的发音变化项$(a,s,c)$，即仅仅包含其标准发音项$(a,b,c)$；然后，将可能的发音变化 $s$ 的声学特性包含进基本的声学模型 $b$ 中，形成一个统一的 HMM 模型，见图 5-2。

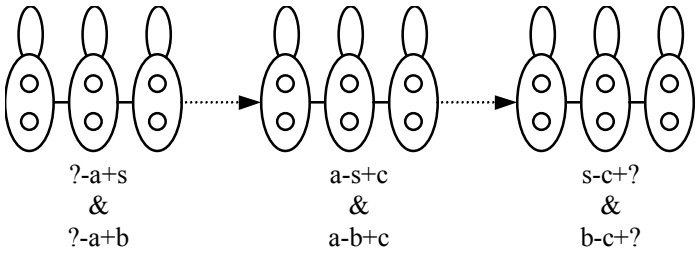


图 5-2. 上下文相关模型的参数共享示意图

与第 2.2.2 节相同，其中基元符号内的“－”表示左相关的上下文，“＋”表示右相关的上下文，“？”表示不关心或任何上下文，这些上下文信息是在识别搜索中构造解码网络时动态生成的。

由于上下文的相关性对于发音变化至关重要，并且 CD 模型参数的精度要远远高于 CI 模型，为此本章将研究的焦点放在 CD 模型参数的共享上，而不考虑 CI 模型的情形，图 5-2 中也仅仅给出了基于 CD 模型的参数共享示意图。

在本文的 CD 模型中，无论是 CD-IF，还是 CD-GIF，都是通过决策树技术来合成未出现的 tri-IF 或 tri-GIF 且避免训练数据的稀疏，并使得各个基元的 HMM 状态完全共享的（见 2.2.2 节）。由于各个共享状态内部的高斯混合参数与声学输出概率分布直接相关，因此这里对模型参数的共享，就是通过对这些高斯混合参数更进一步地进行共享和组合来实现的。而对这些参数进行共享的依据，则是利用了 HMM 状态间的相似度指标。

### 5.2.2 状态相似度的定义

假设有两个不同的 HMM 参数集合$H_b$和$H_s$，其中$H_b$采用标准的单发音词典$L_b$，而$H_s$采用包含可能发音变化的多发音词典$L_s$（下标 $b$ 和 $s$ 分别代表 Base-form 和 Surface-form）。

将它们对已知拼音层标注 $\mathbf{W}=W_1W_2\cdots W_m$ 的相同观测样本 $\mathbf{O}_i=o_1o_2\cdots o_T$ 以强制时间对准的方式进行状态级的 Viterbi 解码，分别得到两个不同的最优状态序列 $\hat{S}_b$ 和 $\hat{S}_s$，即

$$\hat{S}_i=\arg\max_S P(\mathbf{W}\mid\mathbf{O},S;H_i),\qquad (i=b\ or\ s) \tag{5.1}$$

其中 $S$ 为任意可能出现的 HMM 状态序列。两个最优状态序列（$\hat{S}_b$ 和 $\hat{S}_s$）相对其声学模型（$H_b$ 和 $H_s$）与标注 $\mathbf{W}$ 来说，具有最大似然的时间边界信息。

假设 HMM 状态 $b_i$ 和 $s_j$ 分别位于上述两个最优状态序列 $\hat{S}_b$ 和 $\hat{S}_s$ 中，两者之间的相似度 $Q_{ij}$ 定义，为在上述解码过程中 $s_j$ 被混叠为 $b_i$ 的概率，即

$$Q_{ij}=P(s_j\mid b_i)=\frac{T(b_i,s_j)}{\sum_k T(b_i,s_k)} \tag{5.2}$$

其中 $T(b,s)$ 表示状态 $b$ 和 $s$ 在这两个最优状态序列 $\hat{S}_b$ 和 $\hat{S}_s$ 中的重叠时间长度。可以看出，这个定义既依赖于这两个最优状态序列本身的内容，又依赖于其时间边界信息。此外，$H_b$ 和 $H_s$ 中所包含的状态集合 $\{b_i\}$ 与 $\{s_j\}$ 可以是相同的（见下节），也可以是不同的。

### 5.2.3 模型参数共享的过程

如第 2.2.2 节所述，对于利用决策树方案实现的 CD-IF 和 CD-GIF 模型，其共享状态的结构如图 5-3 所示。

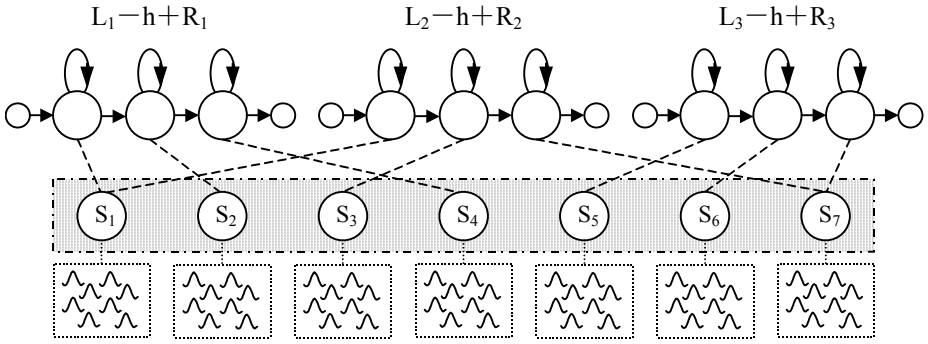


图 5-3. 上下文相关 HMM 模型的共享状态池结构

其中阴影覆盖的部分可以称为“共享状态池”（Shared-State Pool），它包含了所有的公共状态$s_i$（图中$1 \le i \le 7$），供各个 CD 模型内部的状态来绑定（Tying）和共享；与此同时，每个公共状态$s_i$独自拥有一组高斯混合分布参数。假设集合$\mathbf{M}(s_i)$为属于公共状态$s_i$的若干个高斯混合，$w_{ij}$为$s_i$中第 $j$ 个高斯混合的权重，因此状态$s_i$的输出概率分布为：

$$P(o \mid s_i) = \sum_{j \in \mathbf{M}(s_i)} w_{ij} N(o; \mu_{ij}, \Sigma_{ij}) \qquad (s_i \in H_b \ or \ H_s) \tag{5.3}$$

其中 $N(\cdot)$为正态分布（高斯混合）的概率密度函数。

本文提出的基于状态相似度的模型参数共享策略的目标，就是要将这些高斯混合分布参数通过一定的准则也共享起来，从而在标准模型$H_b$内部反映出所需的发音变化。见图 5-4。

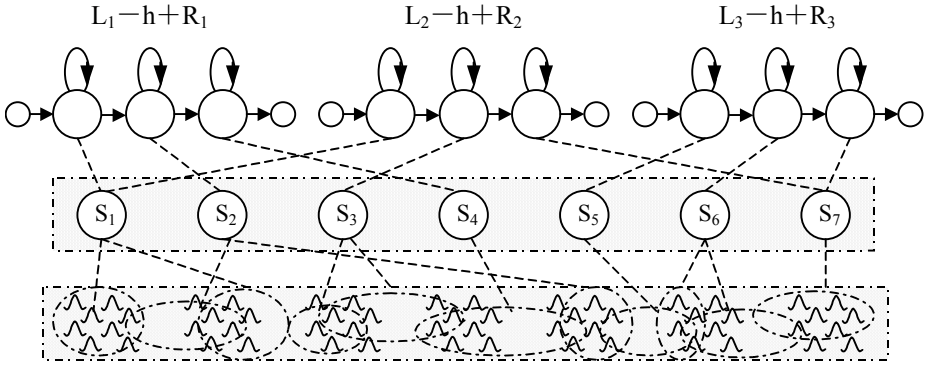


图 5-4. 参数共享后的上下文相关 HMM 模型结构

具体的参数共享其过程为：

1) 将分属于不同共享状态的所有高斯混合提出来，放入另一个完全公用的“共享高斯池”（Shared-Gaussian Pool）中（见图 5-4 下方的阴影覆盖部分），使得这些高斯混合变为公共的。于是初始时，每个公共的高斯混合仅仅与原来的一个共享状态相关联，原有的权重$w_{ij}$保持不变。

2) 利用模型集$H_b$和$H_s$（与 5.2.2 中的相同）对训练数据进行状态级的强制时间对准，分别得到最优的状态序列$\hat{S}_b$和$\hat{S}_s$，序列中的各个状态均分

属于 $H_b$ 和 $H_s$ 的共享状态池；

3) 根据状态相似度的定义，对于 $\hat{S}_b$ 中的每个状态 $b_i$，分别求出它与 $\hat{S}_s$ 中各个可能状态 $s_j$ 之间的状态相似度，将其施加到 $s_j$ 的各个初始权重中，得到 $b_i$ 的新的输出概率分布函数：

$$\begin{aligned} P(o \mid b_i) &= \sum_{j \in H_s} Q_{ij} \sum_{k \in \mathbf{M}(s_j)} w_{jk} N(o; \mu_{jk}, \Sigma_{jk}) \qquad (b_i \in H_b) \\ &= \sum_{j \in \bigcup_{k \in H_s} \mathbf{M}(s_k)} w'_{ij} N(o; \mu_{ij}, \Sigma_{ij}) \end{aligned} \tag{5.4}$$

4) 实际上，对公式 5.4 中的每个 $b_i$，仅仅保留其中具有最大权重的若干个高斯混合，形成新的高斯混合集合 $\mathbf{M}'(b_i)$（见图 5-4 中各个虚线椭圆包围的部分），这些高斯混合依旧放在公共的高斯混合池中，各个状态内部仅仅保留指向它们的指针。然后对其新权重 $w'_{ij}$ 进行归一化，即

$$\sum_{j \in \mathbf{M}'(b_i)} w'_{ij} = 1 \qquad (b_i \in H_b) \tag{5.5}$$

5) 最后，抛掉共享高斯池中未被任何新状态共享的那些高斯混合，并对参数共享后的新模型 $H_b$ 利用 Baum-Welch 算法进行参数重估，然后利用与 $H_b$ 对应的、其原有的单发音词典进行识别。

与现有的 CD-IF 模型和 CD-GIF 模型对应，初始模型集合 $H_b$ 和 $H_s$ 有两种可选的组合方案，因而对应着两种不同的模型参数共享方法：

1) $H_b$ 和 $H_s$ 均为 CD-GIF，称为高斯混合共享方法；

2) $H_b$ 为 CD-IF，而 $H_s$ 为 CD-GIF，称为模型参数组合方法。

它们在实现过程中的差异和对比，将在后续两节中进行介绍。

### 5.2.4 高斯混合共享方法的实现

#### 1. 高斯混合共享方法的特点

初始的 HMM 模型 $H_b$ 与 $H_s$ 均采用 CD-GIF 模型。其特点为：

1) $H_b$ 和 $H_s$ 具有相同的共享状态池和共享高斯池，它们都是来自 CD-GIF 模型；

2) 利用 5.2.3 的方法进行参数共享时，相当于仅仅对 $H_b$ 各个状态所含的高斯混合及其权重进行调整，并没有超出初始模型本身的公共高斯集合（或其子集），因此将这个方法称为高斯混合的“共享”。

**2. 初始参数的获取**

参数共享策略的一个初始假设就是让模型 $H_b$ 采用仅包含音节标准读音的单发音词典，但这里与 $H_b$ 对应的却是基于 GIF 的多发音词典，具体的解决办法如下：

根据 2.1.3 节，GIF 基元集合是 IF 基元集合的一个超集，因此在 GIF 基元集合中总可以为每个 IF 基元找到一个与之对应的、从发音上最接近的 GIF 基元。然后将这个与 IF 基元集合一一对应的 GIF 基元的子集（以及与其对应的 HMM 模型参数）作为 $H_b$，并将基于 IF 的单发音词典中的发音序列替换为对应的 GIF 标识，形成基于 GIF 基元的单发音词典。

然后对训练数据进行强制时间对准方式的 Viterbi 解码以得到用来求解状态相似度 $Q_{ij}$ 的最优状态序列 $\hat{S}_b$ 和 $\hat{S}_s$，其过程见图 5-5。

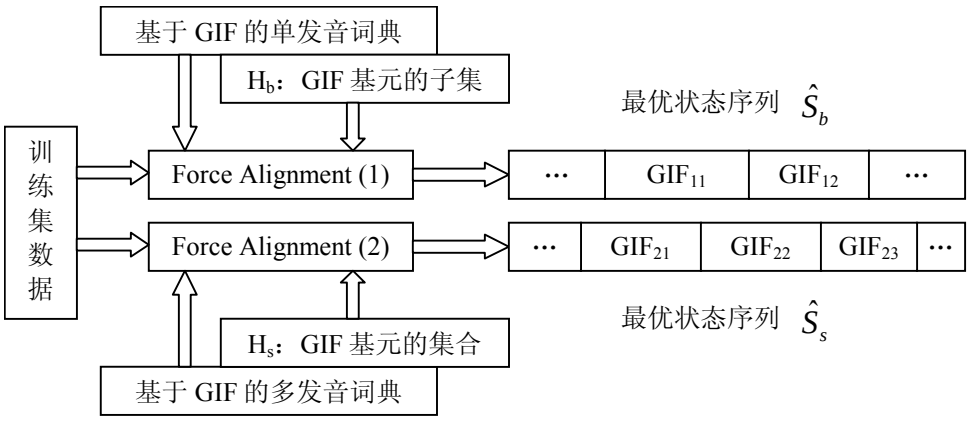


图 5-5. 利用高斯混合共享方法的初始参数获取

**3. 进行参数的共享**

在确定了初始的$H_b$、$H_s$、$\hat{S}_b$、$\hat{S}_s$以后，即可按照 5.2.3 节的方法进行后续的模型参数共享，最终得到新的模型集合$H'_b$。它的基元是以 tri-GIF 形式表示的，而共享状态池中的公共状态和共享高斯池中的公共高斯混合均来自于 CD-GIF 模型；最终的发音词典则采用上面构造的基于 GIF 的单发音词典。

### 5.2.5 模型参数组合方法的实现

**1. 高斯混合共享方法的特点**

初始的 HMM 模型$H_b$采用 CD-IF 模型，而初始的 HMM 模型$H_s$采用 CD-GIF 模型。其特点为：

1) $H_b$使用来自 CD-IF 的共享状态池和共享高斯池，而$H_s$则使用来自 CD-GIF 的共享状态池和共享高斯池；

2) 利用 5.2.3 的方法进行参数共享时，$H_b$各个状态内新的高斯混合集合不同于初始模型本身的公共高斯集合，而是完全来自于$H_s$的公共高斯集合，但对其权重的确定却同时依赖于$H_b$与$H_s$初始的高斯分布，因此将这个方法称为模型参数的“组合”。

**2. 初始参数的获取**

在这里，初始模型$H_b$可以直接采用基于 IF 的标准单发音词典。然后对训练数据进行强制时间对准方式的 Viterbi 解码，以得到用来求解状态相似度$Q_{ij}$的最优状态序列$\hat{S}_b$和$\hat{S}_s$，其过程见图 5-6。

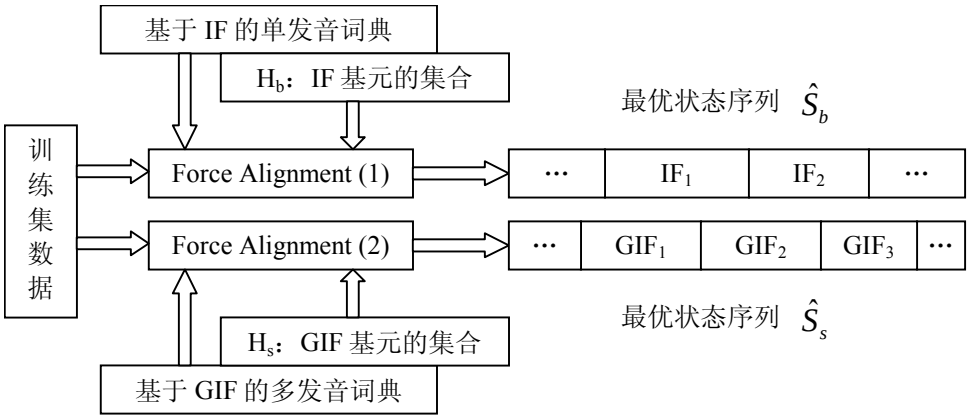


图 5-6. 利用模型参数组合方法的初始参数获取

**4. 进行参数的共享**

同样地，在确定了初始的$H_b$、$H_s$、$\hat{S}_b$、$\hat{S}_s$以后，即可按照 5.2.3 节的方法进行后续的模型参数共享，并最终得到新的模型集合$H'_b$。它的基元是以 tri-IF 形式表示的，而共享状态池中的公共状态来自于 CD-IF 模型，但共享高斯池中的公共高斯混合来自于 CD-GIF 模型（用 CD-GIF 中的共享高斯池来代替 CD-IF 中原有的共享高斯池）；最终的发音词典仍采用标准的基于 IF 的单发音词典。

## 5.3 实验结论与分析

### 5.3.1 识别测试结果及分析

表 5-1 给出了采用参数共享方法并进行 Baum-Welch 参数重估后，训练集的平均似然值。共享状态池中每个共享的状态内部仍含有 8 个高斯混合。

表 5-1. 参数共享后训练集的平均似然值

| 参数共享方案 | 高斯混合共享 | 模型参数组合 |
|---|---|---|
| 训练集平均似然值 | -54.37 | -54.12 |

与表 2-10 中的基准模型 CD-GIF 相比，可以看到进行参数共享后，训练集的平均似然值略有下降。对其进行分析可以给出如下解释：参数共享的目的是采用不包含发音变化项的标准单发音词典并在 HMM 状态内部包含对必要的发音变化的描述，与基准的“GIF 基元＋含发音变化项的多发音词典”方案相比，显然前者模型参数的内聚性要低一些，因而它们与训练集的匹配程度有所下降。

但这个似然值仅仅是一个侧面的参考值，对识别器声学性能的评价应该是综合的。为此对模型参数共享策略再进行识别测试，分别比较了采用上述两种不同初始化方案（即高斯混合共享方案和模型参数组合方案）所得到的新模型的识别性能。结果见表 5-2。

表 5-2. 采用参数共享策略的模型识别结果(%)

| 识别率 / 参数共享策略 | SCR | SER | SER↓ |
|---|---|---|---|
| 参考对比：CD-IF | 44.98 | 55.02 | —— |
| 基准模型：CD-GIF | **46.07** | **53.93** | —— |
| 高斯混合共享的方法 | 46.30 | 53.70 | 0.43 |
| 模型参数组合的方法 | **46.95** | **53.05** | **1.63** |

从中可以看出，无论是采用高斯混合共享的方法，还是采用模型参数组合的方法，参数共享策略的性能均优于基准的 CD-GIF 模型。

分析其原因，正如 5.1 节中对其出发点的讨论，这个策略既降低了由于发音词典中重复的发音变化项所造成的音节间的混淆程度（采用单发音词典），又在最终的 HMM 模型参数中体现出了那些必要的发音变化现象，从而具有更好的对自然语音进行声学建模的能力。

另外，由于 CD-GIF 中分别被每个共享状态所包含的高斯混合集合之间是孤立的、单独训练的，而在参数共享后的模型中，各个共享状态所包含的高斯混合集合之间也是被完全共享的，并且其中某些原来具有极低权重的高斯混合将被抛掉，从而进一步降低了模型的总体规模。这样，在最后利用 Baum-Welch 算法进行模型参数的迭代重估时，每个基元将得到更加充分的训练数据，这对

进一步提高模型的鲁棒性是非常有益的。

采用这两种不同参数共享策略得到的目标模型的声学识别性能间有较大的差异。相对于基准模型 CD-GIF 来说，高斯混合共享方法仅仅有 0.43%的 SER 相对下降率，而模型参数共享方法的 SER 相对下降率较大，为 1.63％。

下面对这两个方法各自具有的优缺点进行分析：

1) 高斯混合共享的方案在解码时采用了 GIF 的子集和由 IF 词典导出的 GIF 单发音词典，因而产生的最优状态序列具有相对大一些的误差；

2) 而模型参数组合的方法采用了由 CD-IF 的决策树生成的共享状态结构，但最终的各个高斯混合却是从 CD-GIF 模型中得到，这二者在一定程度上存在着训练目标函数不匹配的可能性。

### 5.3.2 与相关算法的对比

从直观上看，本文的参数共享策略与半连续 HMM（SCHMM）[Huang90b]以及状态级发音建模（SLPM，State Level Pronunciaton Modeling）[Saraclar99]的方法有些类似，但实际上它们之间的差别是很明显的。

首先，SCHMM 的目标是利用共享的高斯混合来替代离散 HMM（DHMM）的码本从而提高对声学输出的概率描述精度，而本文的参数共享策略则是面向自然语音中大量的发音变化现象，尤其状态相似度 $Q_{ij}$ 指标对非标准发音序列中的插入、删除等音变现象的描述能力是 SCHMM 所不具备的。

其次，与 SLPM 方法相对照，这两种参数共享策略均是面向自然语音中发音变化现象的建模问题，但在本文提出的参数共享策略中，无论采取哪一种初始化参数的产生方法，均利用了可更好地匹配发音变化的 GIF 模型并为之进行了特殊处理（例如与基准的 IF 模型进行参数组合），因而从声学建模的角度来看，具有更好的改善识别性能的潜力。

# 第六章　发音建模方法的整合

## 6.1　各部分思路间的关联

前面各章的研究思路是彼此相关联的，与图 1-2 对应，表 6-1 列出了各部分研究内容的出发点及其相互之间的关联。

表 6-1. 各部分研究思路的出发点及其相互关系

| 研究内容 | 研究动机或目标 | 实现方法 |
|---|---|---|
| GIF 基元集的定义 | 提高基本 IF 基元对自然发音的覆盖面，以涵盖更多的发音变化 | 从 CASS 中初始统计，迭代修正标注、求精 |
| 基于 GIF 的精细建模 | 将由不同 IF 音变得到的相同 GIF 之间的差异引入模型，并解决训练数据的稀疏问题 | 由 IF 或 GIF 模型拷贝至初始 IF-GIF 模型，然后自适应并重估均值 |
| 基于上下文的词典加权 CDW | 考虑音节间的上下文相关性，提高词典加权估值的准确性，并解决训练数据的稀疏问题 | 在半音节层面（IF 和 GIF）对基元的上下文和输出、转移概率综合 |
| HMM 状态内的参数共享 | 减小音节间混淆度，同时不损失声学模型对发音变化的描述能力 | GIF 模型内高斯共享、IF 与 GIF 模型间参数组合 |

## 6.2　各方法的有效性概览

本文对所提出的各个方法，主要采用 SER↓指标进行评价。此外，对于声学模型相关的研究同时采用训练集的似然值作为参考，对于发音词典相关的研究同时采用 PLIC 值作为参考。

图 6-1 给出各个方法相对于其基准模型或基准方法，对声学识别性能的改善能力。这里仅以 SER↓作为指标。

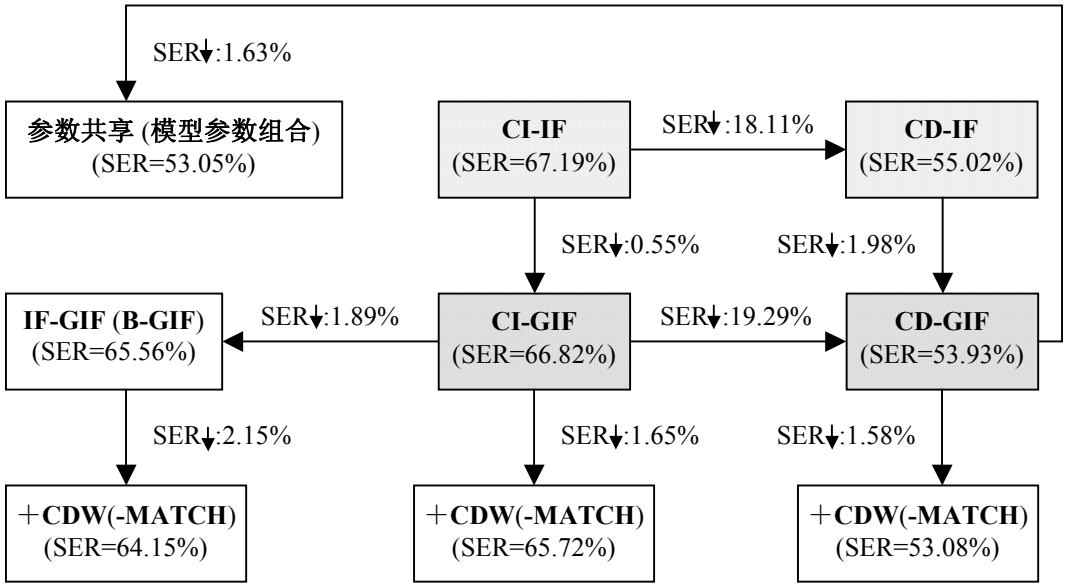


图 6-1. 各方法对声学识别性能的改善能力

## 6.3　语言模型的作用

在本文的所有研究中，均不涉及语言模型（包括汉语的字词层面上的 *N*-gram 和音节层面上的 *N*-gram 等）的作用，这是因为本文关注的焦点是所提出的方法在纯声学意义上对识别性能的改善能力，为此而不采用任何语言模型。但这个目标与语言模型的使用并不矛盾，也不会影响语言模型本身的对识别性能的改善能力。

但作为一个实际的连续语音识别系统，包括基于朗读式发音甚至是自然发音的系统，语言模型是绝对不可缺少的，而且它对识别性能提高的作用也是巨大的。这是因为使用语言模型后（例如统计语言模型），在语言层面上考虑了上下文的关系，因此也同样可以大大降低声学识别的音节混淆程度，甚至纠正一部分声学识别结果中的错误。

为了观察语言模型对整体识别性能的改善程度，尝试在前面的若干个模型中加入语言模型。为简化问题，仅仅采用了从 CASS 的音节标注中统计得到的、采用 Back-off 折扣方案的音节 bigram，并仍在音节层面对识别性能进行评价。实验结果见表 6-2。

表 6-2. 施加音节 Bigram 语言模型后的识别结果(%)

| 声学模型 \ 识别结果 | | 基准识别结果 | | 采用音节 bigram 的识别结果 | | |
|---|---|---|---|---|---|---|
| | | SCR | SER | SCR | SER | SER↓ |
| 基准模型 | CI-IF (单发音词典) | 32.81 | 67.19 | 36.10 | 63.90 | **4.90** |
| | CI-GIF (多发音词典) | 33.18 | 66.82 | 36.33 | 63.67 | **4.71** |
| | CD-IF (单发音词典) | 44.98 | 55.02 | 49.06 | 50.94 | **7.42** |
| | CD-GIF (多发音词典) | 46.07 | 53.93 | 49.42 | 50.58 | **6.21** |
| (IF-GIF) 精细建模 | B-GIF (含自适应和重估) | 34.44 | 65.56 | 37.29 | 62.71 | **5.11** |
| | S-GIF (含自适应和重估) | 33.66 | 66.34 | 36.47 | 63.53 | **4.24** |
| (CDW) 加权方案 | CDW-MATCH＋CI-GIF | 34.28 | 65.72 | 36.79 | 63.21 | **3.82** |
| | CDW-MATCH＋CD-GIF | 46.92 | 53.08 | 50.03 | 49.97 | **5.86** |
| | CDW-MATCH＋B-GIF | 35.85 | 64.15 | 37.77 | 62.23 | **2.99** |
| | CDW-MATCH＋S-GIF | 34.86 | 65.14 | 37.09 | 62.91 | **3.42** |
| 模型参数共享策略 | 高斯混合共享的方法 | 46.30 | 53.70 | 48.51 | 51.49 | **4.12** |
| | 模型参数组合的方法 | 46.95 | 53.05 | 48.35 | 51.65 | **2.64** |

从中可以看出，语言模型对于识别性能的改善是非常显著的，虽然本文并不涉及语音模型对识别器整体性能的提高作用。可以预见，当加入更多的语言层训练语料，或者采用音节的 trigram，甚至采用汉语词层面的 *N*-gram（bigram 或 trigram）后，整体的识别性能会得到更进一步的提高。

# 第七章　总结与展望

## 7.1　论文工作总结与贡献

本文针对自然语音的特点及其识别任务中的若干难点，以汉语的自然语音为研究对象，在其发音建模方面进行了初步的探索研究，提出了若干新方法、新策略，并通过实验证明了其有效性，同时也为继续在自然语音发音建模方面的深入研究积累了一定的经验。

概括来说，本文的工作重点与贡献主要体现在如下几个方面：

1) **提出基于广义声韵集的发音建模技术**

结合汉语自然语音的特点，并针对汉语标准声韵集的不足，定义了广义声韵集（GIF，Generalized Initial/Final），并在此基础上提出对它进行精细建模的方法，可以对更为丰富和精细的发音变化进行建模，同时也为解决训练数据的稀疏问题提供了新思路。

2) **提出上下文相关的发音词典加权策略**

从音节间混淆程度的角度出发，定义了发音词典的固有混淆度（PLIC，Pronunciation Lexicon Intrinsic Confusion）指标，可以有效地评估多发音词典的固有属性，从而为发音词典的构造问题提供直接的、定量的参考依据；提出对多发音词典进行上下文相关加权（CDW，Context-Dependent Weighting）的新方法，与传统的基于直接输出概率（DOP，Direct Output Probability）的加权方法相比，CDW 有效地避免了音节层面统计的稀疏问题，因而更加准确和有效。

3) **提出发音建模的参数共享策略**

该方法同时结合了单发音词典在音节混淆方面的优势以及利用 GIF 对发音变化进行声学建模的优势。通过对传统模型与发音模型进行参数的共享和组合，在保证目标模型具备足够的描述发音变化能力的同时，进一步降低了音节间的混淆度并改善了整体的声学识别性能。

虽然本文的研究是基于汉语的自然语音识别任务，但其中的一些关键思路，例如上下文相关的发音词典加权策略，以及发音建模的参数共享策略等，很容易推广到基于其它语言的自然语音识别任务中。

## 7.2　下一步研究的展望

本文的研究尚处于起步阶段，在许多方面还存在着一些不足。下面将指出这些不足点，以及计划进一步深入开展研究的若干方向。

1) **基元的确定问题**

本文的各项研究主要是基于 GIF 基元。它与标准的 IF 相比，具有更强的描述发音变化现象的能力，但 GIF 的确定在很大程度上依赖于数据库的精细手工标注。由于如此细节的标注需要耗费大量的人力，并且标注者需要具备专业的语音学知识，因此对于其它连续语音识别任务来说，这是件很困难的事情。一种可能的方法，就是以纯数据驱动的方式产生与自然语音较好匹配的识别基元，例如文献[Bacchiani99]介绍的自动产生识别基元的方法，或者从语音学的角度研究利用训练数据引入更丰富的发音变化的方法，例如通过提取 $F_0$ 曲线来判别元音的清化和辅音的浊化等细微音变并进行自动标注等。

2) **发音变化的推广问题**

GIF 基元集（和 CASS 数据库标注）包含了丰富的、很有意义的汉语发音变化现象，这对于其它的汉语自然语音识别研究来说也是非常有价值的。因此，如何将它们所包含的发音变化规律向其它的数据库和识别任务进行推广，也是一个值得研究的问题。本文提出两点思路：对于与 CASS 相同信道的数据库，可以直接采用与 2.3 节相同的方法，利用 GIF 模型和多发音词典对新数据库进行强制时间对准，从而获得包含更多发音变化的新标注；而对于与 CASS 不同信道的数据库，可以利用 CASS 的发音变化规律训练出可依据上下文预测发音变化的决策树或者神经网络，然后从新数据库的标准标注出发预测可能出现的音变并产生相应的初始标注等。这些方法同时也可为现有系统引入新的训练数

据从而减小数据稀疏程度。

3) **跨领域模型的结合问题**

可以利用删除插值（Deleted-Interpolation）技术，将现有的包含了较多发音变化但训练不是很充分的模型（例如 GIF 基准模型）与另一个充分训练的通用模型的参数相结合，从而达到平滑模型参数并进而提高识别性能的目的。这个思路同样也可用于将现有的发音变化规律向新的数据库或识别任务进行推广。

4) **引入跨词的相关性**

自然语音识别的一个通用技术，就是利用多发音词典来反映各种可能的发音变化。但是词典中过多的发音变化项会造成音节或词之间的混淆，并可能因此而降低识别器的性能。此外，简单地包容发音变化序列的多发音词典仅仅体现出了“词内”的发音变化，并不能很好地反映出“跨词”的发音变化。而实际上，大量的发音变化正是由于词或音节之间的协同发音而造成的，而且多发生在两个词的邻接部分。为此，如何在发音建模的过程中引入跨词的相关性，对于进一步提高识别性能是非常有潜力的。在 CDW 的加权方案中，已经引入了跨词相关性的因素，此外还可以通过在发音词典中引入包含了词间发音变化的高频复合词等方法来提高系统对发音变化的建模能力。

5) **寻找新特征和更适宜的建模方法**

自然语音具有语速快且多变、声学特性变化大等特点，而且自然语音的应用环境往往集中于背景嘈杂、电话语音等复杂的环境下，因此其信道差异较大，噪声和回声反射等问题严重。为此，需要进一步研究对现有特征鲁棒性的改进方法或者寻求新的特征参数，以及更好的声学建模技术，使得它们具有更好的噪声鲁棒性、信道的适应性，以及可更好地体现出自然语音的动态特性，从而不断提高自然语音识别器的性能。

# 参考文献


[1] Allen J F, Schubert L K, Ferguson G M *et al* **(Allen94)**. *The TRAINS project: A case study in building a conversational planning agent*. Technical Report 532, Dept. of Computer Science, University of Rochester, Rochester, NY 14627-0226, Sept. 1994

[2] Aust H, Schroer O **(Aust98)**. *Application development with the Philips dialog system*. In: Proc. ISSD. Sydney, Australia, Nov. 30, 1998. 27~34

[3] Bacchiani M, Ostendorf M **(Bacchiani99)**. *Joint lexicon, acoustic unit inventory and model design*. Speech Communication, 1999, 29: 99~114

[4] Bahl L R, Brown P, de Souza P V *et al* **(Bahl88)**. *Speech recognition with continuous-parameter hidden Markov models*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. 1988, 2: 40~43

[5] Bahl L R, de Souza P V, Gopalakrishnan P S *et al* **(Bahl91)**. *Decision trees for phonological rules in continuous speech*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. 1991. 185~188

[6] Beulen K, Ney H **(Beulen98)**. *Automatic question generation for decision tree based state tying*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. 1998, 805~808

[7] Breiman L, Friedman J H, Olshen R A *et al* **(Breiman84)**. *Classification and regression trees*. Wadsworth, Belmont, CA, 1984

[8] Byrne W, Finke M, Khudanpur S *et al* **(Byrne98)**. *Pronunciation modelling using a hand-labelled corpus for conversational speech recognition*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. 1998. 313~316

[9] Byrne W, Venkataramani V, Kamm T *et al* **(Byrne01)**. *Automatic generation of pronunciation lexicons for Madarin spontaneous speech*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. 2001, accepted and to appear

[10] 曹剑芬 **(Cao90)**. *现代语音基础知识*. 北京: 人民教育出版社. 1990

[11] 陈其光 **(Chen98)**. *语言调查*. 北京: 中央民族大学出版社. 1998

[12] Chen X-X, Li A-J, Sun G-H *et al* **(Chen00)**. *An application of SAMPA-C for standard Chinese*. In: International Conference of Spoken Language Processing. Beijing, 2000. 4:652~655

[13] Cremelie N, Martens J P **(Cremelie97)**. *Automatic rule-based generation of word pronunciation networks*. In: European Conference on Speech Communication and Technology. 1997, 5: 2459~2462

[14] Cremelie N, Martens J P **(Cremelie99)**. *In search of better pronunciation models for speech recognition*. Speech Communication, 1999, 29: 115~136

[15] Davis H K, Biddulph R, Balashek S **(Davis52)**. *Automatic recognition of spoken digits*. American Journal of Otolaryngology. 1952, 24: 637~642

[16] Decker A M, Lamel L **(Decker99)**. *Pronunciation variants across system configuration, language and speaking style*. Speech Communication, 1999, 29: 83~98

[17] Dempster A. P, Laird N M, Rubin D B **(Dempster77)**. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B (Methodological), 1977, 39(1): 1~38

[18] Denes P B, Mathews M V **(Denes60)**. *Spoken digit recognition using time frequency pattern matching*. Journal of the Acoustical Society of America. 1960, 32:1450~1455

[19] Dragon Systems **(Dragon)**. Dragon Dictate, 2000. *http://www.dragonsys.com/*

[20] Failenschmid K, Thornton J H S **(Failenschmid98)**. *End-user driven dialogue system design: The REWARD experience*. In: International Conference of Spoken Language Processing. 1998, 2: 37~40

[21] Finke M, Waibel A **(Finke97)**. *Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition*. In: European Conference on Speech Communication and Technology. 1997, 5: 2379~2382

[22] Finke M, Fritsch J, Koll D *et al* **(Finke99)**. *Modeling and efficient decoding of large vocabulary conversational speech*. In: European Conference on Speech Communication and Technology. 1999, 1: 467~470

[23] Fukada T, Sagisaka Y **(Fukada97)**. *Automatic generation of a pronunciation dictionary based on a pronunciation network*. In: European Conference on Speech Communication and Technology. 1997, 5: 2471~2474

[24] Gales M J, Woodland PC **(Gales96a)**. *Mean and variance adaptation within the MLLR framework*. Computer Speech and Language. 1996, 10: 249~264

[25] Gales M J **(Gales96b)**. *The generation and use of regression class trees for MLLR adaptation*. Cambridge University Report, 1996. *ftp://svr-ftp.eng.cam.ac.uk*

[26] Gallwitz F, Aretoulaki M, Boros M *et al* **(Gallwitz98)**. *The Erlangen spoken dialogue system EVAR: A state-of-the-art information retrieval system*. In: Proc. International Symposium on Spoken Dialogue. Sydney, Australia, 1998. 19~26

[27] Gao S, Xu B, Huang T-Y **(Gao98)**. *Class-triphone acoustic modeling based on decision tree for Mandarin continuous speech recognition*. In: International Symposium of Chinese Spoken Language Processing. 1998. ASR-A1

[28] Gauvain J L and Lee C H **(Gauvain94)**. *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*. IEEE Trans. Speech and Audio Processing, 1994, 2(2): 291~298

[29] Goldschen A, Loehr, D **(Goldschen99)**. *The role of the DARPA communicator architecture as a human computer interface for distributed simulations*. Technical report, MITRE Corporation, 1999

[30] Greenberg S **(Greenberg99)**. *Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation*. Speech Communication. 1999, 29: 159~176

[31] Holter T, Svendsen T **(Holter99)**. *Maximum likelihood modelling of pronunciation variation*. Speech Communication. 1999, 29: 177~191

[32] Huang X-D, Ariki Y, Jack M A **(Huang90a)**. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990

[33] Huang X-D, Lee K-F, Hon H-W **(Huang90b)**. *On semi-continuous hidden Markov models*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing. 1990. 689~692

[34] Huang X-D, Alleva F, Hon H-W *et al* **(Huang93)**. *The Sphinx-II Speech Recognition System: An Overview*. Computer Speech and Language. 1993, 7(2): 137~148

[35] Huang X-D, Hwang M-Y, Jiang L *et al* **(Huang96)**. *Deleted interpolation and density sharing for continuous hidden markov models*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing. Atlanta, GA, 1996. 885~888

[36] Huang C, Xu P, Zhang X *et al* **(Huang99)**. *LODESTAR: A mandarin spoken dialogue system for travel information retrieval*. In: European Conference on Speech Communication and Technology. 1999, 3: 1159~1162

[37] Huang Y-F, Zheng F, Xu M-X *et al* **(Huang00)**. *Language understanding component in Chinese dialogue system*. In: Proc. Int. Conf. Spoken Language Processing. Beijing, China: China Military Friendship Publish, 2000, 3: 1053~1056

[38] Hwang M-Y, Huang X, Alleva F **(Hwang93)**. *Predicting unseen triphones with senones*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. Minneapolis, MN, 1993. 311~314

[39] IBM **(IBM)**. *http://www-4.ibm.com/software/speech/*

[40] The International Phonetic Association **(IPA)**. *http://www2.arts.gla.ac.uk/IPA/ipa.html*

[41] Jelinek F, Mercer R L **(Jelinek80)**. *Interpolated estimation of Markov source parameters from sparse data*. In: Gelsema D and Kanal L, eds. North-Holland: Pattern Recognition in Practice, 1980

[42] Jelinek F **(Jelinek98)**. *Statistical methods for speech recognition*. The MIT Press, Cambridge, MA, 1998

[43] Katz S M **(Katz87)**. *Estimation of probabilities from sparse data for the language model component of a speech recognizer*. IEEE Trans. On Acoustic, Speech and Signal Processing, 1987, 35(3): 400~401

[44] Kessens J M, Wester M, Strik H **(Kessens99)**. *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*. Speech Communication. 1999, 29: 193~207

[45] Kim N S, Un C K **(Kim97)**. *Statistically reliable deleted interpolation*. IEEE Trans. SAP, 1997, 5: 292~295

[46] Kriouile A, Mari J F, Haton J P **(Kriouile90)**. *Some improvements in speech recognition algorithms based on HMM*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing. 1990. 545~548

[47] Lamel L, Rosset S, Gauvin J L *et al* **(Lamel98)**. *The limsi ARISE system*. In: Proceedings of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications, Torino, Italy, 1998. 209~214

[48] Lamel L, Rosset S, Gauvain J L *et al* **(Lamel99)**. *The LIMSI ARISE system for train travel information*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing. Phoenix, Arizona, 1999. 501~504

[49] Lee K-F **(Lee89a)**. *Automatic speech recognition: The development of the SPHINX system*. Bosten: Kluwer Academic Publishers, 1989

[50] Lee C-H, Rabiner L R **(Lee89b)**. *A frame-synchronous network search algorithm for connected word recognition*. IEEE Trans. On Acoustic, Speech and Signal Processing. 1989, 37(11): 1649~1658

[51] Lee C-H, Lin C-H, Juang B-H **(Lee91)**. *A study on speaker adaptation of parameters of continuous density hidden Markov models*. IEEE Trans. SP, 1991, 39(4): 806~814

[52] Lee C-H, Gauvain J L **(Lee93)**. *Speaker adaptation based on MAP estimation of HMM parameters*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 1993, 2: 652~655

[53] Lee C-H **(Lee98)**. *Spoken dialogue processing towards telecommunication applications*. In: Proc. ISSD. 1998

[54] Leggetter C J, woodland P C **(Leggetter95)**. *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*. Computer Speech and Language. 1995, 9:171~186

[55] L&H **(L&H)**. *http://www.lhsl.com/*

[56] 李荣 **(Li57)**. *汉语方言调查手册*. 科学出版社. 1957

[57] Li Z, Boulianne G **(Li96)**. *Bi-derectional graph search strategies for speech recognition*. Computer Speech and Language. 1996, 10:295~321

[58] Li A-J, Zheng F, Byrne W *et al* **(Li00a)**. *CASS: a phonetically transcribed corpus of Mandarin spontaneous speech*. In: 6th International Conference on Spoken Language Processing. Beijing, China, 2000. 1:485-488

[59] Li A-J, Lin M-C, Chen X-X *et al* **(Li00b)**. *Speech corpus of Chinese discourse and the phonetic research*. In: Proc. International Conference on Spoken Language Processing. Beijing, China, 2000. 4:13~18

[60] Li A-J, Chen X-X, Sun G-H *et al* **(Li00c)**. *Speech corpus collection and annotation*. In: Proc. International Conference on Spoken Language Processing. Beijing, China, 2000. 45~48

[61] 林焘, 王理嘉 **(Lin92)**. *语音学教程*. 北京: 北京大学出版社. 1992

[62] Lin B-S, Lee L-S **(Lin00)**. *Computer-aided design/analysis for Chinese spoken dialogue systems*. In: International Symposium of Chinese Spoken Language Processing. 2000. 57~60

[63] Liu Y, Fung P **(Liu00a)**. *Rule-based word pronunciation networks generation for Mandarin speech recognition*. In: International Symposium of Chinese Spoken Language Processing. Beijing, 2000. 35~38

[64] Liu Y, Fung P **(Liu00b)**. *Modelling pronunciation variations in spontaneous Mandarin speech*. In: International Conference on Spoken Language Processing. Beijing, China, 2000. 3:630~633

[65] Liu M-K, Xu B, Huang T-Y *et al* **(Liu00c)**. *Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 2000. 2:1025~1028

[66] Liu Y, Fung P **(Liu99)**. *Decision tree-based triphones are robust and practical for Madarin speech recognition*. In: European Conference on Speech Communication and Technology. 1999. 895~898

[67] 罗安源 **(Luo00)**. *田野语音学*. 北京: 中央民族大学出版社. 2000.

[68] Fosler-Lussier E, Morgan N **(Lussier99)**. *Effect of speaking rate and word frequency on pronunciations in convertional speech*. Speech Communication, 1999, 29: 137~158

[69] Ma K, Zavaliagkos G, Iyer R **(Ma98)**. *Pronunciation modeling for large vocabulary conversational speech recognition*. In: International Conference on Spoken Language Processing. Sydney, Australia, 1998. 6:2455~2458

[70] Ma B, Huo Q **(Ma00)**. *Benchmark results of triphone-based acoustic modeling on HKU96 and HKU99 Putonghua corpora*. In: International Symposium of Chinese Spoken Language Processing. Beijing, 2000. 359~362

[71] Meng H-M, Tsui W-C **(Meng00)**. *Comprehension across application domains and languages*. In: International Symposium of Chinese Spoken Language Processing. Beijing, 2000. 69~72

[72] Nilsson N J **(Nilsson71)**. *Problem solving methods in artificial intelligence*. New York: McGraw-Hill. 1971

[73] Odell J J **(Odell95)**. *The use of context in large vocabulary speech recognition*. PhD thesis, University of Cambridge, U.K., March 1995

[74] Os den E, Boves L, Lamel L *et al* **(Os99)**. *Overview of the ARISE project*. In: European Conference on Speech Communication and Technology. 1999. 1527~1530

[75] Paul D **(Paul92)**. *An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model*. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 1992, 1: 25~29

[76] Philips **(Philips)**. *http://www.philips.com/*

[77] Potamianos A, Kuo H-K, Lee C-H *et al* **(Potamianos99)**. *Design principles and tools for multimodal dialog systems*. ESCA Workshop: Interactive Dialogue in Multi-Modal Systems, 1999

[78] Rabiner L R **(Rabiner89a)**. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, 1989, 77(2): 257~285

[79] Rabiner L R, Wilpon J G, Soong F K **(Rabiner89b)**. *High performance connected digit recognition using hidden Markov models*. IEEE Trans. On Acoustics, Speech and Signal Processing. 1989, 37(8):1214~1225

[80] Rabiner L R, Juang B-H **(Rabiner93)**. *Fundamentals of speech recognition*. Prentice-Hall International, Inc. 1993

[81] Reichl W, Chou W **(Reichl98)**. *Decision tree state tying based on segmental clustering for acoustic modeling*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. Seattle, WA, May, 1998. 801~804

[82] Renals S, Hochberg M **(Renals96)**. *Efficient evaluation of the LVCSR search space using the noway decoder*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. 1996, l: 149~152

[83] Riley M D, Ljolie A **(Riley96)**. *Automatic generation of detailed pronunciation lexicons*. In: Lee C H, Soong F K, Paliwal K K, eds. Boston: Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, 1996

[84] Riley M, Brne W, Finke M *et al* **(Riley99)**. *Stochastic pronunciation modelling from hand-labelled phonetic corpora*. Speech Communication, 1999, 29: 209~224

[85] Ronald R **(Ronald00)**. *Two decades of statistical language modeling: where do we go from here?*. In: Proceedings of the IEEE. 2000, 88(8)

[86] Rudnicky A *et al* **(Rudnicky99)**. *Creating natural dialogs in the Carnegie Mellon communicator system*. In: European Conference on Speech Communication and Technology. 1999, 4: 1531~1534

[87] University College London, Department of Phonetics and Linguistics **(SAMPA)**. *http://www.phon.ucl.ac.uk/home/sampa/home.htm*

[88] Saraclar M, Nock H, Khudanpur S **(Saraclar99)**. *Pronunciation modeling by sharing Gaussian densities across phonetic models*. In: European Conference on Speech Communication and Technology. 1999, 1:515~518

[89] Selim S Z, Ismail M A **(Selim84)**. *K-Means-type algorithms: a generalized convergence theorem and characterization of local optimality*. IEEE Trans. Pattern Analysis and Machine Intelligence. 1984, 6(1), 81~87

[90] 宋战江, 郑方, 吴文虎 **(Song00a)**. *汉语连续语音识别系统与知识导引的搜索策略研究*. 自动化学报. 2000, 26(4): 470~477

[91] Song Z-J, Zheng F, Wu W-H **(Song00b)**. *Statistical knowledge based frame synchronous search strategies in continuous speech recognition*. In: International Conference on Acoustics, Speech and Signal Processing. Istanbul, Turkey, 2000, 3: 1583~1586

[92] Strik H, Cucchiarini C **(Strik99)**. *Modeling pronunciation variation for ASR: A survey of the literature*. In: Speech Communication, 1999, 29: 225~246

[93] Sutton S, Cole R *et al* **(Sutton98)**. *Universal speech tools: the CSLU toolkit*. In: Proceedings of the International Conference on Spoken Language Processing. Sydney, Australia, 1998. 3221~3224

[94] Takezawa T, Yamamoto S **(Takezawa98)**. *Dialogue processing in a speech-to-speech translation system between Japanese and English*. In Proc. ISSD. 1998

[95] Viterbi A J **(Viterbi67)**. *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*. IEEE Trans. On IT. 1967, 13(2): 260~267

[96] Wahlster W **(Wahlster93)**. *VERBMOBIL: Translation of face-to-face dialogs*. In: European Conference on Speech Communication and Technology. Berlin, Germany, 1993. 29~38

[97] Willet D, Neukirchen C, Rottland J *et al* **(Willet99)**. *Refining tree-based state clustering by means of formal concept analysis, balanced decision treesand*

*automatically generated model-sets*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. Phoenix, AZ, 1999. 565~568

[98] Witten I H, Bell T C **(Witten91)**. *The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression*. IEEE Trans. On Information Theory. 1991, 37(4): 1085~1094

[99] 吴宗济, 林茂灿, 等 **(Wu89)**. *实验语音学概要*. 北京: 高等教育出版社. 1989

[100] 吴宗济 **(Wu97)**. *试论人－机对话中的汉语语音学*. 世界汉语教学, 1997, 42(4): 3~20

[101] 杨行峻, 迟惠生, 等 **(Yang95)**. *语音信号数字处理*. 电子工业出版社. 1995

[102] Young S J **(Young92)**. *The general use of tying in phoneme-based HMM speech recognizers*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. San Francisco, CA, 1992. 569~572

[103] Young S J, Odell J J, Woodland P C **(Young94)**. *Tree-based state tying for high accuracy acoustic modeling*. In: Proceedings ARPA Workshop on Human Language Technology. 1994. 307~312

[104] 袁家骅, 等 **(Yuan83)**. *汉语方言概要*. 北京: 文字改革出版社. 1983

[105] Zheng F, Wu W-H, Fang D-T **(Zheng96)**. *Speech recognition units in the Chinese dictation machines*. In: 4th National Conf. On Man-Machine Speech Communication (NCMMSC-96). Beijing, 1996. 32~35

[106] Zheng F, Song Z-J, Xu M-X *et al* **(Zheng99a)**. *EasyTalk: A large vocabulary speaker-independent CDM*. In: European Conference on Speech Communication and Technology. Budapest, Hungary, Sept., 1999, 2: 819~822

[107] 郑方, 牟晓隆, 徐明星, 等 **(Zheng99b)**. *汉语语音听写机技术的研究与实现*. 软件学报, 1999, 10(4): 436~444

[108] Zheng F **(Zheng99c)**. *A syllable-synchronous network search algorithm for word decoding in Chinese speech recognition*. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing. Phoenix, 1999, 2: 601~604

[109] Zheng F, Song Z-J, Fung P *et al* **(Zheng00)**. *Mandarin pronunciation modeling based on CASS corpus*. In: Sino-French Symposium on Speech and Language Processing. Beijing, 2000. 47~53

[110] 朱川 **(Zhu86)**. *实验语音学基础*. 上海: 华东师范大学出版社. 1986

[111] Zue V **(Zue97)**. *Conversational interfaces: Advances and challenges*. In: European Conference on Speech Communication and Technology. KN:9~18, 1997

# 附录　博士就读期间完成的其它研究任务

## 附录A　汉语发音水平评价方法的研究

### A.1　研究背景与现状

智能话语水平评价（$S^2E$, Speaking Skill Evaluation）技术也被称为自动评分（Automatic Pronunciation Scoring）技术，是将计算机技术和数字语音处理技术等运用到语言教学中的综合产物。它实际上是一个指导用户学习语言的人工智能专家系统，其研究涉及到语音学、语言学、教育学、人工智能、信号处理、数学建模等多个学科，是一个崭新的研究领域。

近年来，国外已有一些学术研究机构开始研究 $S^2E$ 系统，并在实际中有了一定的应用。真正的智能 $S^2E$ 系统是近几年才出现的，早期的 $S^2E$ 系统一般是与文本有关的，后来又相继研究出与文本无关的、具有发音错误定位和语音化反馈能力的 $S^2E$ 系统，并提出了一系列行之有效的发音自动评分算法。美国政府在研究计算机辅助语言教学的计划中已将其列为重要项目给予支持，一些计算机公司的商品化与文本无关的智能 $S^2E$ 原型系统也已经或将在近期推出。

智能 $S^2E$ 技术的研究意义并不仅仅限于计算机辅助语言教学。它是从传统的连续语音识别（CSR，Continuous Speech Recognition）技术发展起来的一个新兴的、综合的交叉学科。它所聚焦的基元时间对准与评分策略、错误检测与定位技术、基于语音修正的反馈方法等关键技术，不仅大量应用了语音识别、语音合成、语言韵律学、概率统计等多方面的知识和研究成果，而且反过来又会推动这些学科更深入发展。尤其是在连续语音识别技术的发音确认和拒识等方面，智能 $S^2E$ 的研究成果将会对这些技术在连续语音识别中的应用起到不可低估的促进作用。

总的说来，一个智能化的 $S^2E$ 系统应该具有如下几个主要功能模块：(1)标准的教师发音提示与标准基元库的训练；(2)对用户读音的话语水平进行评分，并将这个评估得分映射到与人的主观感觉基本一致的数量级上；(3)对用户读音

中的错误进行准确定位；(4)将综合评估结果以及发音修正建议向用户进行反馈，以便进一步纠正其读音，达到语言学习的目的。

传统 CSR 技术的目标是要将发音准确地映射到识别基元序列上去，这个过程相当于一系列的“是/否”判断，即需要判断每个发音片断是否属于特定的识别基元；而 $S^2E$ 技术不仅要判断出用户的读音片断是否属于规定的识别基元，而且还需要评估用户读音与期待标准发音之间的相似程度，并将这个相似程度映射到与人类主观感觉基本上一致的评分等级上去。由此可见，$S^2E$ 在某些方面的研究目标要比 CSR 更加深入和细致。

## A.2　发音水平评价的关键技术

在 $S^2E$ 相关的研究中，如下三方面的技术是最关键的。

第一，是发音水平的自动评分算法。这包括基于 HMM 的似然对数得分、基于 HMM 的后验概率得分、基于识别分类错误率得分，以及基于语速或韵律匹配得分等的评分方法。

第二，是分数的综合与映射方案。为了使 $S^2E$ 系统的评分结果更加接近于人的主观评测标准，往往需要将多个发音水平评分算法的结果进行综合，并采取线性或非线性的方式将上述概率得分映射到与人的主观评价标准基本一致的数量级上，作为 $S^2E$ 系统对用户跟读语句的最终评估得分。

第三，是发音错误的定位和语音化反馈。利用 CSR 中的发音确认和拒识等技术准确定位用户发音中的错误或不准确读音，并以语音交互的方式回放用户读音与标准发音。在回放过程中，通过适当的语音修正来重点突出错误部分，使用户可以明显体会到自己的发音与标准读音之间的差异，从而有意识地改正。

此外，$S^2E$ 系统一般是基于传统 CSR 技术的 HMM 模型的（也有少量系统采用人工神经网络）。在进行评分之前，首先由用户模仿标准读音进行发音，然后由 $S^2E$ 系统通过 Viterbi 解码过程产生初始的对识别基元边界的划分，之后再进入上述的自动评分阶段。

## A.3　汉语发音水平评价系统的设计

为进行汉语的计算机辅助语言教学及相关的多媒体应用，本人设计并实现了一个针对汉语的发音水平评价系统 $CS^2E$，其功能结构见图 A-1。

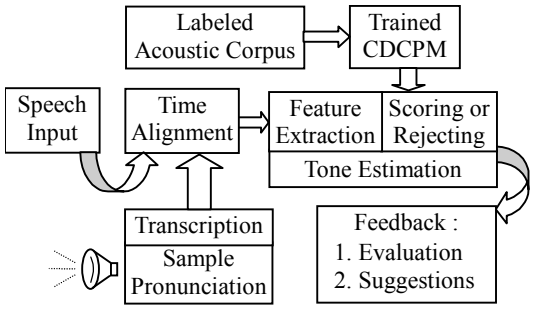


图 A-1 汉语发音水平评价系统 $CS^2E$ 的功能结构

在后续部分将对 $CS^2E$ 中的若干实现技术进行简单介绍。

## A.4　声学模型与时间对准策略

为进行后续的自动评分，并简化问题起见，采用中心距离正态分布 CDN（Center-Distance Normal distribution）来描述系统中声学矢量的输出概率分布。

假设随机变量 $\xi$ 的概率密度函数为正态分布 $N(x;\mu_x,\sigma_x)$，其中 $\mu_x$ 和 $\sigma_x$ 分别为均值和标准方差。定义一个新的随机变量 $\eta=|\xi-\mu_x|$，其概率密度函数为：

$$p(y;\sigma_x)=\frac{2}{\sqrt{2\pi}\sigma_x}\exp(-y^2/2\sigma_x^2)，(y\geq 0) \tag{A.1}$$

其中 $\eta$ 的均值 $\mu_y$ 可以利用公式 $\mu_y=2\sigma_x/\sqrt{2\pi}$ 求出。实际上，$\eta$ 正是正态分布的随机变量 $\xi$ 及其均值 $\mu_x$ 之间的距离，由此公式 A.1 中的正态分布可以称为 CDN 分布。分别定义两个标量之间和两个 $D$ 维矢量之间的加权欧氏距离为：

$$d(x_1,x_2)=|x_1-x_2| \quad 和 \quad d(\vec{x}_1,\vec{x}_2)=\sqrt{\sum_{d=1}^{D}w_d(x_{1d}-x_{2d})^2} \tag{A.2}$$

于是，若将一个矢量$\vec{\xi}$与其均值矢量$\vec{\mu}_x$之间的加权欧氏距离表示为一个随机变量$\eta$（这是一个标量），则这个矢量的 CDN 分布为：

$$N_{CD}(\vec{x};\vec{\mu}_x,\mu_y)=\frac{2}{\pi\mu_y}\exp(-d^2(\vec{x},\vec{\mu}_x)/\pi\mu_y^2) \tag{A.3}$$

事实上，$N_{CD}(\vec{x};\vec{\mu}_x,\mu_y)$不是$\vec{\xi}$的概率密度函数，而是$d(\vec{\xi},\vec{\mu}_x)$的概率密度函数。通过采用 CDN 分布，将语音矢量的概率分布描述进行了降维，从而压缩了模型规模并提高了评分过程的实时性。

有关 HMM 模型间距离方面的研究表明，转移概率矩阵的作用要远远低于声学输出概率矩阵，为此，在进行汉语的发音水平评价过程中，采用了简化的 HMM 模型 CDCPM（Center-Distance Continuous Probability Model）来为汉语发音进行建模。CDCPM 最主要的特征是抛掉了 HMM 中的转移概率矩阵，并利用混合 CDN 分布代替了传统的高斯混合分布。混合 CDN 可以表示为：

$$b_n(\vec{c})=\sum_{m=1}^{M}g_{nm}N_{CD}(\vec{c};\vec{\mu}_{xnm},\mu_{ynm}) \tag{A.4}$$

其中下标 $n$ 表示第 $n$ 个 CDCPM（HMM）状态，$\vec{c}$ 代表语音特征矢量，$g_{nm}$ 为各个 CDN 分布的权重，$\vec{\mu}_{xnm}=(\mu_{xd}^{(nm)})$和$\mu_{ynm}$分布为第 $n$ 个状态内第 $m$ 个 CDN 分布中的均值矢量和中心距离（随机变量）。

模型的训练数据来自于 863 数据库，利用其中的 38 人男声数据和 38 人女声数据分别训练出男声模型和女声模型，采用了 16 维 LPCC 及其自回归矢量作为语音特征，共 32 维。

由于已经抛掉了 HMM 的转移概率矩阵，因此采用了不同于标准 Viterbi 算法的方法进行状态的时间对准。这里首先对输入语音流采用基于归并状态机 MBSDA（Merging-Based Syllable Detection Automaton）的切分引擎进行音节边界检测，之后利用非线性分段 NLP（Non-Linear Partition）来为每个音节内部划分状态边界。对 MBSDA 将在 B.2 节中介绍，这里仅仅介绍 NLP 的原理。

假设某个识别基元（这里为音节）对应的语音特征矢量为$\{\vec{O}_1,\vec{O}_2,\cdots,\vec{O}_T\}$，其中 $T$ 为帧长。定义相邻两帧矢量间的声学差异为$\Delta_t=\|\vec{O}_{t+1}-\vec{O}_t\|(1\le t\le T)$，其距离度量采用公式 A.2。假设每个基元内部包含 $N$ 个状态，且这些状态之间的

累积声学变化量相等，为

$$\Delta_{state} = \frac{1}{N}\sum_{t=1}^{T}\Delta_t \qquad (1 \le n \le N-1) \tag{A.5}$$

令 $L_n$ 为前 $n$ 个状态包含的总帧数。如果存在 $k$，使得

$$\sum_{t=1}^{k}\Delta_t < n \cdot \Delta_{state} \le \sum_{t=1}^{k+1}\Delta_t \tag{A.6}$$

那么 $L_n$ 的值就为 $k$，即 $L_n$ 为状态 $n$ 和状态 $n$+1 之间的分界点，$L_N$ 等于 $T$。在获得了各个基元状态的时间边界信息后，就可以结合 CDCPM 的模型参数，利用下面的方法为用户跟读的发音进行自动评分了。

## A.5　自动评分方法

考虑一个正态分布的随机变量 $x \in (-\infty, \infty)$，其均值为 $\sigma$，从统计意义上看，大约有 95%左右的样本会落在 $[\mu - 2\sigma, \mu + 2\sigma]$ 的区域内，把这种区域定义为关键区 CA（Critical Area）。类似地，对于具有参数 $\bar{\mu}_x$ 和 $\mu_y$ 的 CDN 分布而言，关系式 $\sigma = \sqrt{2\pi}\mu_y / 2 \approx 1.25\mu_y$ 成立，因此包含 95%样本的 CA 就约为 $[0, 2.5\mu_y]$。在此利用这种语音矢量与混合 CDN 分布的 CA 之间的关系来为用户读音进行自动评分，该方法称为关键区百分比 CAP（Critical Area Percentage）方法。

首先选择一个阈值 $TH$ 作为 CAP 中的“百分比”，根据这个阈值就可以度量落入某个关键区 $[0, TH \cdot \mu_y]$ 的样本的百分比。假设 CDCPM 模型的第 $n$ 状态内的第 $m$ 个 CDN 分布的参数为 $\Lambda = \{\bar{\mu}_{xnm}, \mu_{ynm} \mid 1 \le n \le N, 1 \le m \le M\}$，那么对于音节的观测语音矢量 $\mathbf{O} = \{\vec{O}_1, \vec{O}_2, \cdots, \vec{O}_T\}$，基于 CAP 的得分为：

$$\mathbf{S}(\mathbf{O} \mid \Lambda) = \sum_{t=1}^{T}\sum_{m=1}^{M} S(\vec{O}_t \mid n(t), m, \Lambda) \Big/ (T \cdot M) \tag{A.7}$$

其中 $n(t)$是 $t$ 时刻语音帧所属的状态号，它是利用上述的 MBSDA 切分引擎和 NLP 算法得到的。$S(\vec{O}_t \mid n, m, \Lambda)$ 被定义为“部分帧得分”，即

$$S(\vec{O}_t \mid n, m, \Lambda) = \begin{cases} 1, & d(\vec{O}_t, \bar{\mu}_{xnm}) \in [0, TH \cdot \mu_{ynm}] \\ 0, & otherwse \end{cases} \tag{A.8}$$

由于音调在汉语的表情达意中起非常重要的作用，而学汉语的外国人或国内某些方言区人们的发音中，音调不标准是很常见的现象。因此在自动评分中，也同时考虑了对用户读音中音调的估计并计入总分数，以督促用户有目的地去修正不正确的音调。在这里通过对用户发音中的 F0 曲线的变化趋势进行估计，从而得到音调的候选，并根据汉语中的一些连读变调规则进行相应的得分调整。

正如 A.2 节所述，需要对上述自动得分进行映射，以达到与人类主观感觉相一致的等级。这里采用了一个简单的线性函数进行分数映射和发音拒识。

首先根据若干实验，确定一个经验值 $S_{REJ}$ 作为得分的拒识阈值。即如果用户发音的得分 $\mathbf{S}(\mathbf{O}\mid\Lambda)\le S_{REJ}$，那么就认为是发音错误或偏差过大，系统拒绝评分或要求重新跟读，否则，最终的得分 $\mathbf{F}(\mathbf{O}\mid\Lambda)$ 为

$$\mathbf{F}(\mathbf{O}\mid\Lambda)=\left(\mathbf{S}(\mathbf{O}\mid\Lambda)-S_{REJ}\right)/\left(1-S_{REJ}\right) \tag{A.9}$$

显然 $0<\mathbf{F}(\mathbf{O}\mid\Lambda)\le 1$。其它一些 $S^2E$ 系统，采用了更加复杂的自动评分和非线性的得分映射方案，这往往需要根据具体的应用环境和实现需求来确定。

本文在汉语发音水平评价方面的工作属于起步性研究。虽然采用了简化的 HMM 模型以及一些比较简单的评分算法，但是其实时性以及实际的评分测试，仍然达到了比较令人满意的效果。

## A.6 参考文献

有关这部分的详细研究内容及相关引用，请参阅下列文献：

[1] Song Z-J, Zheng F, Xu M-X, Wu W-H. *An effective scoring method for speaking skill evaluation system*. In Proc. European Conference on Speech Communication and Technology (EuroSpeech). 1999, 1: 187-190

[2] 吴文虎，宋战江，王帆。*实用智能话语系统的思路、策略与方法研究*。见：澳门资讯年会论文集。澳门，1998 年 1 月，174~177

[3] 徐明星，宋战江，郑方，吴文虎。*汉语语音水平评价方法研究*。见：第五届全国人机语音通信学术会议论文集。哈尔滨，1998 年 7 月，286~289

# 附录B　汉语听写机中的搜索策略研究与系统集成

## B.1　搜索策略的研究背景

这部分研究内容基于清华大学计算机系语音技术中心开发的大词汇量、非特定人、连续汉语语音识别系统（又称听写机）*EasyTalk*。对于输入的连续语音流，*EasyTalk* 的声学识别层给出由多候选音节串组成的音节网络，然后由语言处理层进行组句分析，得到最终的识别结果。

在声学层面上，用来进行时间对准的搜索策略是一个非常重要的组成部分。采用 HMM 拓扑结构时，搜索算法基本上可以分为两大类：一类是时间同步的，如著名的 Viterbi 解码算法和帧同步搜索算法等；另一类是非时间同步的，如堆栈解码算法和 A*搜索算法等。此外还有一些更加复杂的搜索算法，如双向图搜索算法等。

帧同步搜索算法不但简洁高效，而且可以给出多个声学候选，因此在 *EasyTalk* 中，就采用基于帧同步的搜索算法来产生声学层面的候选音节网络。为进一步提高搜索效率和识别性能，本文提出了基于统计知识的帧同步搜索（SKB-FSS, Statistical Knowledge Based Frame Synchronous Search）算法。它通过基于归并的音节切分自动机产生若干可靠的音节边界点供搜索过程使用，利用状态驻留长度的统计分布和词搜索树来约束搜索过程，通过动态前向预测来进行路径剪枝。在后续部分将首先重点介绍该搜索策略。

## B.2　声学模型、语言模型及切分预处理

图 1-1 给出了 *EasyTalk* 核心识别引擎的功能结构示意图，它的更详细的功能结构见图 B-1。

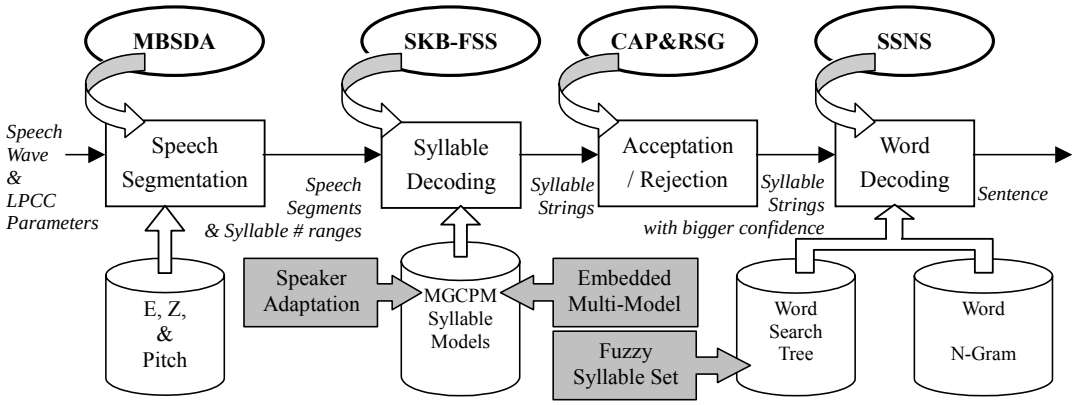


图 B-1 汉语听写机 EasyTalk 的详细功能结构

对模型之间距离度量的研究和实验表明，HMM 中状态转移概率矩阵的作用远不如观测概率矩阵重要。因此，与 A.4 节的模型相同，我们对标准 HMM 模型进行修改，去掉了状态转移概率矩阵，仅保留其观测概率矩阵。各状态内部的特征空间采用混合高斯分布进行描述，其协方差矩阵采用对角形式。系统以汉语的 418 个无调单音节作为识别基元进行建模。

训练数据库采用 863 汉语普通话连续语音数据库，它是在安静环境下以 16 位精度和 16KHz 采样率进行采样的。语音特征选为 16 阶 Mel 倒谱系数 MFCC 及其自回归分析系数，帧长 32ms，帧移 16ms，回归分析宽度为 5 帧。数据库已进行了单音节的手工初始标注。

在模型的训练阶段，首先采用鲁棒性较高的非线性分段算法 NLP 给出每个单音节内部各个状态的初始分点，计算出各个状态内特征的初始观测概率密度函数，然后进行迭代，直到各个模型的状态分点达到稳定；而在识别阶段，采用基于统计知识的帧同步搜索算法 SKB-FSS 来产生声学候选音节网络。

在声学搜索中有两个问题不容忽视，一是搜索路径的组合爆炸问题，二是解码出的状态序列错位问题。事实上，由于汉语的连续语音是以词的边界为瞬间间歇的，若能在声学搜索中加入词边界判决信息，就可以在一定程度上缓解上面两个问题，同时部分地解决语言层的分词问题。因此我们设计了对动态语音数据流进行预处理的切分引擎，它充分利用了声学、语言等方面的统计知识和规则，不断地从语音流中分离出一些相对独立和完整的语音段供后续搜索过

程进一步处理。

切分引擎采用“基于归并的音节切分自动机”思想，即 A.4 节中提到的 MBSDA。它充分利用语音的短时能量、过零率、基音周期和傅里叶频谱等多种特征参数及其差分信息，把特征参数高度相似的相邻多帧语音（它们被认为属于相同的发音状态）进行归并，形成归并类似段（MSS, Merged Similar Segment）。这些 MSS 经过一个包含静音（噪声）、声母、伪噪声、韵母、韵尾等状态的“音节切分自动机”后，输出候选的音节边界点。图 B-2 是它的状态转移示意图。

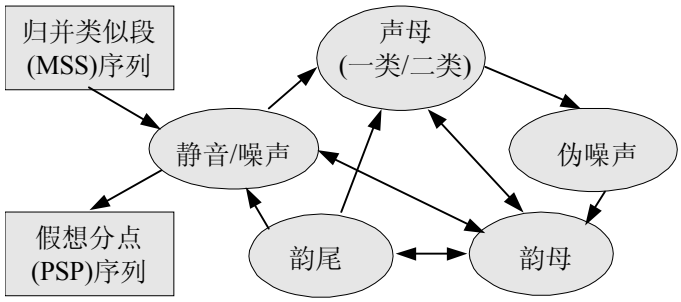


图 B-2 音节切分自动机的状态转移示意图

音节切分自动机充分利用 MSS 中的多种特征和汉语特有的声韵音节结构，以及依据声韵及噪声段的统计特性而制定的若干规则，给出一系列候选音节边界点，它们被称为假想切分点（PSP, Putative Separation Point）。每个 PSP 同时由具有一个信任度值，表明对其正确性的把握程度。信任度值高于某个预设的“接受阈值”的 PSP 被认为是正确的音节分点，定义为真实切分点（TSP, True Separation Point）；否则被认为是错误的或不确定的。

可以把“接受阈值”选得比较苛刻，从而保证认定的 TSP 有接近 100%的正确率。对于那些没有确定为 TSP 的实际音节分点，将由后续的声学搜索过程来发现。另一方面，根据音节切分自动机经历的状态转移过程，并通过一系列规则，可以在每个语音段内准确地估计出它所包含的音节个数范围。切分引擎给出的 TSP 以及有效语音段内的音节个数范围，可以作为提供给后续声学搜索过程的一种导引知识，降低状态解码边界的不确定性，显著地提高搜索效率，使识别速度达到或接近实时。

相邻两个 TSP 之间的有效语音段定义为确定段，每个确定段内包含一个或多个音节，根据汉语语音的特点和发音速度，确定段内不会含有太多的音节数

目。随后的 SKB-FSS 将在每个确定段内部进行，以产生多候选音节串组成的网络，供语言层面进行组句分析。

*EasyTalk* 的基本词汇有 5 万多个词，其中单音节词占 20.81％，双音节词占 65.50％，三音节词占 7.36％，四音节词占 6.33％。此外用户还可任意添加不长于 10 个音节的定制词汇。语言层采用基于 trigram 的统计语言模型，训练语料库包括 1993 和 1994 年《人民日报》的全文，以及《市场报》、《新华社文稿》、《经济日报》的摘编等约 2 亿字的文本，并预先进行了分词和词号标注。在语言模型中，使用一种基于 Turing 概率估计的平滑算法来解决由于数据稀疏而造成的 0 概率 trigram 问题，从而降低语言模型的困惑度。

与声学层面的识别过程相似，语言层面的组句分析也是通过搜索实现的。在 *EasyTalk* 中，根据声学识别层给出的候选音节网络，以及语言模型提供的候选词串的 trigram 概率，采用“音节同步网络搜索”算法，来得到最终的候选语句（目标词串）识别结果。

## B.3　基于统计知识的帧同步搜索

在传统的帧同步搜索算法的实现中，搜索过程是逐帧进行的。对于截止在时刻 $t$ 的每一条部分路径，时刻 $t$+1 的特征矢量被假设为可属于这些路径的任何可能的后续状态。于是搜索进行到 $t$+1 时刻时，$t$ 时刻的每条部分路径都根据其后续的可能状态列表，用 $t$+1 时刻的特征矢量扩展出若干条新的部分路径。经过一些必要的状态合并动作并扔掉一些低竞争力的候选路径后，搜索过程再向前推进一帧。重复这一过程直到语音结尾。

帧同步搜索算法简洁而且高效。然而对于大词汇量、连续语音的识别任务来说，随着时刻 $t$ 的增加，扩展出来的搜索路径会急剧增加。因此需要根据一定的准则随时剪除一些低竞争力的路径。但是，较严的阈值可能会造成一些正确的路径被过早扔掉（这将无法在后续过程中得到恢复）；而较宽的阈值又会大大增加存储空间和搜索过程的负担。

针对上述算法的不足，可以考虑将一些与帧驻留长度有关的统计知识应用于搜索过程的状态转移中，以降低搜索复杂度。有两类统计知识可以利用：

一种是基于纯统计知识的概率描述。比较典型的是对状态驻留长度进行建模，用概率密度来刻划状态驻留长度的分布情况。在搜索时，把系统处在当前状态、当前驻留长度下的条件概率作为惩罚分数加进路径的似然得分中，以此控制搜索路径的取舍。

另一种则是基于统计知识的规则。比如，根据类似的方法统计得到状态驻留长度分布的直方图，搜索时只有驻留长度落在允许范围内的路径才可以进行相应的状态转移或驻留，这可以看成是把第一种方法中的概率分布近似为均匀分布。下面讨论的重点就是这种基于规则的统计知识在帧同步搜索中的应用。

状态驻留分布（SDD, State Dwell Distribution），或被称作统计直方图，是被广泛使用的用以进行搜索剪枝的一种信息。按照 32ms 的帧宽、16ms 的帧移对正常语速的 863 数据库进行统计（采用 6 状态 HMM），得到了如下状态内驻留帧数的统计结果：每个状态内驻留的帧数比例最大的是 2 帧，平均覆盖了 39.6%；而比例最大的驻留帧数区间为 0~5 帧，覆盖了 98.7%以上，或 1~4 帧，覆盖了 95.0%以上的状态。

这里摒弃利用驻留帧数的分布概率来作为惩罚分数的做法，而是确定一个允许的驻留帧数区间$[D_{\min}, D_{\max}]$，该区间确定了在搜索时哪些驻留长度是允许的。很显然，若仅使用 SDD，对语速的变化不会有很好的鲁棒性，也不能保证很好的识别率。因为绝对的区间宽度限制了语速的变化范围，如果发音过长或过短，就不可能得到正确的识别结果。

因此，考虑到状态驻留分布 SDD 对语速变化的低鲁棒性，不再以绝对的驻留区间作为控制状态转移的参考因素，而是采用差分状态驻留分布（DSDD, Differential State Dwell Distribution）作为参考。相邻状态间差分驻留帧数的统计结果为：差分驻留帧数主要集中于 0 帧，平均覆盖了 32.7%；而比例最大的差分驻留帧数区间为-4~4 帧，覆盖了 99.8%以上，或-2~2 帧，覆盖了 95.4%以上的状态。

因此，对于第一个状态，仍然给它分配一个较宽的允许驻留帧数的范围，以保证后续各个状态驻留范围的灵活性；而对于其它的状态，分别为之确定一个允许的差分驻留帧数区间$[d_{\min}^{(s)}, d_{\max}^{(s)}], (s>0)$，于是其有效状态驻留帧数定义为$[D^{(s)}+d_{\min}^{(s)}, D^{(s)}+d_{\max}^{(s)}]$。其中，$D^{(s)}$可以定义为第 $s$-1 状态的驻留帧数（这

被称为 DSDD-LST），也可以定义为前 *s*-1 个状态的平均驻留帧数（这被称为 DSDD-AVG）。显然，采用这种搜索过程中的状态转移控制策略，就很容易与语速的变化相匹配了。

## B.4　SKB-FSS 的路径剪枝策略

虽然采用 SDD 或 DSDD 信息后，可以大大降低搜索时部分路径的膨胀速度，但是其空间消耗仍然很大。因此在 *EasyTalk* 中，又采取了如下措施对搜索进行限制和剪枝：

第一，采用词（音节）搜索树作为路径扩展过程中音节转移时的词法限制。词搜索树中每个叶结点都对应于词表中的一个词，从根结点到叶结点的树枝上的每个非叶结点依次代表了这个词的一个音节读音，且具有起始于根结点的相同“部分发音序列”的词共享相同的“部分非叶结点序列”。当某个部分路径发生音节转移时，仅仅根据词搜索树扩展出可能的后续音节，而不是所有的 418 个音节候选。这就可以大大压缩搜索空间的规模。词搜索树结构见图 B-3。

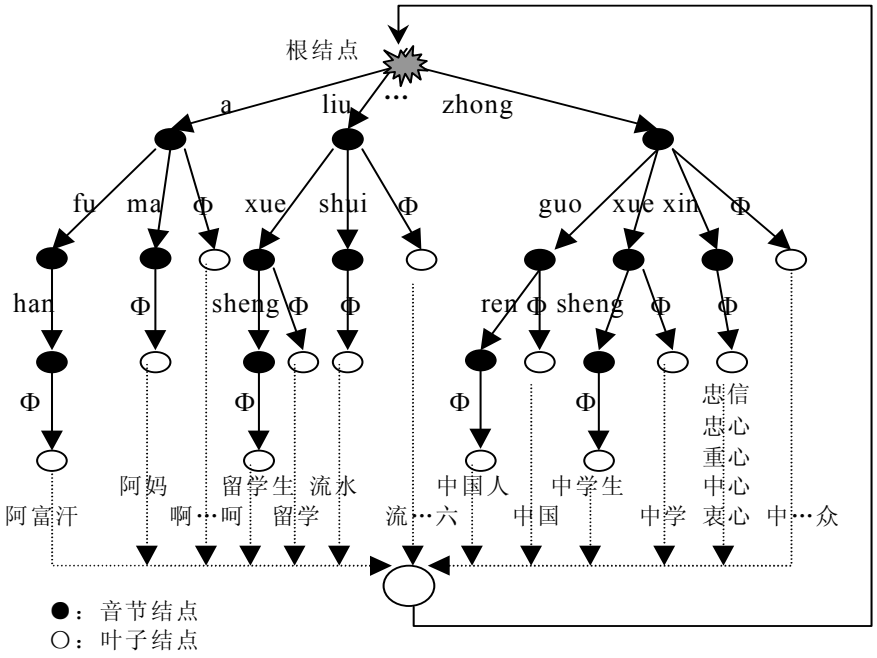


图 B-3　词搜索树示意图

第二，每当完成当前帧的路径扩展时，都要对具有相同候选音节序列和相同当前状态的那些路径进行筛选，仅保留其中具有最高累积声学得分的一条或多条路径。根据 HMM 状态转移的无后效性原理，这种路径剪枝策略是合理的。它可以及时扔掉一些低竞争力的部分路径，以保证后续扩展时较小的空间开销，并且减少多余路径的干扰。

第三，采用动态向前预测路径有效性的方法进行路径剪枝。根据当前确定段的 TSP 和音节个数范围信息，SKB-FSS 必须保证在搜索结束时，有效路径的结尾候选音节的末状态恰好落在确定段的右边界上，否则这条路径就是不合理的。因此，在每帧完成扩展之后，根据 SDD 或 DSDD 中每个后续状态的允许驻留帧数范围，确定每条路径从当前状态出发，总共可以到达的最小和最大的帧数范围$[F_{\min}, F_{\max}]$，然后与当前确定段的总帧数$F_{DS}$相比，若$F_{\min} > F_{DS}$或$F_{\max} < F_{DS}$，则说明这条路径将要扩展的“任何一条路径”中的某音节的末状态都不可能恰好落在确定段的右边界上的，因此它被作为无效路径立即删除。这个过程见示意图 B-4。

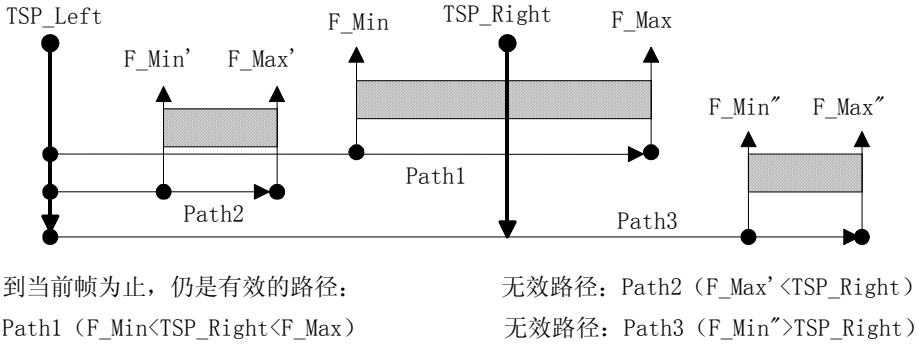


图 B-4 在 SKB-FSS 中采用的动态前向预测剪枝方法

通过上述的路径扩展限制和剪枝策略，即保证了较小的时空开销、提高了搜索效率，也减少了无效路径的干扰，提高了声学层面的整体识别率。

## B.5　增强系统鲁棒性的方法

在 *EasyTalk* 的具体实现中，还采用了说话人聚类和模型自适应的方法来增强声学模型的鲁棒性，以及采用音节的“模糊集”等来增强语言层面上对轻微口音的适应能力等。详见下面列出的参考文献。

## B.6　参考文献

有关这部分的详细研究内容及相关引用，请参阅下列文献：

[1] 宋战江，郑方，吴文虎。*汉语连续语音识别系统与知识导引的搜索策略研究*。见：自动化学报。2000 年，26(4): 470~477

[2] 张继勇，郑方，杜术，宋战江，徐明星。*连续汉语语音识别中基于归并的音节切分自动机*。见：软件学报。1999 年，10(11): 1212~1215

[3] Song Z-J, Zheng F, Wu W-H. *Statistical knowledge based frame synchronous search strategies in continuous speech recognition*. In: International Conference on Acoustics, Speech and Signal Processing. 2000, 3: 1583~1586

[4] Zheng F, Song Z-J, Xu M-X *et al*. *EasyTalk: a large-vocabulary speaker-independent Chinese Dictation Machine*. In Proc. European Conference on Speech Comm. and Technology (EuroSpeech). 1999, 2: 819-822

[5] Zheng F, Wu J, Song Z-J. *Improving the Syllable-Synchronous Network Search Algorithm for Word Decoding in Continuous Chinese Speech Recognition*. Journal of Computer Science and Technology (JCST). 2000, 15(5): 461~471

# 个人简历

宋战江，男，汉族，1972 年 9 月出生于黑龙江省牡丹江市。

1990 年考入南开大学计算机与系统科学系。于 1994 年在南开大学获得计算机软件专业学士学位；于 1997 年在南开大学获得计算机应用专业硕士学位，研究方向为计算机网络与信息系统。

1997 年 9 月考入清华大学计算机科学与技术系，继续攻读博士学位，专业为计算机应用技术，研究方向为语音识别和理解。即将获得工学博士学位。

# 在学期间的研究成果

1. 研究并开发了汉语普通话的语音水平评价系统
2. 设计并编写了动态语音数据流的高效切分引擎
3. 汉语连续语音识别中的声学建模和搜索策略的研究
4. 大词汇量、非特定人、连续汉语语音识别系统的集成
5. 汉语自然语音识别中的发音建模问题的研究

# 论文发表情况


[1] 吴文虎，宋战江，王帆。*实用智能话语系统的思路、策略与方法研究*。见：澳门资讯年会论文集。澳门，1998 年 1 月，174~177

[2] 徐明星，宋战江，郑方，吴文虎。*汉语语音水平评价方法研究*。见：第五届全国人机语音通信学术会议论文集。哈尔滨，1998 年 7 月，286~289

[3] 张继勇，郑方，杜术，宋战江，徐明星。*连续汉语语音识别中基于归并的音节切分自动机*。见：软件学报。1999 年，10(11): 1212~1215

[4] 宋战江，郑方，吴文虎。*汉语连续语音识别系统与知识导引的搜索策略研究*。见：自动化学报。2000 年，26(4): 470~477

[5] Zheng F, Song Z-J, Li L *et al*. *The Distance Measure for Line Spectrum Pairs Applied to Speech Recognition*. In Proc. International Conference on Spoken Language Processing (ICSLP). 1998, 3:1123~1126

[6] Song Z-J, Zheng F, Xu M-X, Wu W-H. *An effective scoring method for speaking skill evaluation system*. In Proc. European Conference on Speech Communication and Technology (EuroSpeech). 1999, 1: 187-190

[7] Zheng F, Song Z-J, Xu M-X et al. *EasyTalk: a large-vocabulary speaker-independent Chinese Dictation Machine*. In Proc. European Conference on Speech Comm. and Technology (EuroSpeech). 1999, 2: 819-822

[8] Song Z-J, Zheng F, Wu W-H. *Statistical knowledge based frame synchronous search strategies in continuous speech recognition*. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2000, 3: 1583~1586

[9] Zheng F, Wu J, Song Z-J. *Improving the Syllable-Synchronous Network Search Algorithm for Word Decoding in Continuous Chinese Speech Recognition*. Journal of Computer Science and Technology (JCST). 2000, 15(5): 461~471

[10] Li A-J, Zheng F, Byrne W, Fung P, Kamm T, Liu Y, Song Z-J *et al*. *CASS: a phonetically transcribed corpus of Mandarin spontaneous speech*. In: International Conference on Spoken Language Processing (ICSLP). 2000. 1:485-488

[11] Zheng F, Song Z-J, Fung P *et al*. *Mandarin pronunciation modeling based on CASS corpus*. In: Sino-French Symposium on Speech and Language Processing (SFSSLP). 2000. 47~53

[12] Byrne W, Venkataramani V, Kamm T, Zheng F, Song Z-J *et al*. Automatic generation of pronunciation lexicons for Mandarin spontaneous speech. In Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). 2001, accepted and to appear

# 致　　谢

本文是在导师吴文虎教授和郑方副教授的悉心指导下完成的。吴文虎教授渊博的专业知识、循循善诱的教导方式和平易近人的待人原则，使我受益匪浅；郑方副教授敏捷的思维、严谨求实的科学作风、一丝不苟的学术态度，使我在丰富多彩的语音研究领域找到了自己的位置和兴趣。在此，谨向两位恩师表示最诚挚的谢意！

感谢方棣棠教授和李树青教授，从我进入实验室时起，就一直在理论问题和生活上给我指导与帮助。作为老一辈科学工作者，他们严谨的科学作风、兢兢业业的工作态度将是激励我不断进步的精神动力。感谢徐明星老师，常与我在一起热烈争论学术问题，共同研究技术细节，他对解决实际问题的浓厚兴趣和认真态度，也将是我学习的榜样。

感谢实验室的全体同学，尤其是李净、张国亮、张继勇、罗春华、何磊、黄寅飞、苏毅、张欣研等同学，在学习、科研、生活等各个方面给予我的无私帮助、合作、探讨和支持。语音实验室是一个团结奋进、乐观向上的集体，在这里，我感受到了家庭般的温暖和快乐。

最后，衷心感谢我的父母、姐姐、哥哥、妻子和其他家庭成员。没有他们的无私付出和鼎立支持，就没有我今天的成功。他们无私的爱和默默的关怀，一直伴随着我的奋斗过程。

谨以此论文献给所有关心、爱护、指导和帮助过我的人！

# 汉语自然语音识别中发音建模的研究

宋战江