密级：公开

汉语语音识别中声学建模及参数共享策略的研究

Study on Acoustic Modeling and Parameter Tying Strategy for Chinese Speech Recognition

(申请清华大学工学硕士学位论文)

院（系、所）：计算机科学与技术系

专　　业：计算机应用技术

研　究　生：张继勇

指 导 教 师：郑　方

二零零一年六月

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得清华大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：_____日　期：_____

关于论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

(保密的论文在解密后应遵守此规定)

签　名：_____导师签名：_____日　期：_____

摘要

声学建模是语音识别领域中的关键问题之一。本文对汉语连续语音识别中的声学建模技术和参数共享策略进行了深入的研究。主要针对两个方面：一、提出并实现了半连续分段概率模型（SCSPM）；二、研究基于决策树状态共享的上下文相关建模方法，并且分别实现了上下文相关音素模型与上下文相关声韵母模型。具体包括：

1、提出并实现了半连续分段概率模型(SCSPM)。该模型在经典 HMM 模型及其修正模型混合高斯连续概率模型（MGCPM）基础上，结合矢量量化技术和连续概率密度描述的方法，以混合共享的方式来描述各状态的概率分布。此外还研究和分析了 SCSPM 模型的各种混合权重精简策略，提出了一种新的在迭代过程中进行权重精简的策略。与原来的 MGCPM 模型相比，SCSPM 模型在保证识别率不下降的情况下，大大降低了模型规模和计算复杂度。

2、对 HTK 平台进行了研究和分析，实现了基于 HTK 平台的声学模型训练和性能评估的有效方法。

3、对上下文相关（Context Dependent, CD）声学建模中基于决策树状态共享策略进行了深入研究。分析了两种不同的决策树构造方法，讨论了问题集的设计和决策树节点的分裂策略。还研究了针对静音模型进行特殊处理的方法，以提高鲁棒性。

4、实现了基于决策树状态共享的上下文相关的音素（CD-Phone）模型。着重研究了其中的基于决策树状态共享的上下文相关建模问题，其中包括根据音素发音特点设计问题集和不同的决策树构造方法。与音节模型相比，CD-Phone 模型能使音节误识率降低大约 10%。

5、研究并实现了基于决策树状态共享的上下文相关声韵母（CD Initial/Final, CD-IF）模型。为了保证声韵母之间的相互搭配关系，在原来的基本声韵母集合上，增加了零声母部分，形成扩展声韵母（Extended Initial/Final, XIF）集合。实验证明 XIF 模型比 IF 模型具有较高的识别率，最终实现的 CD-XIF 模型的音节正确率超过 80%，误识率比基准音节模型降低了大约 25%。

关键词：半连续分段概率模型（SCSPM），参数共享策略，基于决策树状态共享，上下文相关音素模型，上下文相关声韵母模型

Abstract

Acoustic Modeling is one of the key problems in the field of speech recognition. In this paper, the techniques of acoustic modeling and parameter tying strategy are deeply studied. Two main aspects are focused on: the proposition and implementation of the Semi-Continuous Segmental Probability Model (SCSPM); The study of Context Dependent (CD) acoustic modeling with decision tree based state tying, and the implementation of CD phone modeling and Initial/Final (IF) modeling respectively. Including:

1. The SCSPM is proposed and implemented. It is based on the traditional Hidden Markov Model (HMM) and the modified HMM namely Mixed Gaussian Continuous Probability Model (MGCPM), the Vector Quantizaztion (VQ) technique and the feature of continuous probability density distribution are integrated, and the method of Tied Mixture is adopted to describe the probability distribution of each state. Moreover, the mixture weight reduction method is studied and analyzed, and a new effective method which prunes the small tying weight through the iterative training process is proposed. Compared with the MGCPM, SCSPM can reduce the model scale and computational complexity significantly with little degradation in recognition accuracy.

2. The HMM Tool Kit (HTK) platform is studied and analyzed. Based on HTK, an effective method is implemented for acoustic model training and performance evaluation.

3. The Decision Tree (DT) based state tying strategy in the Context Dependent (CD) acoustic modeling is deeply studied. Two different DT design methods are analyzed, the design of question set and the DT node splitting strategy are discussed. Furthermore, the method for treating the silence model is studied to enhance the robustness.

4. The CD-Phone model with decision tree based state tying is implemented. The phone question set and different DT structure are designed. Compared with the baseline syllable model, CD-Phone model reduces the syllable error rate (SER) by about 10%.

5. The CD Initial/Final (IF) model with decision tree based state tying is implemented. To maintain the connection between Initial and Final, the Extend IF (XIF) set is proposed by adding the Zero Initials to the prior standard IF set. Experiments show that the XIF model outperforms the IF model. The syllable correct rate of the implemented CD-XIF model can achieve over 80%, and the CD-XIF model can reduce the SER by over 25% compared with the baseline syllable model.

Keyword: Semi-Continuous Segmental Probability Model, Parameter Tying Strategy, Decision Tree based State Tying, Context Dependent Phone Model, Context Dependent Initial/Final Model

目录

图表目录

第一章　引言

随着计算机科学技术日新月异的发展，人类正迈向一个全新的信息时代。而自然语言作为人类最重要最自然的交流工具，是人类获得信息的最重要的来源之一。使用自然语言与计算机之间进行交流是人类长久以来的梦想。语音识别技术作为实现这一梦想的关键技术，正引起越来越多的人的关注和研究。

1.1 语音识别技术概述

1.1.1 语音识别技术的发展历史

对于机器识别语音的研究，可以追溯到上世纪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。我国在50年代后期，也曾经研制出一套“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，从此开始了计算机语音识别的正式阶段。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，例如线性预测编码（LPC）技术、动态时间规整（DTW）算法[Vintsjuk 1968]、矢量量化（VQ）技术等，都已经在语音识别领域得到了广泛地应用。

自从八十年代中期以来，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（Hidden Markov Model, HMM）的研究和广泛应用，推动了语音识别的迅速发展，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。在90年代IBM公司推出的商业系统ViaVoice也具有很高的水准。

当前，语音识别领域的研究正方兴未艾。在这方面的新算法、新思想和新的应用系统不断涌现。同时，语音识别领域也正处在一个非常关键的时期，世界各国的研究人员正在向语音识别的最高层次应用——非特定人、大词汇量、连续语音的听写机系统的研究和实用化系统进行冲刺。可以乐观地说，人们所

期望的语音识别技术实用化的梦想很快就会变成现实。

1.1.2 语音识别系统的分类

从语音识别系统实现的性能出发，我们可以将语音识别系统作如下的划分：按照被识别人的范围可以分为特定人（Speaker Dependent, SD）和非特定人（Speaker Independent, SI）语音识别；按照词汇量的大小可以分为小词汇量（Small Vocabulary）、中词汇量（Medium Vocabulary）和大词汇量（Large Vocabulary）语音识别；按照说话方式可以分为孤立词（Isolated Word）、连接词（Connected Word）和连续语音（Continuous Speech）的语音识别。在上述分类当中，非特定人、大词汇量、连续语音识别系统即语音听写机系统所包含的技术最为复杂，实现起来也最为困难，因而被公认为代表当前语音识别技术的最高水平。

此外，根据语音识别系统中所采用的模型进行分类，可以分为基于模板的语音识别系统、基于概率统计模型的语音识别系统、基于神经网络的语音识别系统等；也可以根据语音识别系统所完成的任务来分，如语音命令系统、语音听写机系统、关键词确认系统等。

1.1.3 语音识别系统的研究方向

语音识别系统的研究主要集中在如下几个方面：

（1）声学特征。特征提取与选择是语音识别的一个重要环节。特征提取解决了时域语音信号的数字表示问题，而特征选择则通过选取有效的特征为模式划分提供数据。特征提取与选择的好坏直接影响到识别器的性能。常用的声学特征有时域特征、频域特征和倒谱特征。时域特征如短时平均能量、短时平均过零率、共振峰、基音周期等；频域特征有傅里叶频谱等；倒谱特征有有基于线性预测编码(LPC)的倒谱即 LPCC，有基于 Mel 频率弯折的倒谱即 MFCC。Wilpon[Wilpon 1989]等把加权的倒谱和差分倒谱串接起来形成一个大的矢量作为声学特征矢量，取得了好的效果。此外，也有人使用两维的“时-频谱”表示语音

信号的特征[Wilpon 1991]，它考虑了语音信号的时变特征，是频谱的一种高阶时间派生参数。在目前的语音识别系统中，采用倒谱特征来建模最为普遍。

（2）声学模型。随着 HMM 模型的广泛研究与应用，使得语音识别领域中的声学建模技术有了实质性的进展。HMM 能描述不同层次的语音单元，由 Viterbi 解码算法[Viterbi 1967]可以得到与语音序列对应的最佳状态序列，便于解决连续语音识别的问题。另外，人工神经网络（Artificial Neural Network, ANN）领域的研究也给语音识别带来了新的活力。由于人工神经元网络具有刻划各种复杂分类边界的能力，十分适用于语音识别领域。神经网络还可以与 HMM 综合应用于声学建模：由神经网络完成静态的模式划分问题，用 HMM 完成时间对准问题[Franzini 1990][Morgan 1990]，使神经网络更容易地应用于连续语音识别系统。到目前为止，语音识别系统中声学模型的主流仍然是 HMM 模型及其改进模型。此外，根据模型之间的相关性，声学模型还可以分为上下文无关模型和上下文相关模型。上下文无关模型简单，识别率相对较低。而上下文相关模型考虑了连续语音中的发音相关性，因而具有较高的识别率。语音识别系统中的声学建模问题将是本文研究的重点。

（3）语言模型。概括来讲，语言模型可以分为两类，基于统计的语言模型（Statistical Language Model）基于知识的语言模型（Knowledge-based Language Model）。在当前的技术条件下，基于统计的语言模型在实际应用中处于主流地位。它通过对大量实际语料的统计来获得词与词之间的连接信息，从而评价一个词串是否为语言中合理的语句。这在一定程度上回避了基于规则的语言模型其规则集难以严格和完备，以及语义规则难于形式化等困难。因此，现阶段实用语言模型中的规则模型主要用来作为统计模型的补充，对统计模型的结果进行校验和改进。N-Gram 统计模型是最初引入而且应用最广泛的一种语言模型，该模型最初由 Jelinek 等人提出[Jelinek 1983]。但是 N-Gram 模型面临的最大困难是训练语料过于稀疏。针对这一困难，Nadas 给出了图灵估计变形的概率估计方法[Nadas 1985]，Katz 给出了一种基于图灵估计的退化频度估计算法[Katz 1987]等，力求在一定程度上解决训练数据稀疏（即零概率平滑）的问题。

（4）搜索算法。连续语音识别中的搜索，就是寻找一模型序列来描述输入

语音信号，从而得到语音信号的解码序列。搜索的依据是语音信号在声学模型的打分以及加入语言模型的概率。针对 HMM 模型，基本的搜索策略为 Viterbi 解码算法和帧同步算法[Lee 1989]。其基本思路是以帧为单位，任一时刻对每一条路径，都假定当前帧可能是该路径的后续，即每一时刻都在当前所有路径后发展所有可能的路径，以进行一个完备的搜索。但是当这种搜索策略使用到大词表的连续语音识别系统中时，搜索路径会随着时间的增长而急剧膨胀，因此必须使用一定的剪枝策略。

（5）自适应与鲁棒性问题。由于存在不同的说话人、说话方式、环境噪声、传输信道等因素，语音识别系统在实验条件下具有很好的性能，但是应用到实际生活中性能却急剧下降。提高系统鲁棒性，是要提高系统克服这些因素影响的能力，使系统在不同的应用环境下性能稳定。解决的办法可以分为两类：基于语音特征的方法和基于模型调整的方法。前者的目标是寻找更好的、高鲁棒性的特征参数，或是在现有的特征参数基础上，加入一些特定的处理方法，如滤波，去噪，语音增强等。后者的目标是利用少量的自适应语料来修正或变换原有的说话人无关模型，使其成为说话人自适应模型。

1.2 语音识别系统中的声学模型

1.2.1 声学模型在语音识别系统中的作用

语音听写机系统中的核心部分为声学模型和语言模型。二者之间的关系可以用图 1.1 来描述：

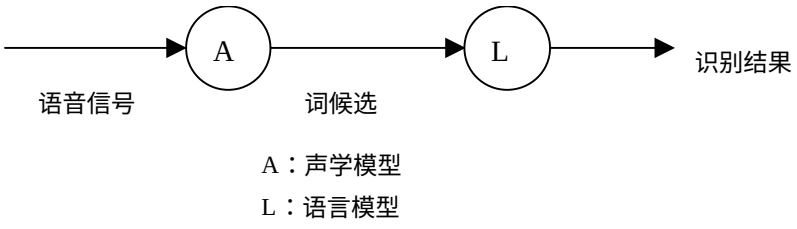


图 1-1 语音识别器的基本组成

假设声学模型的输入（即语音信号）为 A，它的输出（即词序列）为 w，

则整个语音识别系统的任务就是找到一个 w^*，使它满足：

$$w^* = \arg\max_w P(w|A) = \arg\max_w \frac{P(A|w)P(w)}{P(A)} \tag{1.1}$$

其中第二步是根据 Bayes 法则得到的。在这个式子中，$P(A|w)$是由声学模型计算得出的匹配概率，而 $P(w)$是由语言模型计算得出的不同词串的概率。考虑到在计算不同词串 w 时，P(A)是一个常数值，上式也可以写作

$$w^* = \arg\max_w P(w|A)P(w) \tag{1.2}$$

我们可以看到，声学模型是语音识别系统中的枢纽，直接承担着将语音特征数据识别为声学模型串的重任。因此，声学模型的好坏决定了整个语音识别系统的整体性能。语音听写机系统的性能往往以其声学层面上的测试结果来代表。

1.2.2 声学模型的分类

声学模型的主要功能是对识别基元进行模式划分。进行模式划分的方法很多，当前语音识别系统中主流的声学建模技术主要有两类方法：一种是基于隐式马尔可夫模型的概率统计模型的方法；一种是基于人工神经网络的方法。

基于概率统计模型的模式划分方法主要是依据贝叶斯判决准则，因为它使用方便、准确有效而得到了比较广泛的应用，在语音识别领域中占据了重要的地位。在基于概率统计模型的模式划分方法中，HMM 模型比较符合语音信号作为随机信号的变化规律，对语音信号的描述比较准确，具有较高的识别率，所以 HMM 模型及其各种改进模型已经成为语音识别系统中应用最多的方法。

人工神经网络以类比于生物神经系统处理信息的方式，用大量简单的处理单元进行连接而构成一种独具特点的信息处理系统。ANN 的分类能力很强，有自学习、自组织的能力，并行程度高，综合、抽象能力强的优点。理论上讲，人工神经网络能把任意的输入映射到任意的输出上去。只要层数足够多，节点数足够多，训练充分，ANN 能给出最优的划分，无论空间的分布多么复杂[Verhasselt 1998]。

但是 ANN 本身也有一些缺点。ANN 会由于其网络规模过大而造成训练时间太长，另外在训练收敛的过程中存在着局部极小值的问题，因而它的性能发挥受到了限制。由于语音信号是时变的，而且在连续语音信号中协同发音现象的存在，使得相邻语音帧之间是上下文有关的。虽然 ANN 对于静态模式具有很强的分类能力，但若在连续语音识别中仅仅简单地使用 ANN，则不能很好地体现出这些暂态变化特性。

1.3 声学模型中识别基元的选择

语音识别基元的选择在语音识别尤其是连续语音识别中是重要的环节。识别基元的选择应该基于如下两个原则[Zheng 1996]：

（1）具有灵活的可组合性能，即它能够代表语音中的比较独立的一些个性，可以组成其他的语音单位；

（2）具有稳定性，即它应该使得语音中的共性能够得到相当的体现，从而保证识别基元对不同环境的适应能力（即鲁棒性）。

在这两个原则中，灵活性希望基元尽可能地小，如音素；而稳定性则希望基元尽可能地大，如词甚至词组。这两个方面需要综合考虑。不同的识别系统有着不同的性能要求，也可能采用不同的声学特征或语音模型，或采取不同的帧长与帧移策略，所以它们对识别基元的选择标准也不一样。如果识别基元选得不恰当，则很难保证系统的识别性能。

对于西方的语言，常常采用音素(Phoneme) 或上下文相关音素(Triphone)作为识别基元。其中 Triphone 模型由于考虑了上下文相关的影响，极大地提高了语音识别系统的性能，在实际系统中取得了很好的效果。因此，目前基于 Triphone 的声学建模技术是西方语言的语音识别系统的主流发展方向[Yang 1995]。

而对于汉语而言，我们知道，除了能够仿照西方语言采用音素作为识别基元之外，由于汉语是由音节组成的语言，所以可以采用音节作为汉语语音识别基元。此外，每个音节由声母和韵母组成，声韵母作为识别基元也是一种选择。

综上所述，针对汉语语音识别系统，我们有如下的几种识别基元可供选择：

音节(Syllable)基元：汉语的音节结构固定，每个音节对应一个汉字，训练

数据的标注比较容易。采用音节作为识别基元的好处是利用了汉语语音的特点，能取得好的识别效果。实验证明当进行上下文无关建模时，音节作为识别基元能取得很好的识别效果[Zheng 1996]。缺点是音节本身的数目较大（汉语共有 418 个无调音节），不利于上下文相关的建模。此外音节模型还存在训练不充分的问题。

音素(Phoneme)基元：音素是语音的最小单位。其最大的优点是数量很少，在训练数据相对固定的情况下，它们可以得到充分的训练。此外，在改换任务或建立新词条时无需重新建模和训练，甚至在跨语言的识别系统中也能得到很好的应用。由于协同发音（Coarticulation）现象的存在，采用上下文无关的音素模型有较大的误识率。改进的方法是进行上下文相关（Context Dependent, CD）的音素建模，也就是人们通常称作的 triphone 模型。但是如果直接使用 triphone 模型会由于基元数据过多而导致有些模型不能得到充分训练。解决这个问题的较好办法是在建模过程中采用参数共享策略。

声韵(Initial/Final, IF)基元：这是根据汉语的特点建立的模型。其好处是声韵母的数目比较少，声韵结构比较稳定，利用了汉语语音的特点。上下文无关的声韵母模型具有规模小，速度快的优点，但是无法取得很高的识别率[Li 2000]。为了提高系统的识别率，我们需要进行上下文相关的声韵母（Context Dependent Initial/Final, CD-IF）建模。声韵母模型的缺点是只实用于汉语语音识别系统。

1.4 声学模型中的参数共享策略

在实际的语音识别系统当中，除了追求高的识别率外，还需要考虑系统的速度和存储的开销。一个好的语音识别系统应该是识别正确率高，速度快，所需要的模型的存储空间小。参数共享的策略目的是在不降低系统的识别率的前提下，提高系统速度，降低存储空间的大小。

声学模型中的参数共享策略有多种多样，最有效而且最普遍的有两种方式：高斯混合共享和模型状态共享。前者的主要方式是使用半连续 HMM（Semi-Continuous HMM, SCHMM）替代连续 HMM（Continuous HMM, CHMM），后者主要方式是与上下文相关建模技术相结合，通过构造决策树来实现模型的状态共享。

SCHMM 也可以称作共享混合 HMM（Tied Mixture HMM, TMHMM），首先使用混合高斯概率密度函数（Probability Density Function, pdf）来描述整个特征空间的分布，然后再对每个识别基元的每个状态，训练出它的权重，即可得到半连续的混合高斯模型。SCHMM 实现了不同状态间的高斯混合共享，减少了模型存储空间，并且能提高识别的速度。例如：在 EasyTalk 系统[Zheng 1999]中采用的 MGCPM[Zheng 1998]模型是一种 CHMM，它以音节作为识别基元，每个音节又分为 6 个状态，如果每个状态采用 16 个高斯混合来描述，整个系统需要大约 40,000 个高斯混合。而采用 SCHMM 模型，我们只需要大约 5,000 个高斯混合来描述全特征空间的分布，计算复杂度的降低和模型存储空间的减少是不言而喻的。

在上下文相关的建模过程中，由于存在大量的基元，训练数据相对稀疏，存在训练不充分的问题，有些基元甚至在训练数据中不出现。另外模型过于庞大，计算复杂度太高也是一个问题。解决的办法是使用基于决策树的状态共享策略[Reichl 2000]。其中决策树策略提供了一种自顶向下的数据驱动和专家知识相结合的一种有效的分类方法[Breiman 1984]，除了能方便控制模型规模之外，还具有合成那些在训练数据中不存在的基元的能力。

1.5 主要工作和论文安排

本文的研究工作主要是研究汉语连续语音识别系统中的声学建模技术，并且运用参数共享策略针对模型做优化工作。

作者主要完成了如下的工作：(1)提出并实现了半连续分段概率模型（Semi-Continuous Segmental Probability Model，SCSPM），在不降低识别率的前提下降低模型的规模和计算复杂度；(2)对 HTK 平台进行了研究和分析，实现了基于 HTK 平台的声学模型训练和性能评估的有效方法；(3)分别实现了上下文相关音素建模和上下文相关的声韵母建模，研究了基于决策树的状态共享策略，对模型性能进行了实验和分析。与基准音节模型相比，大幅提高了识别率。

本文的篇章安排如下：第二章探讨语音识别中的声学建模技术；第三章介

绍半连续分段概率模型；第四章介绍上下文相关的音素建模；第五章介绍上下文相关的声韵母建模；最后在第六章给出结论。

第二章　语音识别中的声学建模技术

2.1 HMM 模型框架

HMM 是目前语音识别领域最成功、应用最广泛的方法。许多高性能的系统都是基于 HMM 的。这一部分，我们将回顾 HMM 的基本概念及其主要算法，并讨论它的局限性[Rabiner 1989]。

2.1.1 HMM 定义

HMM 是基于马尔可夫链的。马尔可夫过程是一个随机过程 $\{S(t): t \in T\}$，它具备这样的性质，即已知 t 时刻过程所处的状态 $s_t = S(t)$，在 t 时刻以后过程将要到达的状态与 t 时刻以前过程所处的状态无关，这个性质也称为过程的无后效性或马尔可夫性。

马尔可夫过程 $\{S(t): t \in T\}$ 可能取值的全体构成状态空间，可以是连续的或离散的；马尔可夫过程的指标集 T 也可以是连续的或离散的。

对一个状态空间 I 和指标集 T 离散的随机过程 $\{S(t): t = 0,1,2,\cdots\}$，若满足

$$\begin{aligned} &P\{S(t+1) = s \mid S(0) = s_0, S(1) = s_1, \cdots, S(t) = s_t\} \\ &= P\{S(t+1) = s \mid S(t) = s_t\} \end{aligned} \tag{2-1}$$

则称之为马尔可夫链。马尔可夫链在 t 时刻的一步条件转移概率

$$a_{ij}(t) = P\{S(t+1) = j \mid S(t) = i\} \tag{2-2}$$

称为 t 时刻状态 i 到状态 j 的转移概率。显然有

$$a_{ij}(t) \geq 0, \quad i, j \in I \tag{2-3}$$

$$\sum_{j \in I} a_{ij}(t) = 1, \quad i \in I \tag{2-4}$$

隐式马尔可夫模型(HMM)由两个相互关联的过程相互作用而成：一个是状态空间有限的马尔可夫链，一个是随机函数集。HMM 在任何时刻 t 下所处的

状态 s_t 隐藏在系统内部，不为外界所见，外界只能得到系统在该状态下提供的实 R^Q 空间中的一个随机矢量，该随机矢量的发生概率由当前状态相关的随机函数给出。HMM 的状态转移由状态转移概率矩阵 $\left\{a_{ij}\right\}$ 控制。

一个 HMM 由下面一些参数表征。

(1) $\overline{\mathbf{N}}=\left\{1,2,3,\cdots,N\right\}$：模型的状态集合，$s_t=s(t)\in\overline{\mathbf{N}}$ 表示系统在 t 时刻所处的状态；

(2) $A(t)=\left\{a_{ij}(t)\right\}_{N\times N}$：状态转移概率矩阵，其中 $a_{ij}(t)=P\left\{s(t+1)=j\middle|s(t)=i\right\}$；

(3) $B=\left\{b_j(\cdot)\right\}_{N\times 1}$：观察符号输出概率（密度）矩阵，$b_j(x)=P_d\left\{output=x\middle|state=j\right\}$；其中 $P_d\{\cdot\}$ 表示事件概率或概率密度。

(4) $\pi=\left\{\pi_i\right\}_{N\times 1}$：初始概率分布，$\pi_i=P\left\{s(1)=i\right\}$。

状态概率转移矩阵一般是时间的函数，如果与时间无关，那么相应的 HMM 称为齐次的，此时

$A=\left\{a_{ij}\right\}_{N\times N}$：状态转移概率矩阵，$a_{ij}=P\left\{s(t+1)=j\middle|s(t)=i\right\}$。

一个有 N 个状态的齐次 HMM 可以表示为 $\Lambda=\left\{\pi,A,B\right\}$。

2.1.2 HMM 的基本问题及解决

用 HMM 来完成语音识别的研究时，需要解决如下的三个问题。

训练。若有一个 HMM，需要根据该系统所给的若干观察序列 O 确定它的三项特征参数。所有的输出构成一个学习样本集合，其中每个观察序列 O 称为

一个学习样本。设有 M 个样本，此集合可以记为$\{O^{(m)}, m=1\sim M\}$。确定 HMM 特征参数的准则是最大似然准则。

计分。若已知一个 HMM 的三项特征参数，需要对系统可能产生的任何观察序列 O 计算其产生的概率。

状态解码。同样已知三项特征参数，若得到了该系统产生的某个观察序列 O，需要估计该系统产生此序列 O 时最可能经历的状态序列。

“向前－向后”(Forward-Backward)算法或 Baum-Welch 算法[Baum 1972]通过引入向前部分概率函数($1\le j\le N$)

$$\alpha_t(j)=P\{o_1,o_2,\cdots,o_t,s_t=j|\Lambda\}=\sum_{i=1}^{N}\alpha_{t-1}(i)a_{ij}b_j(o_t),2\le t\le T \tag{2-5}$$

$$\alpha_1(j)=\pi_i b_i(o_1) \tag{2-6}$$

及向后部分概率函数($1\le i\le N$)

$$\beta_t(i)=P\{o_{t+1},o_{t+2},\cdots,o_T|s_t=i,\Lambda\}=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j),1\le t\le T-1 \tag{2-7}$$

$$\beta_T(i)=1 \tag{2-8}$$

利用一组迭代公式可以解决第一个问题[Baum 1972][Huang 1989]。

$$\overline{\pi}_i=\frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N}\alpha_T(i)} \tag{2-9}$$

$$\overline{a}_{ij}=\frac{\sum_{t=1}^{T-1}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)} \tag{2-10}$$

$$\overline{b}_i(x)=f(\Lambda,O)\ \text{(是}O\text{和}\Lambda\text{的函数)} \tag{2-11}$$

其中$\overline{b}_i(\cdot)$的迭代公式视 HMM 的不同类型有所不同。

第一个问题解决之后，第二个问题由向前部分概率可以得到：

$$P\{\mathbf{O}|\Lambda\}=\sum_{i=1}^{N}\alpha_T(i) \tag{2-12}$$

Viterbi 解码算法[Viterbi 1967]可以用来解决第三个问题。令：

$$\Phi_1(j) = \pi_j b_j(o_1), \quad 1 \le j \le N \tag{2-13}$$

进行如下递归($2 \le t \le T, 1 \le j \le N$)：

$$\Phi_t(j) = \max_{1 \le i \le N}\left[\Phi_{t-1}(i) \cdot a_{ij}\right] b_j(o_t) \tag{2-14}$$

及

$$\Psi_t(j) = \arg\max_{1 \le i \le N}\left[\Phi_{t-1}(i) \cdot a_{ij}\right] b_j(o_t) \tag{2-15}$$

最后得到：

$$s_T^{(ML)} = \arg\max_{1 \le i \le N}\left[\Phi_T(i) \cdot 1\right] \cdot 1 \tag{2-16}$$

$$s_t^{(ML)} = \Psi_{t+1}(s_{t+1}^{(ML)}), \quad 1 \le t \le T-1 \tag{2.17}$$

记 $S^{(ML)} = \left\{s_t^{(ML)} | 1 \le t \le T\right\}$，此为最大似然(Maximum Likelihood, ML)状态序列，那么该 HMM 产生这个状态序列的概率为：

$$\begin{aligned} P_d^{(ML)} &= \Phi_T(s_T^{(ML)}) = \pi_{s_1^{(ML)}} \cdot b_{s_1^{(ML)}}(o_1) \prod_{t=2}^{T} a_{s_{t-1}^{(ML)} s_t^{(ML)}} \cdot b_{s_t^{(ML)}}(o_t) \\ &= \left(\pi_{s_1^{(ML)}} \prod_{t=2}^{T} a_{s_{t-1}^{(ML)} s_t^{(ML)}}\right) \cdot \left(\prod_{t=1}^{T} b_{s_t^{(ML)}}(o_t)\right) = P\left\{S^{(ML)} | \Lambda\right\} \cdot P_d\left\{O | S^{(ML)}, \Lambda\right\} \end{aligned} \tag{2-18}$$

事实上，由全概率公式，产生该观察序列的概率(密度)为：

$$\begin{aligned} P_d\{O|\Lambda\} &= \sum_S P_d\{O, S|\Lambda\} = \sum_S P_d\{O|\Lambda, S\} \cdot P\{S|\Lambda\} \\ &= \sum_S \left(\pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1} s_t}\right) \cdot \left(\prod_{t=1}^{T} b_{s_t}(o_t)\right) \\ &= \sum_S \left(\pi_{s_1} \cdot b_{s_1}(o_1) \prod_{t=2}^{T} a_{s_{t-1} s_t} \cdot b_{s_t}(o_t)\right) \end{aligned} \tag{2-19}$$

其中 $S = \left\{s_t | 1 \le t \le T\right\}$ 是任意一种状态序列。因此 Viterbi 算法给出的是该和式中的一项，是最大似然状态序列。

语音识别中所有的 HMM 一般都是从左向右结构，即 $a_{ij} = 0, j < i$。从左向右结构又分为无跳跃式($a_{ij} = 0, j < i, j > i+1$)和有跳跃式。

2.1.3 HMM 的分类

根据观察输出概率矩阵中的函数$b_j(x)$是基于 VQ、连续密度还是二者的综合，HMM 又分为离散 HMM (Discrete HMM, DHMM) 、连续 HMM (Continuous HMM, CHMM) 和半连续 HMM (Semi-Continuous HMM, SCHMM)[Huang 1989]。这三种 HMM 的 A 矩阵具有相同的特性，由于：

$$\begin{aligned} P_d\{O|\Lambda\} &= \sum_S P_d\{O,S|\Lambda\} = \sum_S P\{S|\Lambda\}\cdot P_d\{O|\Lambda,S\} \\ &= \sum_S\left(\pi_{s_1}\prod_{t=2}^{T}a_{s_{t-1}s_t}\right)\cdot\left(\prod_{t=1}^{T}b_{s_t}(o_t)\right) \end{aligned} \tag{2-20}$$

因此，本节中我们仅对

$$P_d\{O|\Lambda,S\} = \prod_{t=1}^{T}b_{s_t}(o_t) \tag{2-21}$$

中的$b_{s_t}(o_t)$进行讨论。

2.1.3.1 离散 HMM

DHMM 是基于矢量量化(Vector Quantization，VQ)技术的。它把特征向量空间分成若干个子空间，每个子空间用一个中心向量来表示，表征这个中心向量的是一个码字(codeword)，所有码字的集合构成码本(codebook)。

在计算概率$b_{s_t}(o_t)$时，取而代之的是计算$b_{s_t}\left(V(o_t)\right)$，这里$V(\cdot)$表示把向量量化后所对应的码字代号。

这里对$b_{s_t}(v)$的估计比较容易，通过某种计数的方法就可以实现：

$$\overline{b}_i(v) = \frac{\sum_{t=1}^{T}\alpha_t(i)\cdot\beta_t(i)}{\sum_{t=1}^{T}\alpha_t(i)\cdot\beta_t(i)}\Bigg|_{V(o_t)=v} \tag{2-22}$$

式中 v 表示码字序号。

很显然，由于一个子空间里的所有向量都用一个码字来代替，量化误差会很大，因此 DHMM 的描述误差也就较大。但是，随着计算机处理能力的增强，

通过大规模码本的应用，可以在一定程度上克服这个问题，同时充分利用DHMM 运算复杂度较低的优势。

2.1.3.2 连续 HMM

为了克服 DHMM 对特征空间描述上的不精确性，CHMM 应运而生。CHMM 的主要目的是对特征空间进行比较精确的描述。对一个概率密度进行估计的准确程度取决于训练数据量的多少。当训练数据量足够大时，可以估计得很精确，缺点是选择统计半径的大小比较困难、会浪费很大存储空间和对训练数据量的敏感性。

什么描述方法可以作到既能少占存储空间，又能降低估计复杂度呢？一种好的方法是使用混合高斯密度(Mixed Gaussian Density, MGD)[Wilpon 1989][Huang 1989]，即

$$b_n(x)=\sum_{m=1}^{M} g_{nm}\cdot N(x;\mu_{nm},\Sigma_{nm}) \tag{2-23}$$

其中

$$\sum_{m=1}^{M} g_{nm}=1 \tag{2-24}$$

这是用 M 个混合高斯密度对第 n 个状态的特征空间进行估计。理论上可以证明，当 M 足够大时，MGD 可以比较准确地描述特征向量的概率密度。

2.1.3.3 半连续 HMM

虽然 MGD 描述方法中所要存储的参数不多(每个混合的中心向量 μ_{nm}、协方差矩阵 Σ_{nm} 和混合增益 g_{nm})，但当 M 很大时由于每个 $b_n(x)$ 都需要存储 M 组这样的参数，因此比较浪费空间。SCHMM 结合 VQ 技术和连续密度描述的特点比较好地解决了这个问题。

$$\begin{aligned} b_{s_t}(o_t) &= f\{o_t|s_t\}=\sum_{l=1}^{L} f\{o_t|V_l,s_t\}\cdot P\{V_l|s_t\} \\ &= \sum_{l=1}^{L} f\{o_t|V_l,s_t\}\cdot b_{s_t}^{(D)}\{V_l\}=\sum_{l=1}^{L} f\{o_t|V_l\}\cdot b_{s_t}^{(D)}\{V_l\} \end{aligned} \tag{2-25}$$

其中 $\{V_l|1\le l\le L\}$ 是表征特征空间的码本，$b_{s_t}^{(D)}\{V_l\}$ 是输出离散码字 V_l 的概

率，$f\left\{\cdot\middle|V_l\right\}$为以码字$V_l$为中心的子空间中的特征向量概率密度的高斯逼近。

这种描述方法改变一下形式成为

$$b_n(o_t)=\sum_{l=1}^{L}g_{nl}\cdot f\left\{o_t\middle|V_l\right\} \tag{2-26}$$

这就是共享 MGD(Tied Mixture Gaussian Density, TMGD)[Bellegarda 1990]。

在这样的描述中，所有模型都公用 L 个类似于码字的密度函数，记录一个模型中不同状态的概率密度函数$b_n(x)$只需要一组系数$G=\left\{g_{nl}\middle|1\le n\le N\right\}$即可。虽然 SCHMM 或 TMGD 对特征空间的描述节省了很大的存储空间，而且效果很好，但是由于所有模型的所有状态的特征空间的描述都依赖于这 L 个分布，因此其描述不如 MGD 来得精确，尤其在码本选得不合适时更是如此。

2.1.4 HMM 的局限

尽管基于 HMM 框架的语音识别技术对现代语音识别做出了巨大的贡献，该技术本身仍存在一些固有的局限。这些局限性限制了经典 HMM 在语音识别系统中的实际应用，同时也是其他派生声学模型要重点解决的问题。HMM 的局限性主要表现在：

研究对象——HMM 研究的对象是线性符号串序列，不反映语音现象（自然语言）的结构性，因此不能利用自然语言的结构特性。

一阶假设——关于“系统的当前状态只与前一状态有关，输出概率只依赖于当前状态”的一阶假设对语音信号来讲显然是不恰当的。一阶假设的一种后果是 HMM 模型不能描述协同发音，因为协同发音时各音素或音节间“吃音”、“丢音”现象十分严重，从而每个状态的分布受相邻几个状态的影响，发生了很大变化。另一种后果是状态驻留时间的建模不甚合理，HMM 用指数递减分布来描述状态驻留时间，而统计结果显示，驻留时间大致服从泊松分布。

独立性假设——关于“相邻帧相互独立”的假设也是不对的。在这种假设下，HMM 每次只处理一帧语音。要想用到帧间上下文的相关性，HMM 必须将别的帧的信息吸收到当前帧中来（比如，引入多数据流以处理差分系数，或使

用 LDA 将这些数据流变换成一个数据流）。

HMM 概率密度模型（离散的、半连续的、连续的）建模精度不是最优的。离散参数的 HMM（DHMM）不可避免地会引入量化误差；而连续或半连续的 HMM（CHMM 或 SCHMM）的分布假设（比如高斯混合分布）与实际分布的差异又引入了模型不匹配的问题。

最大似然概率的训练准则导致了声学模型之间可分离度差。提高可分离度可以用最大互信息(Maximum Mutual Information, MMI)法[Bahl 1986]，但这种方法大大提高了运算复杂度，而且实现起来也比较困难。

2.2 MGCPM 模型

针对经典 HMM 时空复杂度高，尤其是 Viterbi 解码过程的复杂性，人们提出了一个基于高斯混合分布和观察序列分段的声学模型，即混合高斯连续概率模型（Mixed Gaussian Continuous Probability Model, MGCPM）[Zheng 1998][Mou 1998]。该模型采用高斯混合分布描述观察矢量在特征空间的分布，同时对 HMM 的训练和 Viterbi 解码过程进行简化，采用对观察序列进行线性或非线性分段代替 HMM 中的状态，用帧同步搜索代替状态解码。经过这样的简化，MGCPM 与传统的 CHMM 相比，识别速度上升，识别率没有多少下降。

设 K 段从左至右高斯混合分段模型中第 k 段的观察输出特征矢量集 $Z_k = \{z_1, z_2, ..., z_T\}$,我们使用如下的概率密度函数来描述特征矢量 z_j 的分布：

$$p(z_j \mid \theta) = \sum_{i=1}^{M} \{\ p_i p(z_j \mid \theta_i)\ \} \tag{2-27}$$

其中 $z_j \in Z$ 是 d 维特征矢量，j =1,2,…,T。M 是高斯混合分布的阶，即高斯混合概率密度函数中单个高斯分布的个数。单个高斯分布的概率密度函数为：

$$p(z_j \mid \theta) = (2\pi)^{-d/2} \mid R_i \mid^{-1/2} \exp\left\{ -\frac{1}{2}(z_j - u_i)^T R_i^{-1} (z_j - u_i) \right\} \tag{2-28}$$

参数 θ_i 包括均值矢量 u_i 及协方差矩阵 R_i。

p_i 是第 i 个高斯分布的权重，满足 $\sum_{j=1}^{M} p_i = 1, p_i \geq 0$。

θ 是高斯混合模型的参数，它包括 p_i 和 θ_i，i =1,2,…, M。

给定集合 Z，使用高斯混合模型描述频谱矢量的分布，需确定模型参数 θ，使得在该参数下矢量集 Z 发生概率 $p(Z|\theta)$ 最大。Dempster 给出了一种 EM 算法[Dempster 1977]求解一般高斯混合分布的参数 θ。本文给出该算法在语音声学建模时的简便形式。

设 Z 中矢量互相独立，有

$$p(Z \mid \theta) = \prod_{j=1}^{T} p(z_j \mid \theta) \tag{2-29}$$

设 $p(Z|\theta)$ 对 θ 可微，当 $p(Z|\theta)$ 取极大值时满足：

$$f = \nabla_{\theta}\left(\ln p(Z|\theta)\right) = 0 \tag{2-30}$$

解上式，即可得到估计 μ_i，R_i 和 P_i 的迭代算法（i＝1,2,…, M）。以 μ_i 为例：

$$\begin{aligned} f &= \nabla_{\mu_i}\left(\ln p(Z \mid \theta)\right) \\ &= \sum_{j=1}^{T} p(z_j \mid \theta)^{-1} \nabla_{\mu_i}\left(\sum_{i=1}^{M} p_i p(z_j \mid \theta_i)\right) \\ &= \sum_{j=1}^{T} p(z_j \mid \theta)^{-1} p_i \nabla_{\mu_i} p(z_j \mid \theta_i) \\ &= \sum_{j=1}^{T} p(z_j \mid \theta)^{-1} p_i p(z_j \mid \theta_i) R_i^{-1}(z_j - \mu_i) \\ &= 0 \end{aligned} \tag{2-31}$$

令

$$P_{i,j} = p(z_j \mid \theta)^{-1} p_i p(z_j \mid \theta_i) \tag{2-32}$$

于是得到

$$\mu_i = \sum_{j=1}^{T}\left(P_{i,j} \cdot z_j\right) \Big/ \sum_{j=1}^{T} P_{i,j} . \tag{2.33}$$

将等式左侧的 μ_i 改写为 $\hat{\mu}_i$，即得到估计 μ_i 的估值 $\hat{\mu}_i$ 的迭代算法。

类似地，可以得到 p_i 和 R_i 的估值 $\hat{p}_i$ 和 $\hat{R}_i$ 的迭代算法。

$$\hat{\mu}_i = \sum_{j=1}^{T}\left(P_{i,j}\cdot z_j\right)\Big/\sum_{j=1}^{T}P_{i,j} \tag{2-34}$$

$$\hat{R}_i = \sum_{j=1}^{T}\left(P_{i,j}(z_j-\mu_i)(z_j-\mu_i)^T\right)\Big/\sum_{j=1}^{T}P_{i,j} \tag{2-35}$$

$$\hat{p}_i = T^{-1}\cdot\sum_{j=1}^{T}P_{i,j} \tag{2-36}$$

需要注意的是，在推导公式(2-36)时，有约束条件 $\sum_{j=1}^{M}p_i = 1, p_i \geq 0.$ 即要在满足此条件的同时进行确定(2-29)式的极值。

2.3 上下文相关的声学建模技术

在连续语音识别系统中，人们在发音时普遍会受到上下文的影响而发生变化，这就是连续语音之间协同发音现象。上下文无关的建模方法对每个识别基元分别独立建模，忽略了这种协同发音的现象，因而上下文模型用于连续语音识别系统中无法取得很高的识别率。解决这一问题的方法是进行上下文相关的声学建模。与上下文无关的建模方法相比，上下文相关建模方法需要考虑如下的几个问题：

（1）如何选取基本识别基元。对于汉语语音识别系统而言，常用的基本识别基元有音节、声韵母和音素。由于汉语有 418 个无调音节，如果考虑上下文相关的变化，则会由于基元数目太多而导致模型无法实现。而声韵母与音素的数目都相对很少（分别只有大约 60 个和 40 个），因此可以用来作为上下文相关模型的基本识别基元。

（2）如何在保证识别率的前提下降低模型的规模。即使采用声韵母或音素作为上下文相关模型的基本识别基元，模型的规模仍然非常巨大。假设基本识别基元的个数为 40，则有 64,000 个可能的上下文相关基元。即使每个基元分为三个状态，每个状态采用单个高斯分布来描述，系统中仍然有 192,000 个高斯分布，如此大规模的模型会导致系统的识别速度下降，而且在训练数据库不是足够大的情况下，有些基元会存在训练不充分的问题。解决的办法是采用参数共享的技术。例如进行状态共享(State Tying)建模，或者混合密度共享（Tied

Mixture）建模（也就是半连续 HMM）。

（3）如何预测在训练数据中没有出现的基元。在上下文相关的声学模型中，由于训练数据的限制，有些基元可能在训练数据中完全不出现，但是可能出现在识别的结果中。为了保证识别解码过程的顺利进行，我们必须采取的补救措施保证每个识别基元都能使用模型描述。通常使用的方法是基于决策树（Decision Tree）的策略，使用那些可见基元的分布来合成在训练数据中不可见的基元。

在实际中的上下文相关声学建模技术中，通常采用决策树与状态共享相结合的策略，这样既可以降低模型规模，避免训练不充分问题，还可以有效合成那些训练数据中不可见的基元。

第三章　半连续分段概率模型

在上一章中我们详细介绍了 HMM 模型理论，以及经典 HMM 模型的改进形式 MGCPM 模型。为了进一步降低模型规模，提高系统的识别速度，本文在 MGCPM 的基础上，提出了一种 SCHMM 与 MGCPM 相结合的建模方法，这就是半连续分段概率模型(Semi-Continuous Segmental Probability Model，SCSPM)。

3.1 SCSPM 原理

3.1.1 分段概率模型(SPM)

通过对传统 HMM 的研和分析究发现，对识别结果影响最大的是 HMM 中的观察输出矩阵，而状态转移矩阵对识别结果的影响并不大[Juang 1985][Zheng 1997]。因此人们提出了基于分段的概率模型（Segmental Probability Model, SPM）。分段概率模型采用了自左向右无跳转的拓扑结构，状态内的特征空间采用混合高斯密度（MGD）来描述，而状态间的转移则采用基于相等特征变化量（Equal Feature Variance Sum, EFVS）[Xu 1999]的非线性分段（None Linear Partition, NLP）的策略[Jiang 1989]来控制。和传统的 HMM 模型相比，SPM 模型的训练和识别计算复杂度降低，而性能几乎没有下降。

3.1.2 与 SCHMM 相结合

在连续密度 HMM(CHMM)和半连续密度 HMM(SCHMM)中，一个状态被描述为一群基本的 pdf（通常是高斯函数）的混合。然而， CHMM 和 SCHMM 之间的区别在于：CHMM 中的每个状态通过一个特定集合（比较小）的高斯 pdf 来建模，而在 SCHMM 中，对于所有状态而言，它们共享一个大的高斯 pdf 的集合，不同的是混合中的高斯函数的权重不一样。因此 SCHMM 也可以被认为是一种共享密度（Tied Mixture）的 HMM[Bellegarda 1990]。SCHMM 具有如下的一些特点[Duchateau 1998]：

在 SCHMM 中的高斯概率密度函数的集合直接对整个参数空间建模，而不

是不同 HMM 状态所覆盖的子空间。在这种情况下，重新预测不同 HMM 状态中的相同的高斯函数的过程可以避免。

在 SCHMM 中，使用全特征空间的数据来训练所需的高斯函数，然后对每个状态只需要用属于它自己的那部分数据来训练它对应的混合权重。而对于 CHMM 状态而言，混合的权重和特定状态的高斯函数都要预测。因此，对于 SCHMM 的状态的训练只需要更少的数据。换句话说，使用同样数量的数据既可以对更多的状态建模，也可以通过增加混合成分的个数来提高状态建模的性能。

3.1.3 SCSPM 模型框架

SCSPM 是结合 SCHMM 模型及分段概率模型两者的特点，由上一章介绍的 MGCPM 模型演化而来。SCSPM 模型的训练分为两步：一是码本的建立；二是模型权重的估计。SCSPM 模型的训练过程如图所示：

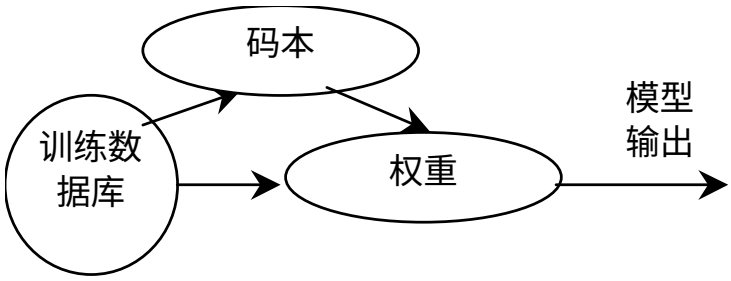


图 3-1 SCSPM 模型的训练过程

3.2 码本生成

SCSPM 中的码本是由混合高斯分布组成的，每个高斯分布由均值和方差来描述。有两种方法可以用来产生 SCSPM 模型所需要的码本，一种是基于分裂的方法，另一种是基于合并的方法。

3.2.1 基于分裂的方法

在介绍基于分裂的方法产生码本之前，我们首先介绍一下 K-均值算法和 LBG 聚类算法(或称 GLA，Generalized Lloyd Algorithm) [Sadaoki 1985]。 K-均值算法的流程如下：

(1) 将训练矢量集 S 调入内存；

(2) 设置最大迭代次数 L；

(3) 设置畸变域值δ；

(4) 设置 M 个码字的初始值 $C_1^{(0)}$，$C_2^{(0)}$ … $C_M^{(0)}$；

(5) 设置畸变初值 $D^{(0)}=\infty$；

(6) 设置迭代初值 $m=1$；

(7) 根据最近邻准则将 S 分为 M 个子集 $S_1^{(m)}$，$S_2^{(m)}$ … $S_M^{(m)}$，即当$X\in S_l^{(m)}$时，下式成立：

$$d(X,Y_l^{(m-1)}) \le d(X,Y_i^{(m-1)}), \forall \mathrm{i} \tag{3-1}$$

(8) 计算总的误差 $D^{(m)}$：

$$D^{(m)} = \sum_{l=1}^{M} \sum_{X \in S_l^{(m)}} d(X,Y_l^{(m-1)}) \tag{3-2}$$

(9) 计算误差增量$\Delta D^{(m)}$的相对值$\delta^{(m)}$：

$$\delta^{(m)} = \frac{\Delta D^{(m)}}{D^{(m)}} = \frac{\left| D^{(m-1)} - D^{(m)} \right|}{D^{(m)}} \tag{3-3}$$

(10) 调整码字 $C_1^{(m)}$，$C_2^{(m)}$ … $C_M^{(m)}$：

$$C_i^{(m)} = \frac{1}{N_i} \sum_{X \in S_i^{(m)}} X \tag{3-4}$$

(11) 判断误差增量的相对值是否满足域值条件，$\delta^{(m)}<\delta$？是，则转(13)；否，则转(12)；

(12) m<L?是，m=m+1，转(7)；否，则转(13)；

(13) 迭代结束，输出码字 $C_1^{(m)}$，$C_2^{(m)}$ … $C_M^{(m)}$，及总的误差 $D^{(m)}$ ；

(14) 结束。

在上述算法中，相对误差域值δ和迭代次数 L 需要依据经验设置。δ一般是一个大于 0 而远小于 1 的数，如果$\delta^{(m)}<\delta$，认为码字已基本稳定，继续迭代获得的相对误差的减少量比起时间消耗来讲，是不划算的；L 指定了最大迭代次数，以免δ设置得太小，迭代次数过多。在我们的实验过程中发现，L 一般来说设为 20 即可。

由于 K-均值算法依赖于初始码本，当没有初始码本时，我们可以改进上面的算法，以分裂的方式得到初始码本，这就是 LBG 算法。LBG 聚类算法可以看作时一种带分裂的 K-均值算法。其实现算法可以描述如下：第一步求出 S 中的全体 X 的质心 X_0，然后对此质心作变换（例如：对质心的每一维都乘以 1.01，或除以 1.01）进行分裂，得到两个新的初值 X_0，X_1；然后以它们作为初始值调用上面的 K-均值算法进行处理。这样经过一次分裂就得到了 2 个码本，并且将全体样本 X 分成 2 个子集。对每个子集分别做同样的处理。这样经过 B 次分裂，就可以得到容量为 $M=2^B$ 的码本。

当每次使用LBG算法分裂码本的时候，得到的码本满足离散度最小的条件，而在半连续模型当中，我们期望得到的码本对于全空间特征具有最大似然估计。因此对于 LBG 算法得到的均值和方差，我们还要进行最大似然估计的调整，我们这里称之为 MGD 迭代。这种操作与 MGCPM 中训练一个状态的流程完全一致，只是它的混合个数在这里为码本的规模。

设 N 为所期望得到的码本的规模，下面我们给出基于分裂产生码本的步骤：

1．计算全体训练数据的均值和方差；当前的码本容量 $n=1$；

2．运用 LBG 算法来将容量为 n 的码本分裂为 n×2 的码本；当前码本容量 n 变为 $n\times2$；

3．运用 MGD 迭代方法来调整码本。目标是使码本中的高斯分布对全特征空间的描述具有最大似然概率；

4．如果 $n=N$，　则结束；否则转至步骤 2。

在上述的基于分裂的方法中，每次迭代都要对获得的码本进行调整以获得最大似然概率，这需要很大的计算量。比如说当 $N=4{,}096$ 时，最后一趟需要同时调整 4,096 个高斯分布的均值与方差，这在当时实验环境下难以完成。

3.2.2 基于合并的方法

在 MGCPM 模型当中，通过对 MGCPM 模型的训练可以得到一些高斯密度。我们可以利用这些高斯密度作为初始的码本。由于 MGCPM 模型中的高斯密度数目较大，我们可以采用合并的方式来逐渐减少码本的规模。这就是基于合并的码本产生方法。步骤描述如下：

1．选择相邻最近的一对高斯密度。在这里定义一个简单的度量高斯密度之间的距离公式如下：

$$D(G_1, G_2) = \| \bar{\mu}_1 - \bar{\mu}_2 \| \tag{3-5}$$

其中 G_t 是第 t 个高斯密度，$\bar{\mu}_t$ 为 G_t 的均值向量；

2．假设 G_i 和 G_j 是上一步中选择出来的两个高斯密度，将 G_i 和 G_j 从码本当中移走，再在码本中加入一个新的高斯分布 G，其均值 $\bar{\mu}$ 与方差 $\bar{\sigma}$ 分别为：

$$\bar{\mu} = (\bar{\mu}_i + \bar{\mu}_j)/2, \quad \bar{\sigma} = (\bar{\sigma}_i + \bar{\sigma}_j)/2 \tag{3-6}$$

3．假设 n 为当前的码本大小，如果 $n = N$，则停止；否则转向步骤 1。

3.3 模型权重估计

得到了码本之后，对于每个识别基元的每个状态，我们需要训练出对应的所有高斯混合的权重。也即是模型的权重估计。在这里采用的是最大似然估计。M 代表码本的大小，θ 代表待训练的模型，θ_m ($m=1,2,..,M$)表示码本中的第 m 个高斯密度，Z 代表训练集的矢量集合，J 表示 Z 中的矢量个数，Z_j 表示 Z 中的第 j(j=1,2,..J)个矢量。g_m 代表第 m 个高斯密度的权重。模型权重的估计公式推导如下：

根据 MGD, 我们可以得到矢量 Z_j 的概率分布：

$$p(z_j \mid \theta) = \sum_{m=1}^{M} \{ g_m (p(z_j \mid \theta_m) \} \tag{3-7}$$

$$\sum_{m=1}^{M} g_m = 1 \tag{3-8}$$

我们定义如下的目标函数：

$$h(\theta) = \ln(p(Z \mid \theta)) = \ln(\prod_{j=1}^{J} p(z_j \mid \theta)) = \sum_{j=1}^{J} \ln p(z_j \mid \theta) \tag{3-9}$$

其中（3-9）式是条件极值，其拉格朗日函数为：

$$f = h(\theta) + \lambda(\sum_{m=1}^{M} g_m - 1) \tag{3-10}$$

对于每一个 $t=1,2,...M$, 我们可以得到如下的微分方程：

$$\sum_{\mathrm{j}=1}^{\mathrm{J}} p(z_j \mid \theta)^{-1} \nabla_{g_t} (\sum_{m=1}^{\mathrm{M}} g_m p(z_j \mid \theta_m)) + \lambda = 0 \tag{3-11}$$

$$\lambda \cdot g_t = -\sum_{j=1}^{J} p(z_j \mid \theta)^{-1} g_t p(z_j \mid \theta_t) \tag{3-12}$$

$$\lambda = -\sum_{t=1}^{M} \sum_{j=1}^{J} p(z_j \mid \theta)^{-1} g_t p(z_j \mid \theta_t) = -J \tag{3-13}$$

根据（3-12）和（3-13）， 我们可以得到 g_t 的迭代公式：

$$\breve{g}_t = J^{-1} \cdot \sum_{j=1}^{J} p(z_j \mid \theta)^{-1} g_t p(z_j \mid \theta_t) \tag{3-14}$$

3.4　SCSPM 的精简策略

在 SCHMM 中对于高斯混合权重的估计，只有一部分高斯混合有比较大的权重。我们可以采取一定的精简策略，使得每个状态只共享一部分的高斯混合概率密度函数，从而大大简化模型的规模。下面介绍一些可以采用的精简策略[Gales 1999] [Fischer 1999]。

方法 1：去掉小权重的高斯混合。如果一个高斯混合的权重小于某个阈值，则将此高斯混合的权重设为 0。

方法 2：只保留一定数目的高斯混合。把权重从大到小来排序，只选取前 N 名并且保留。

方法 3：根据概率阈值选取。把权重从大到小来排列，并且依次选取，直到权重和达到某一个阈值。未选中的权重置为 0。

上述的三种精简策略虽然能极大地降低模型的规模，但是都存在一定的问

题。方法一的结果将不能满足公式(3-8)所限定的条件，方法二和方法三有可能会精简掉一些具有较大权重的高斯密度。

通过考察 SCSPM 模型的训练过程，我们发现对于每个状态的权重估计都是通过 MGD 迭代过程逐步得到的，因此我们将精简策略引入到 MGD 迭代过程当中，从而得到如下新的模型精简策略：

方法 4： 在 MGD 迭代过程中去掉权重小的高斯混合。在迭代过程当中逐渐去掉对描述训练数据贡献小的高斯混合，最终得到一个比较有效的高斯混合。

方法 1 与方法 4 之间的区别在于后者是在迭代的过程中逐渐选择最终所需要的高斯混合，因此阈值可以设置得足够小，并且在迭代结束后得到的权重满足公式(3-8)所限定的条件。

3.5 实验

在下面的实验当中，我们使用了 863 连续语音数据库。我们的实验数据库包含 13 个人的数据，其中每个人有 520 句话。所有的数据都是在低噪声的环境下录入的。其中 10 个人的数据用来作训练，其余 3 个人的数据用来作测试。数据的采样率为 16KHz。我们所使用的特征是 16 维的 MFCC 参数及其一阶差分。汉语的音节作为识别基元，每个音节分为 6 个状态。

3.5.1 码本规模的实验

在 SCSPM 模型当中，码本的规模是影响识别率的一个很重要的因素。下面我们分别给出码本中有 1,024，2048 和 4,096 个高斯密度时模型的识别率。在这个实验中，精简策略使用了方法 4。

由于计算复杂度太高，分裂产生规模为 4,096 的码本没有创建。从表 3-1 可以看出，当把码本的规模从 1,024 增加到 4,096 时，音节误识率会下降大约 36.6%。而且还可以看出由合并的方法得到的码本比使用分裂的方法得到的码本具有更好的性能。

表 3-1 码本规模的实验

码本		音节正确率 (%)		
大小	产生方法	首选	前5选	前10选
1,024	分裂	60.39	85.74	91.56
	合并	61.33	86.08	91.64
2,048	分裂	69.50	88.79	92.76
	合并	70.21	88.93	92.94
4,096	分裂	-	-	-
	合并	75.49	90.48	94.42

3.5.2 精简策略的实验

在这个实验当中，我们选择使用合并的方式得到的规模为 2,048 的码本。对于每种精简策略，分别都有一个经验的阈值，用来保证每种精简后的模型的规模相当，也就是每个状态大致保留 50 个左右的高斯分布。实验结果如下：

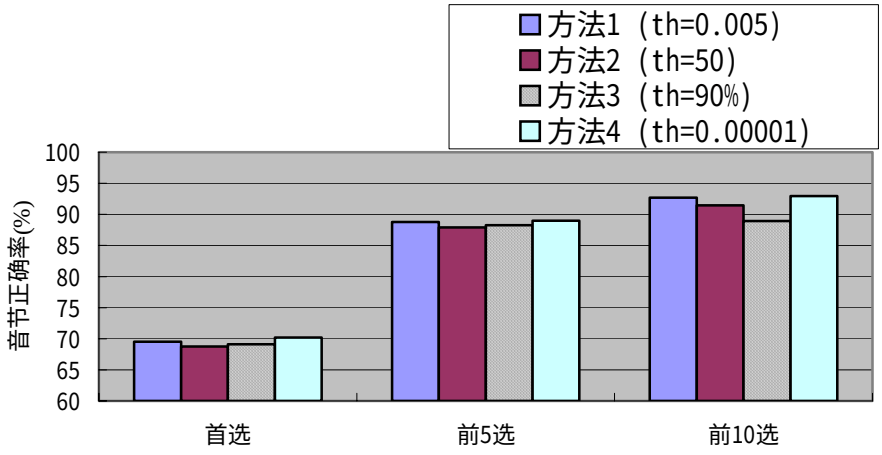


图 3-2 精简策略的实验

从上面的实验结果当中，我们可以看出使用方法 4 能得到最高的音节正确识别率。

3.5.3 SCSPM 与 MGCPM 的对比实验

在第二章我们详细介绍过 MGCPM 模型，在我们对 MGCPM 模型与 SCSPM 模型性能作比较。其中 MGCPM 模型的每个状态分别使用 4 个和 8 个混合高斯

分布来描述。SCSPM 模型的码本使用合并的方法来产生。实验结果如下：

表 3-2 SCSPM 与 MGCPM 的性能对比

模型		音节正确率 (%)		
名称	高斯分布数目	首选	前5选	前10选
MGCPMs (4混合)	9,669	74.35	89.15	93.78
MGCPMs (8混合)	18,685	75.87	90.62	94.55
SCSPMs	4,096	75.49	90.48	94.42

从上面的结果中我们可以看出SCSPM模型的音节正确识别率高于4混合的MGCPM 模型但是低于 8 混合的 MGCPM 模型。此外与 MGCPM 模型相比，SCSPM 模型中的高斯密度数目明显降低。因此，采用 SCSPM 模型在保证了识别率基本不下降的同时能大大降低模型的计算复杂度。

3.6 小结

在这一章里我们讲述了构造 SCSPM 模型的方法。实验结果证明，与 MGCPM 相比，SCSPM 模型能降低模型的计算复杂度，同时识别正确率没有明显下降。码本的规模是影响 SCSPM 模型的一个重要因素，当码本规模变大时，识别率会有所上升，但同时计算复杂度也会增加。为了减少每个状态中共享的高斯分布的数目，我们研究了四种精简策略，其中在模型训练的迭代过程中逐步去掉小的共享权重的策略能取得最好的效果。

第四章 上下文相关的音素建模

在上面介绍的 MGCPM 模型和 SCSPM 模型当中，我们都是采用音节作为识别基元进行建模的。以音节为识别基元的好处是稳定性好，但是不容易考虑上下文相关信息。我们知道，汉语有 418 个无调音节，如果考虑上下文相关信息，则有大约 73,000,000 个可能的识别基元，这会导致模型的规模太大以及训练数据过于稀疏，以至于无法进行有效的建模。因此，当进行上下文相关的声学建模的时候，我们必须选取数目相对较少，持续时间相对较短的识别基元。在西方的语音识别系统中，通常采用的是音素作为识别基元。而在汉语语音识别当中，根据汉语语音的特点，我们还可以选择声韵母作为识别基元。在本章和下一章我们将分别介绍上下文相关的音素建模和上下文相关的声韵母建模。

4.1 基本音素集与音节发音词典

4.1.1 基本音素集

根据汉语的发音，我们可以定义汉语的基本音素集。首先我们给出元音音素的定义[Ma 2000]：

表 4-1 元音音素的定义

音素	定义
/aI/	在韵母“ ai ”,“ an ”中的音素“ a ”
/a/	在其它条件中的音素“ a ”
/Ie/	在韵母“ ie ”中的音素“ e ”
/eI/	在韵母“ ei ”中的音素“ e ”
/eN/	在韵母“ en ”中的音素“ e ”
/e/	在其它条件中的音素“ e ”
/Ci/	在音节“ ci ”，“ si ”，“ zi ”中的音素“ i ”
/CHi/	在音节“ chi ”，“ shi ”，“ zhi ”中的音素“ i ”
/Bi/	在其它条件中的音素“ i ”
/oU/	在韵母“ ou ”中的的音素“ o ”
/o/	在其它条件下的音素“ o ”
/u/	元音音素“ u ”
/v/	音节 “ yu ” 中的元音音素

共有 13 个元音音素。

辅音音素为：

/b/, /c/, /ch/, /d/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /ng/, /p/, /q/, /r/, /s/, /sh/, /t/, /x/, /z/, /zh/

除/ng/表示后鼻音外，其它与汉语语音的声母一一对应。共有 22 个辅音音素。此外再加上一个卷舌韵母/er/和静音/sil/，共有 37 个基本音素基元。

4.1.2 音节发音词典

在基本音素集确定以后，我们还需要确定每个音节与音素之间的组合关系，也就是音节的发音词典。音节由声韵母两部分组成。前面已经说过，音节的声母部分有辅音音素与之一一对应，关系比较简单。因此只要确定韵母部分与音素之间的对应关系，就可以确定音节的发音词典。下面我们给出韵母与音素之间的对应关系：

表 4-2 韵母与音素的对应关系

韵母	音素串	韵母	音素串	韵母	音素串
a	/a/	ai	/aI/Bi/	an	/aI/Bi/
ang	/a/ng/	ao	/a/u/	e	/e/
ei	/eI/Bi/	en	/eN/n/	eng	/eN/ng/
er	/er/	o	/o/	ong	/u/ng/
ou	/oU/u/	i	/Bi/	ia	/Bi/a/
ian	/Bi/Ie/n/	iang	/Bi/a/ng/	iao	/Bi/a/u/
ie	/Bi/Ie/	in	/Bi/n/	ing	/Bi/ng/
iong	/Bi/u/ng/	iou	/Bi/oU/u/	u	/u/
ua	/u/a/	uai	/u/aI/Bi/	uan	/u/aI/n/
uang	/u/a/ng/	uei	/u/eI/Bi/	uen	/u/eN/n/
ueng	/u/eN/ng/	uo	/u/o/	v	/v/
van	/v/Ie/n/	ve	/v/Ie/	vn	/v/n/
io	/Bi/o/				

4.2 构造决策树

4.2.1 问题集设计

决策树的节点的分裂依赖于问题集。基本的原则是发音的相似性。首先我们根据发音相似性定义划分特征，分别为：元音音素划分特征、辅音辅音划分特征和单音素划分特征。其中元音音素和辅音音素的划分特征分别如下表所示[wu 1989]：

表 4-3 元音音素划分特征

划分特征	音素列表
高音(High)	/Bi/, /Ci/, /CHi/, /u/, /v/
中音(Medium)	/e/, /Ie/, /eI/, /eN/, /er/, /o/, /oU/
低音(Low)	/a/, /aI/
上元音(Top Vowel)	/a/, /aI/, /e/, /Ie/, /eI/, /eN/, /Bi/, /o/, /oU/, /u/, /v/
前音(Front)	/aI/, /a/, /Bi/, /v/, /Ie/
尾音(End)	/a/, /e/, /eI/, /eN/, /u/, /o/, /oU/
非合音(Unrounded)	/a/, /aI/, /Bi/, /e/, /Ie/, /eI/, /eN/
合音(Rounded)	/u/, /v/, /o/, /oU/
顶元音(Apical Vowel)	/Ci/, /CHi/
E类元音1(Evowel)	/e/, /Ie/, /eI/, /eN/
E类元音2(Evowel2)	/e/, /eI/, /eN/
I类元音(Ivowel)	/Bi/, /Ci/, /CHi/
O类元音(Ovowel)	/o/, /oU/
U和V音(u and v)	/u/, /v/

表 4-4 辅音音素划分特征

划分特征	音素列表
塞音(Stop)	/b/, /d/, /g/, /p/, /t/, /k/
塞送气音(Aspirated Stop)	/b/, /d/, /g/
非塞送气音 (Unaspirated Stop)	/p/, /t/, /k/
塞擦音(Affricate)	/z/, /zh/, /j/, /c/, /ch/, /q/
塞擦送气音 (Aspirated Affricate)	/z/, /zh/, /j/

非塞擦送气音 (Unaspirated Affricate)	/c/, /ch/, /q/
擦音(Fricative)	/f/, /s/, /sh/, /x/, /h/, /r/
擦音2(Fricative2)	/f/, /s/, /sh/, /x/, /h/, /r/, /k/
清擦音(Voiceless Fricative)	/f/, /s/, /sh/, /x/, /h/
浊擦音(Voice Fricative)	/r/, /k/
鼻音(Nasal)	/m/, /n/, /ng/
鼻音2(Nasal2)	/m/, /n/, /l/
鼻音3(Nasal3)	/m/, /n/, /l/, /ng/
唇音(Labial)	/b/, /p/, /m/
唇音2(Labial2)	/b/, /p/, /m/, /f/
顶音(Apical)	/z/, /c/, /s/, /d/ ,/t/, /n/, /l/, /zh/, /ch/, /sh/, /r/
顶前音(Apical Front)	/z/, /c/, /s/
顶音1(Apical1)	/d/ ,/t/, /n/, /l/
顶音2(Apical2)	/d/, /t/
顶音3(Apical3)	/n/, /l/
顶后音1(Apical End)	/zh/, /ch/, /sh/, /r/
顶后音2(Apical End2)	/zh/, /ch/, /sh/
舌前音(Tongue Top)	/j/, /q/, /x/
舌根音(Tongue Root)	/g/, /k/, /h/, /ng/
舌根音2(Tongue Root2)	/g/, /k/, /h/

为了使得决策树的分裂更加细致，我们将每个音素作为一个划分特征，这就是单音素划分特征。共有 37 个单音素划分特征。

在定义了划分特征之后，每种划分特征都可以转换成 3 种类型的问题：左问题，右问题和中心问题。例如：划分特征“ 停顿（Stop）”与如下的 3 个问题对应：

左问题（Left Question）：

QS “L_Stop”　{b-*, d-*, g-*, p-*, t-*, k-*}

右问题（Right Question）：

QS "R_Stop"　{*+b, *+d, *+g, *+p, *+t, *+k}

中心问题（Central Question）：

QS "C_Stop"　{*-b+*, *-d+*, *-g+*, *-p+*, *-t+*, *-k+*, *-b, *-d, *-g, *-p, *-t, *-k, b+*, d+*, g+*, p+*, t+*, k+*, b, d, g, p, t, k}

4.2.2 节点分裂的评估函数

评估函数是用来估计决策树的节点上的样本相似性[Gao 1998]。我们定义对数似然概率 $L(S) = \log P(X \mid S)$ 为节点 S 分裂的评估函数。其中 $X = \{X_1, X_2, \cdots, X_N\}$ 表示一个父节点总共包含 N 个样本。设 $X^1 = \{X_1^1, X_2^1, \cdots, X_{N1}^1\}$，$X^2 = \{X_1^2, X_2^2, \cdots, X_{N2}^2\}$ 表示由父节点划分的两个子节点所包含的样本，满足 $X = X^1 \bigcup X^2$，$X^1 \bigcap X^2 = \Phi$。父节点和两个子节点的评估函数的值分别表示为 L_{parent}，L_{child}^1 和 L_{child}^2。让 $\Delta = L_{child}^1 + L_{child}^2 - L_{parent}$ 表示其中的增量。在每个叶子节点进行分裂的时候，我们从问题集中选择一个问题，然后根据此问题把节点分成两个子节点并且计算增量 Δ，我们选择具有最大增量的问题，并且根据此问题把节点划分成两部分。当所有问题的增量都低于某个阈值的时候，此节点上的分裂过程将停止。

在具体的实现中，由于 $L(S)$ 不便于直接计算，我们采用如下的辅助函数作替换[Reichl 2000]：

$$Q(s) = \sum_{x_t} \sum_{s \in S} \gamma_s(x_t) \log N(x_t \mid \mu(S), \sum(S)) \tag{4-1}$$

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

$$Q(\hat{S}) \geq Q(S) \Rightarrow L(\hat{S}) \geq L(S) \tag{4-2}$$

因此我们可以使用 $Q(S)$ 来作为评估函数。为了减少决策树分裂过程中的计算复杂度，分裂过程中每个节点上的样本分布都采用单高斯分布来描述。待决策树分裂结束后，再对每个叶子节点采用更加精确的混合高斯分布来描述。

4.2.3 两种决策树构造方法

根据上面定义的 37 个基本音素基元，以这些基本基元为中心，考虑它们上下文相关的情况，我们可以将每个上下文相关的基元表示为 l-c+r 的方式，其中 c 为中心基元，l 为左相关信息，r 为右相关信息。根据汉语之间的搭配关系，我们可以统计出系统中共有 8,757 个可能的上下文相关音素基元。使用决策树的目的是尽可能地将那些发音相似的基元“共享(tie)”到一起，减少最终的状

态数目。这也就是上文提到的参数共享策略。这种做法有三个好处：一是降低模型的规模；二是避免由于训练数据的稀疏性而造成训练不充分的问题；三是合成那些在训练数据中不存在的基元。

对于每个音素基元，我们定义它的拓扑结构如下：

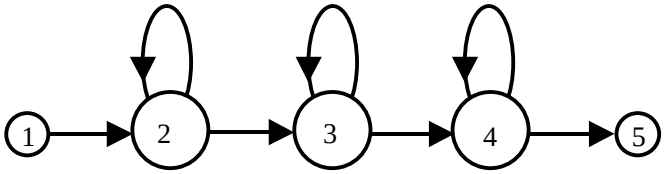


图 4-1 音素基元的拓扑结构

其中状态 1 和状态 5 分别为起始状态和结束状态，它们不能驻留，只在模型中起辅助作用。而状态 2，3 和 4 可以驻留或者转移到下一个状态。因此，在识别中真正起作用的是中间的三个状态。因此在构造决策树的时候，我们只考虑中间的三个状态。

方法 I：对每个中心基元的每个状态分别构造决策树。这种方法假设当基元的中心音素不同时，基元之间相互独立，因此首先根据中心音素对所有的基元进行分类，然后在利用决策树来进行状态共享。图 4-2 给出了中心音素为/aI/的所有基元的状态 2 组成的决策树示意图。

方法 II：对所有基元的同一个状态构造决策树。这种方法假设当中心音素不同时，基元之间仍然有一定的重叠。即使基元的中心音素不同，它们之间的状态仍然有可能共享。基元之间的状态共享情况完全依靠基于决策树的分类策略。图 4-3 给出了所有基元的状态 2 组成的决策树示意图。

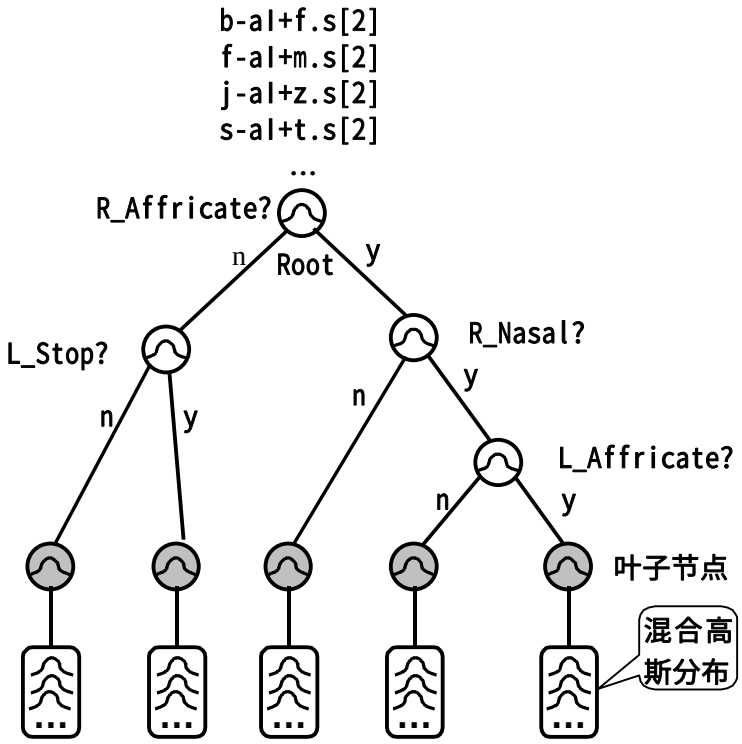


图 4-2 由方法 I 构造决策树示意图

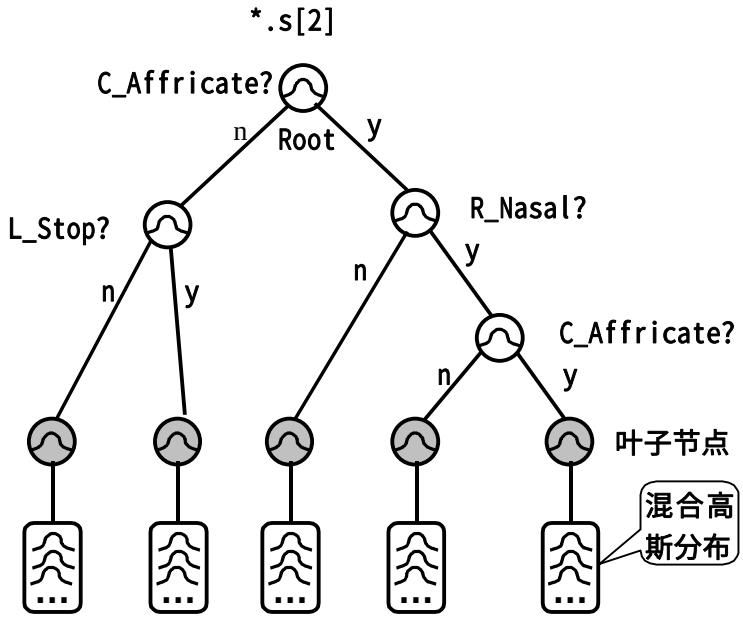


图 4-3 由方法 II 构造决策树示意图

在我们的音素模型实验中，方法 I 共构造了 37×3＝111 棵不同的决策树。

决策树中的节点进行分裂时，我们只使用节 4.2.1 中定义的左问题和右问题。而方法 II 只需构造 3 棵不同的决策树。决策树中的节点进行分裂时，我们除了使用节 4.2.1 中定义的左问题和右问题外，还要使用中心问题。

4.3 实验

我们仍然采用 863 语音数据库进行如下的实验。数据库包含有 80 个说话人的数据，每个说话人有 520 个句子。全部语音在低噪声的办公室环境下录制，采样率为 16KHz。实验使用的特征为 13 维的 MFCC 特征加上 1 维的对数能量，然后再对它们求一阶差分和二阶差分，共同组成 42 维的特征矢量。帧长为 12ms。HTK v2.2[Yong 1999] 被用来进行模型训练（具体的的训练步骤参见附录）。

4.3.1 上下文无关音素模型

在进行上下文相关音素建模之前，我们先训练上下文无关音素（CI-Phone）模型。下表给出了 CI-Phone 模型的实验结果：

表 4-5 CI-Phone 模型的性能

混合数目	音节识别正确率(%)
1	31.30
2	37.69
4	43.42
8	47.37

从上表中可以看出，当没有进行上下文相关建模时，音素模型的识别率非常低，即使混合的数目增加到 8，音节识别率仍然只能达到 47%左右。因此要想获得高识别率，上下文相关建模势在必行。

4.3.2 两种决策树构造方法的比较

在进行上下文相关的声学建模过程中，决策树的构造是很重要的步骤。在这个实验中，我们对 4.2.3 节中的方法 I 和方法 II 进行了对比实验。其中 Δ 是控制决策树节点分裂的阈值。下表给出了模型的音节首选正确识别率：

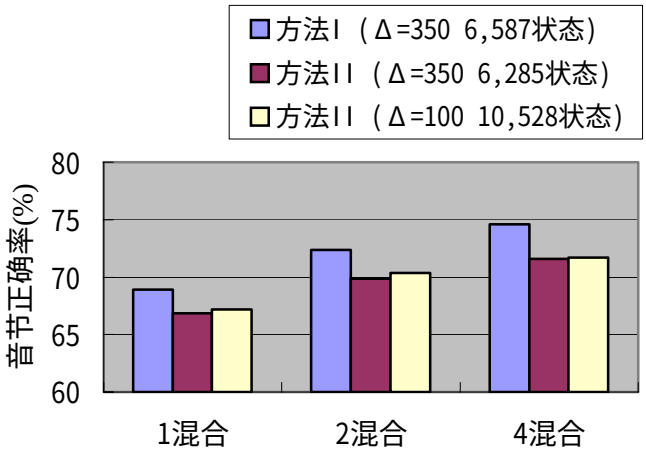


图 4-4 两种决策树构造方法的实验结果对比

从上表我们可以看到，在Δ一定的情况下，由方法 II 训练得到的模型的状态数目比由方法 I 得到的模型的状态数目要少，这说明由方法 II 得到的模型状态之间共享的程度要深。但是测试结果表明，方法 II 得到的模型的音节识别率要低于方法 I 得到的模型大约 3 个百分点。即使我们降低方法 II 的阈值（Δ由 350 降到 100），模型中的状态数目增加了，音节识别率同时也有所上升，但是识别率依然比方法 I 要低。因此，实验结果表明采用方法 I 构造决策树能取得较好的识别率。

4.3.3 CD-Phone 模型与音节模型的比较

为了测试 CD-Phone 模型的性能，我们将它与以音节作为识别基元的上下文无关模型进行了比较。其中音节模型中的每个音节分为 6 个状态，训练步骤参照附录中的“ 上下文无关模型训练步骤 ”。表 4-6 给出了实验对比结果。

从表 4-6 可以看出，CD-Phone 模型的音节正确识别率要明显高于音节模型，在混合数目都为 8 的时候，CD-Phone 模型的误识率比音节模型降低了大约 10%。

表 4-6 CD-Phone 模型与音节模型的比较

混合数目	音节识别正确率(%)	
	音节模型	CD-Phone模型
1	60.51	68.92
2	65.04	72.35
4	69.92	74.59
6	72.37	75.62
8	73.83	76.47

4.4 小结

本章我们详细介绍了上下文相关音素建模方法。主要讨论了其中的问题集设计和决策树的分裂评估函数。此外还分析了两种不同的决策树构造方法。实验结果显示对每个中心基元的每个状态分别构造决策树的方法（方法 I）能取得更好的识别效果。为了显示 CD-Phone 模型的性能，我们将上下文无关音节模型作为基准模型并且与之进行了对比，结果显示 CD-Phone 模型的误识率比音节模型降低了大约 10%，证明了 CD-Phone 模型的性能要高于音节模型。

第五章 上下文相关的声韵母建模

在上一章中我们详细介绍了上下文相关的音素模型的建模方法，并且给出了模型的实验测试结果。由于连续语音中的音素在发音过程中常常会发生音变（增音、减音、浊化等）现象，因此会影响到模型的识别率。而我们知道，汉语是一种基于音节的语言，而且每个音节都是由声母和韵母组成，声韵母结构是汉语语音中的一种稳定的结构。因此选用声韵母作为汉语语音识别系统的识别基元是一种不错的选择。在这一章里我们将讨论上下文相关的声韵母建模技术。

5.1 声韵母基元的选取

根据汉语语音学知识，我们知道汉语语音包含 21 个声母和 38 个韵母[Li 2000]：

表 5-1 标准声韵母(Initial/Final，IF)集合

类型	基元列表
声母（21）	*b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r*
韵母(38)	*a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn*

汉语中的大部分音节是由声母和韵母组成的，而有些音节只包含韵母部分。那些没有声母的音节我们称之为零声母音节。例如：零声母音节“ ang ”只由韵母“ ang ”组成，而音节“ zhang ”由声母“ zh ”和韵母“ ang ”共同组成。由于零声母音节的存在，在连续语音中声韵母的上下文关系比较复杂，声母的上下文只能是韵母，而韵母的上下文既可以为声母，还可以为韵母。因此，在利用上述的标准声韵母基元进行上下文相关建模的时候，总体的上下文相关基元数目会很多。

为了避免上述的由于零声母音节而造成的影响，我们引入了 6 个零声母：

“ _a ”，“ _o ”，“ _e ”，“ _i ”，“ _u ”和“ _v ”。这样零声母音节也可以由一个声母和一个韵母组成。例如：音节“ ang ”可以看作由零声母“ _a ”和韵母“ ang ”组成。下表给出了扩展声韵母的集合。

表 5-2 扩展声韵母（Extended Initial/Final，XIF）集合

类型	基元列表
声母（27）	*b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r, _a, _o, _e, _i, _u, _v*
韵母（38）	*a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn*

当我们采用 XIFs 作为基本识别基元的时候，每个韵母的上下文都只能是声母，因此两个韵母相邻的情况将不再出现。这个规则大大降低了上下文相关建模时的基元的数目。在我们的实验中，采用 IFs 进行上下文相关建模时，存在 122,118 个可能的基元，而使用 XIFs 进行上下文相关建模时，只有 29,047 个可能的上下文相关基元。因此这个改进使基元数目大大降低。

5.2 问题集设计

问题集设计的关键是利用发音方式的相似性。对于每一种发音方式，我们可以得到相应的左问题和右问题。例如：由发音方式“ 塞擦音（Affricate）”可以得到如下的两个问题：

QS “R_Affricate”　{*+z, *+zh, *+j, *+c, *+ch, *+q}

QS “L_Affricate”　{z-*, zh-*, j-*, c-*, ch-*, q-*}

记号“ + ”表示右相关，“ - ”表示左相关。我们还可以考虑发音方式的组合，例如：由发音方式“ 塞擦音（Affricate）”和“ 送气音（Aspirated）”组合可以得到送气塞擦音（Aspirated Affricate）：

QS “R_AspiratedAffricate”　{*+z, *+zh, *+j }

QS “L_AspiratedAffricate”　{z-*, zh-*, j-*}

韵母部分的发音方式同样可以用来构造问题。例如：发音方式为“ a ”音的问题设计为：

QS “R_Type_A”　　{*+a, *+ai, *+an, *+ang, *+ao }

QS “L_Type_A”　　{a-*, ia-*, ua-*}

注意上述的左问题和右问题并不对称，这是由于韵母部分对与它左右相邻的基元有不同的影响。根据上面的方法以及汉语语音知识，我们共设计了 61 个组合问题，其中 32 个左问题和 29 个右问题。

为了让决策树的分类更加细致，我们针对每个声韵母都分别设计了左问题和右问题，例如：对于声母 “ p ” 有如下的单问题：

QS “R_p”　　{*+p }

QS “L_p”　　{p-*}

此外我们还注意到问题集依赖于所选取的识别基元集合。上面给出的例子都是针对 IFs 作为识别基元的。当我们选择 XIF 作为基本识别基元时，我们需要对问题集进行相应的调整。例如，在上面的针对发音方式为 “ a ” 音的左问题需要修改为：

QS “L_Type_A”　　{ _a-*, a-*, ia-*, ua-*}

5.3 静音模型

由于静音持续的时间可能很长（长时间的停顿），也可能很短（连续两个音节之间），因而对于静音模型需要作相应的特殊处理。处理按照如下的步骤进行：

1. 原有的静音模型（即 sil 模型）增加状态 2 到状态 4 的转移弧，设置转移的概率为 0.2；
2. 增加短停顿（Short Pause, sp）模型。此 sp 模型只包含 3 个状态，其中状态 1 和状态 3 分别为起始状态和结束状态，状态 2 的分布与 sil 模型的状态 3 的分布相同；
3. sp 模型增加由状态 1 至状态 3 的转移弧。设置转移的概率为 0.3；
4. 将 sil 模型的状态 3 与 sp 模型的状态 2 进行状态共享。

处理后的静音模型结构示意图如下图所示。

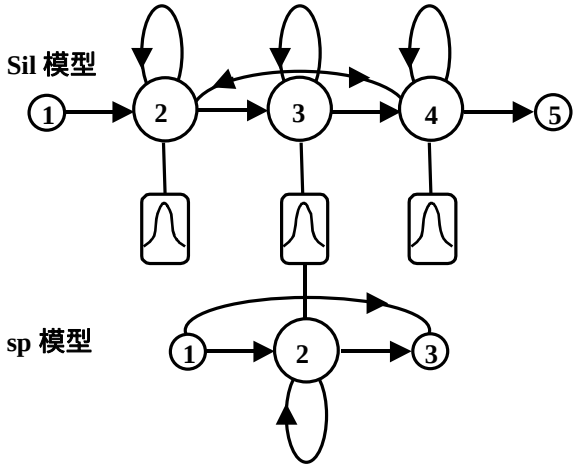


图 5-1 带 sp 的静音模型示意图

通过以上对静音模型的处理，静音模型不但能适应静音持续时间较长的情形，还能吸收那些连续音节之间持续时间非常短的停顿。因而提高了系统整体的鲁棒性。

5.4 实验

为了保证与上一章的上下文相关音素模型的可比性，实验所使用的数据库以及特征都与上一章 4.3 节中介绍的条件相同。模型的训练步骤请参见附录。

5.4.1 IF 模型与 XIF 模型的对比实验

为了对比 IF 模型与 XIF 模型的性能，下表给出了在上下文无关建模的情况下的实验结果。

表 5-3 IF 模型与 XIF 模型的对比

混合数目	音节识别正确率(%)		
	CI-Phone模型	CI-IF 模型	CI-XIF 模型
1	31.30	43.71	47.09
2	37.69	50.13	55.26
4	43.42	54.25	58.67

从上表中可以看出，XIF 模型的音节正确识别率比 IF 模型要高大约 4 个百分点。这说明使用扩展的声韵母基元进行建模能取得较高的识别率。在下面的

实验中，我们选择 XIF 模型作为基本识别基元。

5.4.2 SP 模型对识别率的影响

为了测试加入 sp 模型后对识别率的影响，我们训练了 CD-XIF 模型，并且进行了对比实验结果如下表：

表 5-4 sp 模型对识别率的影响

混合数目	音节识别正确率(%)	
	不带SP模型	带SP模型
1	74.85	75.24
2	76.83	77.28
4	78.77	79.30

从上面的结果可以看出，加入 sp 模型后，模型的识别率提高了大约 0.5 个百分点。这说明这种对静音模型的处理方法是一种有效的方法。

5.4.3 决策树中不同的阈值的实验

影响决策树状态共享的音素有构造决策树的方式，问题集的设计，以及控制决策树分裂结束的条件。其中构造决策树的方式我们在上一章已经进行了讨论，在这里我们选取方法 I 来构造决策树。问题集的设计已经考虑了各种发音方式及其组合，并且也加入了单问题集，因此比较完备，在此也不作讨论。在这一小节里我们主要讨论控制决策树分裂的阈值对模型规模及识别率的影响。下表给出了实验结果（每个状态的高斯混合数目为 1）：

表 5-5 不同的阈值对识别率的影响

阈值 Δ	状态数目	音节正确识别率 (%)
100	21,222	76.23
200	14,630	76.19
350	9,311	75.24

在我们的系统中，进行基于决策树状态共享之前的模型中共包含 42,255 状态，在进行状态共享后，状态数目得到了很大的降低。这说明状态共享策略极大地降低了模型的规模。当阈值 Δ 增大时，状态数目减少，同时音节识别率有

所降低。

5.4.4 各种模型性能比较

在这一小节里我们对上下文无关音节模型(CI-Syllable)、CD-Phone 模型、CD-IF 模型和 CD-XIF 模型进行了对比测试。其中后三种上下文相关模型的决策树构造策略选用方法 I，节点分裂阈值 Δ＝350。实验结果如图所示。

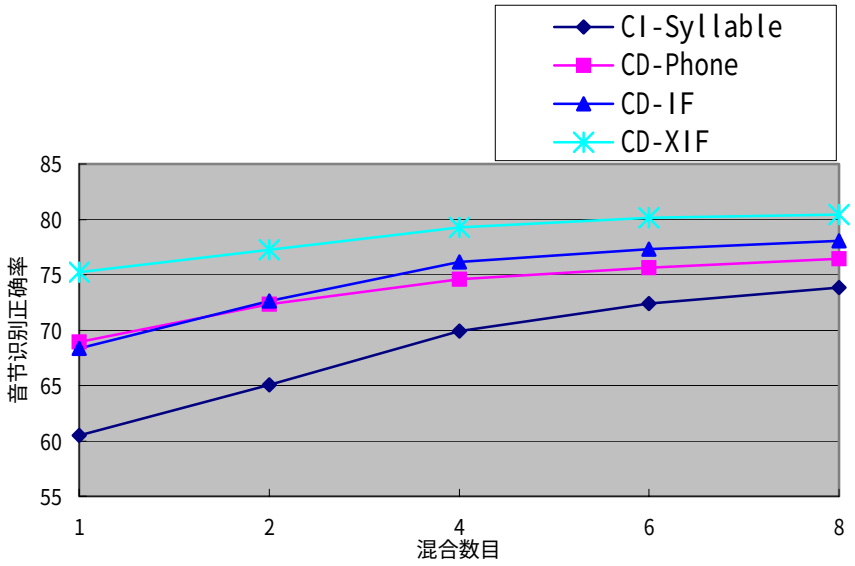


图 5-2 四种模型性能对比

从上图我们可以看出，当状态的混合数增加时，模型的识别率都能得到一定的提高，当混合数达到 8 时，增长的幅度已经很小，因此当混合数为 8 时模型性能基本达到了极限。另外我们还可以看出三种 CD 模型的性能都超过了基准的 CI-Syllable 模型。CD-Phone 模型与 CD-IF 模型性能相近，而 CD-XIF 模型则具有最好的性能，它的音节识别率比 CD-IF 模型要高大约 4 个百分点，达到 80.43%，误识率比基准的 CI-Syllable 模型降低了大约 25%。

5.5 小结

在这一章里，我们建立了基于决策树状态共享的上下文相关的声韵母

(CD-IF)模型。为了保持声韵母之间一致的搭配关系，我们在标准 IF 集合基础上加入了 6 个零声母基元，得到扩展的 XIF 集合。以 XIFs 为基本识别基元而得到的 CD-XIF 模型的音节识别率比 CD-IF 模型要高大约 4 个百分点，CD-XIF 模型的音节正确率达到 80.43%，误识率比基准音节模型降低了大约 25%。

第六章 结论与展望

6.1 论文的主要工作与贡献

论文的主要工作与贡献有以下几个方面：

1、提出并实现了半连续分段概率模型(SCSPM)。该模型在经典 HMM 模型及其修正模型混合高斯连续概率模型（MGCPM）基础上，结合矢量量化技术和连续概率密度描述的特点，以混合共享的方式来描述各状态的概率分布。此外还研究和分析了 SCSPM 模型的各种混合权重精简策略，并提出了一种新的在迭代过程中进行权重精简的策略。与原来的 MGCPM 模型相比，SCSPM 模型在保证识别率不下降的情况下，大大降低了模型规模和计算复杂度。

2、对 HTK 平台进行了研究和分析，实现了基于 HTK 平台的声学模型训练和性能评估的有效方法。

3、对上下文相关声学建模中的基于决策树状态共享策略进行了深入研究。分析了两种不同的决策树构造方法，讨论了问题集的设计和决策树节点分裂策略。此外还研究了对静音模型进行特殊处理的方法以提高鲁棒性。

4、实现了基于决策树状态共享的上下文相关的音素（CD-Phone）模型。着重研究了其中的基于决策树状态共享的上下文相关建模问题，其中包括根据音素发音特点设计问题集和不同的决策树构造方法。与音节模型相比，CD-Phone 模型能使音节误识率降低大约 10%。

5、研究并实现了基于决策树状态共享的上下文相关声韵母（CD Initial/Final, CD-IF）模型。为了保证声韵母之间的相互搭配关系，在原来的基本声韵母集合上，增加了零声母部分，形成扩展声韵母（Extended Initial/Final, XIF）集合。实验证明 XIF 模型比 IF 模型具有较高的识别率，最终实现的 CD-XIF 模型的音节正确率超过 80%，误识率比基准音节模型降低了大约 25%。

6.2 下一步工作展望

上下文相关的声学建模是当前大规模语音识别中的主流技术。与国外的语

音识别技术相比，国内对汉语语音识别中的上下文相关建模技术还处在研究阶段。如何结合汉语语音的特点，利用国际上领先的理论知识和技术方法，提高汉语语音识别系统的整体性能是当前迫切需要解决的任务。对于其中的问题集设计，决策树构造，状态共享策略等方面需要进一步的研究。

为了能在实际中应用语音识别技术，语音识别系统的鲁棒性和自适应性有待于进一步的提高。这方面需要研究的方向有口音建模，模型自适应，噪音去除，发音变化建模等。

附录：基于 HTK 的模型训练步骤

1. HTK 工具简介

HTK 是构建隐马尔可夫模型（Hidden Markov Model， HMM）的工具包。整个 HTK 工具包是由若干带有特定执行功能的程序所组成。按照工具所完成的的功能的性质，我们可以将整个工具包分为三个部分：数据准备、模型训练和优化、识别及性能评估。下面分别进行简要的介绍。

用来进行数据准备的工具有：

- Hbuild：转换各种不同格式的代表语言模型的文件并且输出标准 HTK 网格格式。
- HCopy：数据文件格式的转换。
- HDMan：利用各种数据源来生成发音词典。
- HLEd：编辑标注文件。
- HList：显示 HTK 支持的各种格式存放的数据源中的内容。
- HLStats：从一组 HTK 格式的标注文件中进行各种统计，生成简单语言模型。
- HParse：根据由扩展 Backus-Naur 形式(EBNF)定义的一组可重写的规则描述文件，生成词一级的网格文件。
- HSGen：根据以标准 HTK 网格格式定义的词网络自动随机产生一组句子。
- HSLab：对语音标注文件进行标注的编辑器。

用来进行模型训练和优化的工具有：

- HCompV：统计训练数据中的全局均值与方差。
- HERest：利用 Baum-Welch 算法对 HMM 模型进行一趟嵌入式训练

(Embedded Training)

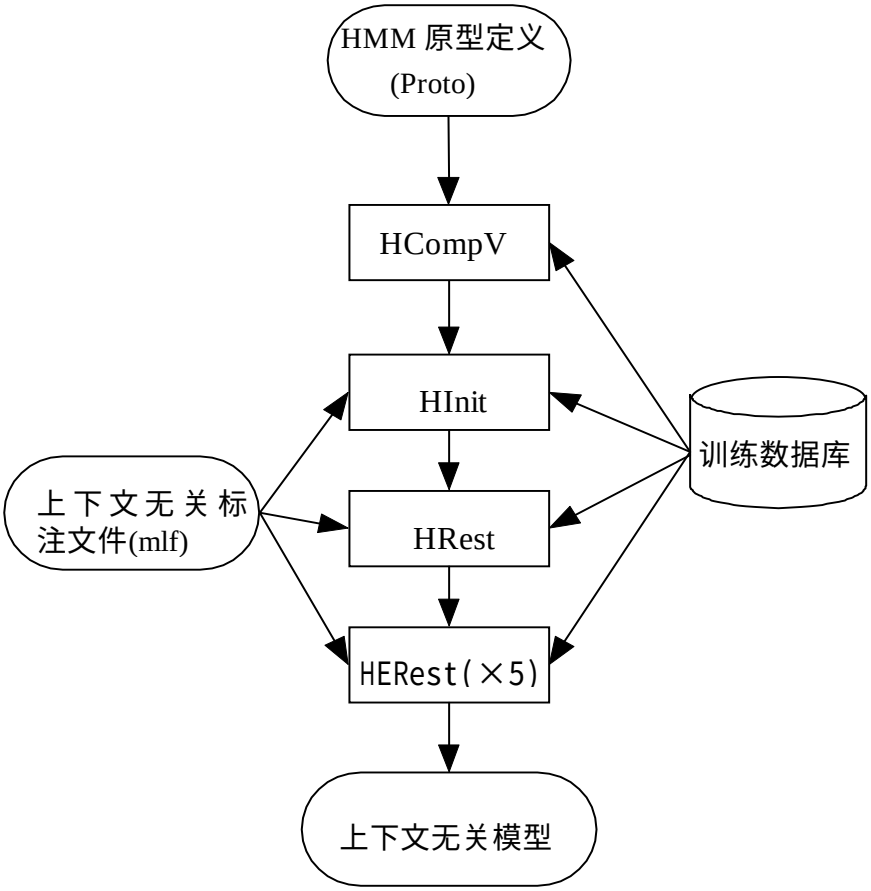
- HEAdapt：利用 MLLR 和/或 MAP 方法来对 HMM 模型进行自适应。
- HHEd：直接对 HMM 模型进行各种编辑和优化操作。例如改变模型类型、上下文相关建模、构造决策树、增加混合数目等。
- HInit：根据一组观察矢量序列对单个 HMM 模型进行初始参数估计。
- HQuant：构造 HTK 格式的 VQ 码表。
- HRest：根据一组观察矢量序列对单个 HMM 模型进行 Baum-Welch 参数重估。
- HSmooth：对一组上下文相关共享混合或离散 HMM 模型进行删除插入平滑。

用来进行识别及性能评估的工具有：

- HResults：HTK 模型性能分析工具。
- HVite：基于 Viterbi 算法的词识别器。

2. 上下文无关模型训练步骤

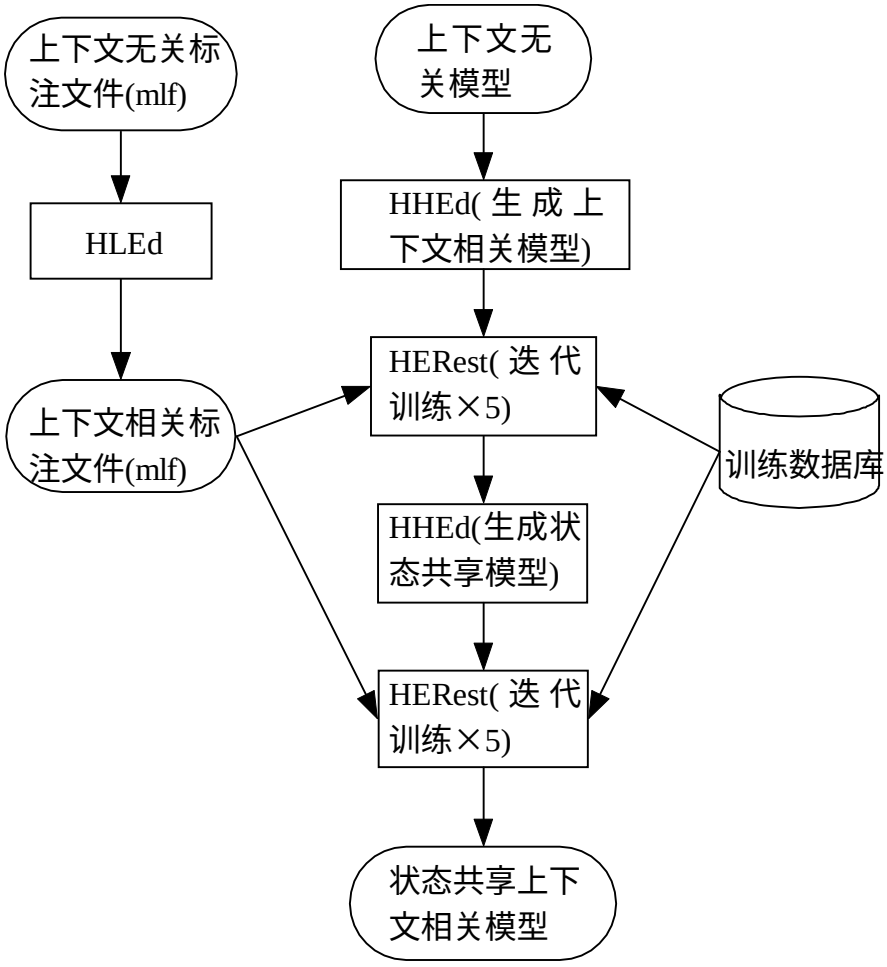
利用 HTK 进行上下文无关建模相对比较简单。通常有两种方法：一种方法是直接利用 HERest 进行嵌入式训练；第二种方法是首先根据基元的标注信息，利用 HInit 和 HRest 训练出初始模型，然后再利用 HERest 作进一步的 Baum-Welch 参数重估。由于第二种方法利用了基元的标注信息，初始模型的训练往往比较有效，因此在我们实际的训练中，采用第二种训练方法。训练的大致步骤如下图所示。



图附-1 上下无相关模型训练步骤

3. 上下文相关模型训练步骤

同上下文无关模型相比，上下文相关模型的训练要复杂很多。首先生成上下文相关的标注文件，可以借助 HLEd 工具来实现。然后利用 HHEd 工具将上下文无关模型转换为上下文相关模型，再用 HERest 进行若干遍迭代训练，得到上下文相关模型。但是由于有一些上下文相关基元没有在训练数据中出现，因此这些基元无法得到训练，在对测试数据进行识别的时候会造成影响。解决的办法是使用基于决策树的状态共享。最终得到状态共享的上下文相关模型。训练的大致步骤如下图所示。



图附-2 上下文相关模型训练步骤

参考文献

[1] **Bahl, L.R., Brown, P.F., Souza, P.V., et al, (Bahl 1986)** "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", *ICASSP*, pp.49-52, 1986

[2] **Baum, L.E., (Baum 1972)** "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," Inequalities, 3, 1972

[3] **Bellegarda, J.R., Nahamoo, D., (Bellegarda 1990)** "Tied Mixture Continuous Parameter Modeling for Speech Recognition", *IEEE Trans. on ASSP,* vol.ASSP-38, No.12, pp.2033-2045, 1990

[4] **Breiman, L., Friedman, J., Olshen, R.A., et al, (Breiman 1984)** "Classification and Regression Trees", Belmout, CA: Wadsworth, 1984

[5] **Dempster, A.P., Laird, N.M. and Rubin, D.B., (Dempster 1977)** "Maximum likelihood from incomplete data via the EM algorithm", J.R.Statist. Soc., vol.39, pp 1-83, 1977

[6] **Duchateau, J., Demuynck, K., Compernolle, D., (Duchateau 1998)** "Fast and Accurate acoustic Modeling with Semi-Continuous HMMs", *Speech Communication,* vol.24, pp.5-17, 1998

[7] **Fischer, V., RoB, T., (Fischer 1999)** "Reduced Gaussian Mixture Models in a Large Vocabulary Continuous Speech Recognizer", *EuroSpeech'99*, Vol.3, pp.1099-1102, 1999

[8] **Franzini, M.A., Lee, K.F., (Franzini 1990)** "A New Hybrid Method for Continuous Speech Recognition", *ICASSP-90,* 1990

[9] **Gales, M., Knill, K., and Young, S., (Gales 1999)** "State-Based Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMM's", *IEEE Trans. On Speech and Audio Processing*, vol.7, No.2, pp.152-161, 1999

[10] **Gao,S., Xu,B., Huang, T.Y., (Gao 1998)** "Class-triphone Acoustic Modeling Based on Decision Tree for Mandarin Continuous Speech Recognition", *International Symposium on Chinese Spoken Language Processing (ISCSLP'* 98), ASR-A1, Singapore , 1998

[11]**Huang, X.D., Jack, M.A., (Huang 1989)** "Semi-continuous hidden Markov models for speech signals", *Computer Speech and Language, vol.*3, pp.239-251, 1989

[12]**Jelinek, F., Mercer, R.L., Bahl, L.R., (Jelinek 1983)** "A maximum likehood approach to continuous speech recognition", *IEEE Trans. On Pattern Analysis and Machine Intelligence,* Vol. PAMI-5, pp179-190, 1983

[13]**蒋力, (Jiang 1989), "**基于概率统计模型的非特定人语音识别方法与系统的研究", *[硕士学位论文]. 北京：清华大学计算机科学与技术系,* 1989

[14]**Juang, B.H., Rabiner, L.R., (Juang 1985)** "A probabilistic distance measure for hidden Markov models", *AT&T Technical Journal,* 64(2), pp.391~408, 1985

[15]**Katz, S., (Katz 1987)** "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoust., Speech, Signal processing*, vol. ASSP-35, pp. 400-401, 1987

[16]**Lee, C.H., Rabiner, L.R., (Lee 1989)** "A frame synchronous network search algorithm for connected word recognition", *Proc. of IEEE,* 77(2), pp.257-285, 1989

[17]**Li, J., Zheng, F., and Wu, W.H., (Li 2000)** "Context-Independent Chinese Initial-Final Acoustic Modeling", *International Symposium on Chinese Spoken Language Processing (ISCSLP'00)*, pp. 23-26, Beijing, 2000

[18]**Ma, B., and Huo, Q., (Ma 2000)** "Benchmark results of triphone-based acoustic modeling on HKU96 and HKU99 Putonghua corpora", *International Symposium on Chinese Spoken Language Processing (ISCSLP'00),* pp.359~362, Beijing, 2000

[19]**Morgan, N., Bourlard, H., （Morgan 1990）**"Continuous Speech Recognition Using Multilayer Perception with Hidden Markov Models", ICASSP, 1990

[20]**牟晓隆，(Mou 1998)** "汉语语音听写机的研究与实现"， *[硕士学位论文]，北京：清华大学计算机科学与技术系，*1998

[21]**Nadas, A., (Nadas 1985)** "On Turing's formula for word probabilities," *IEEE Trans. Acoust., Speech, Signal processing*, vol. ASSP-33, pp.1414-1416, 1985

[22]**Reichl, W. and Chou, W., (Reichl 2000)** "Robust Decision Tree State Tying for Continuous Speech Recognition", *IEEE Trans. Speech and Audio Proc.,* Vol.8, No.5, pp.555-566, 2000.

[23]**Rabiner, L.R., (Rabiner 1989)** "A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition", *IEEE Proceedings,* Vol. 77, No.2, pp.257-286, 1989

[24]**Sadaoki, F., (Sadaoki 1985)** "Digital Speech Processing, Synthesis, and Recognition", *Tokai University Press*, 1985

[25]**Verhasselt, J., Martens, J.P., (Verhasselt 1998)** "Context modeling in hybrid segment-based/neural network recognition systems", *Proceedings of ICASSP*, Vol.1, pp.501-504, 1998

[26]**Vintsjuk, T.K, (Vintsjuk 1968)** "Recognition of Words of Oral Speech by Dynamic Programming", *Kibernetica,* Vol. 81, No. 8, 1968.

[27]**Viterbi, A.J., (Viterbi 1967)** "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Trans. on IT*, 13(2), 1967

[28]**Waibel, A., et al, (Waibel 1989)** "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Trans. On ASSP,* Vol.37, no.12, pp.1888-1898, 1989

[29]**Wilpon, J.G., Lee, C.H., Rabiner, L.R., (Wilpon 1989)** "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech", *ICASSP-89*, Vol.3, pp.254-257, 1989

[30]**Wilpon, J.G., Miller, L.G., Modi, P., (Wilpon 1991)** "Improvements and Applications for Key Word Recognition Using Hidden Markov Modeling Techniques," *ICASSP-91,* pp. 905-908, 1991

[31]**吴宗济，林茂灿，等, (Wu 1989)** *实验语音学教程. 北京：高等教育出版社.* 1989

[32]**Xu, M.X., Zheng, F., Wu, W.H., (Xu 1999)** "A Fast and Effective State Decoding Algorithm", *EuroSpeech'99*, Vol.3, pp.1255-1258, Hungary, 1999

[33]**杨行峻，迟惠生，（Yang 1995）**语音信号数字处理. 北京：电子工业出版社，1995

[34]**Yong, S., Kershaw, D., Odell, J., Ollason, D., et al, (Yong 1999)** "The HTK Book (for HTK Version 2.2)", *Cambridge University,* 1999

[35]**郑方，吴文虎，方棣棠，(Zheng 1996)** "汉语语音听写机中的语音识别基元"，*第四届全国人机语音通讯学术会议（NCMMSC-96）论文集，*pp.32~35, 1996，北京

[36]**郑方，武健，吴文虎，等，（Zheng 1997）** "基于最小分类错误的声学模型间距离度量" *第 3 届全国计算机智能接口与智能应用学术会议 (NCCIIIA). 北京: 电子工业出版社,* pp. 98-103, 1997

[37]**Zheng, F., Mou, X.L., Wu, W.H., et al, (Zheng 1998)** "On the embedded multiple-model scoring scheme for speech recognition", *International Symposium on Chinese Spoken Language Processing (ISCSLP'98)*, ASR-A3, pp49-53, Singapore, 1998

[38]**Zheng, F., Song, Z.J, and Xu, M.X, (Zheng 1999)** "EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine", *EuroSpeech'99,* Vol.2, pp.819-822, Hungary, 1999

致谢

感谢我的导师郑方博士对我的悉心指导。郑老师兢兢业业的工作作风，严谨的科学态度和不断创新的精神是我在今后学习和工作中的榜样。多年来，郑老师为人师表的治学作风，对我的谆谆教诲和严格要求是我成长的动力。在我的研究工作中，方棣棠教授、吴文虎教授、李树青教授和徐明星讲师也给予了我很多详细的指导，帮助我把握了工作的方向。他们渊博的知识和对待科学问题严肃认真的态度给我留下了深刻的印象，这些都是我的宝贵财富。

感谢实验室的所有同学，正是大家所营造的浓厚的学术氛围和轻松的研究环境，是我不断进步的源泉。感谢宋战江、何磊、黄寅飞、燕鹏举、王帆、杨大利、李净、张国亮、罗春华、李芳、陆正中、吴根清等同学，他们的帮助和友谊是我工作不断进步的条件。特别感谢宋战江和李净同学，与他们进行的讨论和合作使我受益匪浅。

感谢所有关心、支持和帮助我的老师，同学和朋友们！

个人简历

张继勇，出生于 1977 年 1 月，1994 年从湖北省黄冈中学保送进入清华大学计算机科学与技术系，1999 年 6 月毕业，获工学学士学位，并于同年继续攻读清华大学计算机科学与技术系计算机应用技术专业工学硕士学位。

发表（已接受）论文

1. Jiyong ZHANG, Fang ZHENG, Mingxing XU, Ditang FANG, **"Semi-Continuous Segmental Probability Modeling for Continuous Speech Recognition"**, *International Conference on Spoken Language Processing (ICSLP'2000)*, Vol.1, pp.278-281, 2000，Beijing
2. Jiyong ZHANG, Fang ZHENG, Mingxing XU, Shuqing LI, **"Intra-Syllable Dependent Phonetic Modeling for Chinese Speech Recognition"**, *International Symposium on Chinese Spoken Language Processing (ISCSLP'2000)*, pp. 73-76, 2000, Beijing
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"**，*软件学报*, Vol. 10, No. 11, pp.1212-1215，1999 年 11 月
4. Jiyong ZHANG, Fang ZHENG, Jing LI, Chunhua LUO, and Guoliang ZHANG, **"Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition"**, EuroSpeech, 2001 (已接收)