南京邮电大学

硕士学位论文

基于HTK的连续语音识别技术研究

姓名：陈泉金

申请学位级别：硕士

专业：信号与信息处理

指导教师：杨震

20100301

摘要

语音识别是让机器能够“听懂”人类的语言、并把人类的语音信号转化为相应的文本或命令的技术[1]。语音识别可以认为是模式识别的一个分支，它与声学、语言学、心理学、数字信号处理、计算机学科等有密切的联系，是一门跨学科的技术。

本文以 HTK 为工具，以 HMM 为模型，实现了母语为英语的非特定人、大词汇量连续语音识别系统。非特定人、连续、大词汇量这三个特点是语音识别走向应用必须要实现的目标。本文中训练和测试的语音库来自 NIST 的 TIMIT，该数据库是一个平衡语料库。在此基础上，完成语音识别关键技术的测试，本文具体工作和创新如下：

1、 研究和分析不同音素模型对于语音识别系统性能的影响；实现了单音素建模和三音素系统，结果证明三音素建模考虑了发音的上下文依赖，识别率相比单音素有显著的提高。

2、 研究和测试不同高斯混合度对系统识别率的影响。在一定训练数据量的条件下，随着高斯混合度的增加，识别率会上升，继续增加高斯混合度，由于训练数据的稀疏性，识别率会不升反降，实验验证了这一结论，识别率在混合度等于 70 时达到最大，继续增加混合度，识别率下降。

3、 研究和分析了不同特征参数对系统识别率的影响。结果证明特征参数的选取对识别率的影响是显著的。本试验中着重比较了 LPC、LPCC、MFCC、MFCC_0、MFCC_0_D、MFCC_0_D_A 等不同特征参数下的识别率，本文的识别系统采用特征参数 MFCC_0_D_A 能达到最高的识别率。

4、 研究决策树状态共享对系统识别率的提升，并设计用于决策树分裂的问题集，实验结果证明决策树状态共享对识别率有一定的提升。

关键字：连续语音识别、隐马尔可夫模型、HTK、TIMIT、语言模型

Abstract

Speech recognition is to enable machines to understand the human voice, which is the technology of transforming the human voice signals into text or commands. Speech recognition can be regarded as a branch of pattern recognition, it has close relationship with acoustics, linguistics,psychology, digital signal processing as well as computer science.

In this thesis, taking HTK as a tool, based on HMM model, a native English speaker-independent, large vocabulary continuous speech recognition system was achieved. speaker-independent, continuous, large vocabulary are three targets to be achieved before applications.Training and testing databases in this article come from NIST, Speech Database TIMIT, which is a balanced corpus. On this basis, the completion of key speech recognition tests are as follows:

1, Research and analysis system recognition rate based on different models: single phoneme model and tri-phoneme model. The tri-phoneme model takes into account the context dependent , experiment results prove that the recognition rate has a significant increase.

2, Research and testing system recognition rate with different Gaussian mixture degree. In the condition of a certain amount of training data, with the increasing of the Gaussian mixture degree the recognition rate will rise, continue to increase the degree of Gaussian mixture, due to the sparsity of training data, the recognition rate would be falling instead of rising, this conclusion was supported by experiment results, the recognition rate reached the maximum when Gaussian mixture degree equals to 70, continue to increase the mixing degree, the recognition rate decreased.

3, Research and analysis the system recognition rate with different characteristic parameters.Selection of characteristic parameters have great impact on the recognition rate. The experiment focused on comparison of LPC, LPCC, MFCC, MFCC_0, MFCC_0_D, MFCC_0_D_A different characteristic parameters of the recognition rate, using characteristic parameters MFCC_0_D_A achieves the highest recognition rate.

4,Researching the decision tree state sharing strategy and design the question set

which can be used to guide the spliting of the decision tree experiment results show that states sharing can improve the recognition rate in some degree.

Keywords: continuous speech recognition, hidden Markov models, HTK, TIMIT, language model

缩略语一览表

ANN	Artificial Neural Networks	人工神经网络
CMU	Carnegie Mellon University	卡耐基梅隆大学
DARPA	Defence of Advanced research projects agency	国防部高级计划研究署
DFT	Discrete Fourier Transformation	离散傅里叶变换
DTW	Dynamic Time Warping	动态时间归整
HMM	Hidden Markov Models	隐马尔可夫模型
HTK	Hidden Markov Model Tookkit	隐马尔可夫工具箱
IDCT	Inverse discrete Fourier Transformation	散余弦反变换
IDFT	Inverse discrete Fourier Transformation	散傅里叶反变换
LP	Linear predict	线性预测
LPC	Linear predict coding	线性预测编码
LPCC	Linear Prediction Cepstrum Coefficient	线性预测倒谱系数
MFCC	Mel Frequency Cepstrum Coefficient	梅尔频率倒谱系数
MIT	Massachusetts Institute of Technology	麻省理工学院
NIST	National Institude of Standards and Technology	国家标准技术研究所
P LP	Perceptual Linear Prediction	感知线性预测
SRI	Stanford Research Institute	斯坦福研究院
TI	Texas Instruments	德州仪器
VQ	Vector Quantization	矢量量化

第1章 绪论

1.1 引言

人类与机器进行语音交流，让机器明白你说什么，这是人们长期以来梦寐以求的事情[2]。语音识别技术就是让机器通过识别和理解过程把语音信号转变为相应的文本或命令的高技术。语音识别是一门交叉学科，它以语音为研究对象，是语音信号处理的一个重要的研究方向，又是模式识别的一个分支，语音识别技术与声学、语音学、生理学、统计学和模式识别理论、信息理论与计算机学科、应用心理学、数字信号处理技术等多个学科的研究领域有关[3,4]。近二十年来，由于计算机技术、信息处理技术、人工智能技术等领域的突飞猛进，语音识别技术取得显著进步，开始从实验室走向市场。人们预计，未来 10 年内，语音识别技术将进入工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域。

伴随着计算机技术的发展，语音识别技术已成为信息产业领域的标志性的技术，在人机交互中逐渐进入我们的日常生活，并迅速发展成为“改变未来人类生活方式的”的关键技术之一。据预测，语音识别将成为继键盘和鼠标之后，人机交互界面革命中的下一次飞跃。让机器接受人类的语言指令，是最简单最快捷的指令发布形式，因而研究人机交互的语音识别技术是人类迈向最终现代化的下一个台阶，市场潜力不言而喻。

1.2 语音识别的发展历史

1.2.1 国外研究历史及现状

语音识别的研究工作可以追溯到 20 世纪 50 年代 AT&T 贝尔实验室的 Audry 系统；它是第一个可以识别十个英文数字的语音识别系统。

但真正取得实质性进展，并将其作为一个重要的课题开展研究则是在 60 年代末 70 年代初。这首先是因为计算机技术的发展为语音识别的实现提供了硬件和软件的可能，更重要的是语音信号线性预测编码（LPC）技术和动态时间规整（DTW）技术的提出，有效的解决了语音信号的特征提取和不等长匹配问题。这一时期的语音识别主要基于模板匹配原理，研究的领域局限在特定人，小词汇表的孤立词识别，实现了基于线性预测倒谱和 DTW

技术的特定人孤立词语音识别系统；同时提出了矢量量化(VQ) 和隐马尔可夫模型(HMM)理论。

随着应用领域的扩大，小词汇表、特定人、孤立词等这些对语音识别的约束条件需要放宽，与此同时也带来了许多新的问题：第一，词汇表的扩大使得模板的选取和建立发生困难；第二，连续语音中，各个音素、音节以及词之间没有明显的边界，各个发音单位存在受上下文强烈影响的协同发音（Co-articulation）现象；第三，非特定人识别时，不同的人说相同的话相应的声学特征有很大的差异，即使相同的人在不同的时间、生理、心理状态下，说同样内容的话也会有很大的差异；第四，识别的语音中有背景噪声或其他干扰。因此原有的模板匹配方法已不再适用。

实验室语音识别研究的巨大突破产生于20世纪80年代末：人们终于在实验室突破了大词汇量、连续语音和非特定人这三大障碍，第一次把这三个特性都集成在一个系统中，比较典型的是卡耐基梅隆大学(Carnegie Mellon University)的Sphinx系统，它是第一个高性能的非特定人、大词汇量连续语音识别系统。

这一时期，语音识别研究进一步走向深入，其显著特征是HMM模型和人工神经元网络(ANN) 在语音识别中的成功应用。HMM 模型的广泛应用应归功于 AT&T Bell 实验室 Rabiner 等科学家的努力，他们把原本艰涩的HMM纯数学模型工程化，从而为更多研究者了解和认识，从而使统计方法成为了语音识别技术的主流。

统计方法将研究者的视线从微观转向宏观，不再刻意追求语音特征的细化，而是更多地从整体平均（统计）的角度来建立最佳的语音识别系统。在声学模型方面，以Markov链为基础的语音序列建模方法HMM（隐式Markov链）比较有效地解决了语音信号短时稳定、长时时变的特性，并且能根据一些基本建模单元构造成连续语音的句子模型，达到了比较高的建模精度和建模灵活性。在语言层面上，通过统计真实大规模语料的词之间同现概率即 N 元统计模型来区分识别带来的模糊音和同音词。另外，人工神经网络方法、基于文法规则的语言处理机制等也在语音识别中得到了应用。

20世纪90年代前期，许多著名的大公司如IBM、苹果、AT&T和NTT都对语音识别系统的实用化研究投以巨资。语音识别技术有一个很好的评估机制，那就是识别的准确率，而这项指标在20世纪90年代中后期实验室研究中得到了不断的提高。比较有代表性的系统有：IBM公司推出的Via Voice和Dragon System公司的Naturally Speaking, Nuance公司的Nuance Voice Platform语音平台，Microsoft的Whisper, Sun的VoiceTone等。

其中IBM公司于1997 年开发出汉语ViaVoice 语音识别系统，次年又开发出可以识别上海话、广东话和四川话等地方口音的语音识别系统ViaVoice’98 。它带有一个32 000

词的基本词汇表,可以扩展到 65 000 词,还包括办公常用词条,具有“纠错机制”,其平均识别率可以达到 95 %。该系统对新闻语音识别具有较高的精度,是目前具有代表性的汉语连续语音识别系统。

1.2.2 国内研究历史及现状

我国语音识别研究工作起步于五十年代，但近年来发展很快。研究水平也从实验室逐步走向实用。从 1987 年开始执行国家 863 计划后，国家 863 智能计算机专家组为语音识别技术研究专门立项，每两年滚动一次。我国语音识别技术的研究水平已经基本上与国外同步,在汉语语音识别技术上还有自己的特点与优势,并达到国际先进水平。中科院自动化所、声学所、清华大学、北京大学、哈尔滨工业大学、上海交通大学、中国科技大学、北京邮电大学、南京邮电大学、华中科技大学等科研机构都有实验室进行过语音识别方面的研究,其中具有代表性的研究单位为清华大学电子工程系与中科院自动化研究所模式识别国家重点实验室。

清华大学电子工程系语音技术与专用芯片设计课题组,研发的非特定人汉语数码串连续语音识别系统的识别精度,达到94.8%（不定长数字串）和96.8%（定长数字串）。在有5%的拒识率情况下,系统识别率可以达到96.9%（不定长数字串）和98.7%（定长数字串）,这是目前国际最好的识别结果之一,其性能已经接近实用水平。研发的5000词邮包校核非特定人连续语音识别系统的识别率达到98.73%,前三选识别率达99.96%；并且可以识别普通话与四川话两种语言,达到实用要求[5]。

中科院自动化所及其所属模式科技(Pattek)公司 2002 年发布了他们共同推出的面向不同计算平台和应用的“天语”中文语音系列产品——Pattek ASR,结束了中文语音识别产品自 1998 年以来一直由国外公司垄断的历史。

1.3 语音识别系统的分类

语音识别系统可以根据对输入语音的限制加以分类[6]。如果从说话者与识别系统的相关性考虑，可以将识别系统分为 3 类：(1)特定人语音识别系统：仅考虑对于特定的专人的话音进行识别的系统；(2)非特定人语音识别系统：识别的语音与什么人无关，通常要用大量不同人的语音数据库对识别系统进行学习；(3)多人的语音识别系统：介于上面两者之间，通常能识别一组特定人的语音，或者成为特定组语音识别系统，该系统仅要求对要识别的那组人的语音进行训练。

如果从说话的方式考虑，也可以将识别系统分为3类：(1)孤立词语音识别系统：孤立词识别系统要求输入每个词后要停顿；(2)连接词语音识别系统：连接词输入系统要求对每个词都清楚发音，一些连音现象开始出现；(3)连续语音识别系统：连续语音输入是自然流利的连续语音输入，大量连音和变音会出现。

如果从识别系统的词汇量大小考虑，也可以将识别系统分为3类：(1)小词汇量语音识别系统。通常包括数个到几十个词的语音识别系统。(2)中等词汇量的语音识别系统。通常包括几百个词到上千个词的识别系统。(3)大词汇量语音识别系统。 通常包括几千到几万个词的语音识别系统。随着计算机与数字信号处理器运算能力以及识别系统精度的提高，识别系统根据词汇量大小进行分类也不断进行变化。目前是中等词汇量的识别系统可以进行实用，到将来可能就是大词汇量的语音识别系统。以上这些不同的限制要求也确定了语音识别系统的技术困难度。

下面介绍语音识别的几种基本方法：

一般来说，语音识别的方法有三种：基于声道模型和语音知识的方法、模板匹配的方法以及利用人工神经网络的方法[7]。

(1)基于语音学和声学的方法

该方法起步较早，在语音识别技术提出的开始，就有了这方面的研究，但由于其模型及语音知识过于复杂，现阶段没有达到实用的阶段。

通常认为常用语言中有有限个不同的语音基元，而且可以通过其语音信号的频域或时域特性来区分。这样该方法分为两步实现：

第一步，分段和标号

把语音信号按时间分成离散的段，每段对应一个或几个语音基元的声学特性。然后根据相应声学特性对每个分段给出相近的语音标号。

第二步，得到词序列

根据第一步所得语音标号序列得到一个语音基元网格，从词典得到有效的词序列，也可结合句子的文法和语义同时进行。

(2)模板匹配的方法

模板匹配的方法发展比较成熟，目前已接近达到实用阶段。在模板匹配方法中，要经过四个步骤：语音特征提取、模板训练、模板分类、判决。常用的技术有三种：动态时间规整(DTW)、隐马尔可夫（HMM）理论、矢量量化（VQ）技术。

◆ 动态时间规整(DTW)[8,9]

语音信号的端点检测是进行语音识别中的一个基本步骤，它是特征训练和识别的基

础。所谓端点检测就是在信号中寻找语音信号的各种段落(如音素、音节、词素) 的始点和终点的位置,从信号中排除无声段。在早期,进行端点检测的主要依据是能量、振幅和过零率等参数。但在比较强的噪声条件下效果往往不明显，现在，进行端点检测的参数不局限于时域参数，频域参数、时频参数等都可以应用到端点检测中，检测更加精确。60 年代学者 Itakura 提出了动态时间规整算法(DTW: Dynamic Time Warping)。算法的思想就是把未知量通过动态规划进行升长或缩短,直到与参考模式的长度一致。在这一过程中,未知单词的时间轴要不均匀地扭曲或弯折,以使其特征与模型特征对正。结合上面对 DTW 和端点检测的论述，我们可以得出结论：端点检测的准确性将会直接影响 DTW 的效果，也可以说 DTW 对端点检测的准确性存在过度依赖的问题。

◆ 隐马尔可夫模型法(HMM) [8,9]

隐马尔可夫模型法(HMM) 是 70 年代引入语音识别理论的,它的出现使得自然语音识别系统取得了实质性的突破。HMM 方法现已成为语音识别的主流技术,目前大多数大词汇量、连续语音的非特定人语音识别系统都是基于 HMM 模型的。HMM 是对语音信号的时间序列结构建立统计模型,将之看作一个数学上的双重随机过程：一个是用具有有限状态数的马尔可夫链来模拟语音信号统计特性变化的隐含的随机过程,另一个是与马尔可夫链的每一个状态相关联的观测序列的随机过程。前者通过后者表现出来,但前者的具体参数是不可测的。人的言语过程实际上就是一个双重随机过程,语音信号本身是一个可观测的时变序列,是由大脑根据语法知识和言语需要(不可观测的状态，指语言的组织和表述方式)发出的音素的参数流。可见 HMM 合理地模仿了这一过程,很好地描述了语音信号的整体非平稳性和局部平稳性,是较为理想的一种统计语音模型。

◆ 矢量量化(VQ)

矢量量化(Vector Quantization) 是一种重要的信号压缩编码方法，它本质上就是进行信号归类——模式匹配，因此也可以用于语音识别。与 HMM 相比,矢量量化主要适用于小词汇量、孤立词的语音识别中。其过程是：将语音信号波形的 k 个样点的每一帧,或有 k 个参数的每一参数帧,构成 k 维空间中的一个矢量,然后对矢量进行量化。量化时,将 k 维无限空间划分为 M 个区域边界,然后将输入矢量与这些边界进行比较,并被量化为“距离”最小的区域边界的中心矢量值。矢量量化器的设计就是从大量信号样本中训练出好的码书,从实际效果出发寻找到好的失真测度定义公式,设计出最佳的矢量量化系统,用最少的搜索和计算失真的运算量,实现最大可能的平均信噪比。

核心思想可以这样理解：如果一个码书是为某一特定的信源而优化设计的,那么由这一信息源产生的信号与该码书的平均量化失真就应小于其他信息的信号与该码书的平均

量化失真,也就是说编码器本身存在区分能力。以离散隐马尔可夫为例，我们需要对提取的特征参数 MFCC 构造矢量量化的码书，该码书的选择会影响到时间和空间复杂度以及系统的识别率。

在实际的应用过程中,人们还研究了多种降低复杂度的方法,这些方法大致可以分为两类：无记忆的矢量量化和有记忆的矢量量化。无记忆的矢量量化包括树形搜索的矢量量化和多级矢量量化。

（3）神经网络的方法

利用人工神经网络进行语音识别的方法是80年代末期提出的一种新的语音识别方法。人工神经网络(ANN)本质上是一个自适应非线性动力学系统,模拟了人类神经活动的原理,具有自适应性、并行性、鲁棒性、容错性和学习特性,其强的分类能力和输入-输出映射能力在语音识别中都很有吸引力,特别是其模仿了目前世界上最完善的识别系统——人脑、以及其具有对空间的非线性划分能力，使得ANN在语音识别等模式分类应用中具有广阔前景。但目前由于存在训练、识别时间太长，缺乏对时序特性的描述等的缺点,目前仍处于实验探索阶段。

由于ANN不能很好的描述语音信号的时间动态特性,所以常把ANN与传统识别方法结合,分别利用各自优点来进行语音识别。

1.4 语音识别系统的结构

目前语音识别领域居于统治地位的是基于统计特性的语音识别系统。一个完整的基于统计特性的语音识别系统可大致分为三部分[10,11]:

(1)语音信号预处理与特征提取

(2)声学建模与模式匹配

(3)语言模型与语言处理

◆ 语音信号预处理与特征提取

选择识别单元是语音识别研究的第一步，也是系统预处理需要解决的任务之一。语音识别单元有单词（句）、音节和音素三种，具体选择哪一种，由具体的研究任务决定。

单词（句）单元广泛应用于中小词汇语音识别系统，但不适合大词汇系统，原因在于模型库太庞大，训练模型任务繁重,模型匹配算法复杂，难以满足实时性要求。

音节单元多见于汉语语音识别，主要因为汉语是单音节结构的语言，而英语是多音节，并且汉语虽然有大约1300个音节，但若不考虑声调，约有408个无调音节,数量相对较少。

因此，对于中、大词汇量汉语语音识别系统来说，以音节为识别单元基本是可行的。

音素单元以前多见于英语语音识别的研究中，但目前中、大词汇量汉语语音识别系统也在越来越多地采用。原因在于汉语音节仅由声母（包括零声母有 22 个）和韵母（共有 28 个）构成，且声韵母声学特性相差很大。实际应用中常把声母依后续韵母的不同而构成细化声母，这样虽然增加了模型数目，但提高了易混淆音节的区分能力。由于协同发音的影响，音素单元不稳定，所以如何获得稳定的音素单元，还有待研究。

语音识别一个根本的问题是合理的选用语音特征参数。特征参数提取的目的是对语音信号进行分析处理，去掉与语音识别无关的冗余信息，获得影响语音识别的关键信息，同时对语音信号进行压缩。在实际应用中，语音信号的压缩率介于 10-100 之间。语音信号包含了大量各种不同的信息（比如除语义外，还有说话人所处的环境背景声、情绪等），提取哪些信息，用哪种方式提取，需要综合考虑各方面的因素，如成本，性能，响应时间，计算量等。非特定人语音识别系统一般侧重提取反映语义的特征参数，尽量去除说话人的个人信息；而特定人语音识别系统则希望在提取反映语义的特征参数的同时，尽量也包含说话人的个人信息。

线性预测（LP）分析技术是目前应用广泛的特征参数提取技术，许多成功的应用系统都采用基于 LP 技术提取的倒谱参数。但线性预测模型是纯数学模型，没有考虑人类听觉系统对语音的处理特点。

Mel 参数和基于感知线性预测（PLP）分析提取的感知线性预测倒谱，在一定程度上模拟了人耳对语音的处理特点，应用了人耳听觉感知方面的一些研究成果。实验证明，采用这种技术，语音识别系统的性能有一定提高。从目前使用的情况来看，梅尔刻度式倒频谱参数已逐渐取代原本常用的线性预测编码导出的倒频谱参数，原因是它考虑了人类发声与接收声音的特性，具有更好的鲁棒性。

◆ 声学建模与模式匹配

声学模型通常是将获取的语音特征使用训练算法进行训练后产生。在识别时将输入的语音特征同声学模型（模式）进行匹配与比较，得到最佳的识别结果。

声学模型是识别系统的底层模型，并且是语音识别系统中最关键的一部分。声学模型的目的是提供一种有效的方法计算语音的特征矢量序列和每个发音模板之间的距离。声学模型的设计和语言发音特点密切相关。声学模型单元大小（字发音模型、半音节模型或音素模型）对语音训练数据量大小、系统识别率，以及灵活性有较大的影响。必须根据不同语言的特点、识别系统词汇量的大小决定识别单元的大小。

声学模型是语音识别系统中最为重要的部分之一，目前的主流系统多采用隐马尔可夫

模型进行建模。 隐马尔可夫模型的概念是一个离散时域有限状态自动机，隐马尔可夫模型 HMM 是指这一马尔可夫模型的内部状态外界不可见，外界只能看到各个时刻的输出值。对语音识别系统，输出值通常就是从各个帧计算而得的声学特征。用 HMM 刻画语音信号需作出两个假设，一是内部状态的转移只与上一状态有关，另一是输出值只与当前状态（或当前的状态转移）有关，这两个假设大大降低了模型的复杂度。

基于统计特性的语音识别模型常用的就是 HMM 模型 $\lambda(N,M,\pi,A,B)$，涉及到 HMM 模型的相关理论包括模型的结构选取、模型的初始化、模型参数的重估以及相应的识别算法等。

◆ 语言模型与语言处理

语言模型包括由识别语音命令构成的语法网络或由统计方法构成的语言模型，语言处理可以进行语法、语义分析。这是语音识别系统的后处理阶段，也是对前面识别结果根据语言规则进行修正的阶段。

语言模型对中、大词汇量的语音识别系统特别重要。当分类发生错误时可以根据语言学模型、语法结构、语义学进行判断纠正，特别是一些同音字则必须通过上下文结构才能确定词义。语言学理论包括语义结构、语法规则、语言的数学描述模型等有关方面。目前比较成功的语言模型通常是采用统计语法的语言模型与基于规则语法结构命令语言模型。语法结构可以限定不同词之间的相互连接关系，减少了识别系统的搜索空间，这有利于提高系统的识别率。

1.5 语音识别技术目前所面临的问题

1、就模型方面而言，需要有进一步的突破。目前现有的各种模型都还存在一些明显不足，尤其在中文语音识别方面，语言模型还有待完善，因为语言模型和声学模型正是听写识别的基础，这方面没有突破，语音识别的进展就只能是一句空话。目前使用的语言模型只是一种概率模型，还没有用到以语言学为基础的文法模型，而要使计算机确实理解人类的语言，就必须在这一点上取得进展，这是一个相当艰苦的工作。此外，随着硬件资源的不断发展，一些核心算法如特征提取、搜索算法或者自适应算法将有可能进一步改进。可以相信，半导体和软件技术的共同进步将为语音识别技术的基础性工作带来福音。

2、就自适应方面而言，语音识别技术也有待进一步改进。目前，像 IBM 的 ViaVoice 和 Asiaworks 的 SPK 都需要用户在使用前进行几百句话的训练，以让计算机适应你的声音特征。这必然限制了语音识别技术的进一步应用，大量的训练不仅让用户感到厌烦，而且加大

了系统的负担。并且,不能指望将来的消费电子应用产品也针对单个消费者进行训练。因此,必须在自适应方面有进一步的提高,做到不受特定人、口音或者方言的影响,这实际上也意味着对语言模型的进一步改进。现实世界的用户类型是多种多样的,就声音特征来讲有男音、女音和童音的区别,此外,许多人的发音离标准发音差距甚远,这就涉及到对口音或方言的处理。如果语音识别能做到自动适应大多数人的声音特征,那可能比单纯提高某个特定人一二个百分点识别率更重要。事实上,ViaVoice 的应用前景也因为这一点打了折扣,只有普通话说得很好的用户才可以在其中文版连续语音识别方面取得相对满意的效果。

3、就强健性方面而言,语音识别技术需要能排除各种环境因素的影响。目前,对语音识别效果影响最大的就是环境杂音或噪音,在公共场合,你几乎不可能指望基于目前识别技术的计算机能听懂你的话,来自四面八方的声音让它茫然而不知所措。很显然这极大地限制了语音技术的应用范围,目前,要在嘈杂环境中使用语音识别技术必须有特殊的抗噪麦克风才能进行,这对多数用户来说是不现实的。在公共场合中,人类能有意识地摒弃环境噪音并从中获取自己所需要的特定声音,如何让语音识别技术也能达成这一点呢?这的确是一个艰巨的任务。

此外,带宽问题也可能影响语音的有效传送,在速率低于 1000 比特/秒的极低比特率下,语音编码的采用将使得输入语音大大有别于正常情况,比如要在某些带宽特别窄的信道上传输语音,以及水声通信、地下通信、战略及保密话音通信等,要在这些情况下实现有效的语音识别,就必须处理声音信号的特殊特征,如因为带宽而延迟或减损等。语音识别技术要进一步应用,就必须在强健性方面有更大的突破。

4、多语言混合识别以及无限词汇识别方面:简单地说,目前使用的声学模型和语音模型太过于局限,以至用户只能使用特定语音进行特定词汇的识别。如果突然从中文转为英文,或者法文、俄文,计算机就会不知如何反应,而给出一堆不知所云的句子;或者用户偶尔使用了某个专门领域的专业术语,如“信噪比”等,可能也会得到奇怪的反应。这一方面是由于模型的局限,另一方面也受限于硬件资源。随着两方面的技术的进步,将来的语音和声学模型可能会做到将多种语言混合纳入,系统因此就可以不必在语种之间来回切换。此外,对于声学模型的进一步改进,以及以语义学为基础的语言模型的改进,也能帮助用户尽可能少或不受词汇的影响,从而可实行无限词汇识别。

5、多语种交流系统的应用:最终,语音识别是要进一步拓展我们的交流空间,让我们能更加自由地面对这个世界。可以想见,如果语音识别技术在上述几个方面确实取得了突破性进展,那么多语种交流系统的出现就是顺理成章的事情,这将是语音识技术、机器翻译

技术以及语音合成技术的完美结合，而如果硬件技术的发展能将这些算法进而固化到更为细小的芯片中，比如手持移动设备上，那么个人就可以带着这种设备周游世界而无需担心任何交流的困难，你说出你想表达的意思，手持设备同时识别并将它翻译成对方的语言，然后合成并发送出去；同时接听对方的语言，识别并翻译成己方的语言，合成后朗读给你听，所有这一切几乎都是同时进行的，只是机器充当着主角，如此的世界将再无语言障碍。

任何技术的进步都是为了更进一步拓展我们人类的生存和交流空间，以使我们获得更大的自由，就服务于人类而言，这一点显然也是语音识别技术的发展方向，而为了达成这一点，它还需要在上述几个方面取得突破性进展，最终，多语种自由交流系统将带给我们全新的生活空间，同时带来社会的巨大变革。

1.6 语音识别技术的前景和应用

在电话与通信系统中，智能语音接口正在把电话机从一个单纯的服务工具变成为一个服务的“提供者”和生活“伙伴”；使用电话与通信网络，人们可以通过语音命令方便地从远端的数据库系统中查询与提取有关的信息；随着计算机的小型化，键盘已经成为移动平台的一个很大障碍，想象一下如果手机仅仅只有一个手表那么大，再用键盘进行拨号操作已经是不可能的。语音识别正逐步成为信息技术中人机接口的关键技术，语音识别技术与语音合成技术结合使人们能够甩掉键盘，通过语音命令进行操作。语音技术的应用已经成为一个具有竞争性的新兴高技术产业。

语音识别技术发展到今天，特别是中小词汇量非特定人语音识别系统识别精度已经大于98%，对特定人语音识别系统的识别精度就更高。这些技术已经能够基本满足通常应用的要求。由于大规模集成电路技术的发展，这些复杂的语音识别系统也已经完全可以制成专用芯片，大量生产。在西方经济发达国家，大量的语音识别产品已经进入市场和服务领域。一些用户交机、电话机、手机已经包含了语音识别拨号功能，还有语音记事本、语音智能玩具等产品也包括语音识别与语音合成功能。人们可以通过电话网络用语音识别口语对话系统查询有关的机票、旅游、银行信息，并且取得很好的结果。调查统计表明多达85%以上的人对语音识别的信息查询服务系统的性能表示满意。

可以预测在近五到十年内，随着技术的进一步进步，语音识别系统的应用将更加广泛。各种各样的语音识别系统产品将出现在市场上。人们也将调整自己的说话方式以适应各种各样的识别系统。在短期内还不可能造出具有和人相比拟的语音识别系统，要建成这样一个系统仍然是人类面临的一个大的挑战，我们只能一步步朝着改进语音识别系统性能的方

向一步步地前进。至于什么时候可以建立一个像人一样完善的语音识别系统则是很难预测的。就像在60年代,谁又能预测今天超大规模集成电路技术会对我们的社会产生这么大的影响。

本文针对非特定人、大词汇量、连续语音识别技术展开研究，内容安排如下：第一章介绍语音识别技术发展的概况和历史现状，并给出识别系统的分类和结构以及应用前景；第二章是识别技术的理论核心，详细描述隐马尔科夫模型；第三章介绍语音识别中涉及的信号处理技术、语音数据库、HTK 工具箱；第四章是识别系统的实现，使用 HTK 搭建大词汇量连续语音识别系统，也是本论文的核心，其中三音素建模和决策树问题集设计是论文的创新点；第五章在第四章的基础上进行了完备的测试对比和结果分析；最后一章是总结和展望。

第2章 隐马尔可夫模型

隐马尔可夫模型（Hidden Markov Models，HMM），作为语音信号的一种统计模型，今天正在语音处理各个领域中获得广泛的应用，也是目前最成功的一种模型，目前各种具有优良性能的语音识别系统几乎都采用了这种模型[12]。有关它的理论，是在 1970 年前后由 Baum 等人建立起来的，随后由 CMU 的 Baker 和 IBM 的 Jelinek 等人将其应用到语音识别中[13]。20 世纪八十年代中期，经过 Bell 实验室 Rabiner 等人对 HMM 深入浅出的介绍，才逐渐使 HMM 为世界各国从事语音处理的人员所了解和熟悉，进而成为公认的研究热点。

2.1 隐马尔可夫模型的定义

2.1.1 信号模型

隐马尔可夫过程描述的是一个双重随机过程：一重用于描述非平稳信号的短时平稳段的统计特征（信号的瞬态特征，可直接观测）；另一重随机过程描述了每个短时平稳段如何转变到下一个短时平稳段，即短时统计特征的动态特性（隐含在观察序列中）。基于这两重随机过程，HMM 即可有效解决怎样辨识具有不同参数的短时平稳信号段，怎样跟踪它们之间的转化等问题。

人的语言过程也是这样一种双重随机过程。因为语言本身是一个可观察的序列，而它又是由大脑里的（不可观察的）、根据言语需要和语法知识（状态选择）所发出的音素（词、句）的参数流。同时，大量的实验证明，HMM 的确可以比较精确的描述语音信号的产生过程。

2.1.2 隐马尔可夫模型的数学描述

通常一个 HMM 可由下面的参数来描述[14,15]：

1) N，隐马尔可夫模型中的状态数。虽然在 HMM 中状态数是隐含的，但在实际应用中它是有确切的物理含义的。本文中，标记模型中的各个状态为 $\{1,2,...,N\}$，在 t 时刻所处的状态为 q_t。

2) M，每个状态中可以观察到的符号数。标记各个观察符号为 $V=\{v_1,v_2,...,v_M\}$，观

察序列为$O=\{o_1,o_2,...,o_T\}$，其中o_t为集合V中的一种观察符号，T为观察序列的长度。

3) 状态转移概率分布$A=[a_{ij}]$，其中

$$a_{ij}=P[q_{t+1}=j\,|\,q_t=i]\quad 1\le i\le N,1\le j\le N \tag{2-1}$$

4) 某个状态下观察符号的概率分布$B=[b_j(k)]$，其中

$$b_j(k)=P(o_t=v_k\,|\,q_t=j)\quad 1\le k\le K,1\le j\le N \tag{2-2}$$

5) 初始状态概率分布$\pi=[\pi_i]$，其中

$$\pi_i=P[q_1=i]\quad 1\le i\le N \tag{2-3}$$

基于这些特征参数，HMM产生观察序列$O=\{o_1,o_2,...,o_T\}$的过程可以描述如下：

1) 根据初始状态概率分布π，选择一个初始状态$q_i=i$。

2) 置观察时间$t=1$。

3) 根据当前状态下观察符号的概率分布B，选择$o_t=v_k$。

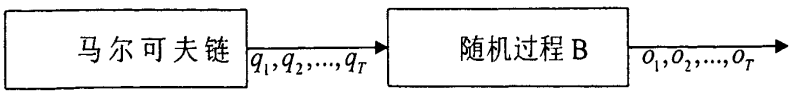
4) 根据状态转移概率分布A，从当前状态$q_i=i$转移到下一个状态$q_{t+1}=j$。

5) 置$t=t+1$，如果$t\prec T$，回到第（3）步，否则结束。

综上所述，一个HMM完全可以由2个模型参数N，M和3个概率分布参数A,B,π来确定。为了方便起见，通常将隐马尔可夫模型定义为

$$\lambda=(A,B,\pi)$$

更形象的说，HMM可以分为两个部分，一个是Markov链，由π,A描述，其产生的输出为状态序列；另一个随机过程B，产生的输出为观测序列，T为观测时间长度。



2.2 HMM的三个基本问题

要使所建立的隐马尔可夫模型能够解决实际问题，以下3个问题必须加以解决[16]。

1) 已知观察序列O和模型$\lambda=(A,B,\pi)$，如何计算由此模型产生此观察序列的概率$P(O|\lambda)$。

2) 已知观察序列O和模型$\lambda=(A,B,\pi)$，如何确定一个合理的状态序列，使之能最佳的产生O，即如何选择最佳的状态序列$q=\{q_1,q_2,...,q_T\}$？

3) 如何根据观察序列不断修正模型参数(A,B,π)，使$P(O|\lambda)$最大？

问题（1）实质上是一个模型评估问题，因为$P(O|\lambda)$反映了观察序列和模型吻合的程度。在语音识别中，我们可以通过计算、比较$P(O|\lambda)$，从多个模型参数中选择出与观察序列匹配得最好的那个模型。

问题（2）关键在于选用怎样的最佳状态准则来决定状态的转移。一种可能的最佳准则是：选择状态q_t^*，使它们在各t时刻都是最可能的状态，即

$$q_t^* = \arg\max_{1\le i\le N}[P(q_1=i|O,\lambda)] \tag{2-4}$$

这里存在一个问题：有时候会出现不允许的转移，即$a_{ij}=0$，那么，对这些i和j所得到的状态序列就是不可能状态序列。所以上述公式得到的解只是在每个时刻决定一个最可能的状态，而没有考虑整体结构、相邻时间的状态和观察序列的长度等问题。

问题（3）实质上就是如何训练模型，估计、优化模型参数的问题。这个问题在3个问题中最难，因为没有解析法可用来求解最大似然模型，只能使用迭代法（Baum-Welch 算法）或使用最佳梯度法。

2.3 HMM 的基本算法

前面提到了 HMM 的三个基本问题，针对实际应用，下面阐述解决这些问题的算法。

2.3.1 前向后向算法

在问题（1）中，给定模型参数$\lambda=(A,B,\pi)$和观察序列O，求解$P(O|\lambda)$。从定义出发计算概率$P(O|\lambda)$，可以得到下式：

$$P(O|\lambda)=\sum_q P(O|q,\lambda)P(q|\lambda)=\sum_{q_1,q_2\cdots q_T}\pi_{q_1}b_{q_1}(o_1)a_{q_1q_2}b_{q_2}(o_2)...a_{q_{T-1}q_T}b_{q_T}(o_T) \tag{2-5}$$

显然，按照公式（2—5）计算$P(O|\lambda)$是不现实的，因为它的计算量相当大。为了有效地解决这个问题，引入前向概率和后向概率来简化运算。它们的定义及有关的递推公式如下。

前向概率定义为

$$\alpha_t(i)=P(o_1o_2...o_t,q_t=i|\lambda) \tag{2-6}$$

即是在给定模型λ下，前t个时刻的观察序列为$\{o_1,o_2,...,o_t\}$，且在t时刻处在状态i的概率。计算公式如下：

$$\alpha_1(i)=\pi_i b_i(o_1),1\le i\le N \tag{2-7}$$

$$\alpha_{t+1}(j)=[\sum_{i=1}^{N}\alpha_t(i)a_{ij}]\bullet b_j(o_{t+1}),1\le t\le T-1,1\le j\le N \tag{2-8}$$

后向概率定义为：

$$\beta_t(i)=P(o_{t+1}o_{t+2}...o_T,q_t=i|\lambda) \tag{2-9}$$

即是在给定模型λ下，从$t+1$时刻开始到观察结束这一段的观察序列为$\{o_{t+1}o_{t+2}...o_T\}$，且在t时刻处在状态i的概率。计算公式如下：

$$\beta_T(i)=1,1\le i\le N \tag{2-10}$$

$$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j),t=T-1,T-2,...,1;1\le i\le N \tag{2-11}$$

根据前向及后向概率的定义可以推导出

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j),1\le t\le T-1 \tag{2-12}$$

或

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{2-13}$$

2.3.2 Viterbi 算法

这个算法解决了给定一个观察序列$\{o_{t+1}o_{t+2}...o_T\}$和 HMM 模型$\lambda=(A,B,\pi)$，在最佳的意义上确定一个状态序列$q_1^*q_2^*...q_T^*$的问题，而如何确定一个最佳状态序列的关键在于选用

怎样的最佳准则。考虑到状态序列的整体特性，Viterbi 算法采用如下的最佳准则。

$$\delta_t(i) = \max_{q_1 q_2 \cdots q_{t-1}} P[q_1 q_2 \ldots q_{t-1}, q_t = i, o_1 o_2 \ldots o_t \mid \lambda] \tag{2-14}$$

即在 t 时刻选择状态 i，使模型 λ 沿状态序列 $\{q_1 q_2 \ldots q_t\}$ 运动产生观察序列 $\{o_1 o_2 \ldots o_t\}$ 的概率最大。根据定义，可以得到 $\delta_t(i)$ 的递推计算公式。

$$\delta_{t+1}(j) = \max_i [\delta_t(i) a_{ij}] * b_j(0_{t+1}) \tag{2-15}$$

$$\psi_{t+1}(j) = \arg\max_i [\delta_t(i) a_{ij}] \tag{2-16}$$

其中 $\psi_{t+1}(j)$ 的物理含义是若 $t+1$ 时刻的最佳状态是 j，则 t 时刻的最佳状态是 $\psi_t(j)$。

基于此最佳准则，我们可以通过以下步骤递推得到最佳状态序列 $q^* = \{q_1^*, q_2^*, \ldots, q_T^*\}$ 和 $P[O, q^* \mid \lambda]$（在模型 λ 下，按照最佳状态序列路径产生观察序列 O 的概率）。

步骤 1：

$$\begin{aligned} &\delta_1(t) = \pi_i b_i(o_1), 1 \le i \le N \\ &\psi_1(i) = 0 \end{aligned} \tag{2-17}$$

步骤 2：

$$\begin{aligned} &\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t), 2 \le t \le T, 1 \le j \le N \\ &\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], 2 \le t \le T, 1 \le j \le N \end{aligned} \tag{2-18}$$

步骤 3：

$$\begin{aligned} &P[O, q^* \mid \lambda] = \max_{1 \le i \le N} [\delta_T(i)] \\ &q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)] \end{aligned} \tag{2-19}$$

步骤 4：

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \ldots, 1 \tag{2-20}$$

通过以上步骤我们可以求出各个时刻系统所处的状态，得到一条唯一的最佳状态转移路径，使得观测序列的概率最大。

2.3.3 Baum-Welch 算法[17,18]

Baum-Welch 算法用于解决 HMM 的训练和参数重估。给定一个观测序列 $O = o_1 o_2 \ldots o_T$，该算法能够确定一个 $\lambda = (A, B, \pi)$，使得 $P(O \mid \lambda)$ 最大。

公式 2-12 给出了 $P(O \mid \lambda)$ 的计算式

$$P(O|\lambda)=\sum_{q}P(O|q,\lambda)P(q|\lambda)=\sum_{q_1,q_2\cdots q_T}\pi_{q_1}b_{q_1}(o_1)a_{q_1q_2}b_{q_2}(o_2)\ldots a_{q_{T-1}q_T}b_{q_T}(o_T) \tag{2-21}$$

通过该式求取 λ 使得 $P(O|\lambda)$ 最大是一个泛函极值问题，计算量巨大，加上训练数据的有限，事实上没有一个最佳的方法来估计 λ。Baum-Welch 算法则利用递归的思想，使得 $P(O|\lambda)$ 达到局部最大，最后得到模型的估计参数 $\lambda=(A,B,\pi)$。

首先引入两个相关的概率定义 $\varepsilon_t(i,j)$ 和 $\gamma_t(i)$。

$\varepsilon_t(i,j)$ 表示在已知观察序列 O 和模型 λ 的情况下，t 时刻处于状态 i，$t+1$ 时刻处于状态 j 的概率，$\varepsilon_t(i,j)=P(q_t=i,q_{t+1}=j|O,\lambda)$。通常 $\varepsilon_t(i,j)$ 采用归一化形式，即

$$\varepsilon_t(i,j)=\frac{P(q_t=i,q_{t+1}=j,O|\lambda)}{P(O|\lambda)}=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \tag{2-22}$$

$\gamma_t(i)$ 为给定观察序列 O 和模型 λ 的条件下，t 时刻处于状态 i 的概率。根据定义有

$$\gamma_t(i)=\frac{P(O,q_t=i|\lambda)}{P(O|\lambda)}=\sum_{j=1}^{N}\varepsilon_t(i,j)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{2-23}$$

通过对 $\gamma_t(i)$ 的时间 t 求和，可以得到观察序列中从状态 i 出发的状态转移次数的期望，而对 $\varepsilon_t(i,j)$ 的时间 t 求和，可以得到观察序列中，从状态 i 转移到状态 j 的状态转移次数的期望。综上，Baum-Welch 算法导出的重估公式为：

步骤 1：

在 $t=1$ 时刻处于状态 i 的概率

$$\bar{\pi}_i=\frac{P(O,q_0=i|\lambda)}{P(O|\lambda)}=\gamma_0(i)=\frac{\alpha_0(i)\beta_0(i)}{\sum_{j=1}^{N}\alpha_T(j)} \tag{2-24}$$

步骤 2：

参数 A 的重估：

$$\bar{a}_{ij}=\frac{\sum_{t=1}^{T}\varepsilon_{t-1}(i,j)}{\sum_{t=1}^{T}\gamma_{t-1}(i)} \tag{2-25}$$

步骤 3：

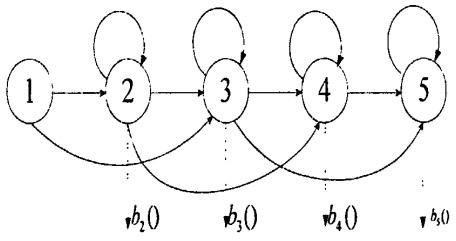
参数 B 的重估：

$$\bar{b}_i(k)=\frac{\sum_{t=1且o_t=v_k}^{T}\gamma_t(i)}{\sum_{t=1}^{T}\lambda_t(i)} \tag{2-26}$$

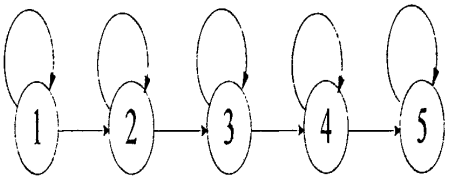
Baum-Welch 算法的基本思想是：按照某种参数重估公式从现有的模型λ'估计出新模型λ，使得$P(O|\lambda')\le P(O|\lambda)$，用$\lambda$代替$\lambda'$，重复上述过程直到模型参数处于收敛状态，即得到了最大似然模型。

2.4 隐马尔可夫模型的类型

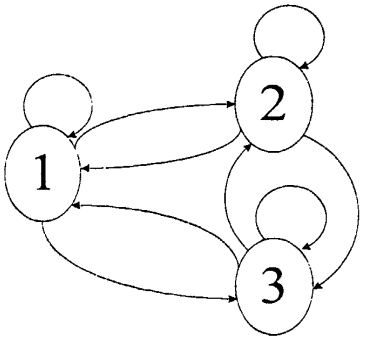
HMM 由两部分组成，其一为马尔可夫链，由π，A描述，显然，不同的π，A决定了马尔可夫链不同的形状。常用的有如下几种：



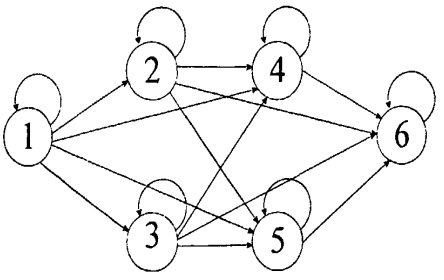
(a) 5 状态由左至右有跨越的马尔可夫链



(b)5 状态由左至右无跨越的马尔可夫链



(c)全连接模型



(d)并行由左向右模型

图 2. 1 几种典型的马尔可夫链示意图

有跨越由左至右模型中，状态转移只能从左至右，而不能反过来，这种模型很适合对语音信号建模，因为语音信号的性质随时间变化。后面实验中训练时使用的模型便是 5 状态、有跨越、从左至右的模型。不同结构的 HMM 模型，各有自己的应用领域。例如全连接的 HMM 可以用于说话人识别；无跨越式从左向右模型符合人的语音特点，可以用来进

行语音识别；而有跨越从左至右模型，其中允许各状态位跳转意味着语音中某些发音在说话中可能被吸收或删除的实际情况；而并行从左至右模型则包含了发同一个语音单位可能出现的变音现象。

实际应用中，为了准确描述模式的状态变化以及构筑更为复杂的模型，常常采用一些拓扑结构的混合。在本实验中，采用 5 状态的 HMM 模型，其中的 1、5 状态分别对应入口状态和出口状态，且均为非发射状态（无输出分布），中间的 2、3、4 状态是由左至右有跳转的，入口状态只能转出，出口状态只能到达，这种结构可以连接成 HMM 序列。

第3章 基于HMM的语音识别系统

3.1 语音信号处理和分析

3.1.1 语音信号的数字化和预处理

语音信号的数字化是数字处理的前提，语音信号的数字化包括两个步骤：采样和量化。Nyquist采样定理要求采样率必须大于或等于信号带宽的2倍，因此一般需要对输入的语音信号作低通（抗混叠）滤波，然后进行A/D转换，如图3-1所示

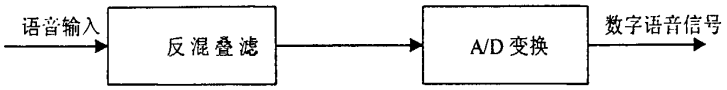


图3-1 A/D变换

预处理是指对数字化后的语音信号的特殊处理：消噪（如果需要的话）、预加重或高频提升，分帧处理等[19]。

由于语音信号的平均功率谱受声门激励和口鼻辐射的影响，高频段大约在800Hz以上按6dB/倍频程跌落，信号过弱，容易在后续处理中失真，为此要在预处理中进行预加重。预加重的目的是提升高频部分，使信号的频谱变得平坦，以便于进行频谱分析或声道参数分析。预加重通过一阶预加重数字滤波器实现，即

$$H(z)=1-\mu\cdot z^{-1} \tag{3-1}$$

式中μ的值接近于1，本文中μ的值为0.97。

由于语音信号是非平稳过程，但可认为它是局部或短时平稳。因此，语音信号分析常分段或分帧处理，一般每秒的帧数约为33~100（对应帧长10ms到30ms），视实际情况而定，分帧既可用前后帧头尾连续的，也可用交叠分段的方法，在语音信号分析中常用“短时分析”表述。

短时分析实质上是用一个窗截取信号。数字信号处理理论告诉我们，两个信号的时域相乘，在频域相卷积，如果采用矩形窗，则矩形窗频谱高频成分必将影响语音信号的高频部分，一般用高频分量幅度较小的窗，以减小这些影响。汉明（Hamming）窗的带宽是矩形窗的两倍，但带外衰减却比矩形窗大得多，因此在信号处理中经常使用。本文中所选取

的窗为汉明窗。

3.1.2 语音信号的时域、频域分析

语音信号包含丰富的信息：时域、频域、声道参数等。根据应用的不同，提取的信息侧重点也有所区别。语音信号分析大概有以下三种方法[20]：

◆ 时域方法

◆ 频域方法

◆ 倒谱域方法

语音信号时域分析的常见参数有：短时能量和短时平均幅度、短时平均过零率、短时自相关和平局幅度差。

a) 短时平均能量

信号流的分帧是采用可移动的有限长度的窗口进行加权的方法来实现，如图 3-2 所示：

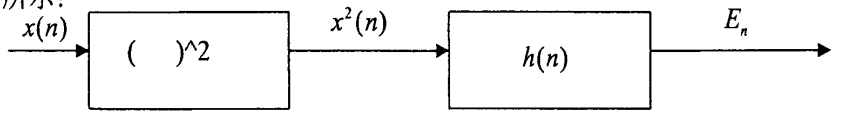


图 3-2 分帧

定义矩形窗函数为：

$$w(n)=\begin{cases}1, 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{3-2}$$

那么，以 n 为标志的某帧语音信号的短时平均能量（Short Time Average Energy）E_n 如下式所示：

$$E_n=\sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2=\sum_{m=n-N+1}^{n}[x(m)w(n-m)]^2 \tag{3-3}$$

根据图 3-2 和公式 3-2，短时能量有两种解释：

◆ 由图 3-2 可知，首先计算原始语音信号各个采样值的平方，然后通过一个冲激响应为 $h(n)$ 的滤波器，最后输出能量序列，这里 $h(n)=w(n)^2$。

◆ 由公式 3-2，首先计算原始语音信号各个采样值的平方，然后用一个移动窗 $h(n-m)$ 选取出一个一个短时平方序列，并将各短段的平方值求和得到短时能量序列。

不同的窗函数的选择将决定短时平均能量的性质。选取的原则是：使得短时能量既能及时跟踪语音能量的缓变规律，同时又要对语音振幅一个基音周期内的瞬时变化有显著平

滑的作用。

b) 短时平均幅度

短时平均能量E_n对于高电平信号，其平方处理方式显得过于灵敏，在处理器字长有限的情况下，容易产生溢出，对于这种情况，可以采用另一种度量语音信号幅度变化的参量，短时平均幅度（Short Time Average Magnitude），定义如下：

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|\, w(n-m) = |x(n)| * w(n) \tag{3-4}$$

c) 短时平均过零率

信号的幅度值从正值到负值要经过零值，从负值到正值也要经过零值，称其为过零，统计在一秒钟内有几次过零，就称为过零率。如果信号按段分割，就称为短时，把各段信号的过零率作统计平均，就是短时平均过零率（Short Time Average Cross Zero Ratio）。

语音信号序列$x(n)$的短时平均过零率z_n定义如下：

$$z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| = |\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]| * w(n) \tag{3-5}$$

公式中 sgn[]是符号函数：

$$\mathrm{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ 0, x(n) < 0 \end{cases} \tag{3-6}$$

短时平均过零率在语音信号分析中应用最多的是清/浊音判决。一般情况下，浊音短时平均过零率为 14 过零/10ms，清音短时过零率的均值为 47 过零/10ms。

d) 短时自相关函数

语音信号短时自相关函数（Short Time Autocorrelation Function）定义：

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-m-k) \tag{3-7}$$

它具有自相关函数所有的性质，由于自相关函数是偶函数，故可以改写为：

$$\begin{aligned} R_n(k) &= R_n(-k) = \sum_{m=-\infty}^{\infty} [x(m)x(m-k)][w(n-m)w(n-m+k)] \\ &= \sum_{m=-\infty}^{\infty} [x(m)x(m-k)]h_k(n-m) = [x(n)x(n-k)] * h_k(n) \end{aligned} \tag{3-8}$$

即自相关函数可以理解为序列$[x(n)x(n-k)]$通过一个冲激响应为$h_k(n)$的数字滤波器的输出，$h_k(n) = w(n)w(n+k)$。

e) 短时平均幅度函数

短时自相关函数是语音信号时域分析的一个重要参量，但是运算量很大，尽管可以利用FFT等快速算法，但在一些场合实现仍是困难的。短时平均幅度差函数AMDF（Short Time Average Magnitude Difference Function）与短时自相关函数有类似的功效，但运算量小很多，所以在语音信号处理中也得到广泛的应用。

$$F_n(k)=\frac{1}{R}\sum_{m=-\infty}^{\infty}|x(n+M)w_1(m)-x(n+m+k)w_2(m+k)| \tag{3-9}$$

语音信号的频域分析

语音的频域分析在语音信号分析处理中具有极其重要的意义，在语音识别中，许多特征参数的提取都是建立在对语音信号的频域分析的基础上。语音信号的频谱对外界的干扰具有一定健壮性，而时域特征参数容易随外界条件变化而发生改变。此外，频谱还具有声学特性，这对语音内容识别或者说话人识别具有重要的意义。

通常，我们需要对每帧语音信号提取特征参数，有 LPC、LPCC、MFCC 等。假设每帧语音信号有 N 个采样点，那么对该帧信号的短时傅里叶变换（DFT）和短时傅里叶反变化（IDFT）为：

$$\text{DFT：}\quad X(k)=\sum_{n=0}^{N-1}x(n)\exp(-j2\pi n/N),(0\le k\le N-1) \tag{3-10}$$

$$\text{IDFT：}\quad x(n)=\frac{1}{N}\sum_{k=0}^{N-1}X(k)\exp(j2\pi n/N),(0\le k\le N-1) \tag{3-11}$$

考虑到运算量，实际中我们通常采用快速傅里叶变换（FFT）提高运算速度。

语音信号的倒谱域分析

倒谱的定义为短时幅度谱的对数傅里叶反变换

$$c_j(n)=IDFT(\lg|X_j(k)|),0\le n\le N-1 \tag{3-12}$$

这里的 j 表示帧号，N 为每帧的样点数，$c_j(n)$ 为第 j 帧信号的倒谱系数。倒谱可以理解为将谱线作傅里叶反变换。频谱和倒谱都是语音识别中非常重要的特征。语音产生的模型可以认为是周期脉冲序列作为激励源的线性滤波器，在语音识别中对信号的处理都是分帧的，在一帧内，而在一帧内可以认为滤波器是时不变的。倒谱的求解过程如下图示：

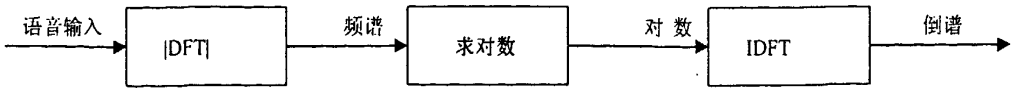


图 3-3 倒谱求解图

频谱的包络特性反映了语音的音韵特性，而频谱的细微结构反映了音源的基本频率，

通过倒谱分析，可以把频谱的包络成分及其细微的结构区分开来，所以在语音识别中，通常选择到谱系数作为语音信号的特征参数，在本文后面的试验中，着重比较了 LPC、LPCC、MFCC 等不同特征参数的选取对系统识别率的影响。

3.2 线性预测及特征参数选取

3.2.1 线性预测信号模型

在随机信号谱分析中，常把一个时间序列模型转化为白噪声序列通过一个数字滤波器的 H(z)输出。在一般情况下，H(z)可以写成有理分式的形式：

$$H(z)=G\frac{1+\sum_{l=1}^{q}b_l z^{-l}}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{3-13}$$

式中 a_i，b_l 以及增益因子 G 就是模型参数，因此，信号可以用有限个数的参数构成的信号模型来表示，如下图示：

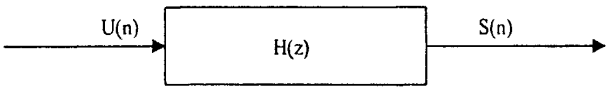


图 3-4 数字滤波器

假设被模型化的信号为 s(n)，模型输入为 u(n)，它的 Z 变换分别为 S(z)和 U(z)，那么，

$$S(z)=H(z)\cdot U(z) \tag{3-14}$$

从时间域上看，信号模型的输出与输入满足下面的差分方程

$$s(n)=\sum_{i=1}^{p}a_i s(n-i)+G\sum_{l=0}^{q}b_l u(n-l),b_0=1 \tag{3-15}$$

上式表明，s(n)可以模型化为它的 p 个过去值 s(n-i)和输入 u(n)及其 q 个过去值 u(n-l)的线性组合；从物理意义上讲就是 s(n)可由过去值及输入信号值的线性组合来预测，所以信号模型化和线性预测有内在的联系。

当输入 u(n)为零均值的随机信号时，系统的输出与输入关系可用相关函数或功率谱关系来表征：

$$R_{ss}(z)=H(z)H(z^{-1})R_{uu}(z) \tag{3-16}$$

$R_{ss}(z)$ 和 $R_{uu}(z)$ 分别表示信号 s(n)和输入 u(n)自相关序列的 Z 变换。在信号模型中，u(n)

是均值为零、方差为σ_u^2的白噪声序列，其自相关$R_{uu}(n)=\sigma_u^2\delta(n)$，所以$R_{uu}(z)=\sigma_u^2$，从而$R_{ss}(z)=H(z)H(z^{-1})\sigma_u^2$，写成功率谱的形式有：

$$|S(e^{jw})^2|=|H(e^{jw})|^2 \tag{3-17}$$

上式假设σ_u^2=1。这表明，信号 s(n)的功率谱$|S(e^{jw})^2|$完全由滤波器的幅度频率响应来决定。按照数字滤波器 H(z)有理式的不同，有以下 3 种信号模型：

◆ 自回归信号模型（AR 模型），此时 H(z)是只含递归结构的全极点模型，由它产生的序列称为 AR 过程序列。

◆ 滑动平均模型（MA 模型），此时 H(z)是只含非递归结构的全零点模型，由它产生的序列称为 MA 过程序列。

◆ 自回归滑动平均模型（ARMA 模型），此时 H(z)含有极点和零点，是上述两种模型的混合结构，相应产生的序列称为 ARMA 过程序列。

理论上讲，ARMA 模型和 MA 模型可以用无限高阶的 AR 模型来表达。对 AR 模型作参数估计时遇到的是线性方程组的求解问题，有多种求解方法。

3.2.2 线性预测误差滤波

信号模型的逼近过程本质上是个线性预测误差滤波问题，线性预测误差滤波是一种特殊的数字滤波，它的传递函数 A(z)由下式确定：

$$A(z)=1-\sum_{i=1}^{p}a_i(z^{-i}) \tag{3-18}$$

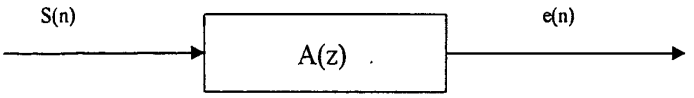


图 3-5 线性预测误差滤波器

如图 3-5 所示，它的输出 e(n)与输入 s(n)满足关系：

$$e(n)=s(n)-\hat{s}(n)=s(n)-\sum_{i=1}^{p}a_i s(n-i) \tag{3-19}$$

式中$\hat{s}(n)=\sum_{i=1}^{p}a_i s(n-i)$称作 s(n)的预测值或估计值，$a_i$为线性预测系数，输出 e(n)为线性预测误差，$p$为滤波器阶数。设计一个预测误差滤波器，就是求解预测系数a_i使得预测

误差 e(n)在某个准则下最小，这个过程称为线性预测过程。求解 a_i 的经典解法有两个：自相关法、协方差法。后来又出现改进的斜格法。

3.2.3 LPCC 参数

线性预测倒谱参数（Linear Prediction Cepstrum Coefficient,LPCC）是线性预测系数（Linear Prediction Coefficient,LPC）在倒谱域中的表示。该特征是基于语音信号为自回归信号的假设，利用线性预测分析获得倒谱系数。LPCC 的优点是计算量小，易于实现，对元音有较好的描述能力，其缺点在于对辅音的描述能力较差，抗噪声性能较差。

利用 LPC 系数，我们可以推导出语音信号的倒谱 c(n)：

$$\begin{aligned} c(1) &= a_1 \\ c(n) &= a_n + \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c(n-k), 1 \prec n \le p \\ c(n) &= \sum_{k=1}^{p}(1-\frac{k}{n})a_k c(n-k), n \succ p \end{aligned} \tag{3-20}$$

LPCC 可由 LPC 递推得到，但同时也继承了 LPC 的缺陷，其中主要的一点就是 LPC 在所有的频率上都是线性逼近语音的，而这与人耳的听觉的特性是不一致的；而且 LPC 包含了语音的高频部分的大部分噪声细节，这些都会影响系统的性能。

3.2.4 MFCC 参数

MFCC（Mel Frequency Cepstrum Coefficient）是一种考虑了人耳的听觉特性的语音特征参数，将频谱转化为基于 Mel 频标的非线性频谱，然后转换到倒谱域上。由于充分考虑了人的听觉特性，而且没有任何前提假设，MFCC 参数具有良好的识别能力和抗噪能力，但其计算量和计算精度要求高。

MFCC 不同于 LPCC，本文在后面的实验也重点比较了这两种不同的特征参数对系统识别率的影响。MFCC 是采用滤波器组的方法计算出来的，这组滤波器在频率的 Mel 坐标上是等带宽的。这是因为人类在对 1000Hz 以上的声音频率范围的感知不遵循线性关系，而是遵循在对数频率坐标上的近似线性关系：

$$F_{mel} = 3322.23\lg(1+0.001)f_{Hz} \quad [21] \tag{3-21}$$

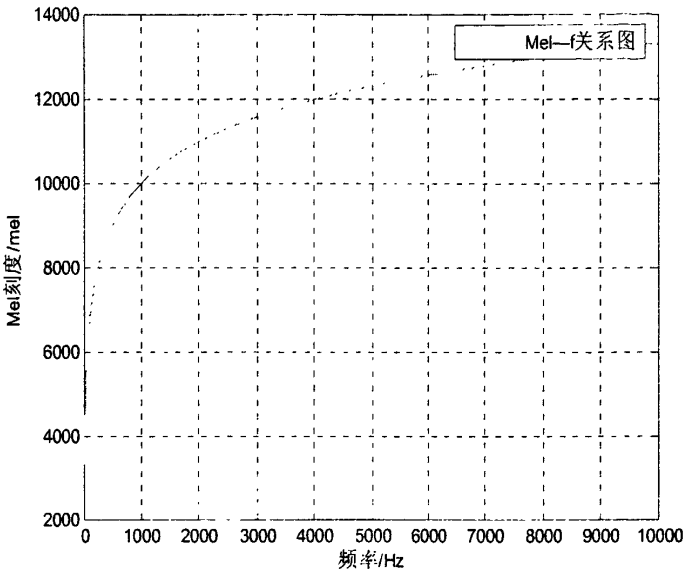


图 3- 6 赫兹频率与美（mel）的关系

MFCC 的计算过程如下图所示，具体计算步骤如下：

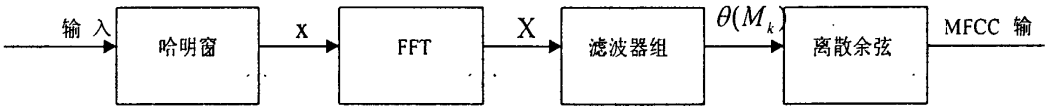


图 3- 7 MFCC 获取过程

◆ 语音信号经过加窗处理后变为短时信号，用 FFT 将这些时域信号 x(n)转化为频域信号 X(m)，并由此计算它的短时能量谱 P(f)。

◆ 将 P(f)由在频率轴上的频谱转化为在 Mel 坐标上的 P(M)，M 表示 Mel 频率，公式 3-21 可完成该转化。

◆ 在 Mel 频域内将三角带通滤波器组加于 Mel 坐标上的 P(M)，然后计算 Mel 坐标上的能量谱 P(M)经过此滤波器组的输出：

$$\theta(M_k)=\ln[\sum_{k=1}^{K}|X(k)|^2 H_m(K)],k=1,2,...,K \tag{3-22}$$

k 表示第 k 个滤波器，K 表示滤波器的个数

◆ 通过一个具有 40 个滤波器（K=40）的滤波器组。前 13 个滤波器在 1000Hz 以下是线性划分的，后 27 个滤波器在 1000Hz 以上是在 Mel 坐标上线性划分的。

◆ 如果 $\theta(M_k)$ 表示第 k 个滤波器的能量输出，则 Mel 频率倒谱 $C_{mel}(n)$ 在 Mel 刻度谱上可以采用离散余弦反变换（IDCT）求得

$$C_{mel}(n)=\sum_{k=1}^{K}\theta(M_k)\cos(n(k-0.5)\frac{\pi}{K}),n=1,2,...,p \tag{3-23}$$

p 为 MFCC 参数的阶数。

3.3 TIMIT 语音数据库

3.3.1 简介

TIMIT 语音库设计的目的是用来获得声学、语音学的知识以及为自动语音识别系统做评估用。该语音库得到美国国防部高级计划研究署信息科学技术办公室（DARPA—ISTO）的支持和赞助。语音库的文本设计由麻省理工学院（MIT）、斯坦福研究院（SRI）和德州仪器（TI）共同完成。录音的地点在 TI，并被 MIT 转录，最后由 NIST（National Institude of Standards and Technology）负责保存、检查、制成 CD。

3.3.2 语音库说话人分布

本文的目的是对非特定人、连续、大词汇量的语音识别技术和系统进行研究和实现。所以在训练过程中一要保证训练的语料涵盖范围尽量广泛，同时要涉及到所有的发音单元。TIMIT 语料库总共有 6300 个句子，这些句子来自 630 位不同的人的录音，其中每个人读 10 个句子；这 630 个人来自 8 个不同的方言区域，方言区域指的是地理区域。不同区域不同性别说话人的分布如下表 3-1 和 3-2 所示[22]。

区域	男性	女性	总共
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
合计	438 (70%)	192 (30%)	630 (100%)

表 3-1 分布图

区域	所属方言区域
1	New England
2	Northern
3	North Midland
4	South Midland
5	Southern
6	New York City
7	Western
8	Army Brat

表 3-2 区域图

3.3.3 语音库文本材料

TIMIT 的文本材料分成三个部分：

1. 由 SRI 设计的用于发掘方言口令的两个句子，标记为 SA
2. 由 MIT 设计的 450 个涵盖完备发音集的句子，标记为 SX
3. 由 TI 设计的 1890 个反映语音学变化性的句子，标记为 SI

其中，用于发掘方言归属地的句子会被所有的人录音一次，每个人还会录音 450 个涵盖完备发音集句子的 5 个、以及反映语音变化的 3 个句子。如下表示：

句子类型	文本句子	说话人次	总共	句子数/每人
SA	2	630	1260	2
SX	450	7	3150	5
SI	1890	1	1890	3
合计	2342		6300	10

表 3-3 TIMIT 语音材料

3.3.4 训练集、测试集划分

TIMIT 是个平衡语料库，这里的平衡是指语料库包含不同的语言风格和文本文字，比如，它可以包括口头和书面风格，公共和私有文本，可作为参考、通用、核心语料库。TIMIT

语料库里面包含了630位不同人的录音，438位男性，192位女性，每人读10个句子，总共包含 6300 个句子的发音。这些句子被划分成两个集合：训练集和测试集。首先介绍划分标准：

1. 大约20%—30%的语音库句子做测试用，剩下70%—80%用于训练
2. 任何一位说话人不能在训练集和测试集中同时出现
3. 训练集和测试集都必须包含所有的方言区域，并且包含男性和女性发音
4. 文本的重叠尽量最小化，理想的情况是没有文本重叠
5. 所有的音素都必须在测试集中出现，最好每个音素在测试集中出现多次

按照上面的5个标准，TIMIT采用了462位人的录音（即4620个句子）作为母语为英文的声学模型训练，训练集总发音时长为3小时49分10秒；完备测试集来自168位人的发音（即1680个句子），时长为1小时23分51秒。一次识别1680个句子通常需要花费好几个小时的时间，所以TIMIT还给出了核心测试集，下面两个表格给出了完备测试集合核心测试集的基本情况[22]：

方言区域	男性编号	女性编号	发音句子数/每人	总共句子数
1	DAB0, WBT0	ELC0	10	30
2	TAS1, WEW0	PAS0	10	30
3	JMP0, LNT0	PKT0	10	30
4	LLL0, TLS0	JLM0	10	30
5	BPM0, KLT0	NLP0	10	30
6	CMJ0, JDH0	MGD0	10	30
7	GRT0, NJM0	DHC0	10	30
8	JLN0, PAM0	MLD0	10	30
合计	16人	8人		240

表3-4 核心测试集

方言区域	男性人数（人）	女性人数（人）	总人数（人）
1	7	4	11
2	18	8	26

3	23	3	26
4	16	16	32
5	17	11	28
6	8	3	11
7	15	8	23
8	8	3	11
合计	112	56	168

表3-5 完备测试集

3.4 HTK工具箱

3.4.1 简介

目前基于HMM模型进行语音识别技术研究最著名的开发平台就是HTK，HTK（Hidden Markov Model Tookkit）是用于自动语音识别的工具箱，由剑桥大学电子工程系语音视觉和机器人研究小组开发，软件是基于C语言结构，可在Unit/Linux和Windows平台上运行[23]。1993年，Entropic Research Laboratory获得HTK的销售权利，并于1995年获得开发权利。1999年，微软买下Entropic，并买下了HTK，然后把授权返还给CUED，这样，HTK得到CUED不断地更新和技术支持。目前HTK被广泛应用于语音识别、语音合成、字符识别和DNA排序等领域[24]。

HTK工具箱使用HMM作为语音识别的核心，当识别系统是孤立字时，它用不同的隐含状态来描述不同的语音发音。而对于连续语音识别系统，多个孤立的HMM子模型按照一定的语言模型组成复合HMM模型序列来刻画连续的语音信号，每一个模型都有进入和退出状态，这两个状态不产生任何输出，主要用于模型的连接和组合。

3.4.2 HTK软件结构

大部分HTK的函数被都编译成一系列的函数库模块。这些模块对外界提供统一的接口，便于学习和使用。它的软件结构图如下图示：

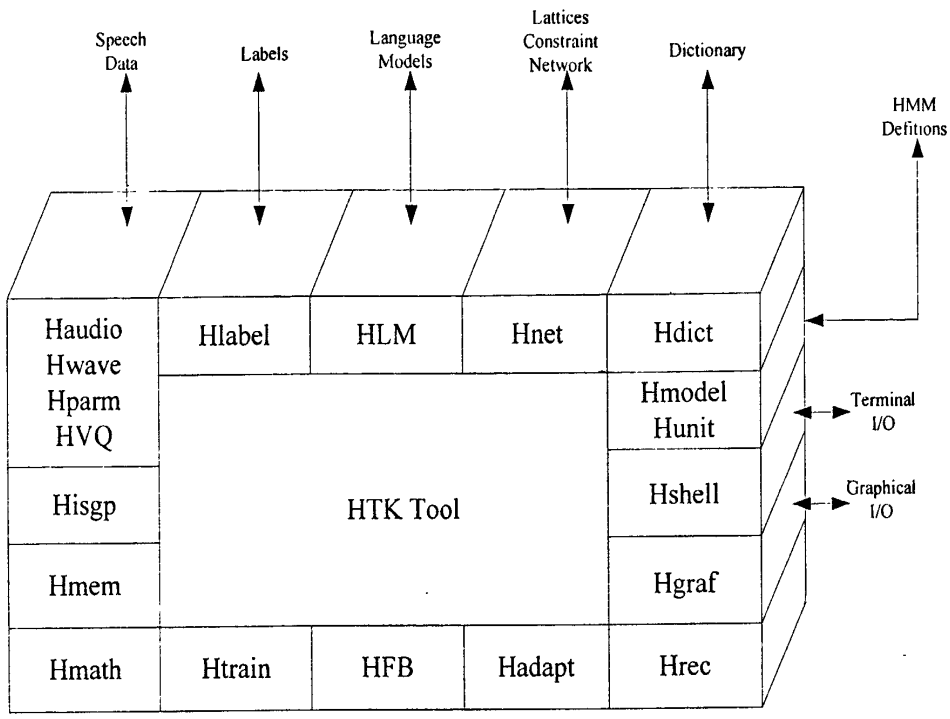


图 3-8 HTK 结构图

其主要功能模块如下：

- Hshell：控制用户的输入输出和与操作系统的接口
- Hgraf：提供用户图形界面
- Hmem：负责内存管理
- HSgip：提供语音分析所需要的处理操作
- HLM：负责语言模型建立
- HLabel：提供标记文件接口
- Hdict：负责建立词汇的发音字典
- HVQ：负责矢量量化 VQ 码本的建立
- HModel：负责 HMM 模型的建立和定义
- HAudio：提供直接音频输入支持
- HWave：提供对各种波形数据格式的支持
- Hparm：提供特征参数文件格式支持
- Hutil：提供有关 HMM 模型的应用例程
- Hadapt：提供 HTK 自适应工具

- Hrec：包含主要的识别处理函数

3.4.3 HTK处理流程

HTK处理流程可分为四个步骤：数据准备、训练、测试、分析。每一个步骤都会用到HTK工具箱不同的函数库模块，其处理流程图导如下图示：

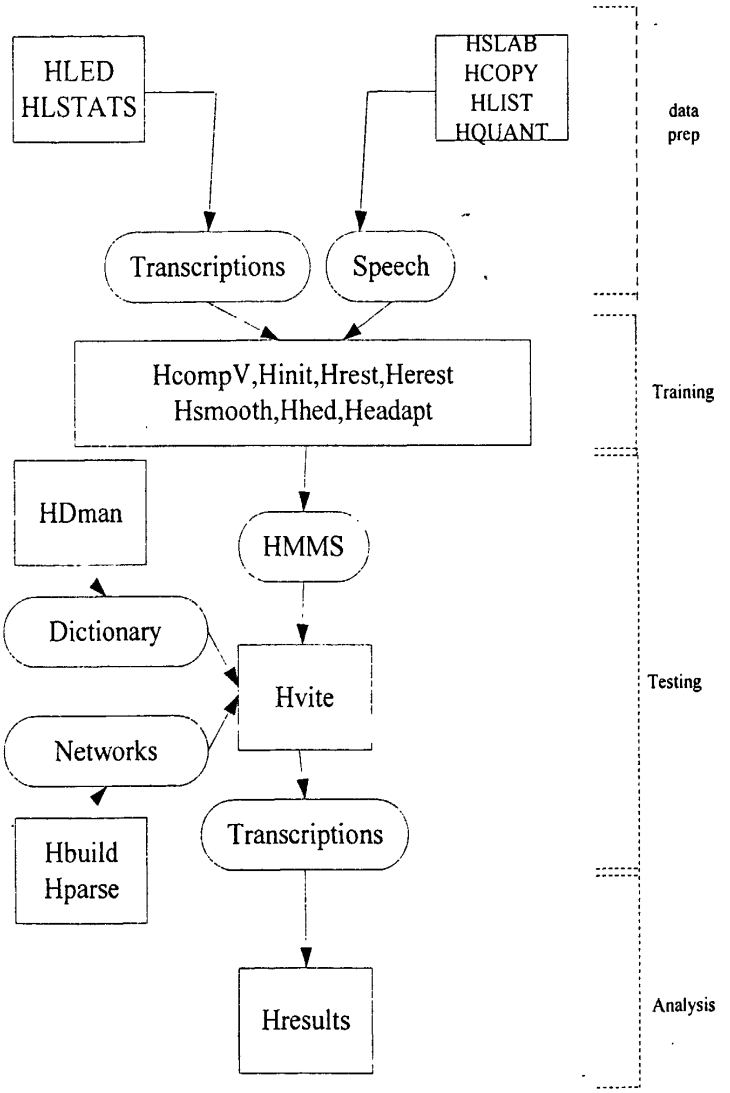


图 3-9 HTK处理流程

数据准备工具：建立一系列HMM模型，首先需要波形语音文件和相应的标注好音段划分的文件（简称标音文件）。通常语音库可以通过购买或者自己录制得到，但是HTK并不直接处理语音波形文件，我们还必须提取这些语音文件的特征参数用于训练和测试。标音文件一般有词级标音文件、字级标音文件、音素级标音文件，不论哪种级别，都要保证

格式的正确。Hslab 可用于录音并标记，Hcopy 可用于提取特征参数，Hlist 可用于查看语音文件和特征参数文件。

训练工具：开始训练之前，我们要定义 HMM 模型的拓扑结构，并以文本文件的格式保存，方便编辑和修改。HTK 工具箱提供了常见拓扑结构的脚本文件，使用这些脚本，可以非常快得到我们想要的原型。这里我们首先关注的是结构，而状态的参数和转移矩阵参数在训练过程中会不断被重估、修改。

测试工具：测试工具也可以解释为识别工具，通过这些函数模块，输入未知语音，系统能给出它的识别结果，并以文本的形式表现出来。本文中用到的识别工具是 Hvite，基于维特比算法。

分析工具：识别系统给出了识别结果，我们还需要分析工具来评估系统识别系能。典型的方法就是对比识别结果和参考模板，然后计算正确率和准确率。本文中用到的分析工具是 Hresults。

第4章 基于HTK的连续语音识别系统实现

4.1数据准备

需要录制训练数据和测试数据，提取特征参数。由于本文采用的是TIMIT语料库，所以录制训练和测试语音文件的任务可以省去；此外，还需要数据的标注文本，并生成任务语法和发音字典。

4.1.1 任务语法

一个语音识别引擎由下面三部分组成：

语言模型：语言模型包含一个字列表以及它们在句子中出现的概率。语法是一个较小的文件，用于限制字的组合，使它们成为有意义的句子。语言模型或者语法中的每一个字都必须有它相对应的发音。

声学模型：包含语言模型和语法中每一个字的确定的发音序列，每个确定的发音我们又称之为音素。在连续、大词汇量的语音识别中，我们也是对音素建模。

解码：可以认为是软件程序，它的输入是人的发音，然后在声学模型中寻找等价的发音。一旦找到匹配的发音，解码器输出与发音相对应的发音序列，这个过程持续到句子的结束或者到达暂停的标记。之后，它根据得到的发音序列，在语言模型中寻找字序列，从而得到句子。

语音识别系统的语法实际上定义了输入的“格式”，这种“格式”也是识别引擎“接受的”或者“希望得到”的输入格式。本文中对任务语法的约束是“宽松”的，认为单词与单词之间并没有彼此的约束关系，在编制语法文件的时候，尤其要注意语法中的文字和训练的声学模型的文字一定要完全一样，否则识别结果可能出错。

由于本实验是大词汇量识别，其中包含的单词（词汇）有6146个，不方便把语法文件放在本论文中，但是，它的基本格式可以描述如下：

$wd=a|abbreviate|abdomen|abides|……

(send-start<$wd>send-end)

这里的$wd是自定义的变量，它可以代表发音字典中的任何一个单词，后面的省略号代表未贴出来的单词，send-start和send-end代表句子的开头和结尾，在建立模型的时候，

可以把它们归结为“sil”（静音）模型。“sil”模型的建立是必要的，因为句子的开头和结尾一般都伴随着若干无声段（静音），用“sil”来描述这些静音是合理准确的。

4.1.2 发音字典

任务语法中出现的是单词（字），而我们需要对音素建模，所以必须有一个发音字典，用来指示每个单词的发音。发音字典有严格的格式要求，每个单词和它相应的发音单独成一行，字典可以手工编写，也可以使用脚本自动得到。由于我们的词汇量有 6146 个，手工编写工作量巨大，所以必须借助 HTK 工具箱的函数 HDMAN 自动得到识别系统的发音字典。

步骤如下：

◆ 创建抄本文件（prompts），抄本文件也就是我们录音时使用的文本文件

◆ 使用脚本 prompts2wlist，从抄本文件得到字列表文件 wlist，wlist 是抄本文件所有单词的有序列表，每个单词一行，并且严格按照字母顺序排列。

◆ 创建 global.ded 脚本，执行 Hdman 得到 wlist 中每个单词的发音。

HDMan -m -w .\lists\wlist -n .\lists\monophones1 -l dlog .\dict\dict1 .\dict\timit .\dict\add

为了表述直观清晰，我把相关文件的一小部分放在论文中。

抄本文件例子如下：

./data/train/dr1/fcjf0/sa1 she had your dark suit in greasy wash water all year

./data/train/dr1/fcjf0/sa2 don't ask me to carry an oily rag like that

./data/train/dr1/fcjf0/sx127 the emperor had a mean temper

./data/train/dr1/fcjf0/sx307 the meeting is now adjourned

./data/train/dr1/fcjf0/sx37 critical equipment needs proper maintenance

./data/train/dr1/fcjf0/sx217 how permanent are their records

./data/train/dr1/fcjf0/sx397 tim takes sheila to see movies twice a week

./data/train/dr1/fcjf0/si1027 even then if she took one step forward he could catch her

./data/train/dr1/fcjf0/si1657 or borrow some money from someone and go home by bus

./data/train/dr1/fcjf0/si648 a sailboat may have a bone in her teeth one minute and lie becalmed the next

.....

每个句子单独一行，每一行中前面的部分是完整路径，后面是相应句子内容。

Wlist 文件如下：

a

abbreviate

abdomen

abides

ability

……

每个单词一行，单词按照字母顺序排列。

字典文件如下：

a	ax sp
abbreviate	ax b r iy v iy ey t sp
abdomen	ae b d ax m ax n sp
abides	ax b ay d z sp
ability	ax b ih l ix t iy sp
able	ey b el sp
ably	ey b l iy sp
abolish	ax b aa l ih sh sp
aborigine	ae b axr ih jh ix n iy sp
aborigines	ae b axr ih jh ix n iy z sp
……	……

global.ded 的内容如下：

AS sp

RS cmu

MP sil sil sp

AS sp 表示在每个单词的发音后面加上短停顿 sp(short pause)，RS cmu 表示去掉重音标记，MP sil sil sp 表示把静音停顿 sil sp 合并为 sil。

执行 Hdman 命令之后会得到一个 monophones1 文件，它是一个发音集，我们会为发音集中的每一个发音建立一个 HMM 模型。本文采用的是 TIMIT 的发音字典，最后得到的发音集如下：

ax	v	m	l	axr	s	eh	hh	dh	em
sp	ey	n	ix	jh	uw	ng	uh	th	sil
b	t	ay	el	aw	ao	y	f	zh	
r	ae	z	aa	ah	er	ch	ow	oy	
iy	d	ih	sh	p	k	w	g	en	

表 4-1 发音集

4.1.3 字级别标音

HTK 软件并不能直接处理前面提到的抄本文件（prompts），我们必须把抄本文件转化成标音文件，标音文件也有严格的格式要求：每个字单独一行，每句的结尾以“.”号结束。为了处理的方便，把所有的标音文件汇总成一个主标记文件（MLF，Master Laber File），借助于 HTK 提供的脚本 prompts2mlf，可以非常迅速完成该工作，本文的字级别主标记文件截取一小部分例子如下：

```
#!MLF!#
"./data/train/dr1/fcjf0/sa1.lab"
she
had
your
dark
suit
in
greasy
wash
water
all
year
.
```

4.1.4 音素级标音

因为是大词汇量识别，我们建模的基本单元是音素，在得到字级别标音文件之后，还

需进一步得到音素级标音文件，HTK提供了从字级标音到音素级标音的转换函数HLEd，结合编辑脚本mkphones0.led，可以直接得到音素级标音文件，本文的音素集如表4-1示，文件一小部分如下：

"./data/train/dr1/fcjf0/sa1.lab"

sil

sh

iy

hh

ae

d

y

uh

r

d

aa

r

k

s

uw

t

.....

4.1.5 提取特征参数

HTK并不直接处理原始波形语音文件，训练和识别用到的都是语音的特征参数，第三章对特征参数做了详细的陈述，这里以MFCC_0_D_A为例，提取的特征参数是梅尔频率倒谱系数，_0表示加上帧能量，_D表示一阶差分，_A表示二阶差分。使用Hcopy提取特征参数的命令如下：

HCopy -A -D -T 1 -C wav_config -S codetrain.scp

文件wav_config是配置文件，内容如下：

```
SOURCEFORMAT = NIST
TARGETKIND = MFCC_0_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
```

这里有必要对配置文件作简要的介绍，SOURCEFORMAT 代表数据源格式，TIMIT 语料库的语音文件均是 NIST 格式的；TARGETKIND 代表提取的参数格式；TARGETRATE 为帧长，SAVECOMPRESSED 表示带压缩（对提取的特征参数压缩），SAVEWITHCRC 表明加上 CRC 循环校验，WINDOWSIZE 为窗长度，USEHAMMING = T 指示使用的是汉明窗，PREEMCOEF = 0.97 代表我们的预加重系数是 0.97，最后三项是滤波器配置参数。

codetrain.scp 是我们手工编写的脚本，为 HTK 指示了输入和输出路径，截取该文件一小部分如下：

```
./data/test/dr1/faks0/sa1.wav    ./data/test/dr1/faks0/sa1.mfc
./data/test/dr1/faks0/sa2.wav    ./data/test/dr1/faks0/sa2.mfc
./data/test/dr1/faks0/sx43.wav   ./data/test/dr1/faks0/sx43.mfc
./data/test/dr1/faks0/sx223.wav  ./data/test/dr1/faks0/sx223.mfc
./data/test/dr1/faks0/sx403.wav  ./data/test/dr1/faks0/sx403.mfc
./data/test/dr1/faks0/sx133.wav  ./data/test/dr1/faks0/sx133.mfc
./data/test/dr1/faks0/sx313.wav  ./data/test/dr1/faks0/sx313.mfc
./data/test/dr1/faks0/si943.wav  ./data/test/dr1/faks0/si943.mfc
./data/test/dr1/faks0/si1573.wav ./data/test/dr1/faks0/si1573.mfc
./data/test/dr1/faks0/si2203.wav ./data/test/dr1/faks0/si2203.mfc
.....                            .....
```

4.2创建单音素模型

4.2.1 基本声学建模单元

基本声学单元的选择是声学建模中一个重要的问题。常见的识别基元有：词（word）、音节（Syllable）、半音节（Semi-Syllable）、声韵母（Initial/Final）、音素（Phone）等[25]。识别基元一般是基于语音学知识的，也可以通过数据驱动的方式产生。不同识别基元有各自的应用领域。

对于词，适应于小词汇表语音识别系统、命令与控制系统。但它不适合连续语音识别，首先，在连续语音识别中，词的数量多，达到几千的数量级，例如本文，有 4000 多个单词，采用词作为声学基元建模，会导致声学模型规模庞大，不仅会使得训练和识别的存储效率低，还会增加识别过程中搜索的复杂度。其次，当词表以外的单词出现时，声学模型处理困难。第三，对如此庞大的基元进行训练，为了使得训练数据覆盖所有的词条，必然需要很大的数据库。

声韵母建模对汉语语音识别非常适用，本文中的数据库是英文，而英语单词是多音节结构的，所以声韵母在这里不详细讨论。

汉语发音和英语发音有个区别，汉字是单音节的，若不考虑声调，汉语中大概有 400 个无调音节，考虑音调的话，有 1300 多个有调音节字。英语单词是多音节的，所以对于大词汇量识别任务，采用音节建模也并不适合本文的连续语音识别。

我们知道，英语单词都有对应的音标，给出一套音标，我们就可以获得所有单词的发音，这是完备的。以国际音标为例，共有 48 个，其中元音 20 个，辅音 20 个，3 个鼻音，2 个半元音，3 个似拼音。如果我们以语音的最基本发音单元—音素（音标）为建模基元，不仅使得模型数量得到合适的控制，而且还能保证每个音素都得到足够多的训练。

4.2.2 定义模型结构

首先定义一个 HMM 模型，因为是初始模型，我们关注的是模型结构，而不是参数值（它可以通过训练进行不断优化），初始模型一般称为“proto”，结构如下：

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "proto"
<BeginHMM>
```

<NumStates> 5

<State> 2

<Mean> 39

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

<Variance> 39

1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0

<State> 3

<Mean> 39

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

<Variance> 39

1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0

<State> 4

<Mean> 39

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

<Variance> 39

1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0

<TransP> 5

0.0 1.0 0.0 0.0 0.0

0.0 0.6 0.4 0.0 0.0

0.0 0.0 0.6 0.4 0.0

0.0 0.0 0.0 0.7 0.3

0.0 0.0 0.0 0.0 0.0

<EndHMM>

这是一个5状态的HMM结构，特征向量是39维的，特征向量的类型是MFCC_0_D_A（即MFCC参数及其一、二阶差分）。我们初始化采用是“flat start”的方式，使用HCompV

来完成对 proto 模型的初始化，然后创建一个 hmmdefs 的文件，把每个音素模型都初始化为和 proto 相同的模型结构，并把模型的名字更改为自己的音素名。特别注意，在 hmmdefs 文件末尾加个空行。HTK 初始化的命名如下：

.\htk\HCompV -A -D -T 1 -C config -f 0.01 -m -S train.scp -M .\hmms\hmm0 proto

此外，还需要创建一个主宏文件 macros，文件内容如下，其中，-f 选项会使得 HcompV 输出一个“floor”方差，它的值是全局方差的 0.01。

~o<VECSIZE> 39<MFCC_0_D_A>

~v varFloor1

<Variance> 39

1.023560e+000　5.922153e-001　8.007562e-001　8.232392e-001　7.904088e-001 7.556760e-001 7.624118e-001 6.574206e-001 6.495683e-001 4.682004e-001 4.769277e-001 3.514470e-001 1.163718e+000 4.548536e-002 3.513668e-002 3.333983e-002 4.192985e-002 4.350298e-002 4.223124e-002 4.364949e-002 4.255898e-002 3.925652e-002 3.239214e-002 2.973102e-002 2.434268e-002 5.539082e-002 6.270031e-003 5.813580e-003 5.071539e-003 6.760838e-003 7.105201e-003 7.140727e-003 7.429008e-003 7.515165e-003 6.921130e-003 5.950140e-003 5.402804e-003 4.455170e-003 8.890656e-003

前面使用 HcompV 完成对 proto 的初始化，然后将 proto 复制给所有的音素模型，这意味着目前所有的音素模型都具有相同的均值、方差以及转移矩阵，我们把所有的音素模型放在一起，组成 hmmdefs 文件。接下来，需要用 HERest 来重估模型参数，也就是使用训练数据，根据 Baum-Welch 算法，改变每个模型的相关参数。重估命令如下：

.\htk\HERest -A -D -T 1 -C .\config\config -I phones0.mlf -t 250.0 150.0 1000.0 -S train.scp -H .\hmms\hmm0\macros -H .\hmms\hmm01hmmdefs -M .\hmms\hmm1 monophones0

重估后的 hmmdefs 会在 hmm1 文件夹下，里面的参数值都会发生改变，train.scp 里面存放的内容是特征参数的列表，它的内容截取如下：

./data/train/dr1/fcjf0/sa1.mfc

./data/train/dr1/fcjf0/sa2.mfc

./data/train/dr1/fcjf0/sx127.mfc

./data/train/dr1/fcjf0/sx307.mfc

./data/train/dr1/fcjf0/sx37.mfc

./data/train/dr1/fcjf0/sx217.mfc

./data/train/dr1/fcjf0/sx397.mfc

./data/train/dr1/fcjf0/si1027.mfc

./data/train/dr1/fcjf0/si1657.mfc

./data/train/dr1/fcjf0/si648.mfc

.....

继续重估两次，对参数进行训练、重估，生成的 hmmdefs 在\hmm3\路径下。重估命令与上面相同，只是文件的路径需要做下改变，从 hmm1 到 hmm2，然后从 hmm2 到 hmm3。

4.2.3 绑定静音模型

前面创建的 hmmdefs 模型文件列表中并没有包含“sp”(short pause)模型，sp 一般用于表示正常发音中单词与单词之间的短暂停，之前已经存在了“sil”模型，用于表示句首和句尾较长的停顿。我们可以把 sil 模型理解为持续时间较长的 sp 模型，现在，需要把 sp 模型绑定到 sil 模型中。具体实现需要把 sil 模型的中间状态拷贝给 sp 模型，然后运行 HHed 函数把它们绑定起来，命令如下：

.\htk\HHEd -A -D -T 1 -H .\hmms\hmm4\macros -H .\hmms\hmm4\hmmdefs -M .\hmms\hmm5 sil.hed .\lists\monophones1

其中，sil.hed 是命令脚本，内容如下：

AT 2 4 0.2 {sil.transP}

AT 4 2 0.2 {sil.transP}

AT 1 3 0.3 {sp.transP}

TI silst {sil.state[3],sp.state[2]}

绑定之后，运行 HERest 继续重估两次，现在我们的模型列表文件在 hmm7 路径下。

4.2.4 重校准训练数据

之所以要重新校准训练数据，是因为在英语中有些单词有多种发音现象的存在，比如单词 live，它作为动词和形容词有不同的发音，之前在实现字级标音文件到音素级标音文件的时候，我们并没有考虑到这种情况，采用的都是单词第一次碰到的发音，经过重校准，我们采用的发音能够更好的匹配声学数据。

重校准需要用到 HVite 函数，命令如下：

HVite -A -D -T 1 -o SWT -b silen -C .\config\config -H .\hmms\hmm7\macros -H .\hmms\hmm7\hmmdefs -i aligned.mlf -m -t 250.0 150.0 1000.0 -y lab -a

-I .\labels\trainwords.mlf -S train.scp \dict\dict1 .\lists\monophones1> log

重校准之后，还需要重估两次，这样，我们的 hmmdefs 文件在 hmm9 路径之下。

4.3上下文相关建模

4.3.1 上下文相关建模必要性

前面我们已经完成了单音素的建模，但是，对于大词汇量的连续语音识别系统来说，只有单音素是远远不够的，识别效果也是不理想的。主要的原因是没有考虑到模型所处的上下文，一个音素的发音会由于上下文音素的不同而产生不同的发音。比如，我们在说话时，往往在某些音还没发出时，就转入下一个音，即俗称的连音问题；还有，发音时存在协同发音的现象、语音信号随着说话人语速差异以及生理条件不同产生变化、语言的歧义性及结构的随意性，等等，这些都是客观存在的问题，如果仅仅采用单音素建模，根本无法解决这些问题中的任何一个，识别效果也会很差；因此，必须建立上下文相关的 HMM。

上下文相关模型的出发点是精确的表示、建立 HMM 模型，使得基于 HMM 模型的语音识别系统达到理想的效果[26]。而影响模型精确性的主要音素就是来自上下文的变化，概括的说，上下文的变化有两类，一类是阶段效果，主要是由于说话人的不同和语音环境的不同而产生的对模型的影响；比如噪音环境，比如说话人不同性别、年龄、口音、说话方式、心情等，解决这类问题的方法有背景消（降）噪、说话人自适应等。另一类是局部效果，主要考虑一段语音内的发音变化，集中体现在协同发音方面。比如，某个音还没有发到位就转入下一发音，对听众来说，可能影响不大，但是对于要求音素模型稳定的识别器来说，就会直接影响它的识别结果。通过这些分析我们可以看出，一个音素在特定的音素上下文中的声学特征比较稳定，但在不同的上下文中，会表现出不同的特点。如果我们在建模过程中充分考虑到协同发音的话，会在很大程度上改善识别效果。

所谓上下文相关音素，就是考虑一个音素与其左或右相邻音素关系后选取的音素。这样，对于 N 个音素，就可能存在 N^2 个左和右上下文相关音素（每个音素都可以与其它 N 个音素之一构建左或者右的双音素，所以 N 个音素序列共 N^2 个可能），成为双音素；同样可能存在 N^3 左和右上下文相关的三音素。通常，我们可以考虑某个音素前面 M 个上文音素和后面 N 个下文音素，但是，M 和 N 越大，模型复杂度越大，模型数量越多，并且，音素间发音的相互影响也会随着 M、N 的增大而减小。一种简单并且有效的方法是考虑每

个音素的前一个和后一个音素，建立三音素模型。

三音素的提出对大词汇量的连续语音识别有重要的意义，它对正确识别率的提升具有显著作用。在语音识别中，有协同发音现象的存在，某些发音非常接近，在训练中又得不到足够的样本数据，会造成识别错误、音素替换等情况的出现，从而导致识别率的降低。三音素并不简单的考虑每个发音本身，它把每个音素的前一个音素和后一个音素作为一个整体，综合考虑。

三音素（Triphone）可以分为两种：逻辑 Triphone 和物理 Triphone，前者指语言上可能的音素组合，可以理解为广义的；后者指训练语音数据中出现的音素组合。本论文中，有 47 个单音素，如表 4-1，按照逻辑 Triphone，将会有 103823 个三音素，而实际训练语音中出现的三音素只有 7000 多个（执行 HLed，系统会生成一个 Triphones 列表，里面包含训练数据中出现的所有的三音素）。对这么多的声学单元进行建模，尤其困难，容易出现数据稀疏问题（即某种组合的训练数据不足）。为了保证模型的精确性，必须保证每个出现的三音素在训练样本中至少出现 10 次；此外，如果每个三音素出现的次数少于 3 次，在 HTK 运行过程就会出现警告信息。所以，减少声学单元的数目显得非常有必要。

4.3.2 从单音素模型创建三音素

三音素模型本质上是考虑了音素的的上下文关系，还可以进一步细分为词内三音素和词间三音素[27,28]。本文建立的英语语音识别系统只考虑词内部的上下文关系，这样做是考虑了以下情况：首先，只考虑词内产生的不同上下文的数目要远远少于词间上下文的数目，这样，也可以使得大多数的上下文关系会出现在训练集里[29]；其次，只考虑词内的上下文会降低识别器的复杂度，假如考虑了词间上下文，在识别过程中，每个词的首个音素和末尾音素会依赖于前一个词和后一个词，这会增加搜索算法的复杂度，使得搜索效率降低[30]。

本文中，单音素模型有 47 个（包括 sil、sp），扩充成三音素的结构为：L-X+R，L 代表音素 X 的左边音素，R 代表 X 的右边音素。从单音素扩充到三音素，模型的数量急剧增加，结合实际的训练语料，实验中三音素模型个数达到了 7364 个。但三音素对系统识别率的提高却是显著的，在后面的章节，会着重做比较。

在数据准备阶段，我们已经有了字级标音文件和音素级标音文件，现在的模型是三音素模型，那么，相应的标音文件也应该是三音素级标音文件，实现方法如下：

创建 mktri.led 编辑脚本，内容如下：

WB sp

WB sil

TC

然后，执行命令：

.\htk\HLEd -n .\lists\triphones1 -i .\labels\wintri.mlf mktri.led .\labels\aligned.mlf，该命令输出两个文件，triphones1 为三音素列表文件，wintri.mlf 为三音素级标音文件。截取文件部分内容如下：

triphones1:	wintir.mlf:
sil	#!MLF!#
sh+iy	"./data/train/dr1/fcjf0/sa1.lab"
sh-iy	sil
sp	sh+iy
hh+ae	sh-iy
hh-ae+d	sp
ae-d	hh+ae
y+uh	hh-ae+d
y-uh+r	ae-d
uh-r	sp
.....	.....

接下来，我们要初始化三音素模型，需要用的命令是 HHEd 和编辑脚本 mktri.hed，执行：

.\htk\HHEd -H .\hmms\hmm9\macros -H .\hmms\hmm9\hmmdefs -M .\hmms\hmm10 mktri.hed .\lists\monophones1

初始化完成之后依然需要重估两次，得到的三音素模型文件最后在 hmm12 路径之下。

4.4数据驱动及决策树状态共享

在模型参数重估的过程中，由于训练量不足，很多分布的方差我们是用主宏文件的 vFloors 替换的。为了改善这种情况，我们现在通过状态来共享数据，对音素进行聚类，解决数据稀疏问题，使输出更加稳健（提高模型鲁棒性）。HTK 有两种方式实现状态的聚类：数据驱动方式（自底向上）、决策树方式（自顶向下）[31,32]。

4.4.1 数据驱动的状态共享

数据驱动的聚类方法，在初始时把所有的状态都作为一个类，每次合并两个最小类，直到最大类的大小达到一个阈值或者类的数目达到聚类的要求。下图给出数据驱动状态约束示意图：

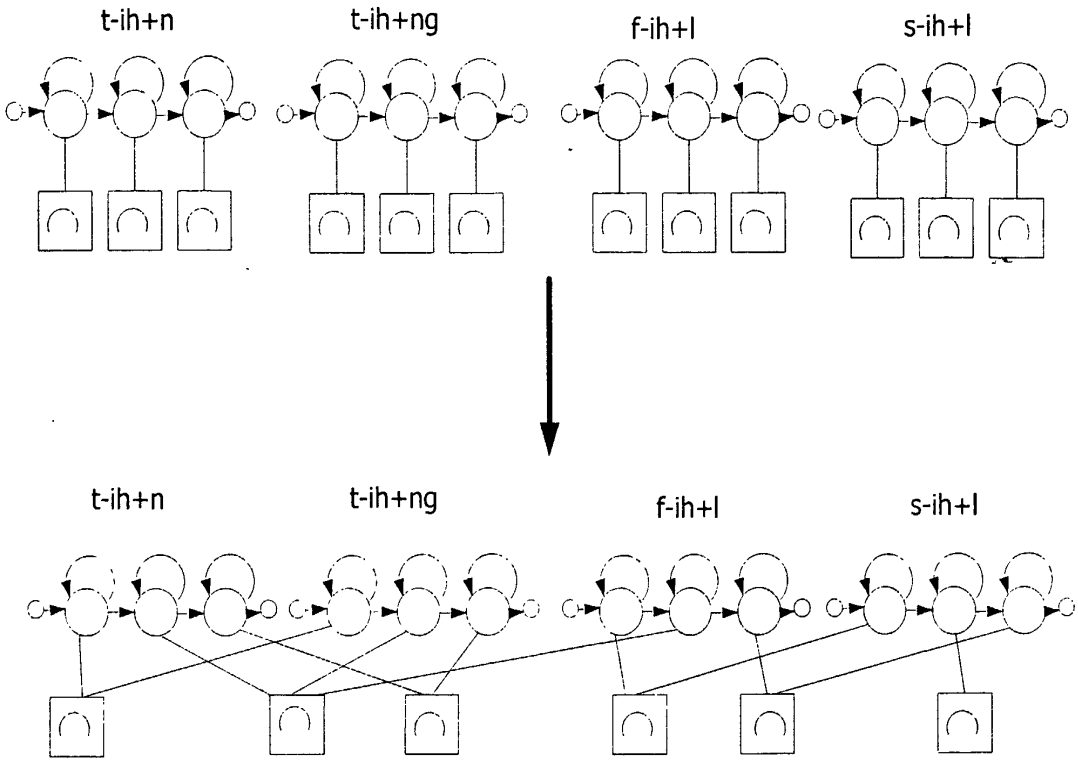


图 4-1 数据驱动状态约束

这种数据驱动的聚类方式受可用的数据的限制，它无法处理在语音数据中没有出现的三音素 Triphone[33,34,35]。对于小词汇表系统，可以通过设计合适的文本库解决这个问题，但是对于大词表 Triphone 系统，这个问题无法回避。所以，必须得有另一种方式既能进行状态的聚类，同时能够处理训练数据中没有出现的 Triphone，基于决策树的聚类方法可以实现这个目标。

4.4.2 决策树状态共享

基于决策树的状态共享是一种自顶向下的聚类过程。决策树状态共享如图 4-2 所示，它将中间状态是“aw”的所有三音素模型的第二个状态合并，然后，根据问题集所列的问

题分裂已合并的状态，从问题树的根节点开始回答问题，直到到达叶节点，所有位于同一叶节点的HMM的第二个状态就会被绑定。

决策树的构建过程如下[36,37]：

1、准备一系列先验语音学问题集；2、从包含相同语音学中心单元的所有上下文无关观测帧的根节点开始，初始化决策树；3、给未扩展节点寻找最优的问题，将节点分裂成两个孩子节点；4、返回步骤3，直到满足阈值。如下图示：

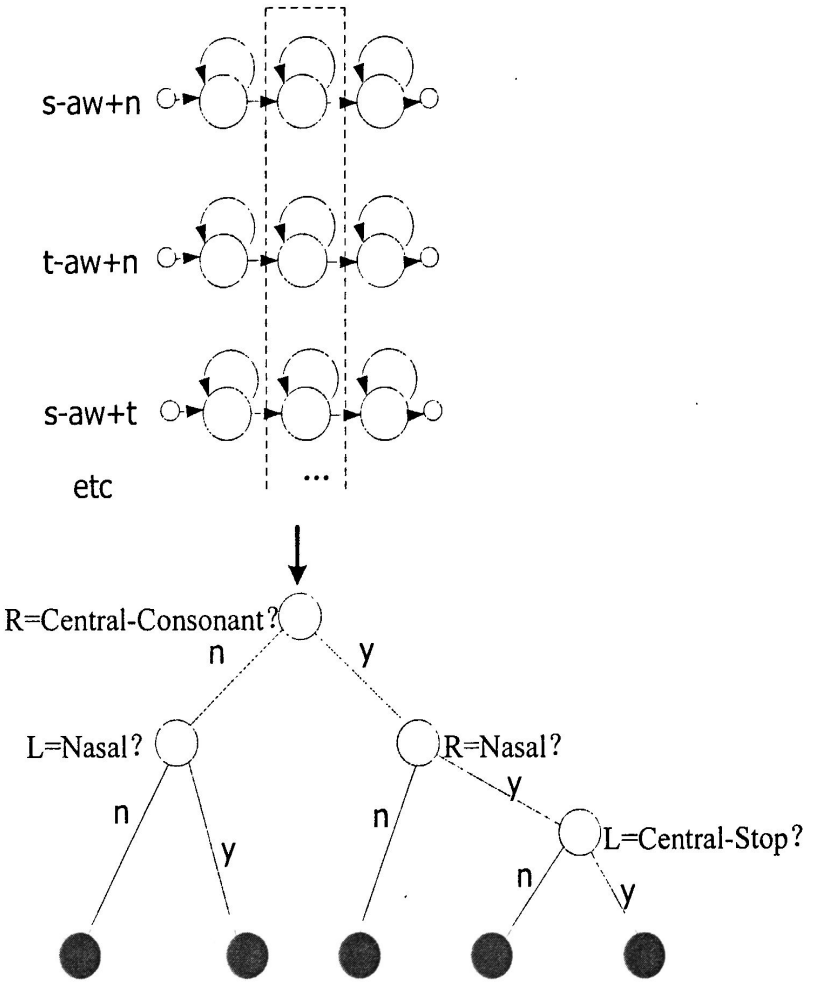


图 4-2 决策树聚类

由于决策树是根据问题集进行分裂的，所以问题集的好坏将直接影响到声学建模的性能，本论文在问题集的设计也进行了详细的研究。

实现命令如下：

HHEd -H .\hmms\hmm12\macros -H .\hmms\hmm12\hmmdefs -M .\hmms\hmm13 tree.hed .\lists\triphones1 > log

最后记得重估两次优化性能，更新后的hmmdefs文件在hmm15路径下。

4.5问题集设计和优化

问题集是供决策树构造使用的问题的集合，其主要依据是发音方式的相似性[38,39]。节点分裂时选中的节点将会与该节点绑定，依此决定哪些基元的哪些状态被共享。所以问题集设计的好坏对系统性能有直接的影响。

4.5.1 问题集设计

问题集的构建首先要对语音音素有系统整体的了解。在此基础上，需要对音素进行分类，基本分类要求有：

1、元音类

2、辅音类

3、前元音、后元音、中元音、长元音、短元音、双元音

4、摩擦辅音

5、鼻音

6、流音

7、爆破音

8、其他

以上是英语音素分类的概要，如果识别对象是中文也可以类似分类。有了具体的分类，给每个类别添加左问题、右问题，依此构建问题集。

以英文国际音标为例，我们采用上面的分类标准，对48个音标分类如下表：

元音	长元音	i:	a:	e:	u:	o:					
	短元音	i	ɔ	∧	C	ʊ	e	æ			
	双元音	ei	ai	Ci	au	ue	eu	ie	ce		
辅音	清辅音	p	t	k	f	s	θ	S	Ts	tr	ts
	浊辅音	b	d	g	v	z	ð	3	d3	dr	dz
	鼻音	m	n	ŋ							
	半元音	w	j								
	似拼音	h	r	l							

表4-2 国际音标分类

问题集是根据发音方式的相似性进行分类的，下面简单介绍发音方式和分类。

第一，　按照发音方法分类（辅音）

1.stop

爆破音--是指发音器官在口腔中形成阻碍,然后气流冲破阻碍而发出的音。

2.fricative

摩擦音--由发音器官造成的缝隙使气流产生摩擦而发出的声音。

3.affricate

塞擦音--有塞音和擦音紧密结合所构成的音，发音时注意最初形成阻碍部位要完全闭塞，然后渐渐打开。

4.nasal

鼻音--口腔气流通路阻塞,软腭下垂,鼻腔通气发出的声音。

5.liquid

流音--舌端齿龈测流音 ,舌端紧贴上齿龈,舌前部向硬腭抬起,气流从舌的一侧或两侧滑出。

6.glide

滑音--指发音器官移向或移离某一发音动作的过度音。

第二，按照嗓音分类（辅音）:

1.voiced

浊音--振动声带所发出的辅音，除气流受到阻碍外，同时振动声带发出乐音。

2.voiceless

清音--发音时不振动声带所发出的辅音，即清音纯粹由气流受阻所构成的且不带乐音。

第三，元音的分类:

1.monophthongs　单元音

2.front　　前元音

3.central　　中元音

4.back　　后元音

5.diphthongs　　双元音，复合元音

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，每一个问题是由一系列上下文来定义的。

以爆破音（stop）为例，可以得到两个问题:

"R_Stop"　　{ *+p,*+pd,*+b,*+t,*+td,*+d,*+dd,*+k,*+kd,*+g }

表示音素的右边是爆破音；

"L_Stop"　　{ p-*,pd-*,b-*,t-*,td-*,d-*,dd-*,k-*,kd-*,g-* }

表示音素的左边是爆破音

4.5.2 加入简单问题优化

决策树节点分裂过程中，会选择最好的问题。在最大似然准则下，最好的问题就是能得到最大对数似然增量的问题[40]。不同的状态，左右上下文的影响很大，我们可以加入简单的问题，更精确的刻画方差，简单问题如下：

"R_aa"　{ *+aa }

"R_ae"　{ *+ae }

"R_ah"　{ *+ah }

……　　……

"R_z"　{ *+z }

"R_zh"　{ *+zh }

在第五章我们会分析问题集的加入对系统性能的改善。

4.6识别和分析

三音素模型创建之后，用训练数据完成模型参数的重估，得到的训练好的模板已经可以用来进行语音识别。这一节简单介绍识别工具 HVite 和分析工具 HResults，围绕着 TIMIT 语料库，下一章将给出详细的识别测试和分析结果。

4.6.1 语音识别函数 HVite

HVite 是 HTK 通用的语音识别工具，基于维特比算法，通过匹配语音文件和 HMM 网络，最后输出每个 HMM 的标音文本。Hvite 还支持参数共享和输出概率分布计算；为了提高处理速度，我们还可以对 Hvite 设置搜索阈值。在优化语音识别性能时，也会用到 Hvite 函数来实现说话人的自适应。

举例说明，上一小节我们把绑定状态的三音素模型建立、重估，存放在 hmm15 路径下，以它为模板，我们执行 Hvite 进行识别，命令如下：

.\htk\HVite -C .\config\config1 -H .\hmms\hmm15\macros -H .\hmms\hmm15\hmmdefs -S

test.scp -i .\results\recognise.mlf -w wdnet -p 0.0 -s 5.0 .\dict\dict2 .\lists\triphones1

识别结果将会存放在.\results 路径下，格式为主标记文件的形式（MLF），文件内容截取部分如下：

```
#!MLF!#
"./data/test/dr1/faks0/sa1.rec"
2200000 3000000 if -684.121582
5800000 8200000 she -1762.760986
8200000 10600000 had -1937.525513
10600000 11800000 your -1020.937622
11800000 14900000 dark -2573.632080
14900000 18100000 suit -2576.117188
18100000 19600000 in -1292.632690
19600000 19900000 a -264.481354
19900000 24000000 greasy -3338.606201
24000000 28200000 wash -3115.263428
28200000 32200000 water -3067.476807
32200000 34400000 all -1716.476807
34400000 38200000 year -2717.433838
.....
```

每一行代表一个识别的单词，并给出了开始和结束时间（单位是 100 纳秒），以及对数似然概率。

4.6.2 分析工具

语音识别最后还需要给出识别率，首先介绍两个计算公式：正确率(cor)、准确率(acc)[41]。

$$cor = \frac{H}{N} * 100\% \tag{4-1}$$

$$acc = \frac{H - I}{N} * 100\% \tag{4-2}$$

公式 4-1 表示词正确识别率，公式 4-2 代表识别精度

H代表正确识别数，N代表总数，I代表插入的单词数。

HTK的分析工具HResults每次把识别结果和参考模板比较，统计出识别的正确率和准确率。一个识别结果的输出例子如下：

```
====================== HTK Results Analysis =======================
  Date: Tue Jan 05 18:18:07 2010
  Ref : coretest.mlf
  Rec : .\results\coremix1.mlf
---------------------- Overall Results --------------------------
SENT: %Correct=7.50 [H=18, S=222, N=240]
WORD: %Corr=68.89, Acc=56.01 [H=1428, D=34, S=611, I=267, N=2073]
===================================================================
```

这个识别结果的含义如下：首先，“SENT”代表句子的正确识别率，测试的句子总数是240，其中正确识别18个句子；其次，“WORD”代表词的识别率，测试句子总共有2073个词，其中正确识别了1428个词，删除34个，替换611个，插入267个。

第5章 实验结果及分析

5.1高斯混合度对识别率影响

每个 HMM 模型都有若干个状态（本文取 5 状态），每个状态都有与其相关联的观察输出概率分布，记为$b_j(o_t)$，它决定了在 t 时刻状态 j 的观察输出o_t，状态 i 与状态 j 之间的联系为转移概率a_{ij}。在本文定义的 HMM 模型中，状态 1 和状态 5 是非发射态，没有输出，只作为模型连接用。下面是一个简单的 5 状态左到右结构的 HMM 模型：

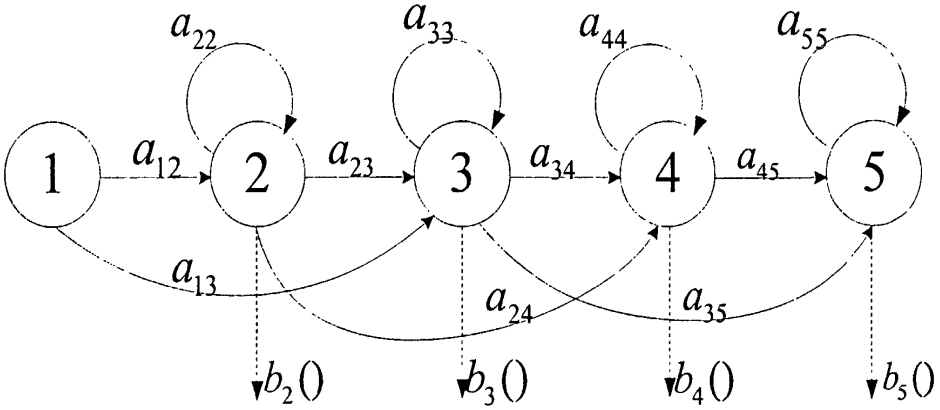


图 5- 1 左到右 HMM 模型

我们的 HMM 模型是连续密度模型，即它们的输出概率可用混合高斯密度表示，以状态 j 为例，$b_j(o_t)$的输出分布可表示如下：

$$b_j(o_t)=\prod_{s=1}^{S}[\sum_{m=1}^{M_{js}}c_{jsm}\mathrm{N}(o_{st};\mu_{jsm},\Sigma_{jsm})]^{\gamma_s} \tag{5-1}$$

S 表示数据流个数，M_{js} 代表了状态 j 对数据流 s 的高斯混合成分数，c_{jsm} 代表第 m 个成分的权值。$\mathrm{N}(;\mu,\Sigma)$ 为多元高斯，均值向量为μ，协方差矩阵为Σ，公式如下：

$$\mathrm{N}(o;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \tag{5-2}$$

其中n是观察向量o的维度，γ_s则代表了数据流的权值，它的默认值是 1。对数据流采取不同的权值代表了数据流重点程度的差异，但是目前 HTK 没有提供函数工具来执行此类操作。关于数据流个数S，一般S=1 代表单参数向量，S=2 表明参数向量包含了能

量选项，S=3，则包含静态系数、能量、一阶差分和二阶差分，S=4 意味着包含静态系数、对数能量、一阶差分和二阶差分[42,43]。

本文实验中重点比较了不同高斯混合度对识别率的影响，随着混合度的增加，识别率会上升，当混合度增加到一定程度的时候，由于训练数据量的不足，识别率会不升反降。

测试的句子来自 TIMIT 建议的核心测试集，有 240 个句子，来自 8 个不同的方言区域，每个方言区域各采纳 3 个人的发音，分别是两男一女，这样的测试集能反映地域、性别等因素，具有实际意义。实验结果如下表：其中第二列和第三列是词的正确率和准确率，第四列是句子的识别率

混合度(mix)	正确率（Cor）%	准确率(Acc)%	句子识别率%
1	68.89	56.01	7.50
2	74.87	65.46	10.42
4	80.13	73.13	13.75
6	84.13	78.87	21.25
7	86.35	82.10	26.25
8	88.18	84.32	28.33
9	89.15	84.61	30.42
10	89.92	85.77	32.08
11	90.30	87.46	35.00
12	90.74	88.37	36.67
14	90.98	89.00	39.17
18	91.51	89.58	40.42
20	92.14	90.64	40.83
25	92.62	91.56	44.58
30	93.25	92.23	48.33
35	93.68	92.81	51.25
40	93.92	93.20	53.33
45	94.16	93.39	54.58
50	94.45	93.92	57.50
60	94.65	94.26	58.75

70	94.84	94.45	58.75
75	94.69	93.92	59.58
80	94.55	93.83	59.17

表 5-1 高斯混合度与识别率统计表

根据表 5-1 做出的关系图如下图示：

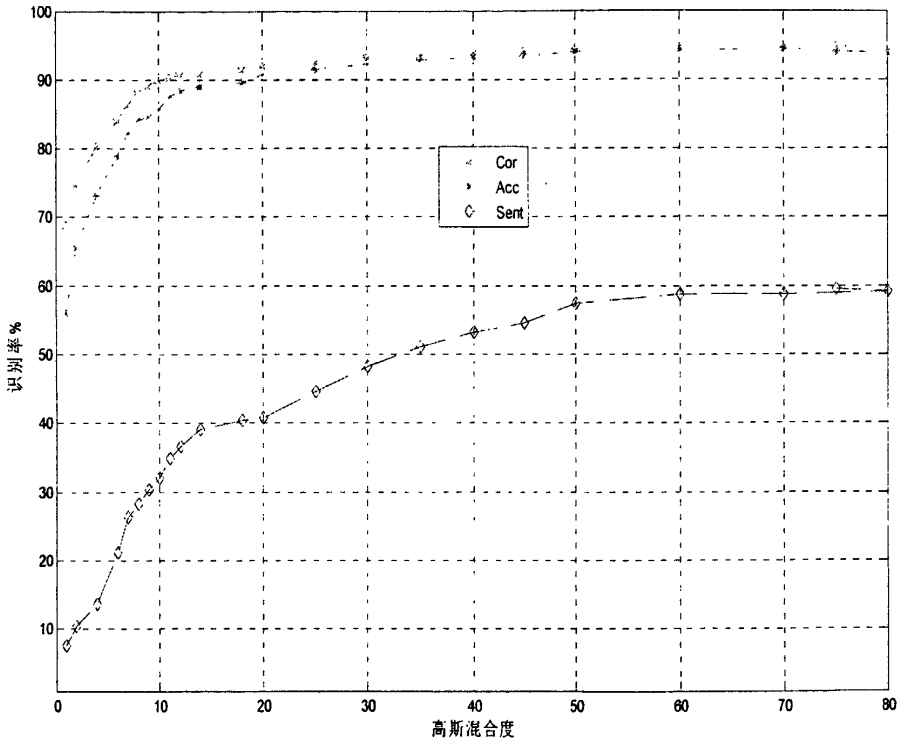


图 5-2 高斯混合度与识别率关系图

从表 5-1 和图 5-2 可以看出，随着混合度的增加，正确率和准确率都会提升，当混合度达到一定程度时（本例子中为 70 左右），由于训练数据量的不足，识别率会不升反降；另外，增加混合度也会带来识别时间的增加，由于每次识别的句子达到了 240 句，当混合度高时，识别时间是必须考虑的音素，下面是我在同一台电脑上记录的完成全部任务需要的识别时间和不同高斯混合度的关系：

混合度	1	2	4	6	7	8	9	10	11	12
识别时间（分钟）	64	79	107	120	130	139	145	153	165	171

表 5-2 混合度与识别时间关系表

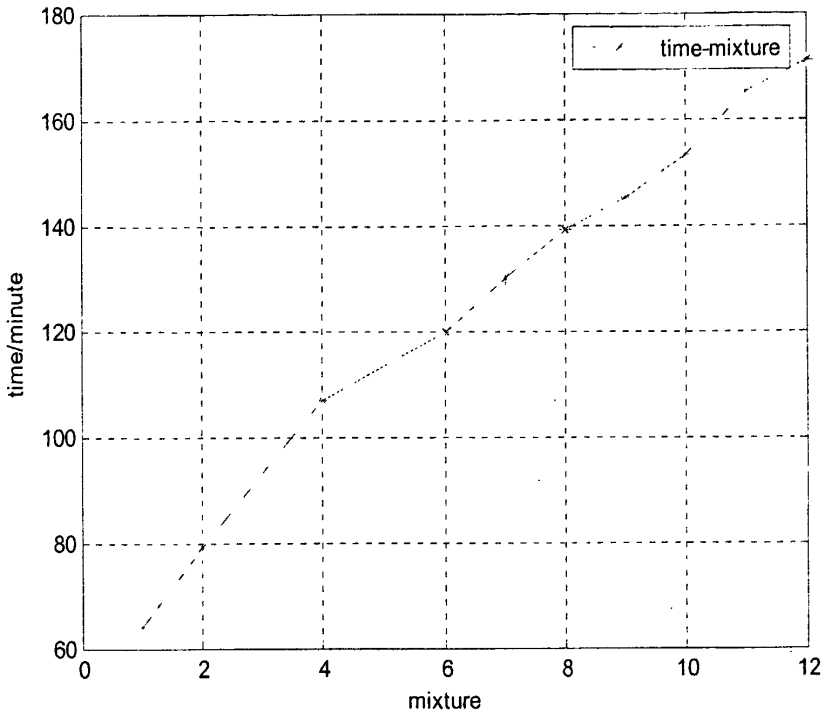


图 5-3 混合度与识别时间关系图

5.2特征参数与识别率

HTK 支持多种文件格式和多种参数格式。HTK 支持的语音文件格式包括：HTK、TIMIT、NIST、SCRIBE、SDES1、AIFF、SUNAU8、OGI、WAV、ESIG、AUDIO、ALIEN、NOHEAD；支持的特征参数有：WAVEFORM、LPC、LPREFC、LPCEPSTRA、LPDELCEP、IREFC、MFCC、FBANK、MELSPEC、USER、DISCRETE、ANON；值得注意的是，特征参数后面可以加上后缀，比如我们采用的特征参数 MFCC_0_D_A，_A 表示二阶导数，_D 表示一阶导数，_0 表示添加零阶倒谱的系数 C0，此外常用的还有_E 表示加载对数能量，_K 表示循环校验、_Z 表示加入倒谱的均值，等等。

实验中着重比较了 LPC、LPCC、MFCC 不同特征参数的选取对识别率的影响，测试集还是 240 个句子，为了节约时间，不同特征参数的训练和识别都选取混合度等于 1，前

面我们已经详细介绍了混合度对识别率的影响，所以这里选择混合度等于1对不同特征参数是公平的，同时也是最节约时间的方法，识别结果如下表：

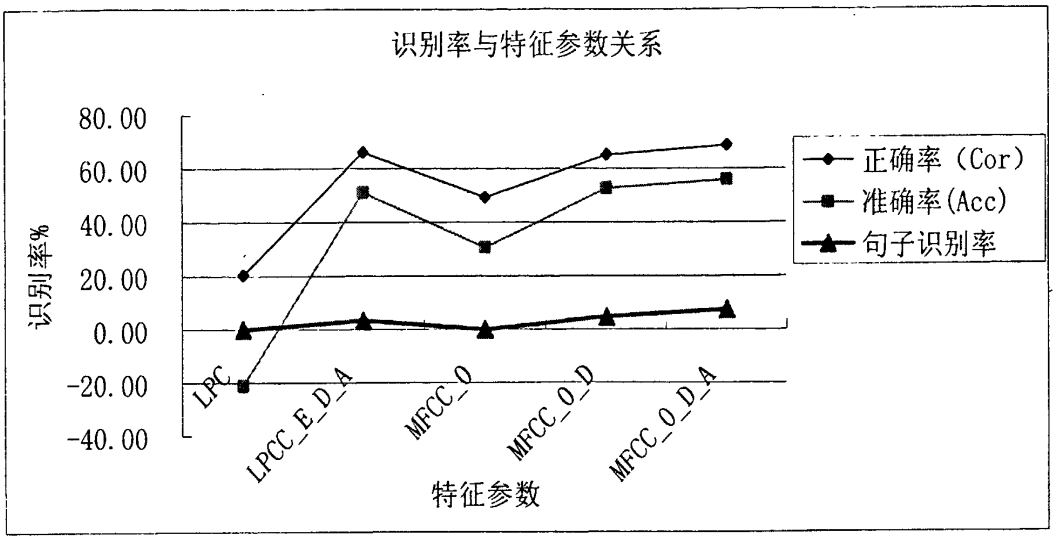


图5-4 识别率与特征参数

从上面的表和图形中可以得到，LPC并不适合作为语音识别的特征参数，MFCC的性能比LPCC好。准确率为负数是因为正确识别的词数目（H）小于插入的词的数目（I），根据公式4-2计算准确率，结果就会为负数。

5.3模型选取与识别率

之前已经介绍了单音素模型和三音素模型，下面在核心测试集上比较这两种不同的模型，识别率有巨大的不同（混合度等于1的情况）：

模型	正确率（Cor）%	准确率(Acc)%	句子识别率
单音素	38.64	-11.36	0
三音素	68.89	56.01	7.50

表5-3 不同模型识别率

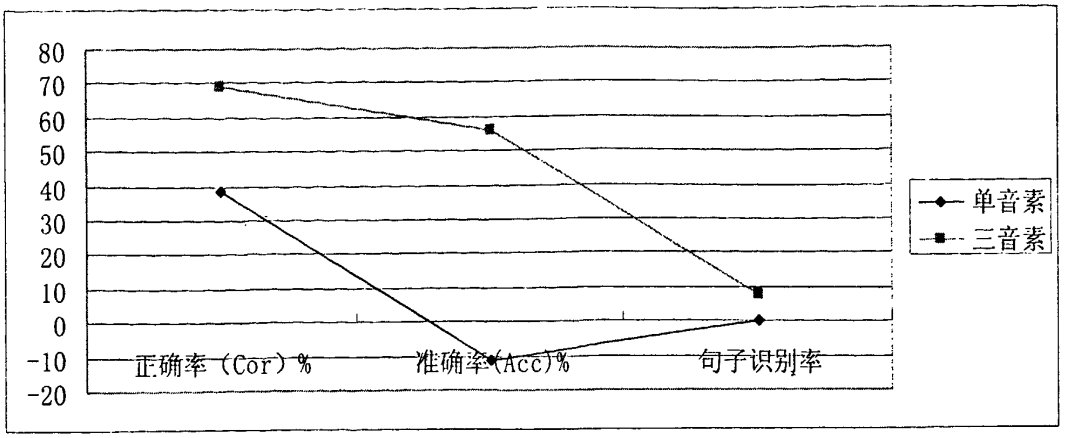


图 5-5 模型与识别率关系

在使用三音素建模的时候，模型数量比较庞大，可能会导致训练数据的稀疏性，为此在第四章介绍了使用决策树状态共享策略，其中决策树分裂的依据便是第四章提到的问题集设计。决策树的使用能给识别率带来明显的提升，而决策树的好坏关键就是问题集设计，试验结果如下表：

	正确率（Cor）%	准确率(Acc)%	句子识别率
三音素	68.89	56.01	7.50
决策树状态绑定 三音素	71.03	58.27	8.30

表 5-4 决策树状态共享提升识别率

这证明了本文设计的三音素模型以及采用决策树，进行基于 HMM 的语音识别，是有效的，改善了系统识别性能。

第6章 论文总结和展望

6.1总结

本论文主要是实现非特定人、大词汇量、连续语音识别系统，系统模型是HMM模型，并以HTK为实现平台，得到的实验结果具有实际的应用和参考价值。

论文的主要工作和贡献可以归纳为以下几个方面：

将语音识别模型从单音素模型提升到三音素模型，使得识别率有了非常明显的提升。连续大词汇量的语音识别中，发音之间通常有相互影响，单音素并不考虑前后之间的关联，而三音素模型考虑了语法和发音的上下文关系，更具有合理性，实验结果也有力的支持了这一观点；采用单音素、单混合度高斯混合模型，识别率只有38.64%，这离实际应用有极大差距，而三音素在单高斯混合度的条件下，识别率能达到将近70%，通过增加混合度，最后能达到90%以上。

系统测试了高斯混合度对识别系统的性能影响。在具有240句测试集上，完整而又系统的测试了识别率与高斯混合的关系，也印证了我们的假设：增加高斯混合度，识别率提高，继续增加混合度，由于训练数据量的稀疏性，识别率不升反降。因此对于具体的系统需要根据训练数据的多少和需要的识别时间（即系统复杂度），选择合适的模型混合度，识别时间也是特别要考虑的因素，当高斯混合度达到20以上时，每次识别时间都需要4个小时以上（这与具体机器有关，此处是相同计算机时结果），并且随着混合度的增加，识别时间也不断增加。对高斯混合度的测试也是本论文中语音识别测试的重点。

比较了不同特征参数对系统识别率的影响。特征参数的选取对识别率的影响是显著，目前主流的特征参数所示MFCC_0_D_A，实验中还测试了LPC、LPCC、MFCC_0、MFCC_0_D等其它特征参数的识别率，证明了MFCC倒谱系数的动态性能对最终的识别效果有较大的提高，而LPC并不适合作为连续语音识别的特征参数。

研究参数共享策略，分析了决策树的分裂过程，并设计可用于本实验的问题集，提升识别率，使得输出更加稳健。

6.2展望

语音识别是智能人机接口的重要工具，具有非常重要的意义和极其广泛的应用前景。

本文对实用的大词汇量、连续语音识别技术进行了系统研究。然而限于时间和精力关系，还有一些领域有待将来继续开展研究，主要如下：

1、语音数据库的丰富和完善问题。任何语音识别系统都要经过训练，而训练的内容就是语音文件，从某种程度上说，语音数据库就是我们的根基，它是否系统、完整、完备最终决定了识别系统的系能。TIMIT 数据库是本文作者从国外服务器上花了两天时间下载下来的良好的英文数据库，可惜我们还缺乏如此完备的中文数据库。

2、识别系统复杂性问题。语音识别系统无论是训练还是识别，系统复杂，计算量大，例如本文实验就是围绕着核心测试集，作者花了将近 10 天的时间进行测试才完成的，测试中用到的台式机有三台，其中有两台是 24 小时运行的。从这可以看出识别速度和识别时间是不得不考虑的因素。HTK 中我们可以使用-B 选项将文档的输入输入设置成二进制格式，这是提高识别效率的一种方式；而我们最希望的是识别算法的改进，目前的算法是 Viterbi Baum Welch 算法，更高速的识别算法值得研究。

3、带噪语音和自适应问题。要使识别系统能够在实际中应用，带噪语音识别和说话人自适应是需要实现的。对带噪语音，可以有两种方式处理，一是噪音消除，二是采用含噪的语音库进行训练。说话人自适应也是可以进一步研究的方向。

4、模型的改进问题。论文中实现了单音素模型、三音素模型。目前有学者提出了 5 音素模型，这想法是绝对合理的，但是要面对的问题是模型数量的急剧增加，以及对训练数据量提出了更为严格的要求，随着计算机技术的进步，结合完善的数据库，我们有理由相信这是可以实现的。

参考文献

[1] 王炳锡,屈丹,彭煊.通用语音识别基础.国防工业出版社[M],2005.1

[2] 汤玲.基于 HMM 模型的语音识别系统研究.国防科学技术大学研究生论文.2005.11

[3] 胡磊.基于隐马尔科夫模型的语音识别技术研究.武汉理工大学硕士学位论文.2007.5

[4] 彭荻.语音识别系统的声学建模研究.北京邮电大学硕士学位论文.2007.4

[5] 语音识别综述

[6] 张爱英.汉语大词汇量连续语音识别系统.科学中国人第 11 期,2008.

[7] 刘潇.语音识别系统关键技术研究.哈尔滨工程大学硕士学位论文.2006.2

[8] 王炳锡,屈丹,彭煊.通用语音识别基础.国防工业出版社[M],2005.1

[9] 王炳锡.语音编码.西安电子科技大学出版社[M],2002.7

[10] 杨尚国.噪声环境下语音识别系统研究.兰州理工大学硕士学位论文.2007.5

[11] 章学勇.连续数字语音识别系统的研究与实现.天津大学硕士学位论文.2006.1

[12] Anant G.Veeravalli,W.D.Pan,Reza Adhami.A tutorial on using hidden markov models for phoneme recogniton. ICASSP 2005,pp533-537

[13] Claudio Neves,Arlindo Veiga,Luis Sa.Efficient Noise-Robust speech recogniton front-end based on the ETSI standard.ICSP 2008 proceddings,pp.609-611.

[14] F.J.OWENS.SIGNAL PROCESSING OF SPEECH[M]. New York: McGraw-Hill, c1993.

[15] Antonio M.Peinado,Jose C.Segura.Speech Recognition over digital channels. IEEE Trans on Speech and Audio Processing,2002,7(3),page:331-346.

[16] Anant G.Veeravalli,W.D.Pan.A turorial on using hidden markov models for phoneme recognition. SSST,2005,page:154-157.

[17] O Siohan ,C Chesta ,C H Lee.Joint maximum a posteriori adaptation on transformation and HMM parameters[J].IEEE Trans on Speech and Audio Processing,2001,9(4),page:418-427.

[18] L.R.Rabiner.A tutorial on hidden markov models and selected application in speech recognition.Advances in Neural Information Processing Systems15,2003,page:10-29.

[19] 游展,肖熙,王作英.连续语音的三音子 DDBHMM 识别方法.清华大学学报,第 49 卷,第 4 期,2009.

[20] Anant G.Veeravalli, W.D.Pan,Reza Adhami,Paul G Cox.Phoneme Recogniton using Hidden Markov Models.Huntsville Simulation Conference,October 2004.

[21] Lawrence Rabiner.A tutorial on hidden markov models and selected applications in speech

recogniton.Proceedings of the IEEE,Vol 77,No2,pp.1307-1315,February 1989.

[22] spkrinfo.txt. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus.1990.10.12

[23] Steve Young,Dan Kershaw,Julian Odell,Dave Ollason,Valtcho Valtchev,Phil Woodland.The HTK Book.copyright 1995-1999 Microsoft Corporation,copyright 2001-2002 Cambridge University Engineering Department.

[24] Nicolas Moreau.HTK Basic Tutorial.2002.2

[25] Montri karnjanadecha,Stephen A.Zahorian.Signal modeling for isolated word recognition.ICASSP 1999,pp.293-296.

[26] Joseph W.Picone.Signal modeling Techniques in speech Recogniton.Proceeding of the IEEE,vol 81,No9,pp.1215-1247.September 1993.

[27] Xuedong Huang,Kai Fu Lee.On speaker-Independent,speaker-dependent and speaker-adaptive speech recogniton.IEEE transactions on speech and audio processing,vol 1,No 2, pp955-958,April 1993.

[28] X.Huang,Y.Ariki,M.Jack.Hidden Markov Models for speech Recogniton.Edinburgh,UK, Edinburgh unviersity press[M],1990.

[29] 倪崇嘉,刘文举.汉语大词汇量连续语音识别系统研究进展.中文信息学报第 23 卷,第 1 期.2009.1

[30] 蔡琴.基于 HTK 的维吾尔语连续数字语音识别研究.新疆大学硕士学位论文.2007.5

[31] P.C.Woodland,J.J.Odell,V.Valtchev,S.J.Young.Large vocabulary continuous speech recognition using HTK.Cambridge University Engineering Department.1994.

[32] M.Ferretti,S.Scarci.Large-vocabulary speech recogniton with speaker-adapted codebook and HMM parameters.Proc.Eurospeech,pp.154-156,1989.

[33] 袁里驰.基于改进的隐马尔科夫模型的语音识别方法.中南大学学报,第 39 卷第 6 期,2008.12

[34] 章学勇.连续数字语音识别系统的研究与实现.天津大学硕士学位论文.2006.1

[35] 杨尚国.噪声环境下语音识别系统研究.兰州理工大学硕士学位论文.2007.5

[36] X.Huang.Phoneme classification using semi-continuous hidden Markov models.IEEE Trans.Signal Processing,pp.1066-1071,May 1992.

[37] P.price,W.Fisher,J.Bernstein.A database for continuous speech recogniton in a 1000-word domain.IEEE conference on acoustics speech and signal Processing,pp.1282-1290,1988.

[38] 张杰,黄志同,王晓兰.语音识别中隐马尔科夫模型状态数的选取原则及研究.计算机工程与应用,2000.1

[39] 杨熙,苏娟,赵鹏.Matlab 环境下的语音识别系统.语音技术,第 31 卷,第 2 期,2007

[40] Rabiner,Juang B H.Fundamental of speech recogniton.New York:Prentice Hall,pp.15-20 1993.5.

[41] S.A.Selouani,D.O'Shaughnessy.Robustness of speech recogniton using genetic algorithms and a mel-cepstral subspace approach.ICASSP 2004,PP.201-203

[42] Makhoul J,Gray A.Linear Prediction of speech.Springer Verlay ,1996.6, pp120-133

[43] Mohamed.M.Azmi,Hesham Tolba.Noise robustness using different acoustic units.ICALIP procedding 2008,pp.1115-1119.

[44] 戴高乐.Perl 5 程序设计.清华大学出版社[M],2001.9

致 谢

首先感谢我的导师杨震教授在这三年中对我学习和生活各方面的关心和照顾。在本课题的研究过程中，自始至终得到了杨老师的悉心指导和帮助，杨老师渊博的学识、严谨的治学态度和乐于助人的品格，使我受益匪浅，并将对我今后的工作和学习产生重要的影响。

感谢我们实验室的崔景伍老师郑宝玉教授，得益于崔老师的热心帮助和管理，教研室一直保持着良好的学术环境，使研究工作事半功倍。感谢徐挺挺、卢亮亮、张兆城、邓文君、王薇、胡婷、黄平凤等同学在课题研究中提供的帮助。

感谢在百忙中抽出时间来审阅论文，参加答辩并给予我批评和指导的各位专家和教授。

我在南邮的近三年时光充满了温馨和欢乐。感谢我的家人，我的每一步成长都离不开他们的关爱和支持，正是由于他们的鼓励和无私的奉献，使我能够安心完成学业。

陈泉金

二零一零年三月

攻读硕士学位期间所发表的论文

陈泉金.基于 SCS 编码的数字双重水印算法.第 24 届南京地区研究生通信年会，2009.12

基于HTK的连续语音识别技术研究

作者：陈泉金
学位授予单位：南京邮电大学