第一章 引 言

随着计算机硬软件技术的发展及人们对计算机系统进一步智能化、多媒体化的需求，通过键盘、鼠标等手段的传统人机对话模式已经不能适应人与机器快速、方便、直接和高效的交互要求．自然地人们希望把语音这个语言的载体作为人机通讯的媒介，进一步提高社会信息化和自动化的程度．在人们对勿需任何停顿的不认人无等待听写、通过电话的自动语音信息查询、语音拨号等一大批应用的推动下，加上语音本身在诸如黑暗、障碍、手忙(如在奔驰的车上打电话，飞行员操作)等环境和条件下的不可替代性(可称为以语音为中心的应用,**Speech-Centric Application**)，推动着语音识别理论和技术的飞速发展，在当今社会这个技术孕育着突破．但象涉及人类智能的所有其他问题一样，由于语音信号的复杂性，它也面临着巨大的挑战．而汉语由于其本身输入的特殊性以及汉语语音的单音节和声调结构特点，使得它在技术实现以及市场容量上具有不完全等同于西文的特点，这吸引着国内及国外许多研究者的兴趣和投入，从而大大推动了汉语语音识别技术的发展．本章将对上述这些问题进行深入的分析．

1.1 主流语音识别技术的发展

在过去几年里，语音识别技术取得了巨大的进展，再明显不过的是在大词汇量、连续语音(LVCSR)和非特定人的识别上．目前世界上最先进的实验室系统(非实时)对大多数说话人的词识别错误率已降低到5%-10%的水平．如果作一些说话人自适应的话，则对大多数人来说其错误率可进一步下降到5%以下[69]．这与作为主流技术的HMM在理论、方法和技术上取得的进展功不可没．

HMM首先由IBM的Baker, Jelinek等人借用统计模型(包括声学层面上的HMM和语言处理用的n-元统计模性)的方法用于语音识别而提出，其中Baker用之于描述声学序列[70]，而Jelinek用于描述语言序列[71]．一个典型的HMM如图1.1所示，它有三个输出状态，一个入口状态和一个出口状态．其中出入口状态用于由单元HMM连接成词或句子的复合HMM描述．HMM可以理解为一个有限状态的机器，在每一个时间点t,如果进入状态j，则以概率$b_j(y_t)$产生一个输出y_t．同时从状态i到状态j的跃迁由概率a_{ij}一个控制．

这样经过状态序列X而产生符号序列Y的联合概率为:

$$P(Y,X|M)=a_{x(0)x(1)}\prod_{t=1}^{T}b_{x(t)}(y_t)a_{x(t)x(t+1)}$$

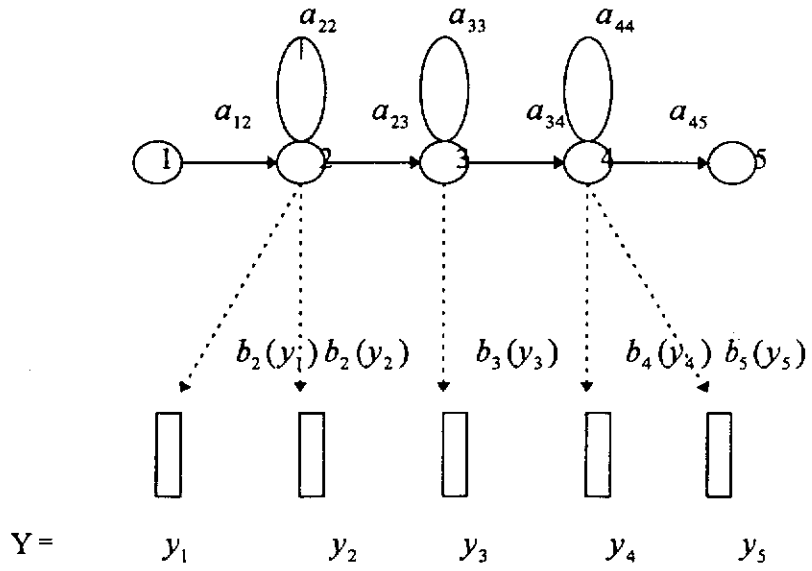


图1. 1：HMM模型示意图

这儿x(0)被约束为模型的入口状态，而x(T+1)被约束为模型的出口状态．在实际过程中，由于我们只知道模型的输出序列Y，而不知道模型的状态转移过程，因而该模型称为隐马尔可夫模型．对于HMM模型用于语音识别，已经有非常好的算法解决模型的训练、评估及状态序列分割等三个问题[26],如Baum-Welch算法，Forward-Backward算法以及Viterbi算法等等．

在HMM模型中，输出分布函数反映了说话人本身和说话人之间的语音信号谱的本质变化，对建模准确度最为重要．离散密度HMM(HMM/VQ) 采用离散的概率分布配上一个矢量量化器(VQ)．由于这种方法计算输出概率只需查表，因而计算量很小，非常高效．但VQ的引入带来的量化噪声，严重影响了建模的精度．在目前计算机计算能力迅速提高的条件下，目前采用较多的为连续密度函数分布的HMM[79,80]．在连续密度分布函数的选择上，B.Juang，L.Rabinar等在保证训练过程收敛的条件下，将连续密度HMM中概率密度函数的约束由对数凹(Logarithmic Concavity)扩展到椭圆对称(Elliptical Symmetry)，进而提出了基于混合Gaussian分布概率密度的HMM模型(CDHMM)[24]，大大扩充了函数选择的自由度．P.Brown,L.Bahl等深入分析了语音识别中基于最大似然准则ML(Maximum Likelihood)训练HMM模型参数理论上的不足，提出了基于最大互信息MMI(Maximum Mutual Information)准则训练HMM模性的新算法[77]．与此同时，许多学者也相继提出了HMM模型训练的新方法，如最小区分信息MDI(Minimum Discriminative Information)，最小误识率MEE(Minimum Emperior Errror)准则，该准则又可称为区分训练(Discriminative Training)[80,81],用于话者自适应的最大后验概率MAP(Maximum A Posterior)准则等[76]．Gapall和Maiarii等通过适当的变换，分别设计了MEE，MDI，MAP和MMI准则下HMM模型的Forward-Backward迭代重估算法．在控制HMM模型计算量方面，日本学者提出了特征参数离散化，

通过引入状态方程描述HMM模型结构，以降低Forward-Backward重估公式中的冗余计算等方法.

HMM理论的发展及计算机硬软件的迅猛发展，使得语音识别系统的研制取得了成功，这大大提高了它在公众中的知名度，反过来也推动了该技术的进一步完善. 美国在语音识别领域取得了令世人瞩目的成就. 1987年BBN公司开发了BYBLOS连续语音识别系统. 八十年代末期CMU推出了SPHINX系统,被认为是语音识别的里程碑[68]. 90年代初，ARPA计划推出了“Wall Street Journal”和“航空订票系统口语理解”(ATIS)两个复杂任务，又掀起了基于大规模语音数据库的上下文有关声学模型的建立和基于大量知识最优的复杂搜索识别算法的研究和实现，并取得了重大进展. 目前已在实验室成功研制了以听写机为应用背景的65,000词书面语连续语音识别系统，以及以民航信息查询为背景的口语识别系统，即ATIS系统. 在德国，法国等欧洲国家，语音识别也得到了足够的重视，建造了许多有特色的连续语音识别系统，如德国的PHILIPS多语种大词汇量连续语音识别系统和法国LIMSI的GAUVIAN等开发的连续语音听写系统[75]，英国剑桥大学的HTK系统等[82].

语音识别上另一个重要的方法为人工神经网络(NN). 人工神经网络的研究始于本世纪四十年. 80年代初，随着Hopfield提出了联想记忆模型，人们开始了对基于神经网络的语音识别的重视. 80年代末90年代初以来，针对语音信号的时序、动态的特点，先后提出了TDNN、预测网络等模型[103]. 近年来，NN技术和HMM技术相结合的研究更成为这一领域的热点[131]. 虽然目前基于NN的系统还无法在识别率上与基于HMM的系统相比，但随着研究的深入，特别是NN和HMM的取长补短，其高度平行性和高区分度的特点是会在语音识别中有所作为的.

此外，在该期间听觉模性的研究得到了进一步的深入. 听觉感知机理的建模不仅只限于用简单的物理过程进行模拟或用基本的信号处理方法实现，神经网络的出现为更加准确地描述人耳的生理结构，感知特性提供了可能. 从实用角度出发研究的声学通道效应、抗背景噪音能力、话者特征、语言模性以及语音特征分析等方面都得到了更深入的研究.

1.2 语音识别对现有技术的挑战

人类在长期的社会实践和复杂环境中逐步建立起了完善的语言听觉功能. 基于HMM主流技术的语音识别虽然取得了重大进展，但比起人类的听觉能力来，尚相差甚远. 在自由交谈中，我们能在不断切换的主题下交流概念和思想，用所熟悉的单词按一定的次序组成自然语言；可以把一个词拆成几个子单元(Sub-unit)，如字，用不同的方式不同的词、语句来表达相同的概念. 我们能领会南腔北调，能在男女老少不同人发出的语音中抽取共同的语义信息；我们还能从语音

中感觉出喜悦、哀伤、悲愤等各种感情，能在嘈杂的环境中将注意力集中在一个人说话上．在自然语流中，还充斥着“啊”、“喔”等大量的中间停顿和重复的不完整句子，但是我们能够在瞬时把握抓住语义特征，并辨认出说话人和其说话风格[104].

这就是自动语音识别(ASR)和理解(Understanding)所要解决的问题.除了目前对人耳定位机理、感情及风格等因素研究较少外，语音识别和理解的研究者们对上述大多数现象已有足够的认识和了解，特别是在对各种知识的利用上的认识．这些知识有语音学和语言学的．其中语音学知识有音位学，发音规则及韵律规则等；语言学知识包括词汇学、句法、语义和语用等．但困难在于，面对着如此错综复杂的可变的数据来完成归类和识别这些单元并在各个层次上运用知识，则是个非常棘手的问题.其可变性具体体现在语音层面上的离散性和语言层面上的模糊性．由于每个人的声道解剖结构上的差异，社会语言学的背景不同，声学环境的多样性(背景，Microphone,A/D通道等)，生理环境的迁移和音素受上下文的影响，这些因素的离散性带来了声学建模的困难；语言的的歧义性和语言结构的随意性在日常语言中随处可见.而句子结构的随意性即词序的随机性在汉语中表现的尤为突出，如“做难人”，“做人难”，“难做人”都是合理的组合，且它们的意思也非常的接近.这种模糊性使得我们无法用简练的框架来表达这些知识.

正是由于语音、语言现象的复杂性，上述知识的获取，知识的表示以及根据特定任务利用和组织这些知识的策略成为我们面临的最大挑战．事实上,只有建立从声学、语音学到语言学的知识为基础的语音处理机制，才有可能获得能与人相比的高性能的计算机语音识别系统.这是个不可避免且必须要解决的问题[105].

从现有的大量研究成果也可定量地说明上述问题的严重性及复杂性.语音识别中最简单的系统为认人的．在安静环境下，有一定语法限制的该类小词表的孤立词识别系统，目前其误识率可做到小于0.5%．在未加人任何新的语音学、语言学知识的条件下，当我们由特定人识别转为非特定人识别时，其误识率可提高2－3倍，即为1％－1.5%；当发音方式由孤立词转为连续语音时，其误识率可提高3倍，即为3％－4.5%；当识别词表有数量级的提高而变为大词表的识别时，其又可有3倍的误识率提高，即为9%-13.5%；而识别声学环境由安静转为变为嘈杂时，又可有3倍的错误提升，此时误识率就可达27%-40.5%．进一步放松对说话方式的限制，即当待识别的语音由自然规范的语音转为无限制的自然语言或口语时，其识别率又可有4倍甚至更多的提升[17]．此时我们可以想象为何目前的识别系统还无法做到类似于人类的处理能力．对于一个识别率低于50%的系统是无法实用化的.

下面是作者的一组实验例子：一个定量地表明了误识率与词汇量的关系实验．实验为某条件下的特定人孤立词识别系统，系统的最大词汇量为46500，采用本章第五章的搜索算法．结果见表1.1 .

从表1.1 可以发现，在起始词汇量增加时，识别率下降比较快，而到了一定程度后，识别率下降速度变得稍微慢一些，因为其混淆度逐步趋向饱和．在这基础上改进后的系统中，35000词的识别率达到88.7%．采用同样的识别方法，当作者的研究由特定人转为非特定人时，其识别率由88.7％下降到了65％，误识率大致是原来的3倍．这也从一个侧面印证了国外对于识别难度的这方面估算．

表1.1识别率和词汇量的关系

词汇量	0	5000	10000	20000	35000	46500
识别率	100％	91.2％	89.3%	84.8%	82.6%	81.9%

防止这种识别率滑坡的唯一解决方法就是需要我们加入更加多的知识源，**并加以合理的利用**．如动态差分特征，新的特征维，更加精确的声学模型，更加高级的语言知识，韵律知识等等．这就是目前语音识别技术所孜孜追求的．

1.3 国际语音识别研究的最新动态

随着语音识别在实验室取得的巨大成功和产业界的加入，当前国际语音识别的研究出现了一些新的动向．**在商用化方面，首先是强调与实际以语音为中心的应用和软件环境的结合**，以便充分体现语音作为一种交互手段的快捷、方便和可靠的特性．这也使它所要解决的问题更具有针对性，更方便地利用特定环境底下的许多语法约束和任务约束来提高准确率，如基于Agent环境的语音操作界面等等．**第二是语音识别的算法标准化问题日益得到重视**，语音识别的API(应用程序接口)也已出现．这些API的出现使得有特殊应用需要的用户通过调用API可以方便地构建自由的语音识别应用系统，这将大大加快语音识别应用步伐．

商用化应用步伐的加快，又对语音识别的研究方面提出了许多更加迫切的需求．**首先是说话人的自适应问题**．说话人无关的系统无疑是最理想的系统,但从目前的技术来讲，其代价是相对高的错误率．对许多应用来说，说话者往往成为一个固定的用户或者可以在使用前输入一定数量的语音．在这种情况下，通过自适应手段来改善声学模型与说话者的匹配情况从而大幅度地降低错误率对未来的系统是非常必要的．**第二是针对应用需求的语音识别的鲁棒性研究得到进一步重视，特别是信号的预处理技术**．鲁棒性包括Microphone无关技术，强噪音底下的识别，远距离Microphone语音输入，背景噪声、音乐与语音信号的分离，对集外命令或词汇的拒识，稳定的识别率等．这些方向值得我们重视，特别是在一定环境下语音识别的应用问题．**第三是任务无关的语音识别**．目前的声学模型相对来说是与任务无关的，但由于统计语言模型往往从任务领域有关的特定语料中训练产生，使得系统的性能与任务关系非常密切．例如一个从办公室信件训练出

来的语言模型不能很好的用在法律文本上. **从长远的观点看，任务无关只能通过语言模型本身描述能力的的增强来完成.** 例如统计方法与语法规则知识的结合等等；从近期观点来看，可采用多领域混合建模的静态语言模型(LM)同动态改变的N-元模型相结合的方式来得到一定程度的解决. 另外的研究热点还包括口语语音的识别、识别系统的实时响应等等.

1.4 汉语语音识别的特点及研究现状

目前的语音识别方法或多或少地跟语种有关. 汉语由字构成词，由词构成句子，所以汉语中的最小单元就是字，也称为音节字. 音节在传统上可划分为声母和韵母两部分(用C和V表示). 声母共有22个(包括一个零声母)，多为清音，浊音只有m、n、l、r四个；韵母共有三十七个，由一到三个元音组成. 22个声母和37个韵母总共构成了汉语约400个无调音节、这400个左右音节再加上五个音调(阴平，阳平，上声，去声，轻声)构成了汉语的1300个左右有调音节. 这1300个有调音节进而可组成成千上万个常用汉字词组,并进一步构成连续语音(C1V1C2V2...CnVn). [86,91,97]. 在汉语中，没有明确的词的定义，词与词之间没有空格.

正是由于汉语的单音节特点，国内90年前的语音识别重点在汉语的孤立音节识别上. 在此期间，清华大学、中科院自动化所、声学所、四达公司、北方交通大学、西安电子工程大学等单位结合汉语语音学的特点，在基础理论、模型和算法、实用系统方面作了大量的工作. 期间取得了一批重大成果[108,109]，发展出了一批基本技术. 台湾在汉语语音识别方面也取得了较高的水平，其中以李琳山教授主持的研究小组最为出色，开发成功了“金钟一号”(Golden Mandarin(1))[110].

理论上，全部孤立音节识别问题的解决就可完成汉语无限词汇量的语音识别,但是由于单音节输入不仅效率低，而且语音输入的自然度极差，所以在市场上并未取得成功(如四达公司和星河公司的产品). 同时受语音数据库尤其是计算机设备等诸多条件的约束，研究重点一直停留在特定人的语音识别上. 由于特定人数据库规模的有限，进而引起一些方法上的局限，造成系统的性能有“运动员水平”和“群众”水平之分[98].

从90年起，自动化所开始研究大词汇量的孤立词识别和中等词汇量的连续语音识别，同时清华大学等单位在非特定人识别研究方面也取得进展. 在“863”计划和科学院“85”重大项目“汉语人机语音对话系统工程”课题的的引导下，在短短的几年内我国人机语音通讯的研究从组织、方法、系统乃至立题等方面全面进展. 94年起开始全面转向大词汇量孤立词识别、连续语音识别和非特定人识别. 在95年底举行的评测中，无论是孤立词的还是连续语音的非特定人听写系统都取得了比较理想的结果. 96年国内有关单位在原有工作的基础上，再接再厉，

分别设计采集了公用的多说话人连续语音数据库，建立了基于大规模语料统计的统一的语言模型，为下一步的工作打下了结实的基础．同时台湾李琳山教授小组研制的“金钟一号”也逐步发展成为能识别连续语音的特定人听写系统[36]．

由于汉字输入的特殊性和中国潜在的巨大市场，国外大公司和研究语音识别的专业公司已经开始进军中文语音识别的研究．他们利用在西文语音识别方面积累的技术和经验，对建模单元和声调处理作一些修正后移植于汉语语音识别，取得了重大进展，例如 IBM,Motorola，APPLE和Dragon等公司．这一方面将大大推动汉语语音识别的研究，另一方面也给国内的研究工作提出了更大的挑战．如何做到与国际研究的接轨，并进而创新，**如何确立和进行有一定超前度的目标和研究**，成为摆在我们面前的一个问题．

1.5 本论文研究内容及安排

汉语非特定人、大词汇量的语音识别从直接意义上可用于汉字的文字录入，另一方面，一旦拥有这个技术，将打开一系列语音识别应用的大门．作为语音识别中的难点，该问题研究在国内应该说刚刚起步．本论文就以该研究为目标，重点围绕建立一个高性能的听写机系统（包括孤立词及连续语音）所需的技术展开，特别是针对汉语语音特点的声学模型的建立、搜索识别算法以及最后系统的集成．论文试图反映作者的部分研究成果，同时也阐述对语音识别的一些看法及体会．第一章将对主流语音识别技术的发展、面临的挑战及努力方向、汉语语音识别的现状作一简述；第二章将从音节困惑度和搜索复杂度出发，深入地分析各种知识源对音节困惑度的影响和词库的组织形式，从而为识别特别是各种知识的递进式使用提供指导．

第三章在非特定人数据库的基础上针对汉语语音的特点研究比较了汉语的建模单元，音节间上下文有关、音节内上下文有关、声调有关、端点有关等声学模型，并对连续密度HMM的一些训练要点及增强其建模能力的方法和途径进行了详尽的研究和实验；在第三章研究的基础上，第四章中分析研究了基于聚类和基于语音学知识和数据驱动相结合的决策树的声学模型建立过程，为包含声调信息的汉语声学模型的建立提供一种可靠的、自动的解决方案．

第五章开始转向孤立词搜索算法的研究．在对各种算法作分析综述后，提出了适合汉语单音节特点的高效启发式识别搜索算法，并在搜索的精度及效率上作了详尽的分析；在基于前几章核心模型和算法的基础上，第六章将对我们建立的非特定人、孤立词汉语听写机的诸模块和系统集成作一介绍．包括信号前端处理、训练策略、语言模型的建立及后处理、系统的实时实现及结构等，并给出当时的的最后文字转换率．

第七章将着重分析连续语音识别的搜索算法．包括m元统计模型条件下搜索最优化的问题，搜索空间的动态扩展方法和汉语大量同音词在搜索中的处理方法．在这基础上，给出了我们建立的一个连续语音识别系统的初步结果．

第二章 语音识别困惑度、复杂度及知识的组织

语音识别是一个利用多种知识从语音串出发来搜索出最优的一个句子或词的过程. 这些知识包括声学知识、语言知识、词库约束等等. 然而如何利用这些知识的策略是不同的. 最优的一种策略无疑是利用所有知识, 一次性地搜索出结果出来, 但由于许多的知识是全局的或者说是长距离关联的, 这会产生巨大的计算量. 分阶段搜索是解决这个问题最有效的手段. 但如何分阶段使用知识源最为合理就构成了语音识别的框架问题. 在这一章中我们将对各种策略作定量和定性的分析. 这涉及到两个问题: 语音识别的困惑度、复杂度和有效的知识组织形式. 它们构成了搜索识别策略的基础.

我们首先将介绍困惑度的概念, 然后针对语音识别的搜索空间的大小明确了复杂度的定性概念, 指出了它们之间存在的一种反比关系. 针对汉语的特点, 从语音识别的角度出发, 我们提出了音节的困惑度指标, 用于衡量不同识别框架下音节的识别难度及混淆度. 采用这个指标, 我们比较了不同识别框架下音节困惑度的大小, 从而比较各种思路的优劣. 我们还从复杂度角度出发讨论了词表和平滑后的二元统计模型的组织形式及知识的分布式表示. 这些讨论使我们对汉语语音尤其是连续语音的识别有了相对比较完整的思路.

2.1 语音识别的困惑度(Perplexity)

由于语音研究的需要, 人们需要构造许多任务. 为了比较这些不同任务之间系统的性能, 需要一种对任务难度的客观描述. 虽然词汇量大小在表述一个任务时经常被提到, 但它本身对描述一个任务的难度却无能为力. IBM的科学家Jelinek等基于信息论中的熵概念在1983年提出了困惑度的概念(Perplexity)[1]. 假设句子H由词串W组成, 即

$$H = w_1 w_2 \ldots w_n$$

该句子可由一个有限状态网络产生, 网络中每一状态节点代表一词. 假设一个状态节点s产生下一个词w的概率为P(w|s), 则对状态s来说其熵Hs(w)为

$$H_s(w) = -\sum_w P(w|s)\log_2 P(w|s) \qquad (2.1)$$

对一个任务来说，其信息熵H(w)即为所有状态信息熵Hs(w)的平均值。假如π_s为在产生句子过程中在状态S的概率，则:

$$H(w)=\sum_s \pi_s(s)H_s(w) \qquad (2.2)$$

则其困惑度可定义为

$$S(w)=2^{H(w)}$$

显然任务的困惑度取决于所有可能出现的句子的概率。对于人工构造的任务，可能产生的句子经常没有赋予概率。Shannon证明了对于一个可能产生N个句子的任务来说，如果其句子的平均长度为l，则不管每个句子出现的概率为多少，其最大熵为$\frac{1}{l}\log_2 N$,所以其困惑度为$N^{1/l}$[111]。设想把所有的句子安排成一颗树，则从每一分支出来的词的平均个数为$N^{1/l}$。因而困惑度可以直观地理解为在某一点后接可能的词个数。

Jelinek根据上述熵的定义，对当时的一些连续语音识别任务进行了比较，并根据实验结果证明了系统误识率与困惑度之间有着直接的关系[1]。见表2.1：

表2.1 错误率与困惑度的关系

词 汇 量			词 错 误 率	
任务	词表	困惑度	识别方法1	识别方法2
CMU-AIX05	1011	4.53	0.8%	0.1%
Raleigh	250	7.27	3.1%	0.6%
Laser	1000	24.13	33.1%	8.9%

从上表可以看到，CMU－AIX05的词汇量虽然比Laser要大，但其误识率则要比后者小得多；Raleigh的词汇量仅是CMU－AIX05的四分之一，但其误识率却要比后者大得多。而从表中却明显地发现了困惑度与误识率之间的正比关系，即困惑度越小，误识率也就越低。这从直观上也是非常容易理解的。

以上的困惑度定义未涉及到词表的音素构成特性。两个任务可以拥有同样的困惑度但如果一个的词表的词汇较长，则该个的识别显然更加容易，这个事实也将在我们下面的讨论中得到证明。此时准确衡量识别难度需要从更基本的单元出发讨论。例如，可以认为句子由音素构成而非词构成来讨论音素的困惑度。同样地在音素级上两个困惑度完全一样的任务，可能由于一个任务的音素之间的混淆度较大，其困惑度显然也较大，此时我们就需要估算条件熵H(w|y)了。对于大词汇量汉语语音识别来说，其基本音节是公用且相同的，**因而我们的讨论在音节层面上衡量就足够了。**

而语音识别中另外一个无法正式定义的概念－复杂度则与所完成任务的搜索空间大小相对应. 我们可以粗略地把要描述任务的各种语言及声学的约束关系用有限状态图表示后的状态数称为计算复杂度[112]，但它还跟其他因素有关系. 困惑度定义了对于某个识别任务或识别阶段的识别难度，而复杂度则定性地描述了识别任务在满足约束条件下搜索出最优的句子所要消耗的计算量.

虽然困惑度和复杂度表示了两个不同的概念，它们之间却存在着直接的关系. 例如，它们都同词汇量、语言知识及知识表示都有关联. 当一个识别任务其语言知识越丰富，则每一词后接的词的个数越确定，而要表示这些词之间的关系就要越多，复杂度也就越大，因而它们之间存在着一种反比关系. 另外非常重要的一点是，计算复杂度还与采用的声学模型和搜索策略、裁剪程度有关. 例如：把音节安排成树状结构，在音节层面上作搜索时，如果不考虑音节间的上下文有关声学模型和任何语言约束，则只需约500个节点和500个连接，如图2.1表示；当我们考虑音节间的上下文有关声学模型时，假如声母根据前面韵尾不同分成两类时，此时图2.1中节点a就要根据韵尾不同区分成a1和a2两个节点，节点数和连接数随之增大到1000个左右. 同样地当我们在词层面上搜索时，仅仅考虑考虑词内上下文有关模型(Intra-Word)和同时考虑词间的上下文有关模型(Cross-word Context dependent)时其复杂度也大不一样. 下表列出了同困惑度和复杂度有关一些因素，这些因素是我们在设计语音识别算法尤其是系统时必须考虑和权衡的.

表2.2 同困惑度和复杂度有关的要素表

困惑度	词汇量，语言知识，知识表示等
复杂度	词汇量，声学模型，语言知识，知识表示，搜索策略

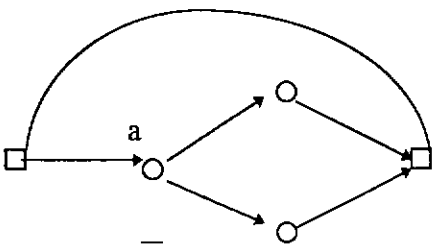


图2.1：音节内CD模型

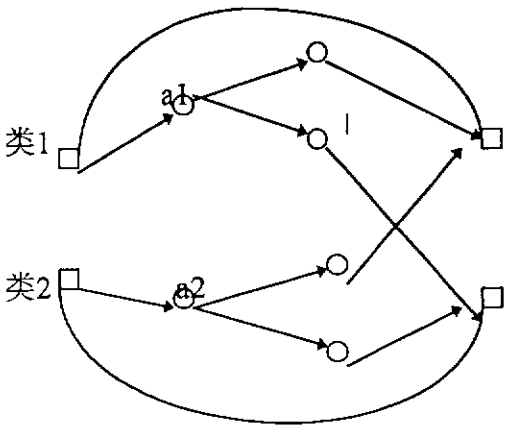


图2.2 右上下文分成两类时音节间CD模型

2.2 汉语音节的困惑度和计算复杂度

上面的概念描述中，并不涉及具体的识别策略．不难看出，困惑度和复杂度实际上只说明了在运用某种知识条件底下的任务难度．但正如上面指出的，由于运算的复杂度的关系，往往需要分阶段逐步递进运用知识．例如在大词汇量识别中，可以把它分成两个阶段进行识别，第一阶段只输出音素序列及候选，第二阶段在音素序列基础上再组合词．在这种情况下，我们会非常关心第一阶段音素的第一候选识别率和前几名包含率．而第一阶段音素的识别率显然同第一阶段的音素困惑度有直接关系．所以从研究该搜索阶段减少误识率的角度出发，我们需要对每一步识别目标的困惑度作大致的估算．**如果某一步困惑度太高如不适度加进一定的知识以减少困惑度，则会造成该阶段识别率损失，最终造成不可恢复的错误．**识别框架正是需要在困惑度和复杂度之间作一平衡．

我们从汉语语音识别角度出发，更加具体地来考察这个问题．由于汉语的单音节结构及词定义的不确定性，特别从计算复杂度出发，人们很自然地可以只在音节层面上进行声学匹配．在此层面上，我们可以利用音节的一元、二元乃至三元的音节连接模型．显然没有任何约束的音节搜索其不确定性非常之大，不考虑声调的话，困惑度即为全音节个数，而加上音节之间的二元连接模型其困惑度会有一定的降低．从另一角度出发，虽然汉语在词的定义上存在一定的模糊性，但理论上说，所有连续语音皆可由有限的词汇连接而成，即词库外词可由词库内的单字词或与双字词连接而成，自然地人们也希望能在词一级进行声学匹配，因为这毕竟能省去组词一步．在词层面上的声学匹配可以利用的知识包括词的一元、二元乃至三元模型，这样的模型显然也将大大降低识别混淆度．纵向上音节的二元连接模型到底能降低多少混淆度，横向上是纯粹的词组一元约束搜索有效还是二元音节搜索有效是我们非常关心的问题．这些都是我们非常关心的问题．

为此我们提出**汉语音节困惑度**的概念．对于所有识别框架，必须从音节困惑度角度出发，再辅以搜索复杂度加以综合考虑和平衡．从音节层面考虑困惑度有两个好处：第一是所有框架都要用到全部音节的识别，这使得我们可以抛开音节间的混淆度；第二是从音节出发就是强调音的识别准确度．没有底层的高识别率，任何高层语言模型都是无济于事的．由于要考虑声学层面匹配的困惑度，因而我们假设句子是由一串音节组成而非由词组成，即：

$$H = S_1 S_2 \ldots S_n \qquad (2.3)$$

从93年全年<<人民日报>>统计中得到了音节的使用频度，二元连接概率，32000词的使用频度，二元连接概率和三元连接概率．由于计算量的原因，音节的三元知识和词的三元知识在目前还无法直接加入到声学匹配中去，因而我们主要考察音节和词的一元、二元模型下的音节困惑度．

2.2.1 音节一元连接

当不考虑音节的使用频率时，汉语中共有407个无调音节，后面所接音节以等概率连接，无调音节使用频率也相同，这样对所有的音节节点都有:

$$H_s(w)=-\log_2(1/407)*1/40\ ,$$

平均信息熵H(w)为$-\log_2(1/407)$,所以其困惑度为407; 显然在没有任何语言模型的条件下，从上式计算出来的困惑度与我们的直观完全一致．从复杂度来说，表示这个音节自由连接语法只需要407个节点就可．当考虑了音节的使用频度时，可以得到困惑度为165.6左右．

2.2.2 音节二元连接

此时我们在计算过程中，信息熵式中的p(w|s)即为在前音节确定条件下音节的二元连接概率，而π_s即为音节的使用频率．根据困惑度定义我们只需要用音节的使用频率及音节之间的二元连接概率．用程序统计计算，困惑度结果为86.15．同2.2.1相比音节的困惑度减少了将近1/2.这说明了音节二元连接概率是极其有效的．我们将在第七章说明，在搜索出二元连接概率最大的句子的约束条件下，它也只需要用407个节点来表示，也就是说其计算复杂度无任何增加．但当需要产生N-Best结果时，其复杂度要增加约5-10倍．

2.2.3 词一元连接

此时在式2.1中，我们可以把音节串看作一串词中相应的音节组成．在词的内部，每一个词的非尾音节其后接音节是唯一的，因而其不确定性为零．考察尾音节，其后接音节节点数可能性为全音节的个数407，因而

$$H_s(w) = -\log_2 1/407$$

而最后一个节点在句子中的出现概率为该词的出现概率除以词的音节长度，也就是:

$$\pi_i = p(w_i)/n_i,$$

因而熵为:

$$H(w)=-\log_2 1/407*\sum p(w_i)/n_i$$

从上式可以发现，若所有的词皆为一字词，则其熵值同全音节一样,困惑度为407;若所有词为二字词，则熵值为原来的一半，而困惑度则为20.17．长词越多，使用频率越高，则其困惑度越小.

在93年人民日报实际统计出的词表中，我们只搜集到四字词为止，五字词以上词认为可有其它基本词汇组合而成．具体个数分布如下表；经计算后 $\sum P(w_i)/n_i$ 值为0.72，熵值为6.24，困惑度则为75.70.

表2.3 词汇音节数分布表

分布/词长	一字词	二字词	三字词	四字词
个数	4236	20915	3976	2125

我们从上表可以看到，由于绝大多数的高频词为单字词和双字词，所以按照词汇自由连接方式其音节困惑度同音节二元连接相比减少了约12％．如果仅从状态空间来说，其复杂度则大大增加了(10万个节点)．但在实际搜索过程中，由于BEAM搜索的裁剪，后面的许多层节点根本勿需遍历，因而搜索量还是很小的．经我们研究观察，95％的搜索时间消耗在头两层节点上，而前二层节点的个数与音节二元连接相当.

2.2.4 词的二元连接

下面让我们来考察一下按照词的二元连接时音节的困惑度．此时词的最后一个音节可能后接音节的概率由词的Bigram根据后接词的头音节换算得到．具体地假设前一词的最后一个音节表示为s,后面可接m个词，后接词w的头音节由Head(w)表示，则

$$P(S_j|S)=\sum_{w}P(W|S,Head(W)=S_j)$$

由此估算 $H_s(w)$，并进一步估算熵和困惑度(同上).根据实际计算，在词的二元连接模型下，音节熵值为2.58,其困惑度则降低到6.01，大大低于词的自由连接模型.

需要说明的是，上述结果是在未对二元模型作任何平滑的条件下取得的，也就是说没有在训练语料中出现的连接置为零．加上平滑后，困惑度会略有上升，但现有数据已充分说明了二元模型对减少音节混淆度的有效性.

通过分析可以看出，困惑度为6.01时相当于词的一元连接模型中平均词长为3.35时的情形，这事实上也同绝大多数是单字词和双字词的二元连接其平均前后距离字长约为3－4一致.

2.3 汉语语音识别中词典的作用

表2.4列出了上述分析的一些结果。从中可以看出，词的一元连接的困惑度低于音节二元连接的困惑度。这意味着词典信息是一种非常有效的信息。从声学层面上看，即使在未外加任何统计的情形下，它的熵值还比音节的二元连接所包含的信息大；从语言分析看，文献[87]从汉字熵的角度研究表明基于词的模型明显地比基于字的模型约束力强。特别是词的一元模型比字的二元模型熵更大，同本文在声学层面的分析结果一致。这充分表明了在汉语语音识别中词典信息极为关键，构词中含有大量的语言约束信息。

表2.4: 不同语言约束下的音节困惑度

语言约束	音节一元连接	音节二元连接	词表一元连接	词表二元连接
音节困惑度	165 · 61	86 · 5	75 · 7	6 · 01

因而本论文将研究基于词汇级的声学层面匹配以及在连续语音识别中与语言模型的融合问题。当然这会带来一定的计算问题，但通过精心组织和有效的搜索策略，我们希望找到一条有比较高的识别精度，其计算量又在一定限度内的算法。比如在第五章孤立词识别中，我们可以首先在音节连接上进行预搜索，然后再在词层面上进行声学匹配做到二者兼得。

2.4 词库和二元知识的组织方式

在小词汇量包括全音节识别中，我们不需担心词表的组织问题。但在大词汇量语音识别中，词库的组织方式对搜索效率极为关键。另外从上面的分析可以看到，在汉语中声学层面的搜索在词汇一级进行对减少音节的混淆度提高识别率至关重要。但在这上要加入词库二元连接概率，搜索会变的异常复杂。此时，其连接概率的表示方式对搜索效率也至关重要了。

2.4.1 两种词表组织形式

在插值二元语言模型中，概率可以表示成为:

$$\Pr(z|y) = \Pr^*(z|y) + \lambda(y)\Pr(z),$$

其中$0 \le \lambda(y) \le 1$，经过上述平滑后，显然任何两词之间都有概率连接，这样一种最直接的连接就是任何两词之间产生一种连接，假设词汇量为V,则共需要V^2个连接来表示二元概率.

Placeway等人[113]在1993年针对上述式子提出了一种更为有效的表示方法，该方法用词之间的直接的连接来表示在训练语料中观测到的连接，而对每一个词用空节点来表示在训练语料中未出现的概率连接估算. 如下图2.3：

在图2.3中，(x,z)在训练文本中未被观测到，因而只需要通过空节点把平滑部分的概率加上即可. 而(y,z)在训练语料中被观测到，则需要两个连接. 假设我们在训练语料中共观测到d个二元连接，系统的词汇量为V，则这种表示共需要d+2*V个连接. 当大词汇量(例如V>1000)时，这个数目比起全连接V^2要小得多.

但是我们应该注意到，上述这种表示法没有把词之间的声学相似性考虑在内. 在大词汇量中，许多词它们的头几个音素事实上是相同的，可以用树的形式把这些共同部分表示出来.

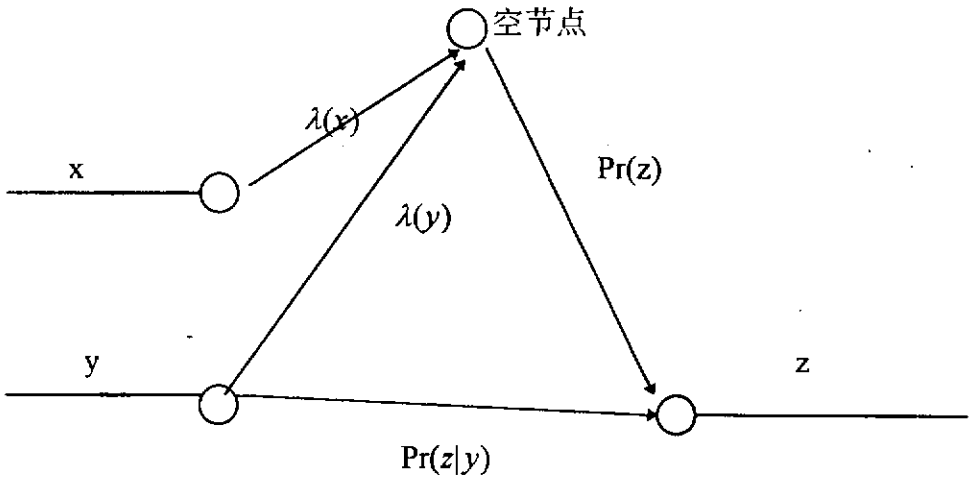


图2. 3: 平滑二元模型表示

Philips实验室的研究人员报道了采用Viterbi-Beam搜索时树状表示的优越性. 它们证实，当词汇用线性组织时，在搜索过程中95%的路径假设发生在头两个音素中，这使得树状表示非常具有吸引力[114]. 同时，词的树状表示能大大节省空间. 例如，对于10000词的听写系统来说，其所需连接数之比是2.69. 然而，不同于线性表示，在树状表示中词的标号(即哪个词)只有到了叶节点才能知道. 这样当二元模型知识加入时，在每一个激活的词的叶节点需要扩展一颗树，然后在扩展的树到达叶节点时再加入语言知识. 这一方面迟缓了语言知识的加入，可能导致多的错误，另一方面使得需要非常大的空间. 但其本身的高效率特别是在头两层节点上的大量裁剪表示抵销了这两个消极的影响.

笔者在1992年始独立地采用了树状的词库组织方式，大大提高了46500词的搜索效率，并一直沿用至今．但此时的工作只是在孤立词搜索中，并且未考虑二元知识的应用．

2.4.2 树状静态和动态扩展

孤立词的树状搜索空间构造相对比较容易[8]．但在连续语音中需要作树的扩展．有二种方式把连续语音的搜索空间组织成树状模式，一种可以事先静态构造，一种在搜索过程中动态扩展．动态扩展过程我们将在第七章搜索中介绍．这儿只介绍其静态构造过程．不管怎样，静态构造由于在搜索过程中不需额外的扩展消耗，同时有可能作离线的树空间压缩，因而具有很大的吸引力．

一种新颖的有效的压缩表示方法如上所示，其基本概念也是加入空节点．对于每一个词来说，其所有后接的词若在训练集中出现，则只需将其安排成一颗树，如果词y是词x的后接词，则将它们的概率Pr(y|x)赋予词x的后接树中词y的叶节点与词y的后接树的根节点上．空节点为全词树的根节点．全词树的叶节点对应所有的词汇，它们至其后接树的连接概率为它们的词频,而所有后接树的根节点皆联一连接至全树的根节点，即空节点，其连接概率为$\lambda(y)$.在每一颗树中，当有两

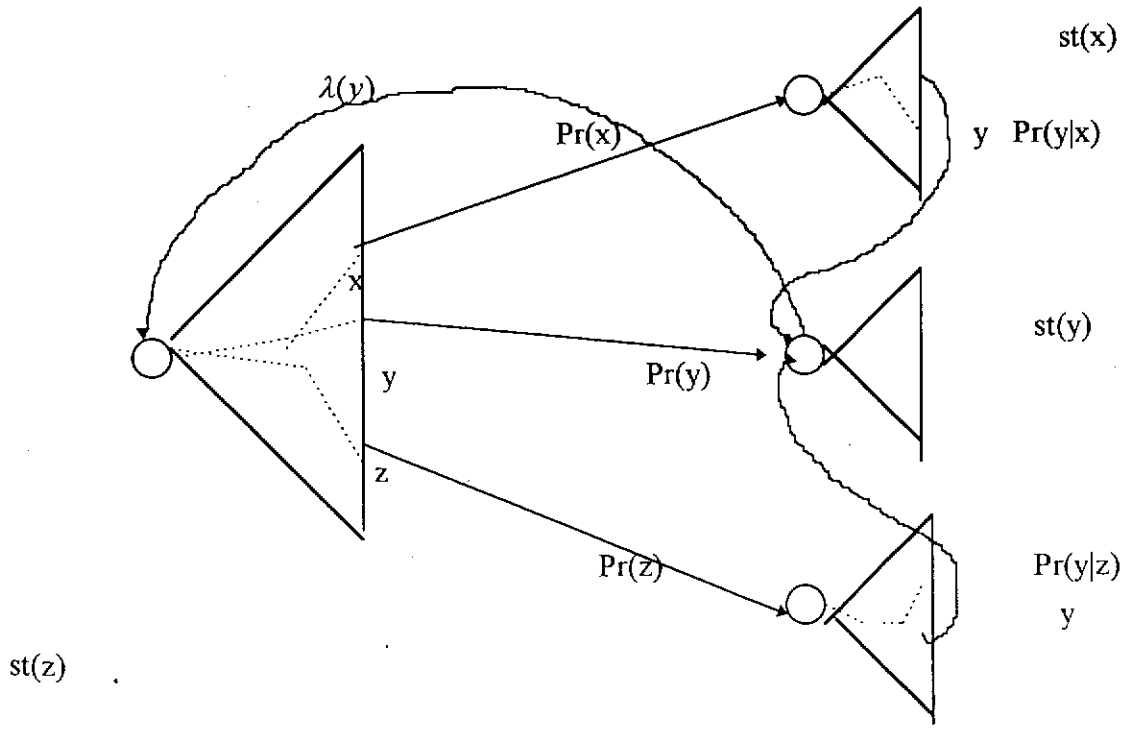


图2.4 静态二元模型搜索图的构造

词它们的前部分路径完全相同时，我们必须把它们分开，以保证每词对应一个叶节点。在上图中，左面的大树表示全词树，而后面的小三角形表示某词的后接树，用(st(.))来表示。

用上述各个独立的后接树而不用V个全词重复来静态表示词的二元连接非常有效。实际上，所有后接树的大小依赖于观测到的Bigram的数目，这往往大大小于整词表的大小。

但这个静态构造的树空间还是超出了我们现有计算机的能力。以我们的32000词识别为例，整树需要大约9万个节点。我们收集到了约150万个连接，也就是平均每个词后接可有50个词，大约需要60个节点。这样共需要180万个节点。按每节点需要8个字节计算，共需120M内存。尽管静态构造的网络可以用一定的算法加以进一步的消减，以减少搜索量，但其压缩量还是比较有限的。

基于上述原因，在我们第七章的研究工作中将采取动态扩展的策略。

2.4.3 树状组织中的分布式概率

在基于树状结构的搜索空间中，包含在树中HMM的声学信息和词间的连接语言信息分别分布在不同的区域中。其中一个不利的因数是我们只有到了叶节点才知道词的序号，这造成了语言知识的延迟应用。但我们知道，在Beam搜索过程中，越早使用这些信息，将越早抛弃那些不可能的路径。通过对词语言概率的因数分解，我们就能做到这一点。语言概率的因素分解步骤如下:

a. 置根节点为所有词的最大概率值;
b. 设a为树中的一个节点，b为其一个子节点，从a出发的所有词的最大概率为Pa,从b出发的所有词的最大概率为Pb,则置节点b的概率因素为(Pb-Pa);
c. 穷举树中所有节点。

显然按照上述求解过程得到的每条路径的累加概率即为该条路径代表的词的实际概率。经过这样的分解后，从根节点到叶节点，每扩展一个节点就可以最大限度地而且安全地使用语言模型了。上述分解也适用于二元概率。此时，对每个词，都需要有一个独立的分解过程。对于大词汇量来说，存贮量太大。

我们在构造连续语音识别音节层面的初搜索时进行了这种分解。按前半音节100个模型分解，原来的400个音节使用频率分解成为(100+400)个节点频率，原来的400x400音节二元连接频率分解为(400x100+400x400)个节点频率。在我们的一个非特定人连续语音识别实验中，加入上述分解模型后，音节误识率及前10名不包含率降低了约30%左右。

2.5 结论

本章从音节困惑度和搜索复杂度出发，定量定性地分析汉语语音识别在不同的知识源条件下(二元音节连接、词一元连接、词二元连接等)汉语音节的困惑度问题、词库的组织形式和树的扩展方法，得出了一些非常有价值的结论:

. 汉语词典约束对减小音节困惑度从而提高语音识别底层识别率具有重大的意义，而为了提高搜索效率，词典应该组织成树状;
. 在连续语音识别中，应该采用动态的树扩展方法. 而采用树状条件下概率的分布式表示能有效地提高语言模型的使用效率.

对于本章得出的一些结论，构成了我们第五章和第七章搜索框架的基础. 如在词法树上直接进行声学层面匹配，邻词－时间双属性树的动态扩展等等。

第三章 汉语建模单元和上下文有关声学模型研究

语音识别用的声学模型必须准确有效地反映不同的人、在不同的时候和不同的上下文环境中发音的时变性及随机性,它是语音识别的最核心部件。为解决连续语流中出现的各种音变现象及由于非特定人说话带来的样本的离散性，必须建立精确的声学模型；这种模型的精确性依赖于两个因素：第一个是训练语料的设计，也就是所选择的语料应该尽可能地包括各种语言学和语音学中的现象；第二是所选择的模型有足够的描述能力和强大的训练算法。同时，由于训练数据永远不能覆盖所有的声学音变现象和说话人，又要求模型有一定的鲁棒性。通常精确性和鲁棒性是一对矛盾，但在大数据量训练的条件底下，我们有可能做到两者的统一。这也是国外前几年在该领域研究极其活跃并且得到迅猛发展的原因[11,12,13,15,18,20]。

众所周知，汉语是一种单音节结构的语言，尤其它具有非常单一的声韵母结构。这些声韵母是一种非常稳定和固定的组合。所以传统上，汉语语音的识别皆采用声韵母为基本的建模单元；另外一方面，象西方语言一样，我们也可以采用音素作为基本的建模单元，在本章中我们将通过实验比较来研究说明汉语的声韵母是比较简单而又合适的单元；在单音节识别中，实验早已表明采用音节内部上下文有关的模型能大大提高建模精度，我们的研究表明在连续语流中必须考虑音节之间的协同发音和声调有关的上下文建模。我们还通过对离散密度HHMM(HMM/VQ)和连续密度HMM(CDHMM)的性能比较来证明CDHMM具有非常强大的建模能力。本章还指出了增强模型分类能力的两种途径和它们各自的应用范围。

3.1 COSDIC数据库

语音信号的多变性使语音识别技术的发展遇到了很大的挑战。语音系统性能的改进有三条途径：更好的模型，更好的匹配技术和更多的数据。匹配搜索技术和模型的比较都依赖新数据。可以说近几年来语音识别的巨大进展有一半的功劳得益于大量精心组织的多说话人的新语音数据库。建立语音数据库是语音识别工作必不可少的基础。Robert Leroy Mercer曾说过：Dere's no data like mo' data. [128]

语音库的建设无论是从投资还是人力、物力的耗费上讲，无疑都是巨大的。为了节省这种投资，应精心选择和组织语料，使得能消耗尽量少的成本来满足研究的需要。我们实验室于1995年秋季建立了一个面向汉语语音识别的汉语综合语音库COSDIC[106]，本论文的大部分工作试验研究都是在作者参与的这个数据库提供的语音上进行的。下面对其语料设计原则及数据情况作一介绍：

3.1.1 语料设计原则

语料库的建设分为两个部分，语料设计和样本采集．语料设计是关系一个语料库是否好用的关键，语料设计与语料库应用目的密不可分．建立COSDIC语音库的目的满足大词汇量、非特定人语音识别的研究开发和性能评测的实际需要。在语料的设计和组织方面我们的原则是:

a. 语料的选择独立于识别方法
b. 尽量满足选用不同识别基元(音素、声母、韵母、音节)，不同的识别条件建模的需要．
c. 语料的选择应充分考虑语流中的前后语音单元之间的相互影响，即语流中的音联现象，以满足建立精确的声学模型的需要。

为此，语料覆盖了常用的几种识别单元(音素、声韵母)的二元音联现象，包括音节内部和音节之间的音联现象。在覆盖全部二元音联现象的同时，还考虑研究的需要，覆盖了相当一大部分的三元音联现象．

3.1.2 COSDIC库内容

COSDIC库由以下五部分组成: 单音节词、多音节词、数字串、试验句和语句．它们相互独立又可混同使用．

a.单音节部分: 单音节组里考虑了全部带调音节和常用单音节汉字，带调的单音节共有1460个，常用汉字选取了100个．1460个音节分成10组，每组基本上包含了所有声韵母，每12人发音一组，男女各半．100个常用汉字分为5组，每24人发音一组，男女各半。

b.词: 解决上下文有关声学模型的建模所需语料．编选原则是全面覆盖、较少冗余、自然的发音和同一语音现象尽量多的语音环境．按照覆盖所有音素级二元音联和声韵母级二元音联的原则共选词986个，分成四组，还补了一些自造的词．以上单音节和词库共120人基本上保证了每一声韵目二元连接有30个样本，对每一音素二元连接有120个样本．

c.数字串: 选用的基本单元是单个数字音节，共有11个；数字串共选取80个，串长从1到7不等．覆盖所有音节间二元音联，共有音节间二元音联现象140个，每个音联现象至少出现二次。本库构成一个相对独立的语音库，可用于汉语连续数字的识别研究和产品开发。

d.试验句: “他去无锡市，我到黑龙江”，供语音分析用

e.语句设计: 设计原则基本同词，出发点是尽量覆盖汉语自然语流中的音联现象。共选出大约120段左右段落文章，大约1100句意思完整的句子．本语句由于时间及人力限制，未录制响应的数据。

上述数据的录音是在声霸卡上进行，环境噪声控制在38-42DB之间。我们的实验基本上是在女声上进行。需要说明的是，在录制COSDIC数据库前，我们已经有一个共35人，每人发音500个词的数据库。此数据库在录音室用数字录音机录制，再经过外接滤波器、放大器，采用TMS320C30开发板作A/D转换。我们分别用DB1，DB2和DB3来标记如下上述库中的部分或全部数据:

DB1: 35人500词数据库(TMS320C30开发板采样，30人训练，5人测试)
DB2: 60人女声单音节库(50人训练，10人测试)
DB3: 60人女声单音节和词库(60人训练，"863"数据测试)

3.2 汉语声韵母与音素建模的比较[132]

在以 HMM 为框架的大词汇量语音识别研究中，为了增强对语音信号的建模能力，模型的复杂度或维数需要增加，这导致所要估算参数的大量增加；另一方面由于训练数据的有限性及搜索空间的急剧膨胀，都需要对基本建模单元的选择作慎重的考虑。

在英语语音识别中，建模单元曾选用过单词、音节、半音节、音素等，现在大多数的大词汇量系统都采用音素作为基本建模单元。汉语发音同英语发音有很大不同。汉语的每个字都是单音节，只有400多个无调音节，每个音节由声韵母二部分组成。由于这样一种特殊的结构，现有的汉语语音识别系统大都以声韵母作为建模单元。更细地分析，声母由一个音素构成，均为辅音。韵母由一个、二个或三个音素共13个音素构成，均为元音。无疑，一共只有36个基本音素的建模在对训练数据和存储的要求上是非常有吸引力的。为此我们进行了考察和比较。

3.2.1 确定建模单元的三个关键要素

考察一个单元的合适与否，可以从一致性、可训练性和可分享性三个角度考虑。一致性表示该模型代表的单元其变化程度的大小，可训练性表示实际训练过程中得到样本数的容易程度，可分享性则指该模型与其它模型有否内在联系。不同的建模单元有着各自的优缺点。建模单元小，模型训练充分，但在语流中的变体多，在不同上下文中不太稳定，从而导致模型不能精确地描述对应的语音；建模单元大，模型变体少，但训练不充分。在一定的训练的数据条件下，选择一个好的建模单元是在模型的可训练性和一致性之间找到一个均衡点。

在汉语中，相对于以整个音节为建模单元，声韵建模的好处是模型个数少，可训练性好。相对于更小的音素，声韵建模的好处是模型稳定性好。同以音节为单位建模相比，声韵建模和音素建模存在的共同问题是模型在不同的上下文中不稳定，变体较多。对于这个问题的解决方法是对模型细化[68]，即把处于不同的上下文的单元作为不同的模型来对待，这样就能使得模型更准确，但同时又带来一个问题，即模型增多，可训练程度下降。但同整个音节为单位建模相比，此时

细化的模型之间具有共享性，细化的模型可用对应的非细化的模型来作平滑处理。基于以上的考虑，本文将主要研究和比较以音素为单位建模和以声韵母为单位建模的性能。下表3.1 列出了汉语可能的建模单元的三性比较:

表3.1 各种建模单元的特性表

建模单元	一致性	可训练性	可分享性
词	好	不好	不好
音节	好	一般	不好
声韵母	较好	较好	好
音素	不好	好	好

3.2.2声韵母建模单元的比较

在下面的试验中，原始语音数据16K采样，16比特量化，以384点为一帧处理，帧之间重叠192点．每帧数据经预加重和海明窗后，计算规一化能量和12阶MFCC系数，再求出一阶,二阶差分，共39维．识别模型采用离散密度HMM．利用多码本技术，对上述特征进行量化处理：第一套码本对应12阶MFCC，第二套对应一阶差分MFCC，第三套对应二阶差分MFCC，第四套对应能量，差分能量和二阶差分能量．前三套码本大小为256，第四套为64．因此最终的一帧语音特征由四个码字来表示．本节中，试验一到四在单音节库COSDIC-DB2中进行，而试验五在COSDIC-DB3中进行。

试验一：声韵母和音素上下文无关模型的比较:

在汉语普通话中，声韵母加上零声母共有59个，我们为此建立59个独立的HMM模型，外加一个静音模型，共60个；音素共有模型35 个，外加一个静音模型，共模型36个模型 ．由于对声韵母模型都比较熟悉，这儿只列出13个元音音素:a,o,e(i),i,u,v,(zh)i,(z)i,n,ng,er,e(n),(i)a(n)。 识别结果见表3.2。

表3.2上下文无关模型识别率比较

模型及个数	首选	前二选	前三选	前四选	前五选
声韵母(60)	59.6%	76.7%	84.4%	88.7%	90.9%
音素(36)	44%	63%	71%	77%	82%

试验二：考虑右上下文有关模型的比较

对于声韵母建模，由于在单音节数据中，韵母的后边总是静寂段，所以采用右相关模型时，韵母实际上没有被细化．对于声母，当采用右相关模型而考虑右边的韵母的影响时，根据韵母的第一个音素把韵母根据分为八组．理论上22个声母应有176个模型，但由于单音节400个发音的限制，实际只产生声母细化模型100个:

第一组: a ai an ang ao
第二组: e ei en eng er
第三组: o ong ou
第四组: i ia ian iang iao ie in ing iong iu
第五组: v van ve vn
第六组: (zh ch sh) i
第七组: (z c s) i
第八组: u ua uai uan uang ui un uo

识别结果如表3．3所列:

表3．3右上下文有关模型识别率比较

模型及个数	首选	前二选	前三选	前四选	前五选
声韵母(138)	70．7%	85%	90．5%	93．1%	93．9%
音素(142)	68．4%	83．8%	89．2%	92．2%	93．3%

试验三: 考虑左右上下文有关模型的比较

当采用左右相关模型时，由于在单音节数据中，声母前边总是静寂段，所以此时的声母的左右相关模型同右相关模型是一样的．对于韵母而言，由于右边总是静寂段，此时的左右相关模型就只考虑左边声母的影响，这时根据声母的发音特点分为八组:

第一组(爆破音): b, d, g, k, t, p
第二组(摩擦音): f, h, x, s, sh
第三组(塞擦音): z, c, zh, ch, q, j
第四组: (浊鼻音) n
第五组: (浊摩擦音) r
第六组: (浊鼻音) m
第七组: (浊边音) l
第八组: 零声母

对于音素建模，考虑声母音素受右边韵母音素影响时，当右边的韵母部分的第一个音素不同时，就建立一个不同的细化模型；韵母部分的第一个音素受左边声母音素的影响时，把声母分为八类处理．这些同声韵母方式中细化方式相同，而对韵母内部的音素建立上下文关系时，则不对元音的音素进行进一步的分类，即是准确的三音子模型。

由以上实验结果看出，在不作任何细化的情况下，无论是声韵模型还是音素模型，结果都很差。当考虑上下文影响而对模型细化后，两者识别率都有了明显改善。音素建模的右相关模型第一名错误率下降了43%，音素建模的左右相关模型第一名错误率下降了40%。声韵模型的两种细化模型的第一名误识率分别下降了25.7%和4.7%。

当使用左右相关模型时，由于模型数过多，无论是声韵母建模还是音素建模，结果比右相关模型反而差，下面将分别采用平滑方法来处理这个问题。

表3.4左右上下文有关模型识别率比较

模型及个数	首选	前二选	前三选	前四选	前五选
声韵母(315)	61.5%	78.4%	85.4%	89.2%	90.9%
音素(325)	66.9%	81%	86.5%	90.5%	92.1%

试验四：对左右上下文有关声学模型的Back-off 平滑处理

平滑是解决模型训练不足的一种方法，即对样本数太少的模型，用细化的模型和其对应的非细化模型或和一个最低门限值作插值[68]，使在训练集中出现很少的码字的概率不至于太低。在上面的训练算法中，我们已对未出现的码字的概率赋了一个Floor值。下面将主要讨论细化模型同非细化模型的平滑对识别率的影响。

设 p1 表示细化模型的某个码字的概率，p2表示对应的非细化模型的概率，则平滑后的细化模型的概率p可由如下公式得到：

$$p = p_1 \cdot \lambda + p_2 \cdot (1-\lambda) \qquad (0 <= \lambda <= 1)$$

在本实验中，细化模型是左右相关模型，对应的非细化模型是右相关模型，我们借助语言模型中的平滑方法，采用BACK-OFF处理：即当细化模型样本数小于某个给定的阈值时，则左右相关模型由相应的右相关模型代替，即$\lambda = 0$。当细化模型数大于某个给定的阈值时，令$\lambda = 1$，即使用左右相关模型。表3.5给出了最好的平滑域值为100时的音素和声韵母建模识别率。

表3.5 ：左右相关模型平滑的识别率比较

模型及个数	首选	前二选	前三选	前四选	前五选
音素(325)	70.7%	83.7%	88.7%	91.5%	92.6%
声韵母(315)	71.6%	85.0%	90.0%	92.8%	94.8%

可以看出，经过平滑，识别率有明显改善，当阈值取100时效果最好，第一名错误率下降了11%。当建模单元为声韵母时，有相似的结论。

试验五：音素建模和声韵母建模在大词汇量识别中的性能

为了进一步比较两种建模方法的性能，文章还用两种建模单元方法在大词汇识别中作了初步的实验。实验所用训练数据为COSDIC数据库中女声库的所有人的单音节和词的数据，测试集为1996年国家“863计划”评测所用的60个句子约450个词。系统词汇量为32000。由于两种建模方法在不细化时性能都很差，因此文章直接对两种建模方法的右相关模型作了比较:

表3.6: 大词汇识别中的性能比较及样本数统计

模型类型	模型个数	首选	前十名	样本数 >100 模型所占比例
音素的右相关模型	351	66.9%	91.7%	68%
声韵母的右相关模型	627	70.6%	92.6%	52%

3.2.3 识别性能分析和两种建模单元的比较

表3.7 是采用不同建模单元和不同细化方式时的模型样本数的统计。

采用左右无关模型时，二种模型平均样本数都大于100，并且大多数模型的样本数大于100，因此可认为两者都得到了充分训练，而此时由于声韵母建模单元大，相对于音素建模，模型在不同上下文中的不稳定性的问题就不太突出，因此识别率比音素建模好.

表3.7不同建模单元时性能分析表

	平均样本数		样本数>50模型所占比例		样本数>100模型所占比例		识别率（第一名）	
	音素	声韵母	音素	声韵母	音素	声韵母	音素	声韵母
左右无关模型	607	245	97%	98%	94%	93%	44%	59.6%
右相关模型	150	102	63%	63%	41%	40%	68.4%	70.7%
左右相关模型	65	45	43%	32%	20%	11%	66.9%	61.5%
平滑处理后	-	-	-	-	-	-	70.7%	71.6%

当采用右相关细化模型时，两种建模方式的样本数大于100的模型数占总模型数的比例分别为41%和40%，此时由于两种方式可训练程度差不多，仍然是由于声韵母单元大的缘故，它的性能比音素要好。

当采用左右相关模型时，两者的可训练程度急剧下降，样本数大于100模型的比例分别为20%和11%，但由于音素建模训练充分的模型要远高于声韵模型，所以此时音素建模好.

当采用平滑的方法对左右相关模型处理后，由于声韵母的右相关模型的性能比音素好，结果平滑后的声韵母左右相关模型的识别率好于平滑后的音素的左右相关模型。

综上所述，在单音节中，声韵母建模的最好结果要略好于音素建模的最好结果．在大词汇量非特定人识别中，由于音素的不稳定性加剧，声韵母建模在大词汇量识别中优势更加明显．从计算量角度看，按照音素拼接出来的词或句子的模型长度要比按照声韵母拼接多出一约倍，这会增加许多计算量．由此我们认为，在汉语语音识别中以声韵母作为基本单元具有一定的优势．

3.3 音节间上下文有关的汉语声韵母声学模型研究[107]

在早期的特定人和现在的非特定人单音节识别中，很多音节内部声母受韵母影响的研究证实声母模型按照韵母韵头分类细化能大大提高音节的识别准确率(包括我们的上述试验)．但是音节间的相互协同发音现象在汉语语音识别中一直被忽略，没有得到系统的研究，特别是在大规模的非特定人的数据库测试条件下．本项研究就是为了建立音节间上下文有关的声韵母细化模型，对声母和韵母之间的相互影响进行分析．

考虑到三音子(Triphone)模型个数巨大，即使整个COSDIC库都无法为每个三音子模型提供充分的训练样本．为了排除因训练数据不足而导致模型没有得到充分训练而导致识别率下降的可能，在本项研究中，我们选择了COSDIC数据库中的DB1数据库进行试验．训练集和测试集中的词汇为相同的500个词．这样我们就能保证测试词汇中的三音子在训练集中至少有30个样本数．

当我们直接按照按照声韵母为单元考虑它对左右声韵母的影响时，Triphone的个数可达12万个，这样数目对初始化和模型训练都是不可能的．同上面一样，我们首先必须根据汉语声韵母的组成及发音特点对此进行分类．我们已经将声母根据发音方式的不同进行了分类，这种分类在连续语流中考虑对韵母的左右影响时要略作改变．例如“紧要”(jin yao)中的零声母”0_i”对韵母IN的影响同“紧握”(Jin wo)中零声母”o_u”对韵母IN的相互作用显然是不同的,也就是我们对零声母要进行进一步的分类．而对韵母的分类也要根据它对声母的左边作用和右边作用区分对待．在前边的实验中，我们讨论了它自右边对声母产生作用时根据它第一个音素分成了八组，而考虑它自左边对声母产生作用时，则又要根据它的韵尾可以分成以下几组:

第一组a结尾: a ia ua
第二组e结尾 : e ve ie
第三组i结尾: i (zh ch sh i) i (z c s)i ai ui uai ei
第四组o结尾: o iao uo ao
第五组u结尾: u iu ou
第六组v结尾: v
第七组n结尾: van vn uan un ian in en an
第八组ng结尾: uang iang ing iong ang eng ong
第九组er结尾: er

按照上述分类法，在汉语连续发音方式中，每个声母考虑左边影响时外加静音类共有10类，当考虑右边的影响时按照音节表的约束，总共有100类；每个韵母考虑左边声母的影响时共有7类，而考虑右边声母的影响时，把零声母分裂成6类，外加可能的静音类共有13类。

我们共安排了四组实验，首先建立了60个模型的上下文无关的模型，作为基准系统。其识别率只有85.7%。象以往一样，当考虑音节内部的右上下文有关模型即把声母模型细化时，识别率有较大幅度的提高，降低误识率约37%。下面两组实验则考虑了音节间的上下文协同发音关系。当我们考虑音节间左相关模型时，识别率比较上下文无关模型也由一定的提高，降低误识率约为20%；当我们考虑音节间右相关模型时，识别率最高，降低误识率约为52.44%，见表3.8。

表 3.8 音节间上下文有关模型的识别率变化情况

模型类型	模型个数	词识别率
实验一：上下文无关	22 + 37 + 1	85.7%
实验二：只考虑音节内部的右上下文有关模型	100+37 + 1	91.0 %
实验三：考虑音节间和内的左上下文有关	21*10 + 38*7 + 1	88.6%
实验四：考虑音节间和内的右上下文有关	100 + 37*13 + 1	93.2%

从上面的结果，我们可以得出一些重要的结论。首先，音节间和音节内的左上下文有关模型和右上下文有关相关模型对识别率皆有提高，其中右相关模型提高得更多。这说明每个音素皆会受到来自左面和右面的音素发音影响。其中来自右面的影响要大于来自左面的影响，尤其是实验二和三的比较更说明问题。即使考虑了音节间的左相关模型也不及只考虑了音节内的右上下文有关模型；而考虑了音节之间协同发音的右上下文有关模型比我们常用的只考虑音节内右上下文有关模型的模型误识率减低得更多，**这说明在连续语流中，必须考虑音节之间的这种协同发音，尤其要重视右上下文有关模型。**

3.4 声调有关声学模型研究[107]

汉语是有声调的语言，对于普通话来说，共有四声(阴平，阳平，上声，去声) 和一轻声，且声调信息一般加载在韵母部分。声调信息在人们的日常识别和理解中具有很重要的意义[86]。在声学上，它影响着音高(基频)走向、能量值和音的长短等超音段信息。

有四个原因需要我们研究探讨声调有关声学模型的研究。首先由于汉语的声调在汉语中有非常重要的辨意作用，它的可靠识别能大大减少语言处理的难度。而在连续语音中声调的单独识别是一件很困难的事情，而其某种形式的识别结果与与音识别结果的混合判决更是不好处理，况且声调的识别又与分割有关，

因而我们迫切需要一种把音和调同步识别的方法；第二个原因是声调之间的相互调制现象是非常明显的，当我们考虑声调识别时当然不能忽略这个现象。从某种意义上，它的处理方法应同上下文音素有关模型类似，我们希望找到一种二者合一的框架；其三，我们一般认为目前的倒谱系数只同通道特性有关，而与激励关系不大，所以韵母的建模往往忽略声调的影响。但从倒谱的角度看，它只是对实际语音生成模型的一种近似，它不可能非常准确地反映通道特性，反映的也不可能仅仅就是通道特性。比如在LPC的谱中，就有学者发现高阶谱中含有激励信息。如果我们直接根据韵母谱特征按声调进行模型分类，能得到一定准确度的声调识别率，据此就可以验证这种设想。第四如果直接加入反映声调特性的基频特征，则在词识别中能大大增加它们之间的区分度(在全音节中，它不能增加区分度，因为基本相似的音节有同样的声调)，从而降低搜索的模糊性，而且免除了声调的后处理这一步。

基于上述思想，我们按照声调不同对每个韵母建立独立的5个声调有关模型，此时模型个数为(100+37*5+1),当然有些模型不一定在集中出现。实验结果同音节内部右上下文有关的模型相比，可以看到识别率有明显的提高，甚至比音节间右上下文有关模型的识别率还要高一些。

表3.9 声调有关的模型误识率变化情况

模型类型	模型个数	识别率
音节内右上下文声母模型	100+37+1	91.0%
声调有关模型	100+37*5+1	93.8%

这个实验初步证明了按照声调建立韵母模型的有效性。事实上，我们在上面的实验中并未加入直接反映声调特性的基频及差分基频特性。我们有理由相信，通过在特征空间上加入基频有关的分量，将大大提高词之间，尤其是中小词表此间的可区分度，从而进一步提高识别率。我们将在下一章中对这问题进行进一步的研究和探讨。

3.5 孤立词识别中端点有关声学模型研究

当我们从全音节识别转移到词识别时，如果不对模型个数分类作任何变动，词中单音节的误识率会有较大幅度的提高(约为17%)。在孤立词识别中，有两个理由需要对这个现象作特别的关注。首先我们注意到汉语中的一些高频词都是单音节词，这些词的识别率对整体系统的识别率举足轻重；其次，对于孤立词识别的树搜索来说，我们总是希望在树的第一层模型的方差较小，这不但能提高识别精度，而且能作大的裁剪，加快搜索过程。如果该层模型离散度较大，势必造成裁剪困难和路径的大幅度增加。

单音节的识别率在词中的下降，究其原因是当它们与词中间的音混同建模时，单音节中那些原来没有受左右影响的比较纯的声韵目模型被泛化。分离出这

种影响使得单音节建模在孤立词中相对独立，有望提高识别率。基于此考虑，我们建立了端点有关的声学模型。这种模型就是将声韵母根据是否处于词首或词尾或词中加以区分。从数量上看，每个模型数目翻了一番，总共为275个。从形式上看，它是左右上下文有关的声学模型的特例。

采用DB2训练，进行全音节识别时，单音节的识别率为70％左右；当采用DB2,DB3混合训练，实验转移到非特定人孤立32000词识别任务集中时，其中的单音节准确率降低为65％左右。而采用端点有关的声学模型后，识别率又基本回升到原来的水平。实验结果见表3.10：

表 3.10：端点有关声学模型

模型类型	模型个数	识别率	
		单音节	总识别率
右上下文有关声母模型	100+37+1	65.2%	62.3%
端点有关声韵母模型	100*2+37*2+1	69.2%	66.7%

3.6 从离散密度HMM到连续密度HMM

我们在上面的实验中皆采用了离散密度HMM。离散密度的HMM模型计算输出概率时只需要查表，与连续密度HMM计算Gaussian函数相比，其计算量要小得多。但由于HMM/VQ方法对所有特征矢量量化成一个固定数目的码本，期间会引入很大的量化误差，从而影响最终的识别率。而连续密度HMM除了计算量大外，由于其建模能力较强，因而需要较大的训练数据量。由于早期计算机容量及计算能力的限制和没有大规模的语音数据库可以利用，限制了这种模型在孤立词和连续语音识别中的应用。随着计算机硬件技术的的迅速更新换代和COSDIC数据库的建立，我们有理由采用更加具有发展潜力的复杂模型。

3. 6.1 采用的连续密度HMM(CDHMM)的类型

Baum等人在对观察概率密度函数加以一定限制前提下，将HMM的算法扩展到连续观察值情形。Liporace和Juang先后放宽了对观察序列概率密度函数的约束条件使之适用范围拓广为：概率密度函数为严格log凹函数或椭园对称函数的有限线性组合，从而使HMM在语音处理中的广泛应用成为可能。CDHMM有很多形式，最常用的是混合高斯密度函数，形式如下：

$$b_j(y_t)=\sum_{m=1}^{M}c_{jm}N(y_t;u_{jm},\sigma_{jm})$$

这而$b_j(y_t)$为P维正态分布函数，μ_{jm}为j 状态m类平均样本矢量，σ_{jm}为j状态m类协方差对角线矩阵，M为分布函数的分量个数。由于满协方差矩阵所需估算参数多，其训练识别时计算量巨大，因而目前一般对之简化为对角化矩阵。对角化后忽略了特征维之间的相关性，必然引入误差。这个问题一方面由多个高斯密度函数混合得到了一定程度的补偿，另一方面我们可以通过特征处理消除这种相关性，以减少这种对模型的依赖。例如正则自相关分析和MFCC的余弦变换都在一定程度上起到了作用。

常规的HMM采用状态来划分输出。另外一种则采用跃迁来划分输出。有时候，我们需要通过HMM的Toplogy结构来约束其时长，即对较长的音设置比较多的跃迁。但如果每一个跃迁有各自的输出概率密度函数(pdf)，则模型的pdf会很多，而过多的pdf意味者更多的参数待估算。但事实上，长音并不意味着一定有大的频谱特征变化，有些相邻状态完全可以共享一个pdf，这就是我们所用的跃迁捆绑式模型。

3.6.2 模型的初始化和训练要点

为研究CDHMM的性能对训练过程的敏感程度，Rabiner和Juang等人专门用一个“Toy”模型来产生训练序列，并用之估算模型的参数，然后将这些参数同“真实”的模型参数进行比较，观察吻合程度[129]。他们得出的主要结论有:

1.混合高斯密度的CDHMM对每个均值附近的误差非常敏感。若初始化时的均值误差足够大，则模型很难有好的估算
2.模型对协方差的误差敏感度比对均值的为小
3.模型对转移概率和混合加权系数的误差不敏感。
4.当参数估算中运算精度不够时，模型参数对初始化值变得敏感
5.协方差矩阵可近似为对角化。

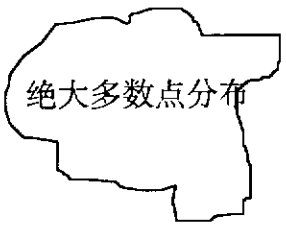
在我们的实验中，一种初始化方法可以根据文献[14]提供的把所有训练集中的矢量求均值和方差作为均匀初始化所有模型、所有输出的分布，其结果是全音节的识别只有45%左右的准确率；然后我们手工切分了60个人的数据至声韵母段，把过渡段几帧划分为声母段，而在模型内部我们尝试了等分割均匀初始化和倒谱累计差非线性分段法，在没有作任何F-B迭代的情况下，识别结果如下表3.11:

表3.11 初始化实验结果:

初始化方法	无初始化	线性分段	非线性分段
识别率	45%	69.56%	67.92%

以上结论的模拟实验和我们的实验证实了均值初始化时必须有足够的准确性.

在训练算法中，我们虽然对前向和后向概率进行了归一化处理，但还有一些训练数据最后概率或中间结果下溢和上溢，导致训练中间的退出和失败．经过分析有以下两种情况。一种是模型的参数一开始比较接近奇异点，造成其估算出的方差趋向为零；另外一种情况是奇异点离实际分布太远，导致其中间向前概率因子为零。



奇异特征点 A

图3.1 训练奇异情况

为了克服这种情况，我们有必要对估算出的方差进行低限限制．下面的实验表明这种限制不但解决了上面出现的上下溢问题，而且有效地提高了识别率．如表3.12所示．我们的这两组实验证明了CDHMM确实对均值及方差较为敏感，需要妥善处理才能有比较好的结果.

表3.12 方差限制对识别率的影响

方差限值	1.0e-10	1.0e-5	1.0e-4	1.0e-3	1.0e-2	1.0e-1
识别率	68.18% (91.84%)	67.98% (92.02%)	69.1% (95.28%)	72.46% (98.4%)	72.46% (98.46%)	73.38% (97.94%)

3.6.3 特定人和非特定人孤立词识别实验

利用上述的一些实验结果，我们进行了用CDHMM识别特地人和非特定人孤立词的比较实验．在特定人实验中，训练集为3000词的发音，测试集为1000词的发音，我们采用已经得到的离散密度HMM分割训练集中的所有数据再初始化相应的CDHMM模型；而非特定人实验在数据库COSDIC-DB3中进行．首先利用我们手工切割的单音节数据初始化模型，再在单音节数据库上迭代二遍后作为词训练的初始化模型．训练过程采用严格的Forward-Backward算法，在上述实验中我们用的Mixture为4．。在未对模型的拓扑结构、M值大小、识别搜索过程参数控制等作优化的条件底下，特定人第一名误识率降低了60%之多．非特定人的实验也有较大幅度的提高，约为37%。见表3.13。

表3.13 离散密度HMM和连续密度HMM比较

HMM类型	离散密度HMM		连续密度HMM	
识别率	Top-1	Top-10	Top-1	Top-10
特定人数据库	87%	98.8%	94.8%	99.5%
非特定人数据库	65%	92.3%	78%	98.5%

我们的进一步实验表明(见表3.14),CDHMM的混合数为1时其建模能力已经达到离散密度HMM的水平。而CDHMM的混合数根据训练数据的大小可以自由的调节。这些充分说明了CDHMM建模能力的强大。

3.7 增强模型分类能力

在基于大规模数据库的非特定人连续语音识别中，提高HMM的模型描述能力至关重要。考虑诸多的因素，HMM模型至少需要对下面带来的语音信号的一些变体进行建模:

性别(Gender)信息: 由于男女声的谱分布区域有较大的距离，因此需要模型能将特征空间分为性别有关的子空间;
口音(Accent)信息: 由于不同语言区域而造成的口音差异，需要模型能反映口语引起的特征离散分布;
个人信息: 同一性别、同一口语的不同说话人甚至同一说话人会由于每人声道结构的不同、自身心里生理状态的变化产生变体;
上下文信息: 由于语言的系统发音(co-articulation)，同一语音单位在不同的上下文环境中发音会有所不同，由此需要得到上下文相关的模型分布；除此之外，还有背景噪音，通道变迁(Microphone,A/D转换电路等因素。

有两种方法来对上述这些变化进行处理。一种是直接增加模型对各种分布的复杂刻划能力；例如在HMM中，我们可以增加模型的状态和分布(例如逼近状态连续的Markov过程),增加特征的数量，增加混合数(Mixture)，甚至我们可以采用二阶Markov链来反映前后帧之间的相关性。另外一种方法是人为地事先根据变体的情况对HMM进行预分类建模。对于后者来说，我们必须事先能清楚如何对这种变体进行分类。例如性别信息，我们可以清楚对说话者进行分类，因而我们能非常简单地把模型分成男女两类(gender-dependent model)，事实也证明这种分类对提高准确率是有益处的[68]；又如对上下文有关的音素变体，我们也能根据文本构成信息确切地知道某个发音所处的语境，从而进行分类，就象我们上述实验中对声韵的分类一样。而对于某些不能说清楚的离散分布，例如个人信息，又如说话者生理和心理的变化我们就只能通过提高模型的建模能力和对这些变化的处理能力来下功夫了。就目前水平来说，我们似乎还不太可能根据说话人高

兴、悲伤、感冒等因素来建立模型，此时只有通过提高模型自身的描述能力这条途径。

大量的实验已经证明增加混合数是提高识别率，增加模型描述能力最简单最有效的办法。表中是作者对上述非特定人孤立词识别作的一组实验，可以发现随着混合数的增加，识别率稳步上升，M从1到4大概降低误识率31%. 不仅如此，文献[14]中用实验证明了事实上只要简单地通过增加混合数其识别率就可以达到同建立了上下文有关和性别模型的系统水平.

表3.14 非特定人孤立词识别率同混合数的关系

混合数	M=1	M=2	M=3	M=4
识别率	67%(94.7%)	73.4%(95.9%)	75.7%(97.2%)	77.3%(97.5%)

对这种现象的解释最简单的当然是模型描述能力的提高，但还有一个深层次的原因是多个Gaussian分布的混合，实际上起到了一种数据的平滑作用，这种平滑作用虽然只是局限在一个输出分布内部进行，但它提供了处理训练集外数据的一种能力.

我们可以推想，如果一个模型考虑到上下文关系需要分成n类，而根据说话人的差异要分成m类，则该模型如果不事先分类，则至少需要m*n个混合数才能比较准确地对这些变体进行建模. 但是这样建立出来的模型在识别和训练时不管上下文的识别结果一概都要计算m*n个混合Gaussian概率分布，大大增加计算量. 更为糟糕的是由于这两种主要由上下文和说话人的不同带来的数据分布离散性很难用简单的聚类方法加以分离. 其粗放的聚类模型结果是由于方差较大，带来进一步的搜索变慢和混淆度增大. 而在真正的连续语音中，某些音确确实实发生了严重的协同发音现象，甚至同原来的注音有非常大的出入，用这种方法生硬地加以相同考虑是不合适的.

由于上述原因，作者认为目前语音识别的声学模型建模可以通过事先区分模型来建立性别有关、上下文有关的模型，而通过增加模型的描述能力来刻划诸如口音、说话人、生理、心里及其它不可预见的变化. 在这一章中，我们已经显示了一些实验结果来说明二者的有效性，而在下一章中我们将对第一种预分类建模方式进行更加具体和综合的讨论.

3.8 总结

在本章中我们比较了声韵母和音素两种建模单元，建立了音节间上下文有关、声调有关和端点有关声学模型，研究分析了语音信号的变体不同来源及其处理办法. 概括地我们可以得出以下结论:

. 通过分析和实验比较，从一致性、共享性和可训练性以及最后的实验结果来说，声韵母是汉语语音识别比较合适的一种建模单元.

. 在连续语流中，汉语语音识别仅仅考虑音节内部的协同发音现象是不够的，必须考虑音节之间的相互作用. 而来自右边的协同发音比来自左边的协同发音要严重的多，必须加以重点处理.

. 在连续语流中，协同发音现象的研究必须考虑声调. 本章建立的声调有关的模型不但能提高识别准确率，而且为研究音与调的一体化识别方案提供了一种新的思路.

. 端点有关的声学模型作为上下文有关模型的一种特例，在孤立词识别中非常有效而简单的分类方法.

. 分析了语音信号变体的不同来源，明确提出了提高建模精度的两条途径. 确立了采用上下文有关模型建模协同发音现象，用增加混合数刻划说话人分布的CDHMM建模方案.

上述的一些结论对实际声学模型的建立提出了新的挑战，其突出表现在大量的模型个数和有限训练数据之间的矛盾，训练集中未出现或未充分训练模型在识别时的处理等. 针对这些问题，我们在下一章将探讨建立一个基于语音学知识和数据驱动相结合的声学模型建立过程和框架，来比较完善地解决这个问题.

第四章 基于语音学知识及数据驱动的建模研究

上一章的讨论中，我们已经看到建立上下文有关的声学模型非常有用。但在实际中，要反映这种协同发音现象所需的模型个数是巨大的。例如在英语中，它共有 48 个基本的音素，就需要 11000 个模型。对汉语来说，即使我们按照上一章的分类，再加上声调之间的相互影响，其数目也是巨大的。但在实际的训练过程中，由于数据的有限性及局限性，又不可能建立这么多的模型。即使存在满足需要的巨大语音数据库，也会遇到存储量和计算量的问题。因而最大限度地利用有限的数据，训练出可靠而又准确的模型就成为一种必要。那么，我们如何对一个特定的语音数据库建立适度规模的声学模型呢？如何处理在训练集中未出现的三音子(Triphone)呢？

本章首先对国外在声学模型建立过程的上述问题处理办法进行一下回顾，然后我们采用连续密度的 HMM 作为基本模型，结合汉语语音学的特点及我们在上一章的一些研究结果，实现比较了两种不同的声学模型建立过程－自底向上的聚类过程和自顶向下的决策分裂过程。这两种过程皆实现了对应 Triphone 的同一状态输出的分布共享，然后根据输出分布相似情况和可训练度决定模型个数。但从结果和过程看又具有各自的特点。自顶相下的分裂过程是基于汉语语音学知识和数据驱动相结合的一种框架。在建立这种框架后，我们自然地提出了把汉语的声调作为其中决策要素的一部分加以融合，比较好地解决了汉语语音识别和声调识别的同步问题和判决问题。本章还给出了作者设计的一套基于汉语语音学知识的分裂决策过程关键控制因素－基于语音学知识的上下文问题。

4.1 国内外声学模型建立的发展

Bahl 等人首先在 1980 年提出了建立上下文有关的音素模型的方法。上下文广义地可指这个音素左右的所有声学现象和语言现象。包括说话人发音大小、说话速度、韵律、调、发音位置等，但狭义地可以理解为两个主要因素，及左右的音素对中间音素的影响。此时的音素就称为三音子(Triphone)。 BBN 公司的 Schwartz 等人在 1984 年首次发表了 Triphone 模型的比较性研究，其实验结果表明，基于 Triphone 模型的小规模语音识别系统的误识率比基于上下文无关的独立音素的降低了 50%[130].

1989 CMU 的 K.F.Lee 发表了对 Triphone 更为详细而彻底的研究，该研究由于在 DARPA 计划的 RM 数据库上进行，因而更具有权威性，同时掀起了用不同模型不同策略研究声学模型的高潮。 K.F.Lee 的工作首先从上下文无关的 48 个音素的 BaseLine 作起，动用一切已有的技术，通过建立 Triphone,聚类建立泛化 Triphone(GeneralizedTriphone),再经内插、平滑，在无语言知识的条件下，词的

识别率从 55.1%提高到了 72.7%. 而在词二元概率的条件下，词识别率则从 91.2%提高到了 95.4%. 针对识别中一些功能词误识率较高的特点， K.F.Lee 还建立了上下文有关的功能词模型，进一步提高了识别率. 这个首先在大词汇量、非特定人、连续语音识别背景下的工作大大鼓舞了人们对声学模型的探索，对 HMM 固有的潜力也有了更加乐观的预测，并成为声学模型研究的一种方法论. **这种方法论的本质是在大规模语音库的条件下实现对模型的优化.** 应该指出的是，所有这些方法的聚类和平滑皆是基于在模型层面上的相似度比较.

在 HMM 中，和具体状态相关联的输出分布(Emission/Output Distribution)描述了模型对应的语音中不同的声学事件[26]. 为了提高模型对这些声学事件的描述能力，一般需要增加模型的复杂度和参数数量，同时要求用于模型训练的语料足够丰富. 为了解决模型具体结构和有限的训练语料之间的矛盾，先后发展了模型参数的平滑算法和共享技术. 共享技术不仅能够改善模型参数估计的充分程度，且能大幅度的减少训练和识别的计算量. 在这个问题上， X.Huang 和 IBM 的科学家同时独立地提出了半连续密度 HMM(SCHMM)和捆绑分量 HMM(Tied-Mixture HMM). 在这以后，许多学者不断发展上述建模思路，又先后发展出了更加精细的结构，如模型状态或输出分布层次的共享(Distribution/State-Level Clustering)[15]，输入混合概率密度函数中混合分量层次的共享(Mixture-Level Clustering)[11,48]等.

上述基于状态或分布的共享思路和上下文有关模型的结合，大大推动了建模的精确性和鲁棒性. 通过模型的合并和共享，我们说产生了 Generalizaed Triphone,而通过输出分布函数的合并和共享，则发展出了 Senone,Fenone 等概念. 1994 年 Digalakis 等基于 Tied Mixture 算法提出了名为 Genone 的语音单元. 该算法中，混合分量的聚类不再是基于语言学知识，而是基于概率密度函数中各混合分量间的相似度测量，从而提高了混合分量捆绑的程度，避免了语言学特征和语音学特征间不一致带来的偶然性错误.

在声学模型的发展历史过程中，半连续 HMM(SCHMM)的出现可以说是承上启下. 在 X.D.Huang 把 CMU 的 Sphinx I 系统用 SCHMM 建模， Mei-Yuh Hwang 考虑了词间 Triphone 以后 Sphinx I 演变成 Sphinx II 系统，词的识别率有了进一步的提高,降低误识率 11%到 20%. 它作为连接 DHMM 和 CDHMM 的桥梁，首先实现了不同模型之间参数的共享和模型输出的平滑，大大提高了参数估算的可靠性,从而启发了人们的思想. 而随着 ARPA 数据库规模的扩大， CDHMM 重新浮出冰面，逐步显示出其应有的生命力. 事实上，通过分布和模型之间的相似性度量， CDHMM 也完全可以把某些分布和输出加以捆绑，而其分布个数(Distribution)却能根据实际需要及数据库的大小灵活配置. 而 SCHMM 中码本的大小只能用实验调整，因而 CDHMM 在建模能力上更胜一筹. 几乎在同时，Cambridge 的研究小组用连续密度的 HMM 也提出了状态分布的聚类,他们在 TIMIT 数据库上音素的识别率从 67 % 左右上升到 72 %， RM 数据库上识别率也从 70 % 上升到 73 %.

综上所述，我们可以看到，上下文有关的模型导致了对语音单元的刻划越来越细(从词模型，音素模型，上下文有关音素模型)，对该单元的分布表示则越

来越多。而模型间参数的共享则为这种技术提供了可操作的手段。

国内对声学模型的研究也作了一些工作。主要是对声母根据右上下文有关即后接韵母的韵头作细化，取得了非常明显的识别率改进，并基本成为一种定式。但就总体来言，未对声学模性的建立过程做过非常完整和充分的研究，特别是在词组识别或连续语音识别条件下，对上下文有关模型的建立、模型参数的估算方法及过程缺乏系统的和定量的的研究实验。本文就试图采用国外比较成功的典型方法和我们在声学模型方面已经有的一些结果，针对汉语语音学的一些特点，为汉语声学模型的建立提供一个完整的解决方案。同时尝试把汉语固有的声调处理方法也融入到该框架中去。

4.2 音节间上下文有关模型的训练和输出估算

为研究上下文有关模型和共享输出估算，我们必须注意文本的音素上下文与实际发音中的音素上下文的不一致情况。这主要产生于某两个音素中间可能出现的停顿。这也是设计音素平衡语料时面临的一个问题。因而我们所说的 Triphone 应该是实际发音的上下文关系。这就需要我们在进行模型训练时，假设音节中间有可能产生停顿，因而必须插入静音模型，这使拼接变得较为复杂。下面列出了两个具体实例，拿词汇“中国”为例。

在音节之间不考虑 Silence 时，其拼接图如下:

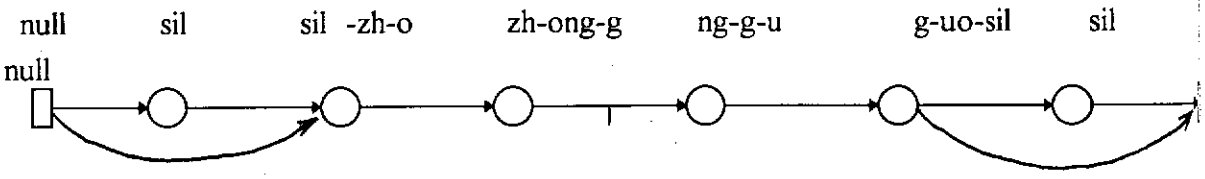


图 4.1 上下文有关模型训练 FSN 图

当考虑静音时，拼接变为:

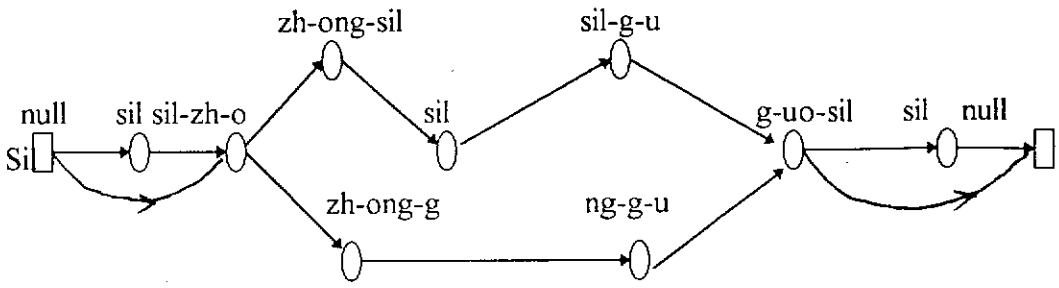


图 4.2 上下文有关模型训练 FSN 图

当我们试图单独建立功能词模型时，拼接变的更为复杂．在此就不再详述了．对于具有状态或输出共享的模型参数估算，本质上同基于模型的参数估算没有区别，只是模型之间的估算通过共享分布变得相关了．这样我们就只需要以分布为单位累计概率，而转移概率的估算则仍然以模型为单位．

4.3 基于聚类的输出分布建模

对于基于聚类的输出分布建模，其出发点是首先估算所有考虑了上下文关系的模型的分布，然后按照一定的相似度准则合并这些输出，进而对所有输出分布相同的模型进行模型级的合并。这是一种由细到粗的、自底向上的策略。在具体实现上，又有许多可能的选择。

首先是输出分布合并的范围。我们可以在所有模型的所有输出这样的范围内进行合并，如 Genone 就是这样产生的。**也可以只考虑同一音素下的 Triphone 中所有输出分布进行合并；再一种就是在同一音素下的所有 Triphone 对应状态之间进行合并，如下图所示:**

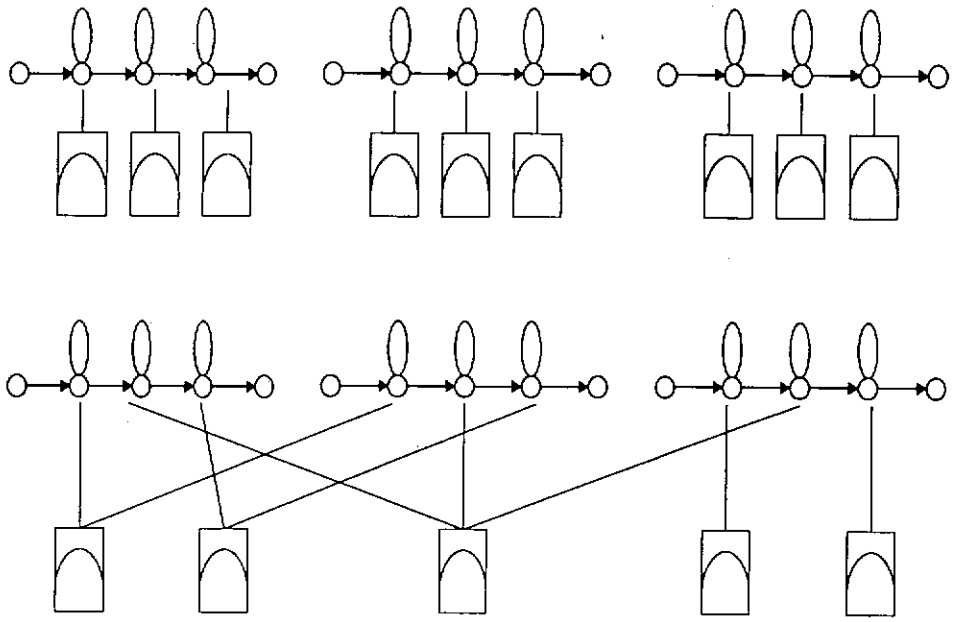


图 4.3　输出分布的聚类示意

在全局范围内进行输出的合并，在一定意义上类似与半连续密度 HMM 的思路，只是其分布的数目可以根据训练数据及相似度门限进行控制。**但这种合并法有可能造成两个不同音素之间享有完全共同的输出分布**，这将大大降低声学模型的区分能力。所以从某种意义上来讲，完全采用这种声学的聚类而不加以一定的语音学控制，会产生预想不到的结果．而在同一音素下的三音子模型的输出分布聚类，可以避免上述问题，但也会出现一个模型的相邻几个分布聚成一个分布的情况，从某种意义上来讲减弱了语音信号的时序特征．正因为这个原因，在

基于聚类和分裂的声学模型建立过程中，**我们采用了同一音素的三音子模型对应状态输出分布的聚类方法。**

其次是聚类的测度和合并的方式。我们已经有了三音子模型及其输出分布，我们可以直接采用模型参数距离来近似分布之间的距离。例如在文献[4]中对二个高斯分布$N_1(\mu_{ik},\sigma_{ik})$和 $N_2(\mu_{jk},\sigma_{jk})$就采用了下面的距离测度:

$$d(i,j)=\left[\frac{1}{v}\sum_{k=1}^{v}\frac{(\mu_{ik}-\mu_{jk})^2}{\sqrt{\sigma_{ik}^2\sigma_{jk}^2}}\right]^{1/2}$$

这种直接以模型参数作为衡量相似度的方法，其合并时并不能直接对二个模型的参数进行合并，此时两类之间的距离测度就只能采用两类之间所有分布的距离的最大值来表示了。这种测度和合并方式，其一没有把距离测度直接同识别时所采用的概率挂钩，所以不尽准确，尤其是不能把两个分布直接合并，造成估算两类之间距离的测度逐步粗糙化。

在我们的研究中，我们直接采用了**基于概率测度的聚类准则和直接数据集合融合**的分布合并方式。所谓基于概率测度的聚类准则就是把我们在识别中采用的概率计算作为衡量二个模型之间的相似测度，而数据的直接融合就是考虑二组数据合并后直接用之估算出一个新的模型来。它们的算法可以描述如下:

a.根据所设计的 Triphone 建立原则，创建所有 Triphone．并用上下文无关的模型对之进行初始化；然后用 Forward-Backward 算法训练这些模型;

b.用上述训练得到的 Triphone 分割所有训练语音库，对其中的每一帧标记上它的属性：这些属性包括其所属的基本音子，三音子模型以及其输出分布编号;

c.对隶属于同一基本音子的三音子模型的对应输出分布，进行聚类；计算任意两类的距离，采用距离最小的二类进行合并。两类的距离准则直接采用概率测度:

c1.计算属于分布 D1 所有帧数据的概率之和 P1

c2.计算属于分布 D2 所有帧数据的概率之和 P2

c3.合并 D1 和 D2 的所有数据，计算它们的均值和方差，记为分布 D12

c4.计算属于分布 D12 的所有帧数据的概率之和 P12

c5.D1 和 D2 的距离为 d(D1,D2)=P12-P1-P2;

d. 当所有类的数据集内总帧数皆超过某一值 R0 或合并后概率值增加小于某一域值 T0 时，聚类结束;

e. 对于隶属于同一单音子的三音子模型，检查它们的分布；如果有两个模型的对应分布皆捆绑在一起，则合并这两个模型。这一步不影响任何识别率，但能大大降低识别时的搜索复杂度。

f. 利用上述得到的分布聚类和模型，重新运行 Forward-Backward 更新模型和分布的输出估算。

在实验中，我们注意到采用上述数据直接合并的方法同模型参数聚类和并的方法相比还有其它的优势．在基于模型参数的聚类中，未加合并前的每个三音子模型必须训练充分，否则其本身的方差和均值就没有意义，并直接影响合并精度．而在基于数据直接合并的方法中，则避免了这个问题．采用上述过程得到的每个分布有着一定样本数目的保证，可以得到充分的训练．

但上述算法有一个问题就是某些在训练集中未出现的 Triphone 在识别时的处理问题．一般我们把它 Backoff 到右上下文有关的模型，若右上下文有关模型也不能得到充分的训练，则可以 Backoff 到左上下文有关模型进而上下文无关模型．从上面的过程也可以看到，除了建立最细的 Triphone 过程外，其聚类过程是一种纯数据驱动的架构．

4.4 基于语音学知识及数据驱动的输出分布建模

除了语音识别研究者外，语音学家们早就对音素受上下文的影响作了许多研究，并提出了一些语音学的一些规则．但是最初这些规则在语音识别中的应用并不是很成功．首先这些规则大多是基于人类的感知而不是根据声学的实际数据而得到的，所以这些规则反映的是一些全局的变异而忽略了一些细微的变化．这种忽略可能对人类感知无关紧要，但对语音识别器来说可能是致命的．另外这些规则有比较笼统的，也有比较具体的；有非常重要的，也有不十分确定的个别现象，这造成在这些规则的具体运用上的困难．我们也从上面分析可以看到，完全依赖声学的相似度会带来某种程度上的模型区分性的降低．一种好的思想就是在语音学知识的宏观指导下再进行根据声学相似度进行模型的分裂和细化．基于这种思想的决策树方法应运而产生了。

1991 年 L.R.Bahl 提出了基于语音学规则的决策树建模过程[18]．这种建模过程借助语音学中对某个音素受何种上下文影响最大等知识，对音素模型根据上下文及模型之间的实际声学距离进行分裂，从主要的到次要的，从粗逐步到细．因为它从最粗的 CI 模型自顶向下由语音学上的一些知识和实际数据分布控制进行分裂，因而我们称它为基于语音学知识和数据驱动向结合的一种声学模型建立方式．

这种方法有几个步骤．首先是根据语音学的一些知识设计一个**控制模型细化过程的问题集**．假设 P 为所有音素集， Np 为个数，则问题可以描述为:

{ 音素 Pi ∈ S 吗 ？}，其中 i=-k,...0, ...k,代表左右音素，也就是我们要考虑左右 K 个音素对本体音素的协同发音现象． S 为 P 中的一个子集合。

如果有 Ns 个子集，则所有可能的问题有 Nq=2KNs．而对于集合 P 来说，可能的子集个数为2^{N_p}．这样问题个数就非常之多．此时我们就需要语音学的知识，例如根据发音方式，发音部位来对一些音素合并成几类，以减少问题的个数．

第二是根据这些问题对所拥有的语音数据库进行标记．这个过程同聚类方法相似，但所记录的内容不尽相同．在本方法中，我们可以在标记过程中对训

练集中的每一帧数据事先回答问题集。作为二元决策树，只需回答“是”还是“不是”即可。我们对所有音素所对应的输出分布创建一个语音数据文件。因为分裂过程是针对相应的分布进行的，所以每个文件可以独立地进行装入。

第三就是决策树的生成和构造。这个算法可以描述如下:

a.从根节点（所有数据）开始进行分裂。根节点包含某个音素所有Triphone对应状态的语音数据及这些数据的模型估算。

b.对没有分裂过的非叶节点n

b.1 计算按问题q进行分裂的概率增加值P(q,n)

b.2 如果满足终止条件，例如P(q,n)小于某一个门限TC,或其数据帧数小于某值RC，则标记该节点为叶节点;

b.3 根据P(q,n)最大的一个问题$q_{\max}$，创建节点n的两个子节点。左节点为不满足问题$q_{\max}$的数据帧集合，右边为满足问题$q_{\max}$的数据帧集合。分别估算它们的分布参数;

c. 直至所有节点都被分裂或成为叶节点。

d.对于隶属于同一单音子的三音子模型，检查它们的分布；如果有两个模型 的对应分布皆捆绑在一起，则合并这两个模型。这一步不影响任何识别率，但能大大降低识别时的搜索复杂度。

e.利用上述得到的分布聚类和模型，重新运行Forward-Backward更新模型和分布的输出估算。

在上述算法中，除了问题集的设计至关重要外，其他方面例如分裂的评估函数和终止条件我们采取了与上面的聚类算法一样的思路。图4.5表示了决策分裂过程的示意。这个方法同基于聚类的方法一样，实现了不同Triphone之间状态分布的共享，有时候甚至出现模型的个数要大于输出分布个数的情况。同时与聚类方法不同的是，它比较好地解决了训练集中未出现的Triphone处理问题。通过回溯分裂决策树，我们总能找到一个最接近目标上下文模型的近似上下文模型，而不仅仅简单地用RCD， LCD或CI模型来替换。

在文献[20]中， K.F.Lee 比较了由底向上块聚类与上述由顶向下分裂方法，在识别集中的Triphone得到充分训练时，二者结果差别不大。但在更大词汇量的识别中，这种所有的Triphone都能在训练集得到训练的几乎是不可能的。在这种情况下，文献[19]研究表明后者较前者可有10%到20 %的误识率降低。

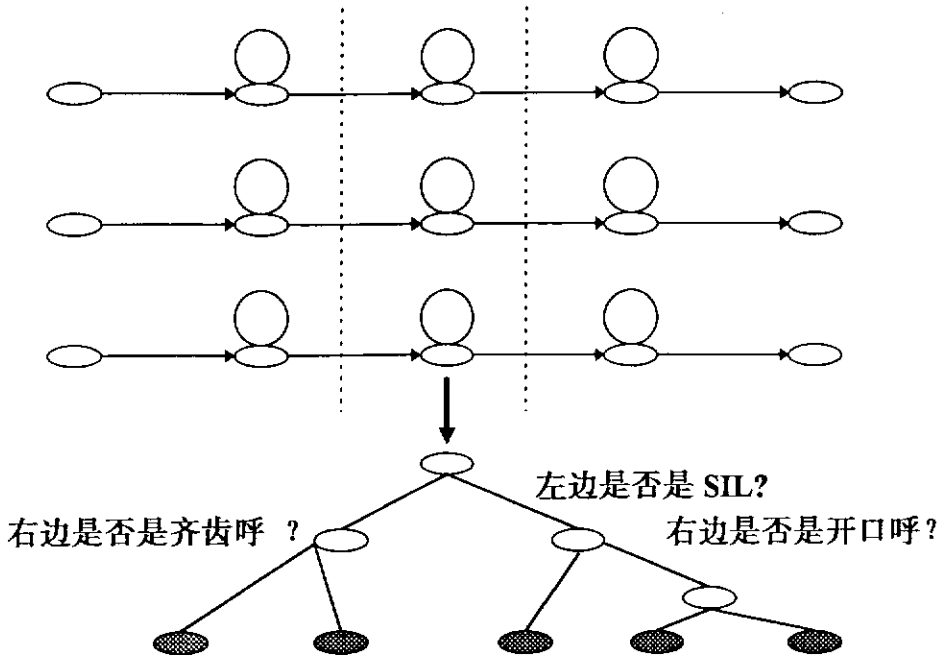


图 4.5 基于语音学知识与数据驱动相结合的决策树

4.5 基于汉语语音学的控制模型细化问题集设计

同单纯建立 Triphone 模型不同，控制模型细化的问题集需要对语音学上的一些分类方法进行组合，由于没有系统的该方面资料可以借用，我们只能对上一章所用的一些分类方法加以组合和分解。例如有关爆破音本身，我们进一步分解出两个问题：即是否是送气爆破音还是不送气爆破音？又如对韵头音，我们根据发音方式和发音部位又进一步组合成撮口呼、开口呼和闭口呼等。基本的分类可参考表 6.1 和 6.2 。

针对这些分类体系，加上我们在上一章获得的一些实验结果，我们对左右上下文有关模型设计了约四十八个问题，这些问题罗列如下:

Q1 右边是否是SIL(R:SIL)?
Q2 右边是否是零声母(R:LSM)？
Q3 右边是否是清音(RL:QY)？
Q4 右边是否是爆破音(R:BPY)？
Q5 右边是否是不送气爆破音(R:BSQBPY)？
Q6 右边是否是送气爆破音(R:SQBPY)？
Q7 右边是否是鼻音(R:BY)？
Q8 右边是否是摩擦音(R:MCY)？

Q9 右边是否是边音(R:BY)？
Q10 右边是否是塞擦音(R:SCY)？
Q11 右边是否是不送气塞擦音(R:BSQSCY)？

表 4.1 声母分类表

发音方式	清/浊	是否送气	成阻部位					
			唇	齿	齿根	卷舌	腭	软腭
爆破音	清	不送气	b		d			g
		送气	p		t			k
摩擦音	清		f	s		sh	x	h
	浊					r		
鼻音	浊		m		n			
边音	浊				l			
塞擦音	清	不送气		z		zh	j	
		送气		c		ch	q	

表 4.2 韵母分类表

发音方式	单元音韵母	复合元音韵母	复鼻尾音韵母
开口呼	a o e er	ai ei ao ou	an en ang eng
齐齿呼	i (zh)i (z)i	ia ie iao iu	ian in iang ing
合口呼	u	ua uo uai ui	uan un uang ong
撮口呼	v	ve	van vn iong

Q12 右边是否是送气塞擦音(R:SQSCY)？
Q13 右边是否是开口呼(R:KKH)？
Q14 右边是否是齐齿呼(R:QCH)？
Q15 右边是否是闭口呼(R:BKH)？
Q16 右边是否是撮口呼(R:CKH)？
Q17 右边是否是 a(R:A)?
Q18 右边是否是 o(R:O)?
Q19 右边是否是 e(R:E)?
Q20 右边是否是开口呼(R:KKH)？
Q21 右边是否是齐齿呼(R:QCH)？
Q22 右边是否是闭口呼(R:BKH)？
Q23 右边是否是撮口呼(R:CKH)？
Q24 右边是否是 a(R:A)?

Q25 右边是否是o(R:O)?
Q26 右边是否是e(R:E)?
Q27左边是否是SIL(L:SIL)?
Q28左边是否是开口呼(L:KKH)?
Q29左边是否是齐齿呼(L:QCH)?
Q30左边是否是闭口呼(L:BKH)?
Q31 左边是否是撮口呼(L:CKH)?
Q32 左边是否是a(L:A)?
Q33左边是否是o(L:O)?
Q34左边是否e(L:E)?
Q35左边是否er(L:ER)?
Q36左边是否是鼻尾音(L:BWY)?
Q37左边是否是n(L:N)?
Q38左边是否是ng(L:NG)?
Q39 左边是否是清音(L:QY)?
Q40 左边是否是爆破音(L:BPY)?
Q41 左边是否是不送气爆破音(L:BSQBPY)?
Q42 左边是否是送气爆破音(L:SQBPY)?
Q43 左边是否是鼻音(L:BY)?
Q44 左边是否是摩擦音(L:MCY)?
Q45 左边是否是边音(L:BY)?
Q46左边是否是塞擦音(L:SCY)?
Q47左边是否是不送气塞擦音(L:BSQSCY)?
Q48左边是否是送气塞擦音(L:SQSCY)?

除了左右音素有关的模型，我们还要研究声调有关的模型，因而我们也必须根据现有的一些实验和观测结果来设计同声调有关的模型。对于本体的声调，一种处理办法是在基本单元一层把它区分开，例如我们上一章所用的方法就是这种。另外一种方法是也把它作为一种 Context，然后根据数据的情况在分裂过程中用软区分的方法。按照硬声调区分方法建模，由于 COSDIC 数据库无法保证其充分的训练，所以我们把自身的声调也作为一种 Context 的影响。左右声调作本体模型的影响，无疑可以采用上下文音素有关办法。但声调的组合可以非常之多$((C_5^1+C_5^2+C_5^3+C_5^4+C_5^5)^2$，因而参考了有关声调相互作用的一些文献[91，134，135]。这些文献基本要点可以归纳如下：阴平基本不受前后音节的影响；去声主要受音节的位置影响；阳平和上声比较复杂，受前左音节的影响，但受前影响，上声和去声可以归一类，阴平和阳平可以归成一类可以；受右影响，阴平和去声可以归为一类，阳平和上声可以归为一类。为此，我们对声调设计了 16 个问题。

Q49 自身为阴平或轻声(T:15)?

Q50 自身为阳平(T:2)?
Q51 自身为上身(T:3)?
Q52 自身为去声(T:4)?
Q53 左边是阴平或轻声(LT:15)?
Q54 左边是阳平(LT:2)?
Q55 左边是上声(LT:3)?
Q56 左边是去声(LT:4)?
Q57 左边是上声或去声(LT:34)?
Q58 左边是阴平或阳平(LT:12)?
Q59 右边是阴平或轻声(RT:15)?
Q60 右边是阳平(RT:2)?
Q61 右边是上声(RT:3)?
Q62 右边是去声(RT:4)?
Q63 右边是阴平或去声(RT:14)?
Q64 右边是阳平或上声(RT:23)?

值得指出的是，上述的问题设计由于缺乏系统的语音学知识的指导，是一种非常初步的分类。尤其在实际语音中不同音素其受协同发音的影响有各自的特点。我们期望随着实验语音学研究的深入，更有效的、更有针对性的分类规则能够得到归纳，从而推动上下文建模的进一步发展。

4.6 聚类和决策树方法在非特定人音节识别中比较

我们采用上述两种方法在汉语非特定人单音节识别中进行了实验。首先我们建立了 65 个上下文无关的模型,其中包括 6 个零声母。对零声母需要提前细化，主要是考虑到在进行词和连续语音建模时，零声母对前面韵母的影响时我们不能把它当作一个音素，而是要根据实际韵母的第一个音素考虑上下文之间的影响。

4.6.1 自底向上的聚类实验

对于纯数据驱动的聚类方法，我们建立了音节内部左右相关的声学模 315 个。同上一章我们用 HMM/VQ 方法研究建模单元一样，由于训练数据的不充分，与只对声母作细化的右上下文有关模型相比，识别率有较大幅度的下降。而通过聚类的方法，然后按照上述 4.3 中所述进行聚类处理，当我们对每一输出分布要求总帧数数声母为 100，韵母为 800 时，结果见表 4.3 。

表 4.3 采用聚类方法的识别率

实验条件	模型数	分布数	平均识别率
上下文无关模型	65	257	60.8%(97.27%)

传统声母细化模型	138	549	76.10%(98.72%)
声韵母左右相关模型	315	1257	69.91%(97.13%)
聚类后的模型	172	580	76.05% (97.9%)

4.6.2 自顶向下的决策分裂实验

在本实验中，我们以 65 个模型作为基本的模型进行分裂决策．为了搞清声韵母对训练数据的不同要求，我们首先对声韵母独立地进行决策分裂．在研究只对声母进行决策分裂时，选择了 400、 200、 120、 60 四个门限．下面列出了在此四种情况下，模型的个数和分布的个数及相对应的识别率．从上述数据可以看到，采用这种分裂决策树方法已经超过了经典的声母细化模型，而其分布数却比前者减少了近 100 .

表 4.4 只分裂声母时的情形:

实验条件	模型数	分布数	平均识别率
上下文无关模型	65	257	60.8%(97.27%)
传统声母细化模型	138	549	76.10%(98.72%)
分裂决策: 400	100	329	73.05%(98.78%)
分裂决策: 200	124	405	75.56%(98.57%)
分裂决策: 100	136	457	76.81%(98.72%)

然后我们独立地对**韵母进行分裂决策实验．我们**选择了 800、 400、 200、 100 四个门限作为分裂决策的条件．下面列出了模型的个数和分布的个数及相对应的识别率.从实验结果可以看到，韵母的左相关模型对识别率也有一定的提高，但幅度不大．在第三章我们已经指出，左边对右边的影响本身要小些，而在模型的初始化过程中，影响韵母发音的过渡段已经划分到声母那里去了，所以对韵母的细化意义不是很大.

在上面实验的基础上，我们对声韵母同时进行决策分裂，考虑到声母最优的门限在 100 左右，韵母最优的终止条件在 800 左右，我们选择(100,800)这一对门限进行比较实验,此时模型数为 165，分布数为 502，而平均识别率为 76.46%(98.8%)见表 4.6 .

表 4.5 只分裂韵母时的情形:

实验条件	模型数	分布数	平均识别率
上下文无关模型	65	257	60.8%(97.27%)
传统声母细化模型	138	549	76.10%(98.72%)

800	94	302	63.8%(98.22%)
400	170	484	63.7%(97.98%)
200	247	729	60.36%(97.89%)
100	294	983	61.04%(97.74%)

4.6.3 方法分析与比较

上面的结果可以综述成表 4.4．从上面可以看出，无论是声母和韵母的决策分裂比较起 65 模型的识别器来说，都有一定的识别率提升，尤其是声母的分裂决策。同传统的 138 个模型相比，不但识别率有一定的提高，而其分布数据却将近减少 100 个。基于聚类方法也取得了几乎与决策树方法相同的识别结果。

表 4.6 性能分析比较表

模型种类	模型个数	分布数	平均识别率
上下文无关模型	65	257	60.8%(97.27%)
传统声母细化模型	138	549	76.1%(98.72%)
模型聚类方法	172	580	76.05% (97.9%)
决策树分裂模型	165	502	76.46%(98.8%)

在 4.3 和 4.4 分析中，我们已经看到决策树的方法处理训练集中未出现的 Triphone 更加准确些。而从聚类方法和决策树方法的过程来看，决策树方法更加流畅，扩展性也更加好一些。　　尤其把汉语的声调协同发音现象考虑进入后，其最底层可能出现的 Triphone 的个数大量增加，直接列举出所有的 Triphone 然后再进行聚类其计算量非常大，甚至几乎是不可能的。因而我们认为从最后的识别率、实现效率、对未见 Triphone 的处理以及扩展性方面来讲基于决策树方法优于聚类方法。我们将把该方法作为继续研究的主要框架。

4.7 决策树过程用于孤立词和连续语音的建模

对于单音节这样任务来说，其本身的协同发音现象比较简单，而本身模型的个数也并不是很多。因而讨论它只是为了证明这种方法的可行性。其最终的目的是同时研究音节之间和音节之内的协同发音现象，试图对孤立词和连续语音的建模提供一个完整的解决方案。在由单音节识别转向词或连续语音识别时，上述的问题和框架不需任何改变就可直接应用。由于在孤立词和连续语音识别建模中需要比较长的时间，我们根据前面一些结果作了几组实验的数据:

实验一：只对声母进行左决策时的结果：分裂终止的条件是帧数小于 100，此时得到的模型数目 331，输出分布数 1122 个