华南理工大学

硕士学位论文

基于分段的说话人语音转换技术的研究

姓名：洪穗

申请学位级别：硕士

专业：通信与信息系统

指导教师：金连文;尹俊勋

20040601

摘 要

随着语音信号处理技术的不断发展和人们对人工智能的不断追求，说话人语音转换技术成为了一个新的研究课题。说话人语音转换技术是把源说话人说的语音转换为象是目标说话人说的语音的技术。说话人语音转换具有广泛的应用领域，比如文语转换（Text－to－Speech，TTS）系统、配音系统和翻译系统等。本文提出了一种基于分段的说话人语音转换方法，这种方法适用于单语种和跨语种的说话人语音转换，本文主要工作包括：

(1).在基于分段的说话人语音转换中，训练语句和转换语句需要进行切分。为了完成对语句的切分，本文采用隐马尔可夫模型的方法，利用 HTK 工具包分别实现了特定人语音切分系统和非特定人语音切分系统。

(2).本文提出了一种基于分段的说话人语音转换方法。和以往的方法比较，这种基于分段的说话人语音转换不要求源说话人和目标说话人是同样的训练语句，所以同时适用于单语种和跨语种的说话人语音转换。在这种基于分段的说话人语音转换中，本文采用“pitch＋mel 倒谱＋MLSA 滤波器”语音编码器，提出了一种基于修改 mel 倒谱和基音周期参数的说话人语音转换方法。在对频谱的转换中，先对每段基本语音的 mel 倒谱参数训练高斯混合模型，求出一个转换函数，然后用转换函数对 mel 倒谱参数进行转换。而基音周期的转换则采用一个全局的转换公式，对基音周期的数值和范围进行修改。

(3).本文运用所提出的基于分段的说话人语音转换方法实现了单语种(英语）说话人的转换。在单语种（英语）说话人语音转换中，采用的语音段库是 41 个单音素库（包括一个静音）。通过分析元音转换前后的 FFT 频谱，本文得出结论：转换后的语音的 FFT 频谱更接近于目标说话人语音的 FFT 频谱。而且，通过主观听觉判断，转换后的语音更象是目标说话人的语音。因此说明这种基于分段的单语种（英语）转换是有效的。

(4).为了实现跨语种（中英）说话人语音转换，本文研究了中英文的语言特点，特别是两种语言的单音素之间的异同点。通过比较，发现英文中大部分英语音素可以在中文中找到相对应的音素，有小部分的英文音素找不到中文对应的音素。为了实现这小部分中英文不对应的音素的转换，本文提出了二叉树的方法来进行跨语种说话人语音转换。实验表明，二叉树的方法可以解决中英文不对应的音素的转换问题，跨语种（中英）的说话人语音转换得以实现。

关键词： 说话人语音转换；分段；单语种；跨语种；高斯混合模型（GMM）；二叉树

ABSTRACT

With the development of the speech processing technology and the human's constantly pursuing AI (artificial intelligence), voice conversion become a new popular topic in research areas. Voice conversion is a technology that modifies the speech signals uttered by a source speaker to sound as if a target speaker had spoken it. Voice conversion technology has many applications, such as TTS (Text-to-Speech) system, dubbing system and translating system. A segment-based voice conversion method that can be applied to both the single-language and the cross-language voice conversion system is proposed in this thesis. The main works of this thesis are:

(1). The training speech and the modified speech must be segmented in the segment-based voice conversion system. So, in order to segment the speech, a speaker-dependent speech-segmented system and a speaker-independent speech-segmented system are implemented by using HMM and HTK.

(2). A segment-based voice conversion method is proposed in this thesis. Compared with the past methods, the proposed segment-based voice conversion method doesn't require a "parallel" speech corpus in the voices of source and target speaker to be the training speech corpus, so this method can be applied to both the single-language and the cross-language voice conversion system. By using "pitch + mel-cepstral + MLSA filter " speech coder , a voice conversion system based on modified parameters of pitch and mel-cepstral.In the conversion course of spectrum, GMM for mel-cepstral vectors of each speech units are trained, the conversion function for mel-cepstral vectors of each speech units are computed, and then mel-cepstral vectors are modified using the conversion functions. To modified pitch, a global conversion function is used to modified pitch value and range.

(3). The proposed segment-based voice conversion method in this thesis is applied to a single-language (English) voice conversion system. In the system, the speech unit database consists of 41 monophones (including a silence unit). A conclusion is got from experiments in this thesis that FFT spectrum of the modified speech is similar to the FFT spectrum of speech uttered by the target speaker. And, the modified speech sounds as if the target speaker had spoken it by subjective tests. Therefore, a single-language (English) voice conversion system can be carried out by using the segment-based voice conversion method proposed in this thesis.

(4). To carry out the cross-language (Chinese-English) voice conversion system, the language characters of Chinese and English, especially the similarities and differences of Chinese and English monophones, are studied in this thesis. It is found that a majority of English monophones have their corresponding Chinese monophones and the rest haven't their corresponding Chinese monophones. In order to modify the monophones which haven't their corresponding Chinese monophones, a bitree method is proposed in this thesis. Experiment shows that the bitree method can modify the monophones which haven't their corresponding Chinese monophones, and a cross-language (Chinese-English) voice conversion system can be carried out by using the segment-based voice conversion method proposed in this thesis.

Keywords: voice conversion, segment-based, single-language, cross-language, GMM, bitree

华南理工大学
学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：[illegible]　　日期：2004年6月8日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华南理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密□，在____年解密后适用本授权书。

本学位论文属于

不保密☑。

（请在以上相应方框内打“√”）

作者签名：[illegible]　　日期：2004年6月8日

导师签名：[illegible]　　日期：2004年6月8日

第一章 绪论

1.1 课题的意义

近年来，人工智能已逐渐成为人们生活中的主题，因此“如何使机器更加拟人化”是科研领域的一个主要方向。“说话人语音转换技术”[1]是“人工智能”中较新的研究课题，指的是一项改变或修饰说话人个性特征的技术，即说话人A说的语音被转换为象是说话人B说的语音。说话人语音转换技术又分成两种：单语种说话人语音转换和跨语种说话人语音转换。单语种说话人语音转换是指A和B说同一种语言，跨语种说话人语音转换是指A和B说不同的语言，比如A说的是英语，而B不会说英语，只会说中文。跨语种说话人语音转换就把A说的英语进行处理成像是B说的，让原本不会说英语的B可以说出英语。说话人A称为源说话人，说话人B称为目标说话人。

说话人语音转换技术具有广泛的应用领域。比如在电影配音中，人们希望配音的声音象是某人的声音，这就可以采用说话人语音转换技术来达到这个目的。在语音翻译系统中，翻译出来的声音不同于原来的声音，人们希望翻译后的语言仍可以用原有的声音讲出来，这就需要说话人语音转换技术。在语音合成系统中，人们希望听到丰富的合成声音，这也需要说话人语音转换技术。而且，在文语转换（Text-to-Speech, TTS）系统中，还常常会碰到这样的问题：比如在一段中文句子中偶尔会出现一个或多个英语单词。对于这种情况，在现有的 TTS 系统中，一般有两个解决方法：一种就是直接把英语单词的字母一个个读出来，另一种办法就是合成汉语的时候用汉语的合成系统，合成英语时用英语的合成系统，但是在一般情况下，合成汉语的声音不同于英语的声音，这就使得一句话中同时出现两种声音，这是人们所不能接受的。为了克服出现两种声音这个问题，这就要进行汉英之间的跨语种说话人语音转换，就是找到一种有效的办法，使得合成出来的英语和汉语的声音是同一个人。

近年来，从事说话人语音转换技术的学者一般都是针对单语种，跨语种的说话人语音转换技术的研究相对很少。本文的目的是找到一种有效的方法进行说话人语音转换，特别是希望这种方法可以运用到跨语种说话人语音转换中。

1.2 语音信号的基本特性

为了更好的理解说话人语音转换技术，这里简单地介绍一下语音信号的一些基本特性和一些基本概念[2,3]。

1.2.1 语音信号的时域特性

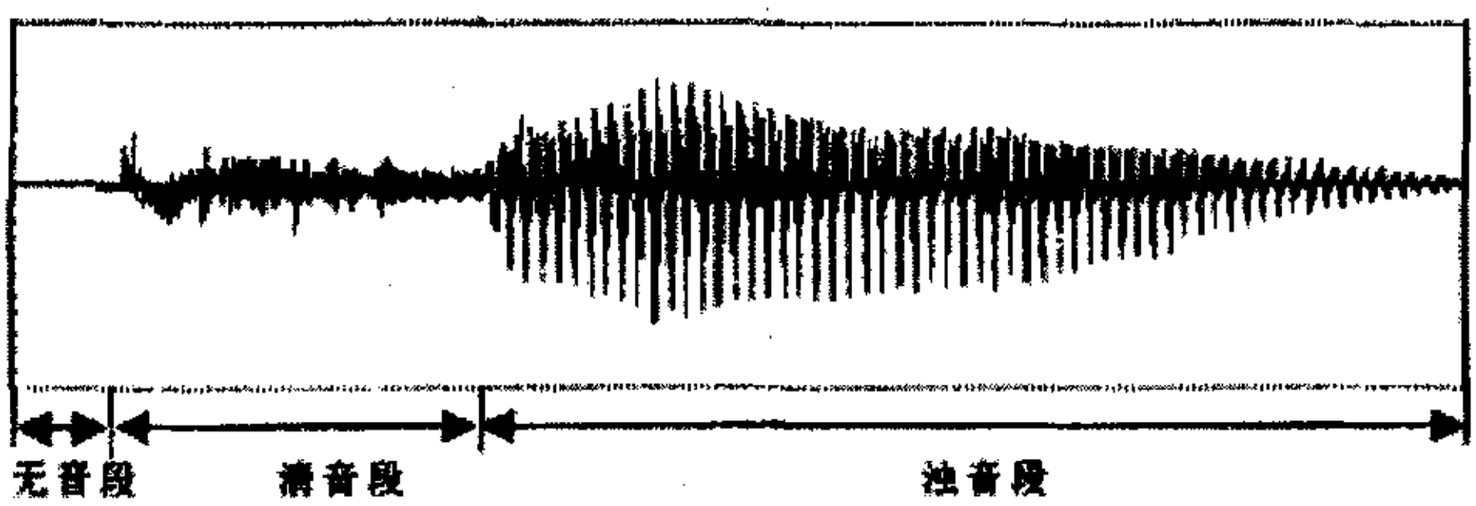


图1-1 语音波形图

Fig.1-1 waveform of speech

由图1-1可以看出语音可分为三部分:无音段、清音段和浊音段。无音段没有语音信号的存在，在背景噪声较低的情况下，幅度近似为零。清音信号的幅度很小，而且没有规律，类似于随机噪声。而浊音信号波形的幅度较大，波形的上下起伏近似呈现周期性，称之为准周期。语音信号在时域上有两个重要的物理参量，即短时平均能量和短时平均跨零率。

从图1-1中可以看出语音信号的幅度随着时间变化有着较大的变化，可用短时平均能量来反映语音信号这一特性，短时平均能量通常定义如下：

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

w(n)是个窗函数，N为窗长。不同的窗对于短时平均能量有不同的影响，一般情况下会用到两种窗：一种是直角窗，一种是hamming窗。

短时平均能量的主要用途：

（1）可以从清音中区分出浊音来，因为浊音时短时平均能量要比清音时大得多。

（2）可以用来区分清音与浊音的分界，无声与有声的分界等。

（3）作为一种超音段信息，用于语言识别中。

在离散时间信号下，如果相邻的采样点具有不同的代数符号就称为发生了跨零。跨零率指的是每秒内信号值通过零值的次数。跨零数可以作为信号序列的“频

率”的一个简单度量。对于窄带信号，这种度量是较精确的，由于语音信号是一类宽带的局部平稳信号序列，所以用平均跨零率来表示语音频率特性不确切，但用短时平均跨零率可以粗略估计语音信号的频率特性。

从语音信号可以直观地看出，清音段的跨零率最高，而浊音段次之，无音段的跨零率最低。频域上各个信号的特性也证实了这一点，清音语音能量集中在较高的频率上，浊音语音能量集中在3kHz以下。

1.2.2 语音信号的频域特性

如果从语音流中利用加窗的方法取出其中的一个短段，再对其进行傅立叶变换就可以得到该段语音的短时谱。清音的短时谱类似于随机信号的频谱。而浊音信号的短时谱有两个特点:第一，有明显的周期性起伏结构，这是因为浊音的激励源为周期脉冲气流。第二，频谱中明显地具有几个凸起点，它们的出现频率与声道的谐振频率相对应。这些凸起点称为“共振峰”（Formant），其频率称为共振峰频率。共振峰按频率由低到高排列为第一共振峰，第二共振峰，…，相应的频率用F1, F2,…来表示，一般浊音中前三个，尤其是前两个对于区别不同语音是至关重要的。

语音信号还有一个很重要的特性“短时性”，即在某个短时段，可以认为语音信号的特征是不变的，这段时间一般可取为5-50ms，因此我们可以采用各种短时处理方法，例如“短时能量”，“短时频谱”等。

1.3 语音的个人特性及其相关参数

要进行说话人语音转换，首先要研究语音的个人特性。如今，国内外关于语音个人特性的研究已取得一定成果[3]。

1.3.1 语音的个人特性

语音波形携带了很多信息:语言学、音段、超音段、超语言学等等，在这些信息中，语言学是当今语音技术领域的主要方向。但同时，语音信号中的非语言学信息，如音质、声音的个人特性，在语音识别、理解领域以及日常人们的对话交流中也起到很大的作用。声音的个人特性(Voice individuality)不仅帮助我们确认说话者，而且使我们的生活丰富多彩。不过，对于非特定人语音识别工作来说，Voice individuality却是一个需要克服的障碍，说话人规整、自适应就是为了解决该问题而发展起来的。语音分析和合成技术的发展已使分析与语音音质有关的

声学特征成为可能，这样将来有一天我们就可以精确地控制计算机语音的个人特性。

与声音个人特性相关的因素可分为社会心理学和生理学两个方面。一个人的说话风格与他的年龄、社会地位、方言等因素有关。说话人风格从声学的角度上看主要体现在韵律特征上，如基频曲线、时长、速率、节奏、停顿、能量，等等。而声音的音质主要是由发音器官的生理、物理特性决定的，但同时受说话人的情绪状态的影响，语音的音色则主要体现在由声门波频率及频谱，声道谱能量等方面。现有的语音技术还不能精确的处理反映说话人风格的韵律特征，而把重点放在改变反映发音器官物理特性的声学参数上。

1.3.2 体现声音个人特性的声学特征参数

声音的个人特性由两大类声学特征共同作用影响的:声源和声道共振。被认为与此相关的声学参数，主要有:

1. 声源参数:(1)平均基频。(2)基频曲线轮廓。(3)基频变化范围。(4)声门波形状。

2. 声道共振参数:(1)频谱包络形状和谱倾斜。(2)共振峰值。(3)共振峰走向。(4)长时平均频谱。(5)共振峰带宽。

对说话人特征的研究已有很长的时间，早期的心理学和语音学的研究揭示了声学参数和说话人年龄、性别等身体特性的关系。最近的研究主要从语音技术和说话人识别的角度来考虑。Mastsumot等人研究了基频(F0)，共振峰，谱包络等声学参数对男声元音的贡献，得出结论，即基频(F0)是说话人特性的最重要的参数，其次是共振峰，再次是F0变动范围及声源谱倾斜。Furui研究了不同说话人的心理和物理上的差异的关系，发现通过倒谱系数光滑的长时平均谱起很大的作用。Nakatsui et al.通过转换三个说话人元音的声源和声道共振参数，认为F0声道共振参数起更大的作用。Itosh和Satio则通过研究元音，音节等的语音合成参数，认为频谱包络是最重要的参数，其次是基频。从以上的研究我们可以得知，不存在唯一特殊的声学参数携带所有的个人特征信息，语音音质是许多语音参数共同作用的结果。

1.4 单语种说话人语音转换的发展

从上世纪八十年代中有人提出基于 VQ 的说话人语音转换[1]的方法以来，单语种说话人语音转换技术已经经过了将近二十年的发展[1,4-10]。总结这二十年的研究成果，单语种说话人语音转换技术主要分成三种：基于 VQ 的方法[1]，基于统计模

型的方法[5]，和基于分段转换的方法[7]。

1.4.1 基于 VQ 的方法

基于 VQ 的方法是一种码本对应的方法，即分别求出源说话人 A 和目标说话人 B 语音的码本，然后求出源说话人 A 与目标说话人 B 码本的对应关系，也就是说求出映射码本（mapping codebook），转换的时候利用映射码本来进行转换。这种方法要求源说话人和目标说话人说相同的训练语句。方法主要分为训练过程(图 1-2)和语音转换合成过程(图 1-3)。训练过程是生成各类参数的映射码本，语音转换合成过程就是利用映射码本实现转换并生成语音的过程。

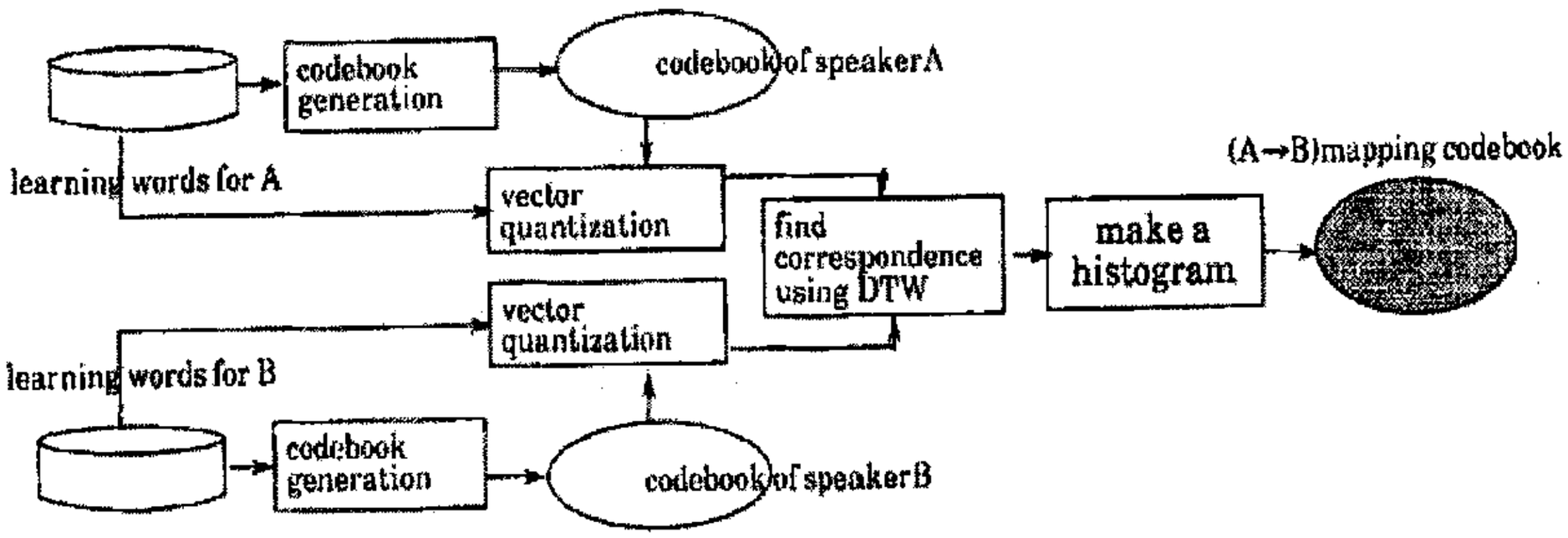


图 1-2 生成码本的方法[1]

Fig. 1-2 methord for generating a mapping codebook[1]

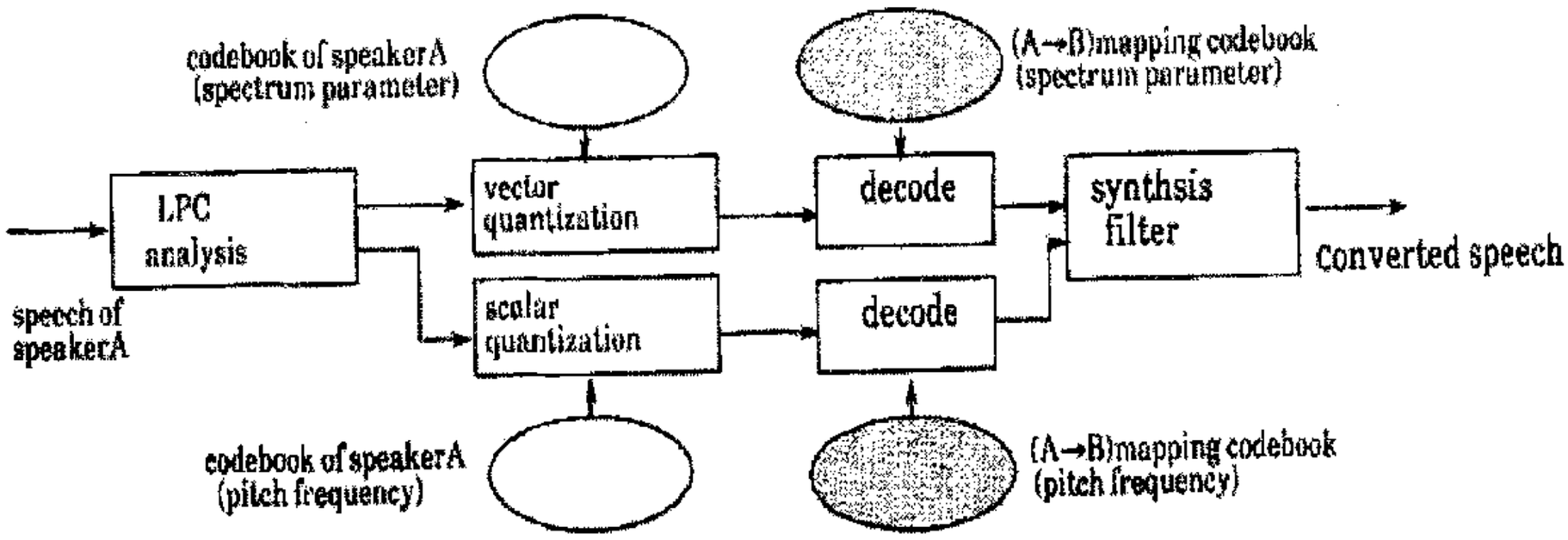


图 1-3 从源说话人 A 向目标说话人 B 语音转换的框图[1]

Fig. 1-3 Block Diagram of a Conversion from Speaker A to Speaker B[1]

1.4.2 基于统计模型的方法

基于统计模型的说话人语音转换分别对源说话人和目标说话人的语音进行

统计，用统计模型来分别表示源说话人和目标说话人语音特征，然后找出两个模型之间的转换函数。在转换的时候用转换函数来把源特征参数转变成新的特征参数。这种方法也要求源说话人和目标说话人说相同的训练语句。目前最经常用的统计模型是高斯混合模型（GMM）。基于统计模型的说话人语音转换方法包括训练过程（图 1-4）和转换过程（图 1-5）。

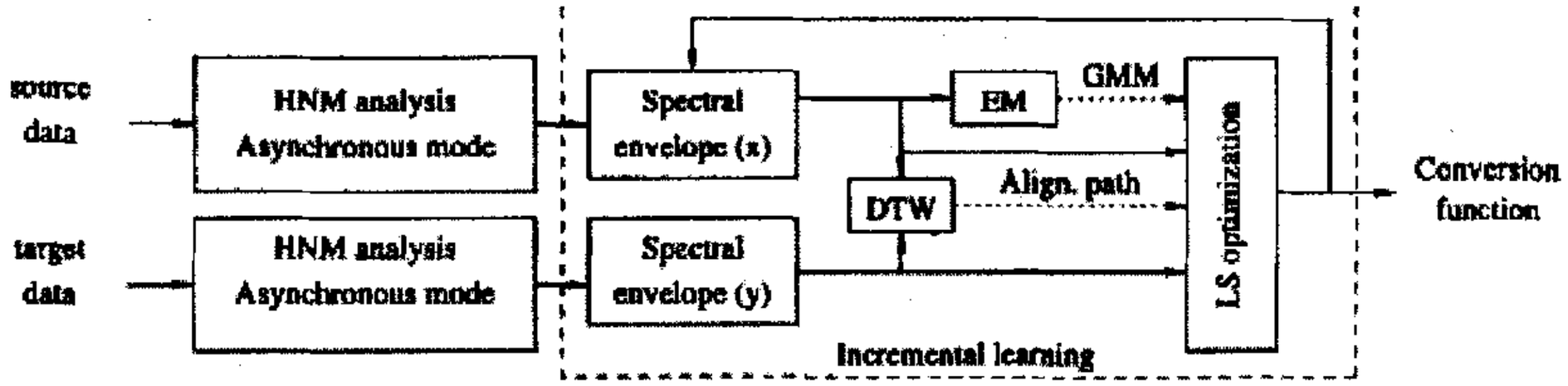


图 1-4 GMM 训练过程框图[5]

Fig. 1-4 Block diagram of leaning procedure[5]

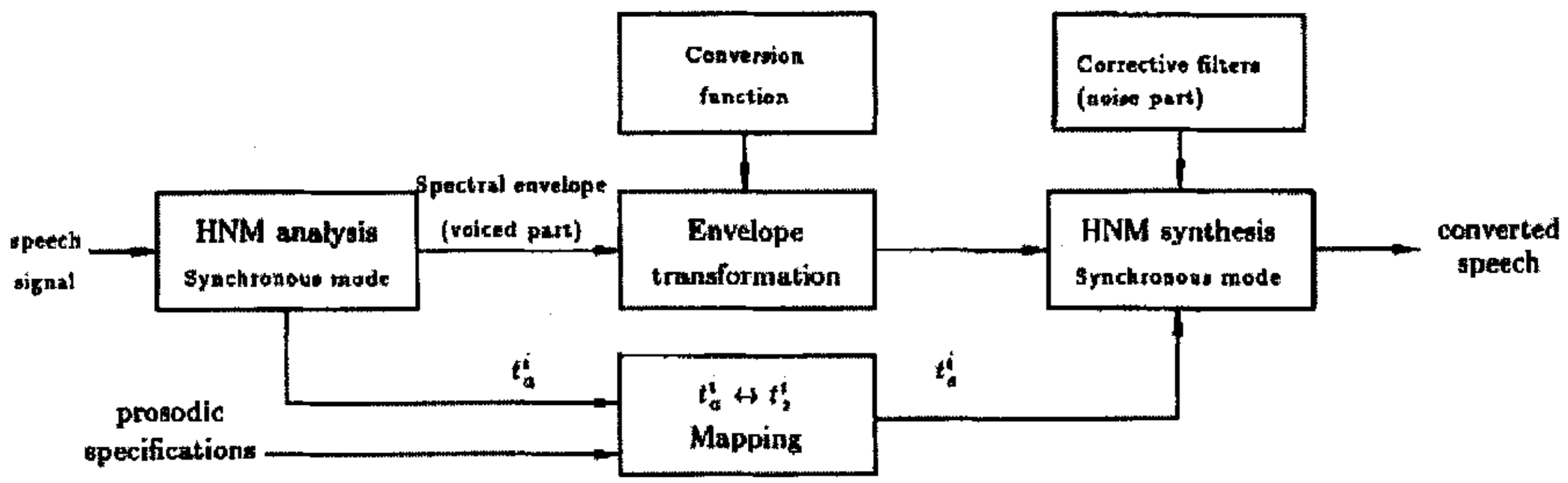


图 1-5 转换过程框图[5]

Fig. 1-5 Block diagram of conversion procedure[5]

1.4.3 基于分段转换的方法

基于分段转换的说话人语音转换是指转换的单元不是帧，而是一段语音。它主要的优势也是有别于前两种方法的地方是源说话人和目标说话人可以说不同的训练语句。这种方法先把源说话人的语音进行切分，这些语音段被目标说话人的相应的语音段进行替换。[7]中的说话人语音转换方法如图 1-6。

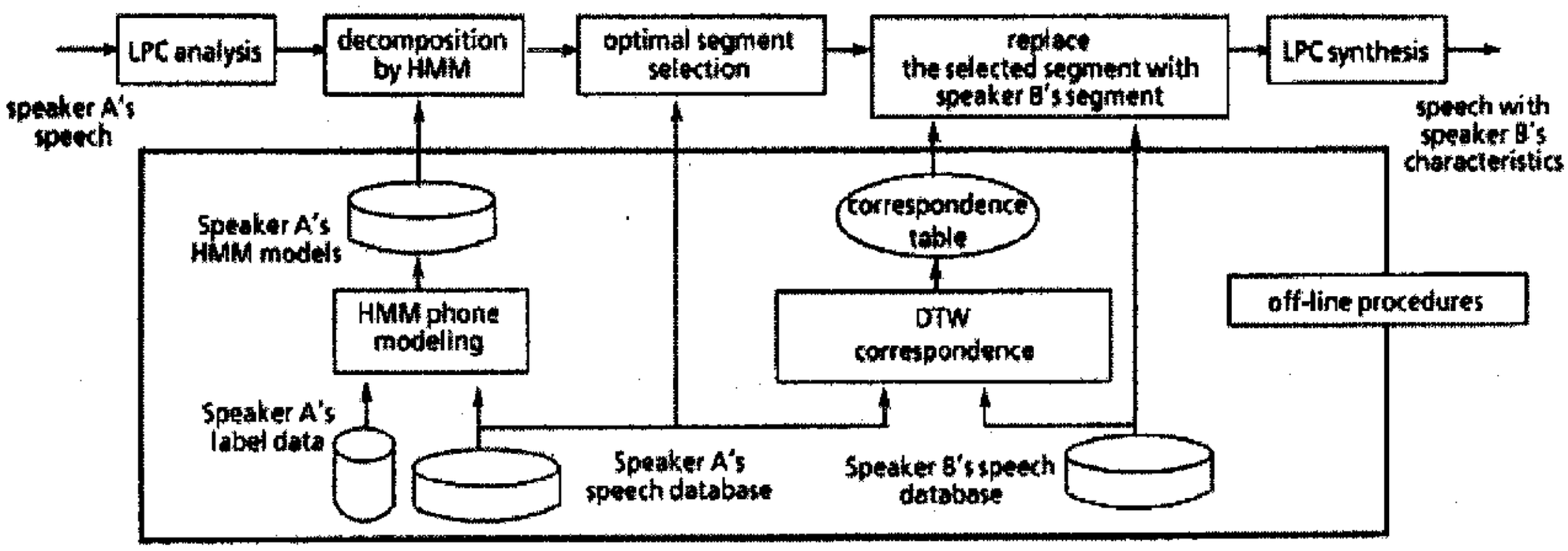


图 1-6 基于分段说话人语音转换[7]

Fig.1-6 Block diagram of segment-based voice conversion[7]

1.5 跨语种说话人语音转换的发展

跨语种说话人语音转换相对于同语种说话人语音转换来说，具有一定的难度，所以研究也相对比较少，现有的方法主要有基于映射码本的方法[11]和基于统计模型（GMM）的方法[12,13]。跨语种说话人语音转换最早是上个世纪九十年代初由日本人提出的，他们进行了日语和英语之间的转换实验，采用的是基于映射码本的方法[11]。他们的方法是先对目标说话人（说日语）建一个码本，合成英语的时候就直接用这个由日语建成的码本进行合成，如图 1-7。

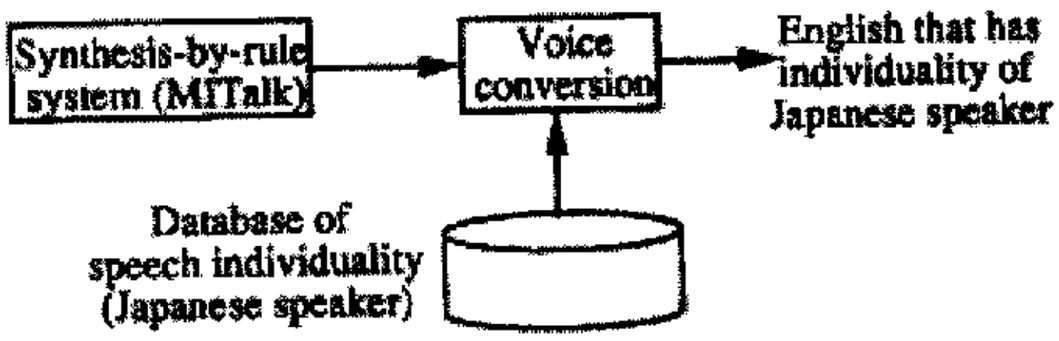


图 1-7 跨语种说话人语音转换模型[11]

Fig.1-7 Cross-languge voice conversion model[11]

但是，不同语言的码本肯定有差别，这种直接替换的方法出来的效果有一定的极限，对于实现质量高的语音是有一定困难的。

到了近期，又有人提出了一种基于统计模型的方法[12,13]。这种方法要求源说话人会说两种语言。先求出源说话人和目标说话人说的日语之间的转换函数，在把源说话人说的英语转换成象目标说话人的语音的过程中，就直接用这个转换函数进行转换。但是这种方法对源说话人的语言能力有要求，当源说话人无法说两门语言时，这种方法就不适用了。

1.6 本文的主要工作与章节安排

本文着重研究了基于分段的说话人语音转换，并和基于统计模型的转换方法结合起来，提出一种基于分段的说话人语音转换方法，这种方法不像[7]中直接用目标说话人的语音段直接替换源说话人的语音段，而是把源说话人的语音段进行转换，得到象是目标说话人的语音段，从而达到转换的目的。本文实现了基于这种说话人语音转换方法的单语种说话人语音转换。而且，针对跨语种说话人语音转换的特点，本文提出了二叉树的转换方法，实现了跨语种说话人语音转换，这种跨语种说话人语音转换的方法不象[12,13]中要求源说话人精通这两门语言。为了需要，本文还研究并实现了用于切分语音的语音切分系统。本文主要工作包括：

（1）语音切分系统的实现：利用 HTK 工具包，在 Linux 操作系统下实现了英语的语音切分系统，用于对训练语句和转换语句的切分和标注。

（2）基于分段的说话人语音转换方法的研究：研究了用于转换的语音编解码模型、频谱参数的转换方法、基音频率的转换方法。

（3）基于分段的单语种说话人语音转换的研究：在 Matlab 平台下实现了 GMM 模型的训练和基于分段的单语种说话人语音转换的方法。

（4）基于分段的跨语种说话人语音转换的研究：比较了中英文的语言特征，重点是比较中英文音素的异同点，并针对跨语种说话人语音转换的特点，提出二叉树的转换方法，最终实现跨语种说话人语音转换。

本文提出的基于分段的说话人语音转换方法克服了训练语句的限制，并采用“pitch+mel-倒谱＋MLSA 滤波器”作为说话人语音转换中的语音编码器，而且为了解决跨语种（中英）中英文音素不对应的转换，提出二叉树的转换方法，使得这种基于分段的说话人语音转换方法适用于单语种和跨语种说话人语音转换。

本文的章节安排如下：

第一章是这整篇论文的绪论，重点介绍了项目的意义、单语种说话人语音转换和跨语种说话人语音转换在国内外的发展情况。目的是为了让读者对这个研究方向有初步的印象。

第二章分别设计了一个特定人和非特定人的语音切分系统。语音切分系统在基于分段的说话人语音转换中是必不可少的，要用它对训练语句进行标注切分，以及对需要转换的语音进行标注。

第三章讨论了高斯混合模型（GMM）的基础知识以及训练高斯混合模型参数的方法——EM（Expectation－Maximization 最大期望）算法。

第四章提出了一种基于分段的说话人语音转换方法。本章论述了在试验中用到的“pitch+mel-倒谱＋MLSA 滤波器”语音编码器，以及论述了本文提出的基于

分段的说话人语音转换方法中频谱和基音频率的具体转换。

第五章实现了基于分段的单语种（英语）说话人语音转换。这个方法克服了训练语句的局限性，即不需要源说话人和目标说话人的训练语句是相同的。

第六章比较了中英的语言特点，重点比较中英文音素的异同点，并对基于分段的说话人语音转换方法提出基于二叉树的转换方法，解决了中英文音素不对应的转换问题，最后实现了基于分段的跨语种（中英）说话人语音转换。

本文紧接着给出了通过整个研究课题所获得的结论，并给出将来的工作方向和改进方法。

最后文章给出了参考文献和致谢。

第二章 语音切分系统

语音切分系统实际上是一个连续语音识别系统，它可以把连续的语音切分出所含有的语音段以及标明语音段的起始和终止时间。语音切分系统在基于分段的说话人语音转换中是必不可少的，要用它对训练语句进行切分标注，以及对需要转换的语音进行标注。

隐 Markov 模型（Hidden Markov Models，简称为 HMM）是最常用的连续语音识别的方法，本文的语音切分系统就采用了隐 Markov 模型（HMM）作为切分标注方法。

2.1 HMM 基本理论

隐 Markov 模型，作为语音信号的一种统计模型，今天正在语音识别系统中应用广泛，这是因为隐 Markov 模型是一种从统计学的角度比较好的描述语音波形特性的模型[14]。隐 Markov 模型是在 Markov 链的基础上发展起来的。由于实际问题比 Markov 链模型所描述的更为复杂，观察到的事件并不是与状态一一对应，而是通过一组概率分布相联系，这样的模型就称为隐 Markov 模型。它是一种双重随机过程。其中之一是 Markov 链，这是基本随机过程，它描述状态的转移。另一个随机过程描述状态和观察值之间的统计对应关系。这样，站在观察者的角度，只能看到观察值，不像 Markov 链模型中观察值和状态一一对应，不能直接看到状态，而是通过一个随机过程去感知状态的存在及其特性。因此称为“隐” Markov 模型，即 HMM[15]。

2.1.1 HMM 的定义

在对 HMM 下定义之前，先来看一个著名的说明 HMM 概念的例子：球和缸（Ball and Urn）实验[15]。

设有 N 个缸，每个缸中装有很多彩色的球，球的颜色由一组概率分布描述。实验是这样进行的，根据某个初始概率分布，随机地选择 N 个缸中的一个，例如第 i 个缸，再根据这个缸中彩色颜色的概率分布，随机地选择一个球，记下球的颜色，记为 O_1，再把球放回缸中，又根据表述缸的转移概率分布，随机选择下一个缸，例如，第 j 个缸，再从缸中随机选一个球，记下球的颜色，记为 O_2，一直进行下去。可以得到一个描述球的颜色的序列 O_1，O_2，...，由于这是观察到的事

件，因而称之为观察值序列。但缸之间的转移以及每次选取的缸被隐藏起来了，并不能直接观察到。而且，从每个缸中选取球的颜色并不是与缸一一对应，而是由该缸中彩球颜色概率分布随机决定的。此外，每次选取哪个缸则是由一组转移概率所决定。

有了球和缸的实验，我们可以给出 HMM 的定义，或者说，一个 HMM 可以由下列参数描述：

1. N：模型中 Markov 链状态数目。记 N 个状态为 θ_1，...，θ_N，记 t 时刻 Markov 链所处状态为 q_t，q_t 属于（θ_1，...，θ_N）。在球与缸实验中的缸就相当于状态。

2. M：每个状态对应的可能的观察值数目。记 M 个观察值为 V_1，...，V_M，记 t 时刻观察到的观察值为 O_t，其中 O_t 属于（V_1，...，V_M），在球和缸实验中所选彩球的颜色，就是观察值。

3. π：初始状态概率矢量，$\pi=(\pi_1,...,\pi_N)$，其中 $\pi_i=P(q_1=\theta_i)$，$1\le i\le N$。在球和缸实验中指开始时选择某个缸的概率。

4. A：状态转移概率矩阵，$A=(a_{ij})_{N\times N}$，其中 $a_{ij}=P(q_{t+1}=\theta_j/q_t=\theta_i)$，$1\le i,j\le N$。在球与缸实验中指描述每次在当前选取的缸的条件下选取下一个缸的概率。

5. B：观察值概率矩阵，$B=(b_{jk})_{N\times M}$，其中 $b_{jk}=P(O_t=V_k/q_t=\theta_j)$，$1\le j\le N,1\le k\le M$。在球与缸试验中，$b_{jk}$ 就是第 j 个缸中球的颜色 k 出现的概率。

这样，可以记一个 HMM 为

$$\lambda=(N, M, \pi, A, B) \tag{2-1}$$

或简写为 $\lambda=(\pi, A, B)$

更形象地说，HMM 可以分为两部分，一个是 Markov 链，由 π、A 描述，产生的输出为状态序列，另一个是一个随机过程，由 B 描述，产生的输出为观察值序列。

通常认为语音信号是一个短时平稳的随机过程，在 10－30ms 的短时段内，语音信号是平稳的，而从整体来看，语音信号是时变的。HMM 模型成功地描述了这种短时平稳的信号。由两个相互关联的随机过程共同描述了语音信号的统计特性，平稳段的信号由对应的状态观察值的随机过程描述，而短时平稳段向下一短时平稳段的转变则由隐含的马尔可夫链的状态概率来描述。

2.1.2 HMM 基本算法

HMM 应用于语音信号处理需要解决三个基本问题：

1. 对给定的观察值序列 $O=(O_1, O_2,..., O_T)$ 和模型 $\lambda=(\pi, A, B)$，如何计算观察值序列 O 的概率 $P(O/\lambda)$？经典算法是前向－后向算法。

2. 如何调整模型参数 $\lambda=(\pi, A, B)$ 使得观察值序列 O 的产生概率 $P(O/\lambda)$ 最大？这实际上是 HMM 模型的训练问题，即从已知模式中获得描述该模型的

HMM 模型参数。经典算法是 Baum-Welch 算法。

3. 对给定的观察值序列 $O=(O_1, O_2,..., O_T)$和模型 $\lambda=(\pi, A, B)$，如何获得相应的状态转移序列 $Q=(q_1, q_2,..., q_T)$？经典算法是 Viterbi 算法。

2.1.2.1 前向－后向算法

这个算法是用来计算给定的观察值序列 $O=(O_1, O_2,..., O_T)$ 和模型 $\lambda=(\pi, A, B)$，观察值序列 O 的概率 $P(O/\lambda)$。但是如果用直接的方法求 $P(O/\lambda)$，计算量是十分惊人的。Baum 等人提出的前向－后向算法是计算 $P(O/\lambda)$的有效算法。

1. 前向算法

定义前向变量

$$\alpha_t(i)=P(O_1,O_2,...,O_t,q_t=\theta_i/\lambda),\quad 1\le t\le T \tag{2-2}$$

有以下递推公式：

初始化：

$$\alpha_1(i)=\pi_i b_i(O_1),\quad 1\le i\le N$$

递归：

$$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(O_{t+1}),\quad 1\le t\le T-1,1\le j\le N \tag{2-3}$$

结束：

$$P(O/\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{2-4}$$

其中 $b_j(O_{t+1})=b_{jk}\,|\,o_{t+1}=v_k$

2. 后向算法：

定义后向变量为

$$\beta_t(i)=P(O_{t+1},O_{t+2},...,O_T,q_t=\theta_i/\lambda),\quad 1\le t\le T-1 \tag{2-5}$$

其中 $\beta_T(i)=1$

有以下递推公式：

初始化：

$$\beta_T(i)=1,\quad 1\le i\le N$$

递归：

$$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}(j),\quad t=T-1,T-2,...1,1\le i\le N \tag{2-6}$$

结束：

$$P(O/\lambda)=\sum_{i=1}^{N}\beta_1(i) \tag{2-7}$$

2.1.2.2 Baum-Welch 算法

Baum-Welch 算法解决了 HMM 模型的训练问题，即 HMM 模型的参数估计问题。由公式（2-2）和公式（2-5）定义的前向和后向变量，可得

$$P(O/\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)\,,\quad 1\le t\le T-1 \tag{2-8}$$

对给定的观察值序列 O=（O_1，O_2,...，O_T），Baum-Welch 算法求得使 P（O/λ）最大的 HMM 模型参数 λ=（π，A，B）。

定义 $\xi_t(i,j)$ 为给定训练序列 O 和模型 λ=（π，A，B）时，时刻 t 时 Markov 链处于 θ_i 状态和时刻 t+1 时处于 θ_j 状态的概率，即

$$\xi_t(i,j)=P(O,q_t=\theta_i,q_{t+1}=\theta_j/\lambda) \tag{2-9}$$

结合公式（2-8）式可得

$$\xi_t(i,j)=\left[\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)\right]/P(O/\lambda) \tag{2-10}$$

时刻 t 时 Markov 链处于状态 θ_i 的概率：

$$\xi_t(i)=P(O,q_t=\theta_i/\lambda)=\sum_{j=1}^{N}\xi_t(i,j)=\alpha_t(i)\beta_t(i)/P(O/\lambda) \tag{2-11}$$

因此，$\sum_{t=1}^{T-1}\xi_t(i)$ 表示从 θ_i 状态转移出去的次数的数学期望，$\sum_{t=1}^{T-1}\xi_t(i,j)$ 表示从 θ_i 状态转移到 θ_j 状态的次数的数学期望。

因此，导出了 Baum-Welch 算法中著名的重估公式：

$$\overline{\pi}=\xi_t(i)\,,\quad 1\le i\le N \tag{2-12}$$

$$\overline{a}_{ij}=\sum_{t=1}^{T-1}\xi_t(i,j)\Big/\sum_{t=1}^{T-1}\xi_t(i)\,,\quad 1\le i,\ j\le N \tag{2-13}$$

$$\overline{b}_{jk}=\sum_{\substack{t=1\\O_t=V_k}}^{T}\xi_t(j)\Big/\sum_{t=1}^{T}\xi_t(j)\,,\quad 1\le j\le N,1\le k\le M \tag{2-14}$$

2.1.2.3 Viterbi 算法

对给定的观察值序列 O=（O_1，O_2,...，O_T）和模型 λ=（π，A，B），Viterbi 算法可以获得使 P（Q,O/λ）最大时相应的状态序列 $Q=(q_1,q_2,...,q_T)$，实际上是一种路

径搜索算法。

定义 $\delta_t(i)$ 为 t 时刻沿一条路径 $q_1,q_2,...,q_t$，且 $q_t=\theta_i$，产生观察值序列 O=（O_1，O_2,...，O_t）的最大概率：

$$\delta_t(i)=\max_{q_1,q_2,\cdots,q_{t-1}} P(q_1,q_2,...,q_t,\ q_t=\theta_i,\ O_1,O_2,...,O_t/\lambda) \tag{2-15}$$

递归求最佳状态序列 Q：

初始化：

$$\delta_1(i)=\pi_i b_i(O_1),\quad 1\le i\le N$$

$$\varphi_1(i)=0,\quad 1\le i\le N$$

递归：

$$\delta_t(j)=\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(O_t),\quad 2\le t\le T,1\le j\le N \tag{2-16}$$

$$\varphi_t(j)=\arg\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right],\quad 2\le t\le T,1\le j\le N \tag{2-17}$$

结束：

$$P=\max_{1\le i\le N}\left[\delta_T(i)\right] \tag{2-18}$$

$$q_T=\arg\max\left[\delta_T(i)\right] \tag{2-19}$$

状态序列求取：

$$q_t=\varphi_{t+1}(q_{t+1}),\quad t=T-1,T-2,...,1 \tag{2-20}$$

2.2 HTK 工具包介绍

HTK（HMM Toolkit）是一个专门用于建立和处理 HMM 的实验工具包[16,17,18]。它由一序列的 C 函数模块和应用工具构成。最初它被专门用于语音识别的开发研究（连续分布的 HMM），但自它发布以后，很快就受到广泛的重视，并且变成了一个通用的开发实验工具包。尽管它也被大量地应用于其他方面，例如语音合成，字符识别和 DNA 排序等，但它基本的，主要的应用领域还是在语音识别方面。

HTK （HMM 工具包）的第一版本 V1.0 发布于 1989 年，当时主要的研究开发者是 Steve Young，他在剑桥大学电子工程系（Cambridge University Engineering Department （ URED ） ） 的 Speech Vision and Robotics Group 一直致力于大词汇量语音识别系统的开发研究。

1992 年早期，研究人员 Phil Woodland 开始加入与 Steve 合作开发 HTK，当时修改和完善 HTK 软件缺陷的工作主要由他们 2 人完成。随着软件使用人数的

增加，这种软件维护和完善的工作量也逐渐变得越来越大。到 1993 年初，他们与 Entropic Research Laboratory 达成协议，由 Entropic 负责软件的发布和维护。自此 Entropic Research Laboratory 获得了 HTK 的使用权，之后不久即 1995 年，HTK 的开发授权也转移到 Entropic，同年 Entropic Cambridge Research Laboratory Ltd. 宣布成立。

1995 年，随着 Entropic Cambridge Research Laboratory Ltd.的成立，Steve 和 Phil 全面负责该实验室的技术工作。当年 10 月，HTK V2.0 发布，它在前面版本的基础上增加和完善了一系列新的函数模块和应用模块。

1999 年 11 月，微软合并了 Entropic 公司，HTK 的开发研究实力得到增强。自此，微软和 CUED 开始共同致力于 HTK Toolkit 的研究开发。微软在给它提供极大帮助的同时，把 HTK 的注册权返还给了 CUED。

从 2000 年，HTK 工具包的源代码可免费从 CUED 的网站获得。此举主要是为了进一步发展语音识别研究和完善 HTK。此时，HTK V3.0 也应运产生了。随着 HTK 源代码和应用工具的免费公开发布，HTK 工具包也不断地得到完善。微软保留了现存 HTK 代码的版权，但它鼓励每一个 HTK 的使用者对源代码作进一步的修改和完善，共同致力于 HTK 工具包和语音识别的研究。

2.2.1 HTK 的软件结构

HTK 的许多功能被建立和封装为一序列的函数模块。这些模块可以使用相同的接口方式和外界进行交互。图 2-1 说明了 HTK 的软件结构和它的输入输出接口[17,18]。

2-1 图中各主要函数模块的功能如下：用户的输入输出和与操作系统的接口由函数模块 HShell 控制，内存的管理由 HMem 负责。语音分析所需的处理操作由 HSigP 提供，数学函数的支持由 HMath 提供。Hlabel 提供标签文件的接口，HLM 负责语言模型的建立，HNet 负责语法网络的建立，HDict 负责建立词汇的发音词典，HVQ 负责矢量量化 VQ 码本的建立，HModel 负责 HMM 模型的定义和建立。

针对所有语音数据的输入输出，HWave模块提供了各种波形数据格式的支持。HParm提供了几种特征参数文件的格式支持。HUtil提供了许多有关HMM模型的应用例程，HTrain和HFB提供了对多种HTK训练工具的支持，HAdapt为多种HTK的自适应工具提供了支持，HRec包含了HTK中主要的识别处理函数。

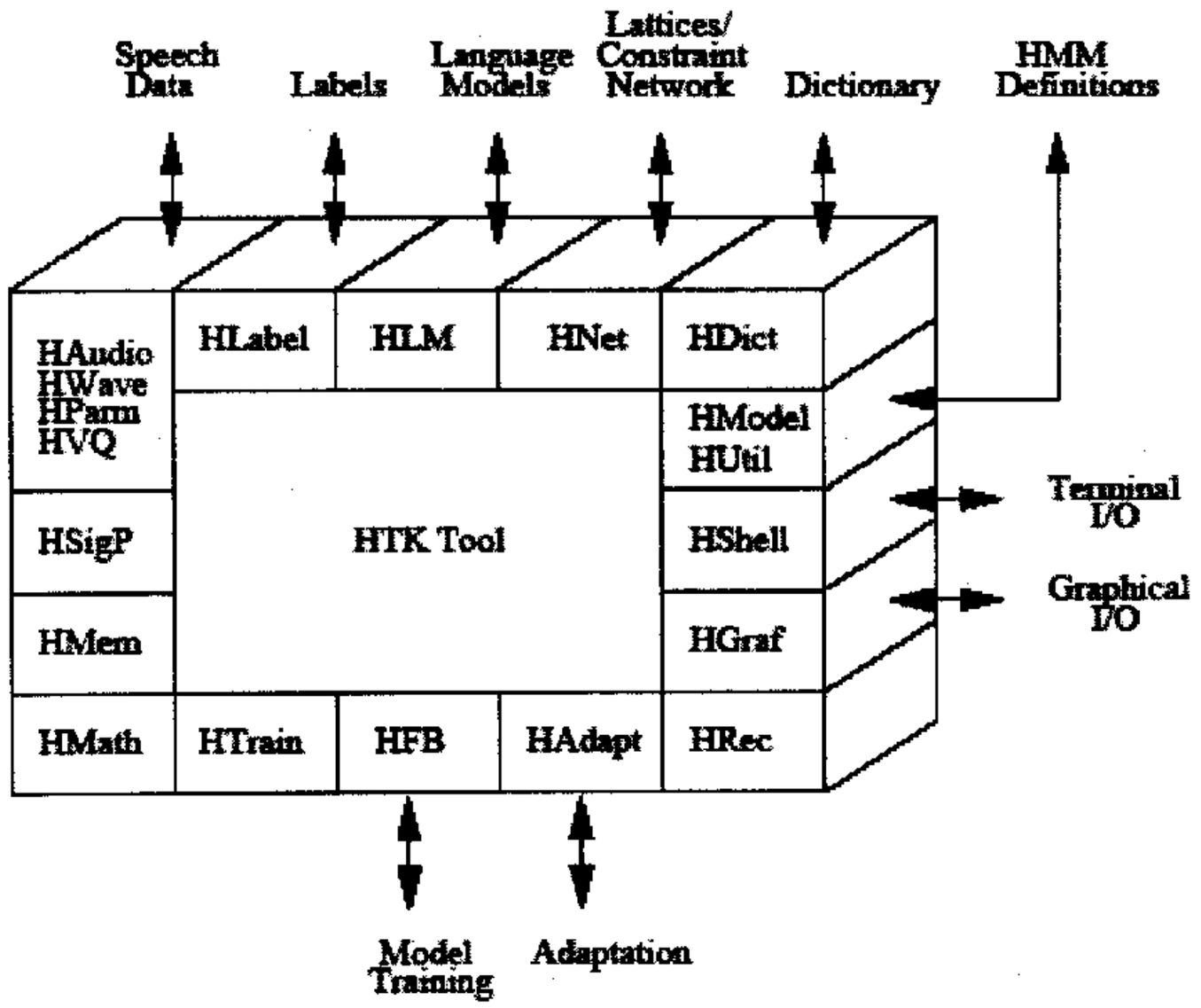


图 2-1 HTK 的软件结构

Fig. 2-1 Software Architecture of HTK

2.1.2 HTK 主要工具介绍

如果按照语音识别系统的建立和实现过程来分类，HTK 的应用工具分为 4 个阶段：数据的准备，模型的训练，数据的测试，结果的分析评价。如图 2-2 所示[17,18]。

1. 数据准备工具

为建立一套 HMM 系统，需要准备一序列的语音数据和相关的语音标签文件。同时语音数据还必须从波形数据转换为相应的特征参数数据，语音标签文件也必须统一为 HTK 支持的文件格式。这一阶段的主要工具为 HCOPY 和 HLED。HCOPY 主要用来把一序列原始的语音波形数据转换为我们需要的特征参数。HLED 可以用来建立实验过程中所需的各种语音标签文件。其他的工具 HLSATAS 用来获取并显示标签文件的一序列统计特性，当我们需要建立一离散的 HMM 系统时，HQUANT 可以为我们建立一个矢量量化的码本。

2. 训练工具

语音识别系统的第二步是建立 HMM 模型的原形（拓扑结构），并利用已有的语音数据训练建立的 HMM 模型，训练的方式有 2 中。当训练用语音数据中每个发音单元（词或音素）的边界确定时，可以采用 HINIT 和 HTREST 提供的训练方

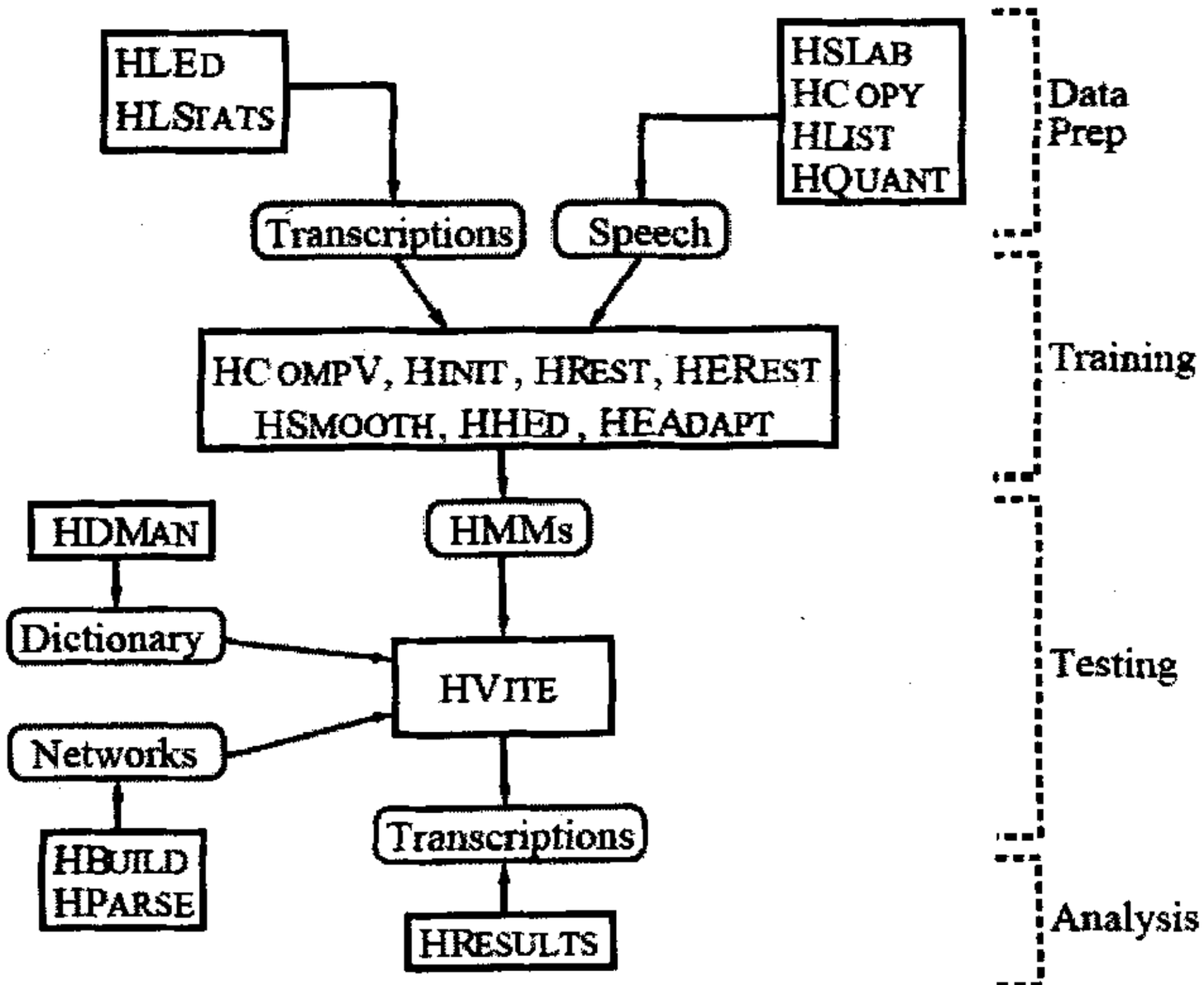


图 2-2 HTK 工作步骤

Fig. 2-2 HTK Processing Stages

式，即孤立词方式的训练。如果训练用语音数据中没有任何的边界信息，这时我们采用 HEREST 提供的嵌入整体训练方式。

3．识别工具

HTK 提供的识别工具是 HVITE，它采用符号传递算法来完成基于 Viterbi 搜索的语音识别。HVITE 的使用必须携带 2 个输入参数，一个是允许的发音词的语法网络，一个是为所有词建立的一本发音字典。发音词的语法网络可以通过 HBUILT 或 HPARSE 来建立。发音字典的建立由 HDMAN 来完成。

4．分析工具

一旦基于 HMM 的识别系统建立后，下一步的任务就是如何评价该系统的识别性能了。HTK 提供了一个实用工具 HRESULT。它可以统计出识别器的各种识别率，包括词的正确识别率，词的错误插入率，词的错误替换率，词的错误删除率。

2.3 MFCC（Mel Frequency Cepstrum Coefficients）参数

在本文实现的语音切分系统采用的语音参数是 MFCC 参数，下面讨论一下

MFCC 的定义和采用 MFCC 参数的原因[17,19]。

人的听觉系统，在没有人的主观倾向的情况下，可以说是一个比较好的话者识别系统，具有很高的准确性。虽然人的听觉系统分辨说话人的机理不一定是最佳的说话人识别方法，但是在目前的技术条件下，如果能达到人类分辨说话人的水平，也是相当可观的。比如，人的听觉系统作为说话人识别系统时，具有很强的抗干扰特性。

MFCC 是 Mel Frequency Cepstrum Coefficients 的缩写，中文含义为 Mel Frequency 倒谱系数。在求得一般的倒谱系数的过程中，FFT 的谱线在频率轴上是等间隔分布的，而在求得 Mel Frequency 倒谱系数的过程中，FFT 的谱线在频率上是不等间隔分布的，而在一个特殊的频率上，即在 Mel 频率上，是等间隔分布的。

MFCC 与倒谱的差别在于：MFCC 模拟了人的听觉特性。在有噪音和频谱变形的情况下，采用 MFCC 作为特征参数的语音识别和说话人识别，正确率比采用 LPC 等特征参数有比较大的改善。为了模拟人的听觉特性，所以采用了非线性的频率刻度（scale）。

心理学的研究进一步证明，无论对于纯音还是语音，人类对于声音音调的感觉不是线性的，而是一种特殊的非线性系统，它响应不同的频率信号的灵敏度是不同的。这必然导致人们去定义新的频率单位。这样频率单位划分方法，应该考虑到人耳听觉系统的特性，而不同于物理学对频率描述。物理上的频率是以 Hz 为单位的，符合人的听觉特性的频率则以 Mel（唛）或 Bark 为单位。

1. 临界带宽

临界带宽（critical bandwidth）是解决上述频率划分问题的一个概念，它是划分 Mel 频率的重要标准。在各个不同的临界带宽内，人们对信号的反应（比如响度）是大不相同的。

临界带宽概念的引入是为了描述噪音对纯音的掩蔽效应（masking effect）：一个纯音可以被以该纯音的频率为中心频率并且具有一定频带宽度的噪音所掩蔽，条件是临界带宽噪声的功率超过了纯音。在一个相同的临界带宽内，如果噪音的声压保持恒定，不论噪音的带宽是否扩展到了整个临界带宽，其响度是相同的。但是一旦超过了这个临界带宽，就可以感觉到响度的变化。

再进一步可以引出一个我们需要的结论：如果总功率相同，在一个相同临界带宽内，多个不同频率的声音组成的混合声音，跟位于此临界频带中央的单频声音是具有相同响度的。如果混合声音所处的频带范围超过了相应的临界带宽，混合声音会比纯音听起来更加响亮。

通过实验，人们已经知道：当中心频率在 1000Hz 以下时，临界带宽一般保持恒定，约为 100Hz。当中心频率超过 1000Hz 时，随着中心频率增长，临界频

带的带宽呈现线性增长，如图 2-3[19]。

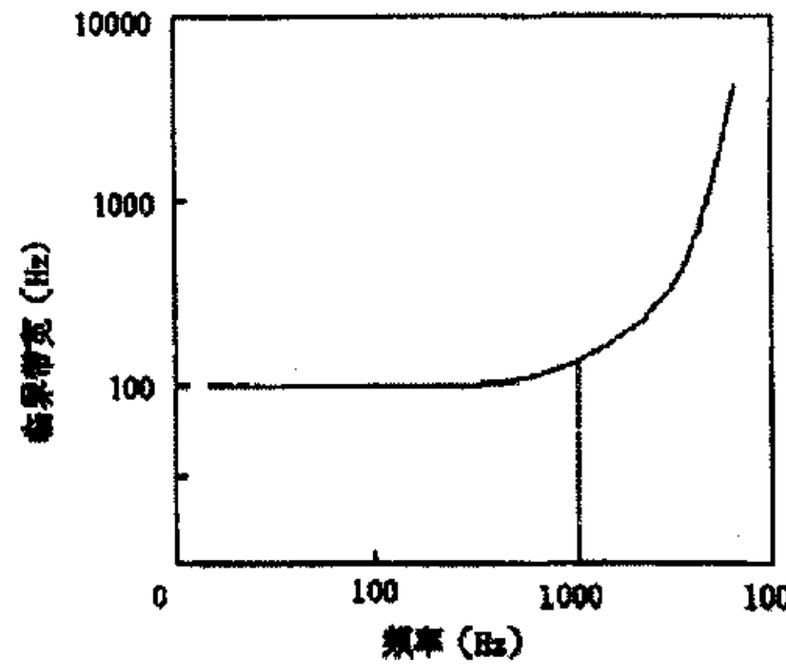


图 2-3 临界频带的带宽

Fig.2-3 Bandwith of Critical frequency band

所以符合人的听觉系统的频率刻度划分方法，应该满足在低频上具有较高的分辨率（resolution），在高频上具有较低的分辨率，符合临界带宽的特性。同时，如果以新的频率刻度为度量单位，临界带宽应该时恒定的。

2. Mel 频率

以 Mel 为单位的频率刻度就是符合这种特性的一种频率刻度，如图 2-4[19]。

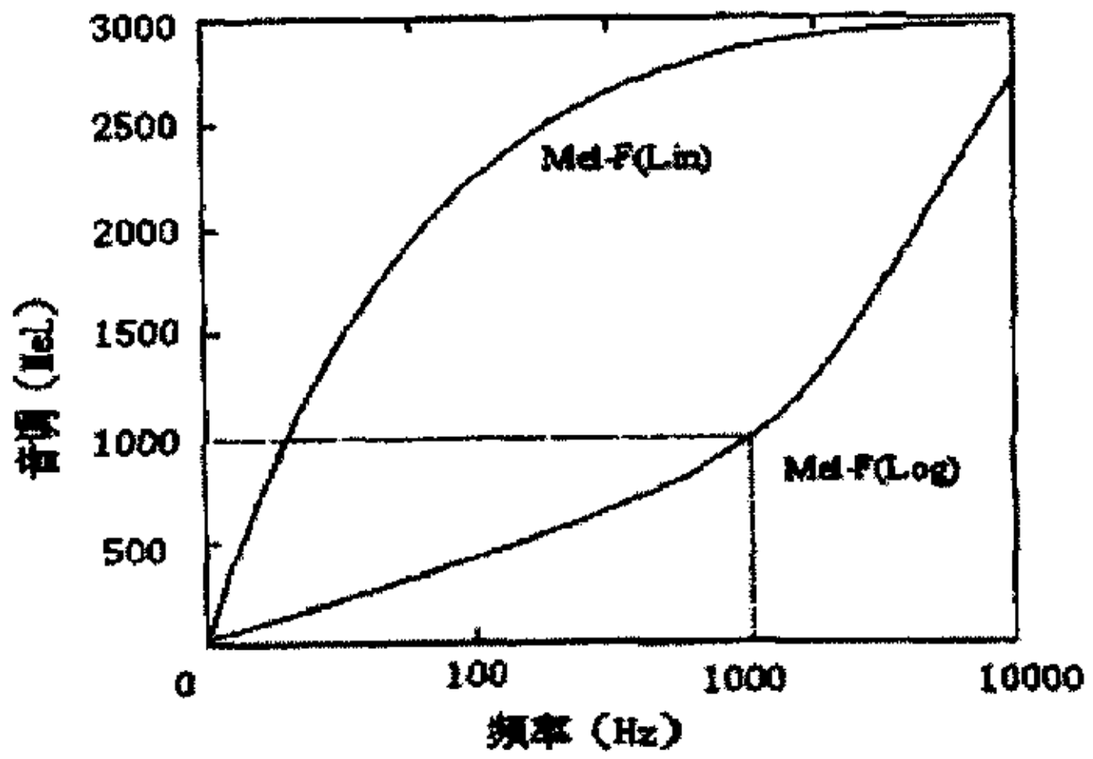


图 2-4 mel 音调和频率的对应关系

Fig.2-4 the mapping of the mel tonality and frequency

通常以频率 1000Hz 作为 1000Mel。在图 2-4 中有两条曲线，即 Mel－F（Lin）和 Mel－F（Log）。两条曲线的纵轴均以 Mel 为单位。曲线 Mel－F（Lin）的横轴为线性频率轴，曲线 Mel－F（Log）的横轴为对数频率轴。

对 Mel－F（Lin）曲线而言，当频率增加时，Mel 的值增加的越来越慢。

对 Mel－F（Log）曲线而言，当频率低于 1000Hz 时，Mel 与频率的对数值呈

线性关系：当频率高于 1000Hz 时，Mel 与频率的对数值也呈线性关系。但是，前者的斜率较小，而后者的斜率则较大。

以 Mel 单位的频率刻度划分与临界带宽在细节上并不精确相等，但是这个差别是很小的。

利用公式（2-21），就可以实现以 Hz 单位到以 Mel 单位的转换。

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \tag{2-21}$$

这个公式较好的满足了上述特性（上文提到的分段线性特性以及临界宽带为等宽度（以 Mel 为单位））。

MFCC 参数也是按帧计算的。首先要通过 FFT 得到该帧信号的功率谱，转换为 Mel 刻度下的功率谱。这需要在计算之前先在语音的频谱范围内设置若干个带通滤波器：

$$H_m(n),\ m = 0,1,\ldots,M-1,\ n = 0,1,\ldots,N/2-1 \tag{2-22}$$

M 为滤波器的个数；N 为一帧语音信号的点数。滤波器在频域上为简单的三角形，其中心频率为 f_m，它们在 Mel 频率轴上是均匀分布的。在线性频率上，当 m 较小时，相邻的 f_m 间隔很小，随着 m 的增加，相邻的 f_m 间隔逐渐拉开。另外在频率较低的区域，f_m 和 f 之间有一段是线性的。带通滤波器如图 2-5[17]。

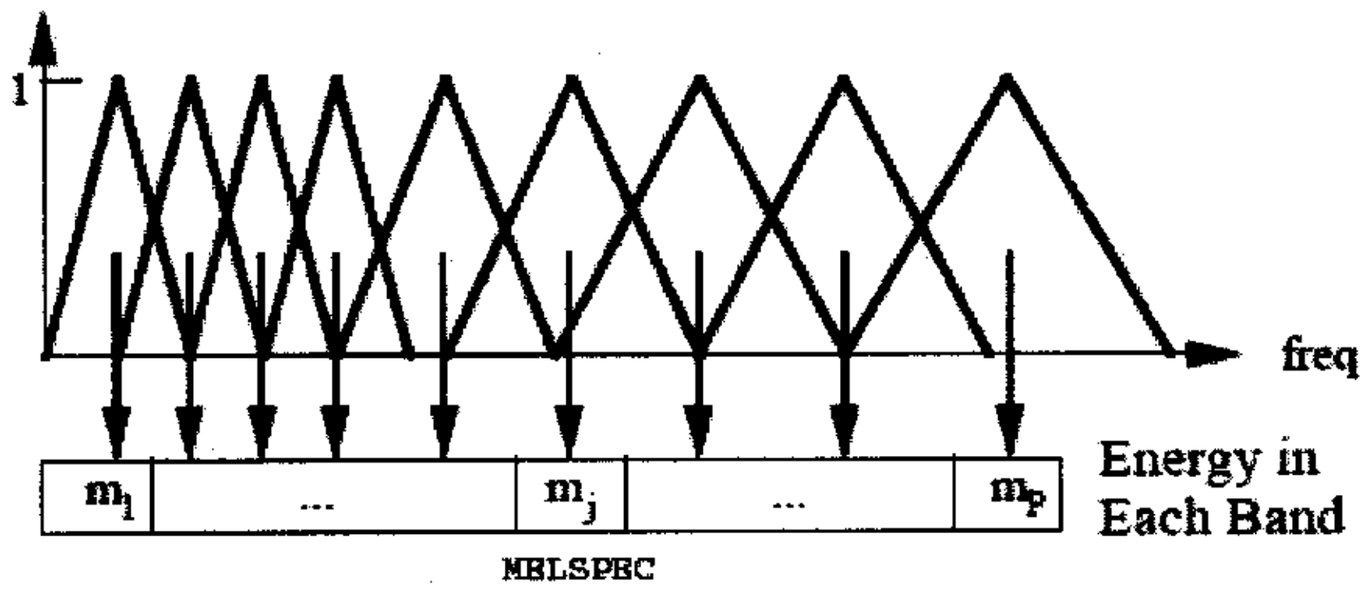


图 2-5 mel-Scale 滤波器组

Fig. 2-5 Mel-Scale Filter Bank

2.4 语音切分系统的实现

语音切分系统实际上是一个连续语音识别系统，它和传统的识别系统一样，其实现过程有两个阶段：（1）样本模型的训练，（2）未知模型的识别。如图 2-6 所示。

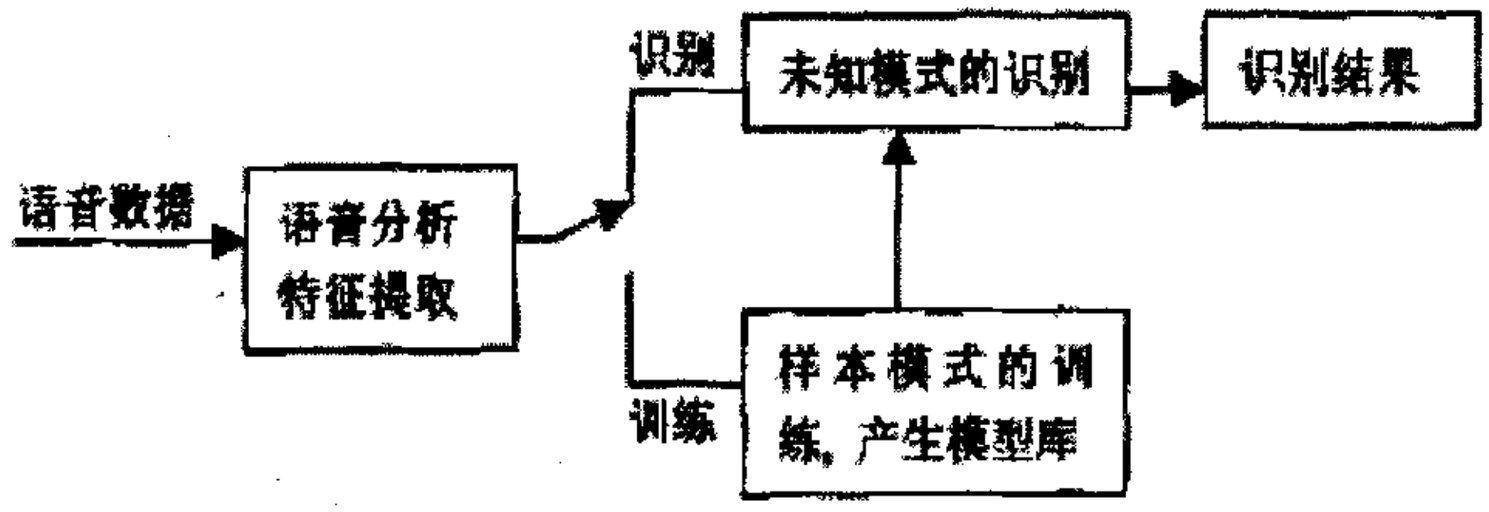


图 2-6 语音切分系统框图

Fig.2-6 Block diagram of speech-segmented system

2.4.1 建立英语切分系统的发音语法网络

发音语法网络是表示语音的音素组成结构的网络。假设要想表示如图 2-7 表示的语法，则语法文本如图 2-7[17]。

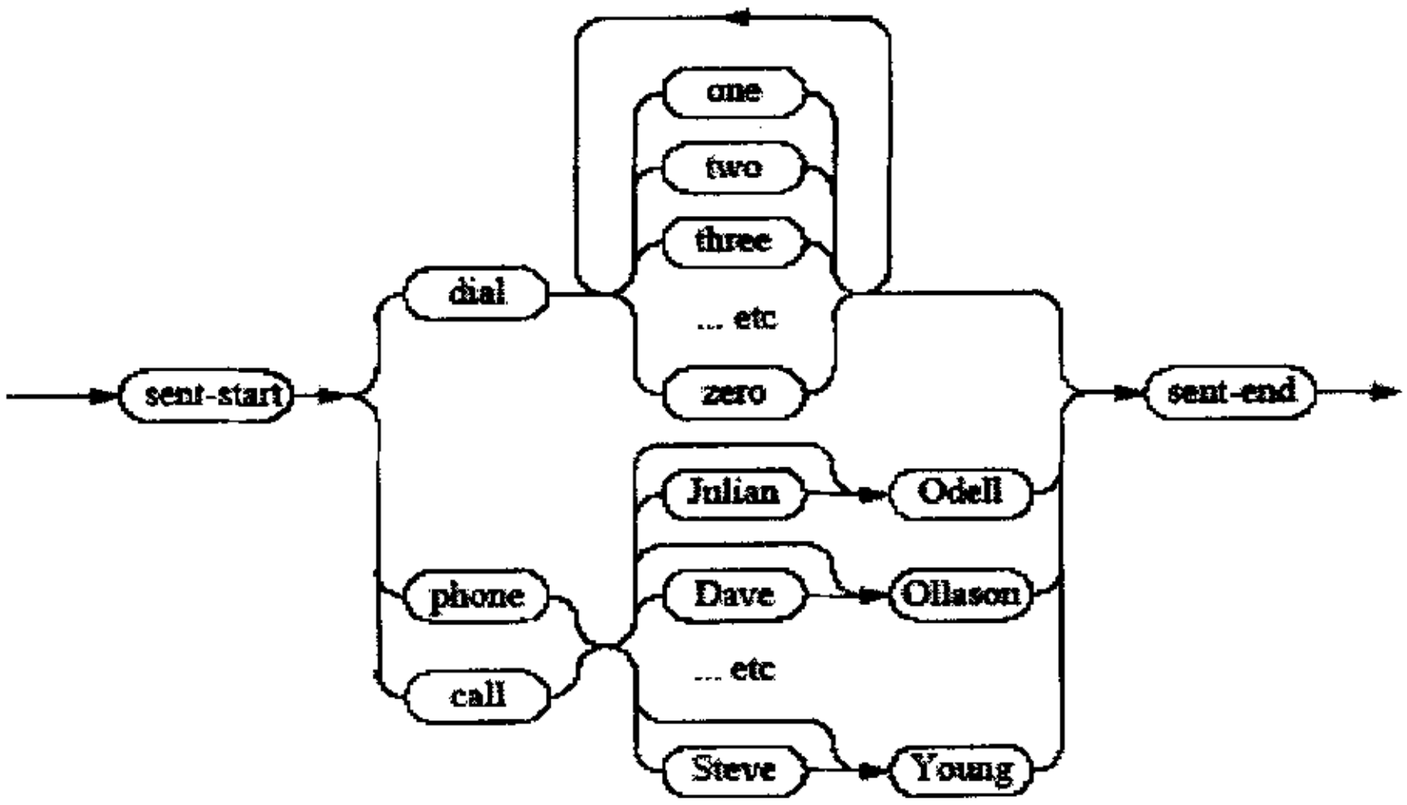


图 2-7 语法结构

Fig.2-7 program

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name  = [ JOOP ] JANSEN |
         [ JULIAN ] ODELL |
         [ DAVE ] OLLASON |
         [ PHIL ] WOODLAND |
         [ STEVE ] YOUNG;
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )
```

图 2-8 语法描述

Fig.2-8 program description

生成语法网络要用到 HParse 模块，如图 2-9[17]。

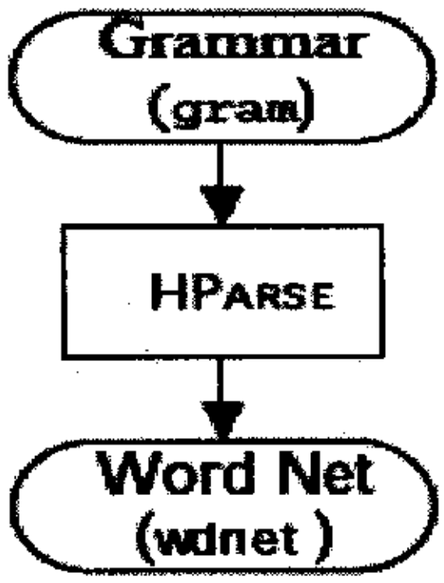


图 2-9　由语法描述生成语法网络

Fig.2-9 program net from the program description

2.4.2 建立字典

为了把词和音素联系起来，就要建立字典，把每个词都用基本音素的组成表示出来。形式如：word　p1 p2 p3 表示单词 word 是由 p1、p2 和 p3 这 3 个音素组成的。

2.4.3 为训练语音建立相应的标签文件

要对训练的语句进行标注，即生成相应的标签文件（*.MLF），这些标签文件要包含基本音素或基本词的符号，以及对应的起始时间和中止时间。

2.4.4 建立特征参数数据文件

针对原始数据文件，我们为每一个原始语音文件建立一个特征参数数据文件（MFCC 文件）。HTK 工具包是基于模块进行操作的，特征参数的提取是采用 HCOPY 模块进行的，如图 2-10[17]。

图 2-11 给出了在本切分系统中用到的特征参数配置。从配置文件可以看到，采用帧长为 25ms，帧移为 10ms 的滑动窗对信号进行短时分帧，分帧后的数据采用 hamming 窗处理，以消除信号边沿锐变。特征参数选用 39 阶 MFCC_0_D_A 系数（静态特征系数加上 0 阶系数，及它们对应的一阶差分值和二阶差分值，共 39 阶）。该参数采用了滤波器组分析方法（在语音频谱范围内设置若干个带通滤波器），应用了人耳听觉感知方面的一些研究成果。

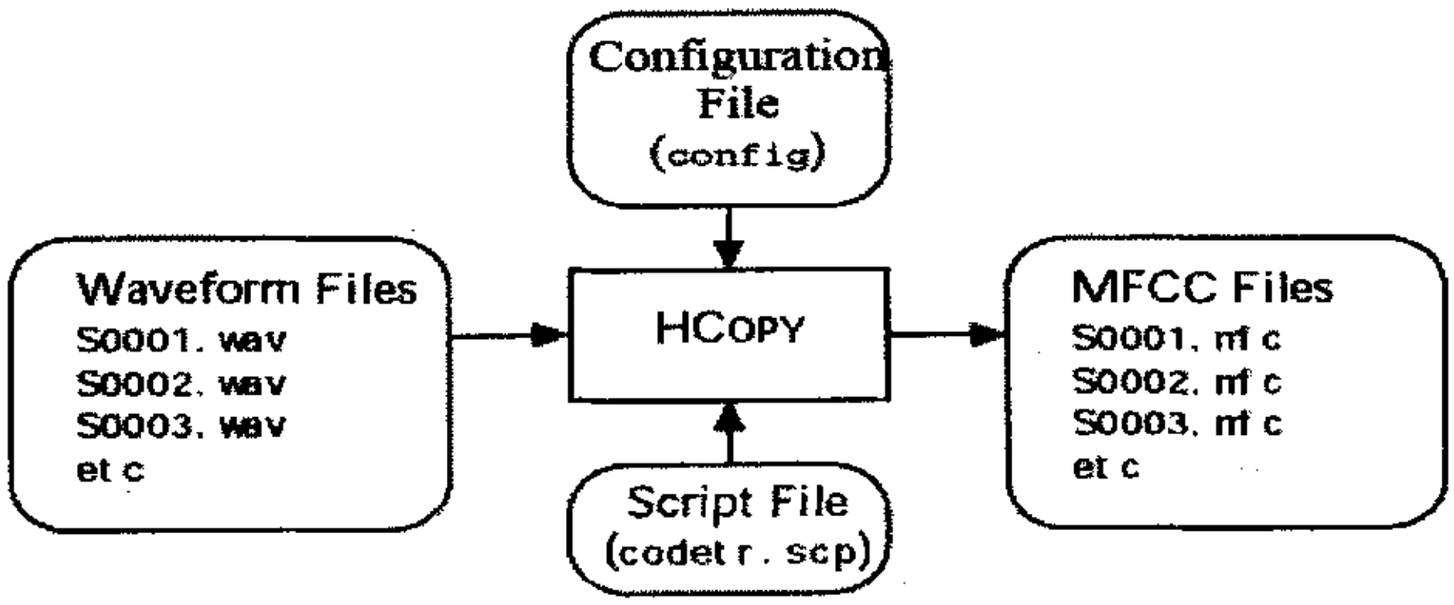


图 2-10 用 HCopy 模块提取特征参数

Fig. 2-10 using HCopy model to get MFCC files

```
SOURCEKIND      = WAVEFORM
SOURCEFORMAT    = WAV

TARGETKIND      = MFCC_0_D_A
TARGETRATE      = 100000

SAVECOMPRESSED  = TRUE
SAVEWITHCRC     = FALSE

WINDOWSIZE      = 250000.0
USEHAMMING      = TRUE
PREEMCOEF       = 0.97

NUMCHANS        = 24

NUMCEPS         = 12


ENORMALISE      = TRUE
```

图 2-11 HCopy 模块的配置文件

Fig. 2-11 Config file for HCopy model

2.4.5 建立与上下文无关的单音素 HMM 模型

1. 建立 HMM 模型的原型

HMM 模型的原型涉及到 HMM 模型的类型，拓扑结构，特征矢量维数等。图 2-12 为本实验中选择的 HMM 模型原型。根据图 2-12，HMM 模型拓扑结构设计为从左到右连续型，状态间的跨度为 0 或 1，模型隐含 5 个状态，1 状态和 5 状态仅仅为开始状态和结束状态，它们不产生特征矢量，故没有概率分布参数。其他的状态的特征参数均值矢量维数为 39 维，协方差矩阵为 39×39 的对角阵。

```
   ~o <VecSize> 39 <MFCC_0_D_A>
<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
    <Variance> 39
      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
    <Variance> 39
      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
    <Variance> 39
      1.0 1.0 1.0 1.0 1.0 ..... 1.0 1.0 1.0 1.0 1.0
  <TransP> 5
   0.000e+0   1.000e+0   0.000e+0   0.000e+0   0.000e+0
   0.000e+0   6.000e-1   4.000e-1   0.000e+0   0.000e+0
   0.000e+0   0.000e+0   6.000e-1   4.000e-1   0.000e+0
   0.000e+0   0.000e+0   0.000e+0   7.000e-1   3.000e-1
   0.000e+0   0.000e+0   0.000e+0   0.000e+0   0.000e-1

<EndHMM>
```

图 2-12　HMM 模型的初始化

Fig.2-12 Initialize HMM

2. HMM 模型的训练

模型训练采用 HERest 模块，它采用常用的训练算法 Baum－Welch 迭代算法。训练过程中，将不断地用前面准备好地特征数据来训练上面产生的各个单音素原始模型，最终产生我们需要的 HMM 单音素子模块，如图 2-13[17]。

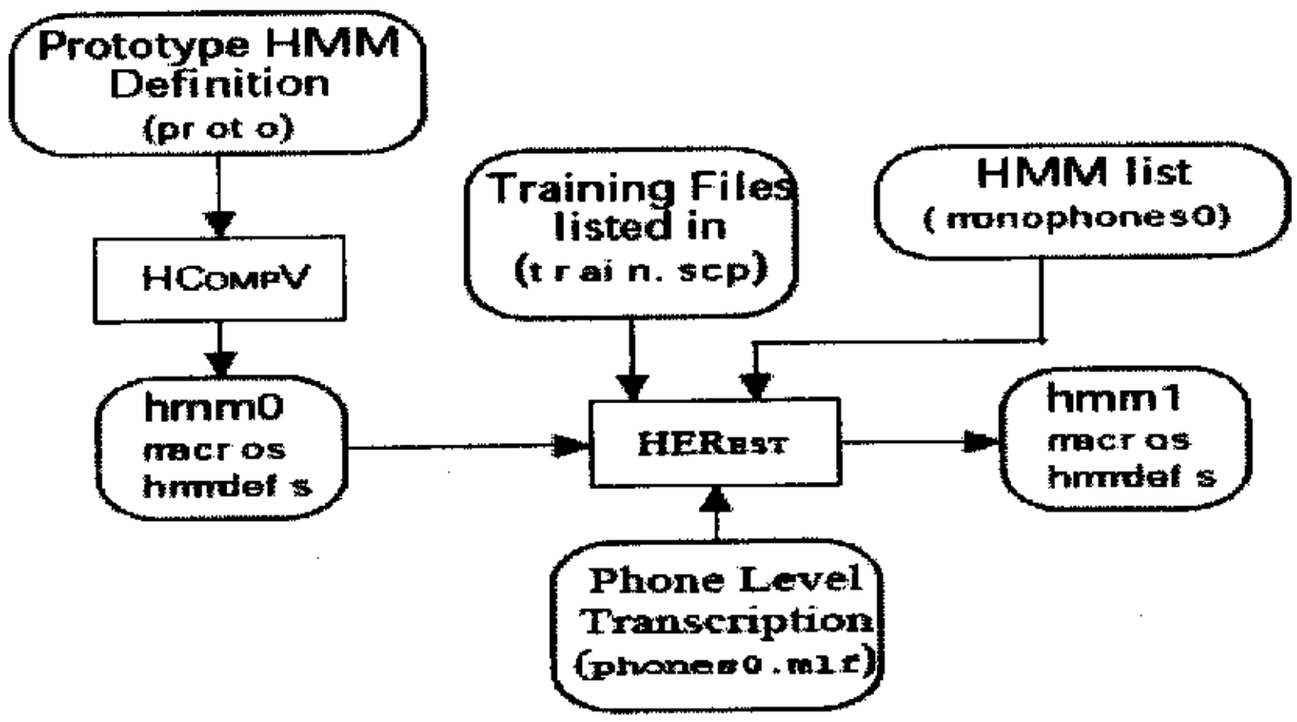


图 2-13　HMM 模型的训练

Figure 2-13 Training of the HMM

2.4.6 语音数据识别切分

语音数据的切分使用 HVite 模块。如图 2-14[17]，未知语音模式的测试采用 HVite 模块。根据前面准备好的声学 HMM 模型库，发音字典，语法网络，和未知语音特征参数，采用 Viterbi 搜索算法进行识别运算。

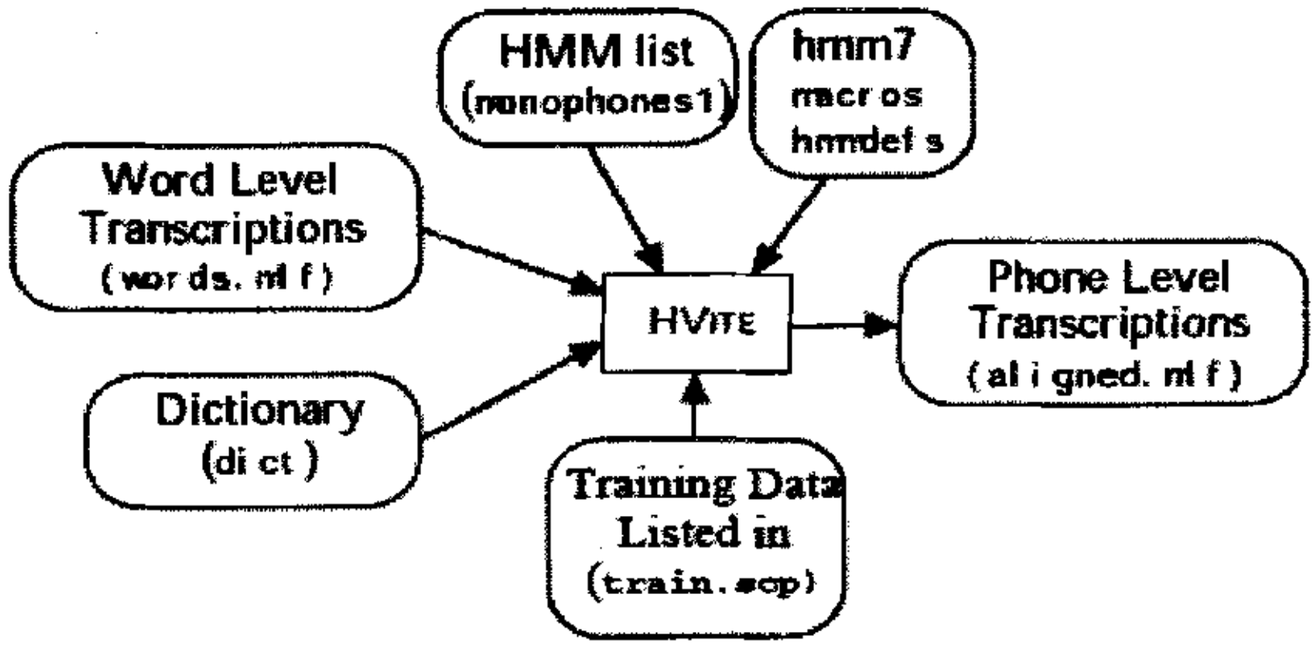


图 2-14 语音信号的测试框架

Fig.2-14 speech recognition

2.4.7 实验结果和讨论

1．特定人的英语语音切分实验

对于特定人的英语语音识别，训练数据采用 CMU(Carnegie Mellon University)的 bdl 的语音库，共 1131 句英语语音。Bdl 语音库是 16k 采样，分辨率为 16bit 的英语语音，而且语音库本身已经带有标注信息。本文采用大部分的语音（900 句）作为训练语句，而剩下（231 句）的语句作为测试语句。下面给出其中一句的切分结果。图 2-4 表示的是语音库提供的标注文件 arctic_a0005.lab，图 2-5 表示的是用特定人切分系统切分的结果 arctic_a0005.rec。

比较这两个标注文件，可以看到，用切分系统生成的标注文件和语音库还是有些差别，但是大多数的音素标注还是比较准确。出现标注误差的原因主要是本切分标注系统中的语法网络过于简单。在本系统中，语法为：(h# <p1 | p2 | p3 | .....> h#)，没有考虑到双音素和三音素的信息，所以切分结果有一定的误差。

```
"arctic_a0005.lab"
        0      500000 pau
   500000     1500000 w
  1500000     1800000 ih
  1800000     2700000 l
  2700000     3700000 w
  3700000     4700000 iy
  4700000     5700000 eh
  5700000     6300000 v
  6300000     7100000 er
  7100000     8400000 f
  8400000     9100000 er
  9100000    10100000 g
 10100000    11000000 eh
 11000000    11400000 t
 11400000    12100000 ih
 12100000    12400000 t
 12400000    12800000 pau
```

图 2-15　语音库提供的标注文件 arctic_a0005.lab

Fig.2-15 the label file from bdl speech database

```
"arctic_a0005.rec"
0 600000 pau
600000 1500000 w
1500000 1800000 ih
1800000 2900000 ow
2900000 3600000 w
3600000 4900000 iy
4900000 5800000 eh
5800000 6300000 v
6300000 7200000 er
7200000 8400000 f
8400000 9100000 r
9100000 10000000 g
10000000 11100000 eh
11100000 11400000 n
11400000 12400000 ih
12400000 12800000 pau
```

图 2-16 用特定人语音切分系统切分的结果 arctic_a0005.rec

Fig.2-16 the label file using the speaker-dependent speech-segmented system

2.非特定人英语语音切分实验

实验中选取多人的 TIMIT 语音库作为训练库，语音库中共有 2580 句语音，258 人（女 129 人，男 129 人）各说 10 句话。下面给出切分一句语音的结果。图 2-17 表示语音库提供的标注文件 faem0_si2022.lab，图 2-18 用非特定人切分系统切分的结果 faem0_si2022.rec。比较这两个标注文件，可以看到，用语音切分系统生成的标注文件和语音库有一定的差别。

```
0 6161875 h#
6161875 6658125 w
6658125 7275000 ah
7275000 7575000 dx
7575000 8760000 aw
8760000 8945625 q
8945625 10400000 f
10400000 10975000 ix
10975000 12036250 tcl
12036250 12245000 d
12245000 13000000 uh
13000000 14325000 sh
14325000 14850000 iy
14850000 15375000 dcl
15375000 15618750 d
15618750 16353750 r
16353750 17656250 ay
17656250 18067500 v
18067500 19518750 f
19518750 20925000 ao
20925000 22500000 h#
```

图 2-17 语音库提供的标注文件 faem0_si2022.lab

Fig.2-17 the label file from the speech database

```
"faem0_si2022.rec"
0 1300000 h#
1300000 2100000 pau
2100000 4300000 pau
4300000 5700000 pau
5700000 6700000 w
6700000 7300000 ay
7300000 7600000 nx
7600000 8500000 aw
8500000 8800000 ax
8800000 9100000 kcl
9100000 10100000 f
10100000 10500000 dh
10500000 10900000 ix
10900000 11900000 pcl
11900000 12300000 d
12300000 12800000 ix
12800000 14300000 sh
14300000 14800000 ix
14800000 15300000 dcl
15300000 15700000 d
15700000 16300000 r
16300000 17800000 ay
17800000 19400000 f
19400000 20900000 ao
20900000 22500000 h#
```

图 2-18 用非特定人语音切分系统切分的结果 faem0_si2022.rec

Fig.2-18 the label file using the speaker-independent speech-segmented system

3. 特定人语音切分系统和非特定人语音切分系统的比较

为了比较特定人语音切分系统和非特定人语音切分系统的切分结果，本文分

别用 10 句特定人 bdl 说的语音和 10 句非特定人 faem0 说的语音进行测试，结果如表 2-1。

表 2-1 两种切分系统的比较

Table 2-1 Comparison of the two speech-segmented system

	10 句 bdl 说的语音	10 句 faem0 说的语音
特定人（bdl）语音切分系统	80.2%	58.7%
非特定人语音切分系统	68.4%	67.1%

由表 2-1 可以知道，当对特定人 bdl 的语音进行切分时，特定人语音切分系统准确性比非特定人语音切分系统的好，当对 faem0 的语音进行切分时，非特定人语音切分系统准确性比特定人语音切分系统好。这是因为非特定语音切分系统相当是一个个人语音特征平均化的切分系统，它不再表现明显的个人语音特征，它对任何语音切分可以保持一定的准确率以上，但是一般都不能得到很好的准确率。为了提高非特定人语音切分系统的准确性，有人提出了说话人自适应的方法[20, 21]。

前面也提过，切分系统的准确率还不是很高，出现切分误差的原因主要是本切分标注系统中的语法网络过于简单。在本系统中，语法为：(h# <p1 | p2 | p3 | .....> h#)，没有考虑到双音素和三音素的信息，所以切分结果有一定的误差。所以将来要对语法进行改进，希望其切分准确率得到提高。

2.5 本章小节

本章重点论述了用 HMM 实现英语语音切分系统的基本原理，并具体说明了用 HTK 工具生成英语切分系统的步骤。最后给出特定人语音切分系统和非特定人语音切分系统的识别结果，并对结果作了分析。

（1）语音切分系统是基于分段的说话人语音转换中必须的部分，它主要是用来对训练语音和预转换语音进行切分标注。

（2）本文采用 HMM 的方法，运用 HTK（HMM 工具包）分别实现一个特定人和一个非特定人的英语语音切分系统。

（3）对于特定人 bdl，特定人的语音切分系统的准确率比较高，但是对于非特定人，非特定人的语音切分系统的准确率比较高。

第三章 高斯混合模型

在说话人语音转换中，为了把源说话人的语音特征转换成目标说话人的特征，希望可以用一种概率模型来描述说话人的语音特征矢量空间，然后求出源说话人语音特征矢量在目标说话人语音特征矢量空间的对应映射矢量，即找到一个转换函数，通过它把源说话人语音特征矢量转换成目标说话人语音特征矢量。高斯混合模型（GMM）是一种常用的用于识别的概率模型[22,23,24]，该模型用多个高斯分布的概率函数的组合来描述矢量在概率空间的分布情况。在本文中，说话人的语音频谱特征矢量空间就用高斯混合模型（GMM）来描述。为了更好的说明本文提出的基于分段说话人语音转换方法，本章重点讨论了高斯混合模型以及用 EM（Expectation—Maximization 最大期望）算法如何求出 GMM 参数[25,26]。而如何求出两个 GMM 之间的转换函数，以及如何进行转换，将在第四章中具体说明。

3.1 高斯混合模型及最大期望算法（EM）

3.1.1 高斯混合模型

一个高斯混合模型的概率密度函数是由 m 个高斯概率密度函数加权求和而得到的。设 X_t（t=1, …, T）是一个 d 维的特征矢量，$X_1, X_2, ..., X_T$ 是说话人语音的特征矢量集。高斯混合模型用下面的概率密度函数描述 X_t 的分布[25]：

$$p(X) = \sum_{i=1}^{m} \alpha_i N(X; \mu_i, \Sigma_i) \tag{3-1}$$

其中：m 是高斯模型混合的个数，

α_i，i=1, 2, …, m 是高斯分布混合时的权重值，

$N(X; \mu_i, \Sigma_i)$ 是第 i 个单高斯分布概率密度函数。

混合权重值满足：

$$\sum_{i=1}^{m} \alpha_i = 1 \quad \text{而且} \quad \alpha_i \geq 0$$

在上式中，$N(X; \mu, \Sigma)$ 表示 p 维正态分布概率密度函数，μ 是均值矢量，Σ 表示协方差矩阵，即

$$N(X; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right] \tag{3-2}$$

设 λ 为高斯混合模型的参数，则 λ 由各单个高斯模型的均值矢量、协方差矢量、

和混合权重值共同组成，如下式：

$\lambda=\{\alpha_i,\mu_i,\Sigma_i\},i=1,2,...,m$

对于一个正态分布的随机变量 X 的概率密度函数，其参数为：$\theta(\mu,\Sigma)$。

均值矢量由下式表示：$\mu=E[X]$；

协方差矩阵由下式表示：$\Sigma=E[(X-\mu)(X-\mu)^T]$

设 σ^2_{mn} 是协方差矩阵的第 m 行 n 列元素，表示第 m 维和第 n 维的协方差，则：$\sigma^2_{mn}=E[(x_m-\mu_m)(x_n-\mu_n)]$，$x_m$ 表示第 m 维元素，x_n 表示第 n 维元素。

事实上，特征矢量集用高斯混合模型来表示实际上就是把这些特征矢量进行分类（高斯模型混合的个数 m，也就是表示把特征矢量集分成 m 类），每一类特征矢量的概率密度函数是一个高斯分布，这个高斯分布的均值矢量表示这一类的中心，协方差表示这一类分布的分散度。第 i 个高斯分布混合时的权重值 α_i 表示第 i 类的统计概率。特征矢量 X 属于第 i 类（C_i）的概率可以表示为：

$$p(C_i\mid X)=\frac{\alpha_i N(X;\mu_i,\Sigma_i)}{\sum_{j=1}^{m}\alpha_j N(X;\mu_j,\Sigma_j)} \tag{3-3}$$

在本文中，对于分段的说话人语音转换，源说话人和目标说话人的每一个基本语音段分别用一个混合高斯模型来表示。给定特征矢量集 $X_1,X_2,...,X_T$，假设各特征矢量是独立的，要构造 GMM 参数就要估计合适的 GMM 参数 λ，使得 $p(X)$ 的值最大，即似然最大。

EM（Expectation－Maximization 最大期望）算法[25,26]是一种常用的、有效的最大似然估计算法，尤其适合于由一个训练集估计出高斯混合模型参数。EM 算法可以迭代的对 GMM 参数进行最优估计。

3.1.2 最大期望算法(EM)

EM（Expectation－Maximization 最大期望）算法是最大似然（Maximum Likelihood 简称 ML）和最大后验概率（Maximum a posteriori 简称 MAP）估计的常用方法，适合由非完整数据最优的估计概率模型参数。在 Dempster，Laird 和 Rubin（1977）对该算法进行介绍后，EM 算法开始被广泛的应用。EM 算法常被用来估计混合高斯模型参数。另外，EM 算法作为一种最大似然估计的通用算法，也被用于估计 HMM、PMN（概率影射网络）等其他模型参数。

EM 算法从初始的对模型参数的猜测开始，利用最大似然的原则，迭代地估计模型参数。每次迭代分两个步骤，E 步根据已知学习样本和当前估计（初值由初始化得到）得到学习数据的分布，M 步在假设 E 步所得到的分布正确的情况下，最大似然地计算模型参数，并不断重复直到局部最大。可以证明，每次迭代都增

大或不改变似然度（当得到局部最大时，似然度将不会改变）。

（1）最大似然估计

最大似然估计是把待估的参数看成固定但未知的量，然后求出能够使学习样本出现的概率最大的参数值，并把它作为参数的估值。把待估的 p 个参数 $\theta_1,\theta_2,...,\theta_p$ 记做 $\theta=(\theta_1,\theta_2,...,\theta_p)$，假设服从该分布的样本集 $X=X_1,X_2,...,X_n$ 有 n 个样本，概率密度函数表示为 $P(X|\theta)$。求出 θ 的最大似然估计值，就是把 $P(X|\theta)$ 看成 θ 的函数，并求出使其最大时的 θ 值。

因为假设各学习样本是独立的从样本集中抽取的，所以

$$P(X|\theta)=\prod_{i=1}^{n}p(x_i|\theta)=\mathrm{L}(\theta|X) \tag{3-4}$$

$\mathrm{L}(\theta|X)$ 被称为对样本集 X 的参数 θ 的似然度或似然函数。似然度可看作样本集给定时，模型参数 θ 的函数。在最大似然问题上，我们的问题是找到一个 θ 使 L 的值最大化。即希望找到 θ^* 满足

$$\theta^*=\arg\max_{\theta}\mathrm{L}(\theta|X) \tag{3-5}$$

为了分析方便，通常用 $\log\mathrm{L}(\theta|X)$ 即 $\log P(X|\theta)$ 来分析。由于对数函数是单调的，所以使对数似然函数最大的 θ 也会使原来的似然函数最大，而不至于影响结果，于是有下式：

$$\log P(X|\theta)=\sum_{i=1}^{n}\log p(x_i|\theta) \tag{3-6}$$

用对 θ 求微分并令它为 0 的求极值法则，可知 θ 的最大似然估计必然满足方程：

$$\begin{pmatrix}\frac{\partial}{\partial\theta_1}\\ \cdots\\ \frac{\partial}{\partial\theta_p}\end{pmatrix}\sum_{i=1}^{n}\log p(x_i|\theta)=0 \tag{3-7}$$

根据概率密度函数 $P(X|\theta)$ 形式不同，最大似然问题可以简单也可以很困难。例如，如果 $P(X|\theta)$ 是一维单个的高斯分布，则 $\theta=(\mu,\ \sigma^2)$。通过取 log（$\mathrm{L}(\theta|X)$）的微分并令其为 0，可以直接解出 μ，σ^2 值：

$$\hat{\mu}=(1/n)\sum_{i=1}^{n}x_i \tag{3-8}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{3-9}$$

但很多时候 log（$\mathrm{L}(\theta|X)$）不易于分析，这时只能求助于更有效的方法。EM

算法就是这样一种更有效的方法之一。

（2）EM 算法原理

EM 算法是一种通用的方法，它能够最大似然地估计非完整数据集的概率分布模型的参数。它是一种迭代算法，每一次迭代都包括两个步骤：计算期望值（Expectation）和最大化期望值（Maximization），所以称为 EM 算法[25,26]。

EM 算法有两种主要的应用。第一种应用用于估计丢失的数据。由于观测过程中的各种问题和限制，数据集中有一些丢失。例如，获得的一幅图的某些象素点的值丢失了，就属于这种情况。另一种应用就是用于估计概率分布的参数。在最大似然估计中，直接对似然函数求极大值往往很困难。但是，如果假设某些隐藏的或没有观测到的数据存在可以简化似然函数，使对似然函数求极大值成为可能，这时可使用 EM 算法估计概率分布的参数。EM 算法的应用主要是后一种。

假设 X 是观测到的样本集，服从某种概率分布 θ，这里称 X 为非完全样本集。这里假设一个完全样本集 $Z=(X,Y)$ 存在，则有联合密度函数：

$$p(Z|\theta)=p(X,Y|\theta)=p(Y|X,\theta)p(X|\theta) \tag{3-10}$$

对这个新的概率密度函数，可以定义一个新的似然函数：

$$L(\theta|Z)=L(\theta|X,Y)=L(X,Y|\theta) \tag{3-11}$$

式（3-11）中，X,θ 是常量，假设 Y 是隐藏信息，是未知的，随机的，服从概率分布 θ 的。

EM 算法首先要做的是，在已知 X 和 θ 的情况下，找到完全样本集 Z 的对数似然度函数 $L(X,Y|\theta)$ 的期望数学。定义为：

$$Q(\theta,\theta^{(i-1)})=E[\log p(X,Y|\theta)|X,\theta^{(i-1)}] \tag{3-12}$$

这里，$\theta^{(i-1)}$ 是指当前参数的估计，用于计算期望。θ 是新的参数值，由 $\theta^{(i-1)}$ 增加 Q 得到。式（3—12）中，X 和 $\theta^{(i-1)}$ 是常量。

上面对数学期望的估计被叫做 EM 算法的 E 步。

EM 算法的第二步称为 M 步，用于最大化 E 步中得到的数学期望。

$$\theta^{(i)}=\arg Q(\theta,\theta^{(i-1)}) \tag{3-13}$$

以上两个步骤根据需要将被不断重复，可以证明，每一次重复都使似然值增大，并且算法能保证收敛于似然函数的局部最大点。

以上只是 EM 算法的基本原理。EM 算法的具体应用很大程度上依赖于具体的概率模型。

（3）用 EM 算法估计高斯混合模型参数

在计算机模式识别中，高斯混合模型参数估计是 EM 算法用的最多的应用之一。高斯混合模型的概率密度函数为：

$$p(X|\theta)=\sum_{i=1}^{m}\alpha_i p(X|\theta_i) \tag{3-14}$$

公式(3-14)中 $\theta=(\alpha_1,...,\alpha_m,\theta_1,...,\theta_m)$，有 $\sum_{i=1}^{m}\alpha_i=1$ 并且每个高斯概率密度模型，其参数为 θ_i。$p(X|\theta)$ 由 m 个高斯模型按照混合系数 α_i 混合得到。

非完整样本集 X 的对数似然函数表示为：

$$\log(\mathrm{L}(\theta|X))=\log\prod_{i=1}^{N}p(x_i|\theta)=\sum_{i=1}^{N}\log(\sum_{j=1}^{m}\alpha_j p_j(x_i|\theta_j)) \qquad (3\text{-}15)$$

其中，N 为 X 中样本个数。m 为混合模型中高斯模型混合数。由于公式（3—15）包含对数和，该似然度函数难以求极值。然而，如果 X 是非完整的，假设存在未观测到的数据项 $Y=\{y_i\}_{i=1}^{N}$，Y 的值将告诉我们每个 X 集中的样本是由混合模型的哪个单个模型产生的，这时，似然函数的表达式将简化。假设 $y_i\in 1,2,...,m$。如果第 i 个样本由混合模型中第 k 个模型产生，则取 $y_i=k$。

引入 Y 后，完整样本集的似然函数可以写为：

$$\log(\mathrm{L}(\theta|X,Y))=\log(P(X,Y|\theta)=\sum_{j=1}^{m}\log(\alpha_j p_j(x_i|\theta_j)) \qquad (3\text{-}16)$$

对高斯混合模型，EM 算法的 E 步计算上式似然函数的数学期望：

$$\mathrm{Q}(\theta,\theta^g)=\mathrm{E}[\log(\mathrm{L}(\theta|X,Y))] \qquad (3\text{-}17)$$

公式（3-17）可展开为：

$$Q(\theta,\theta^g) \qquad (3\text{-}18)$$
$$=\sum_{l=1}^{m}\sum_{i=1}^{N}\log(\alpha_l)p(l|x_i,\theta^g)+\sum_{l=1}^{m}\sum_{i=1}^{N}\log(p_i(x_i|\theta_l))p(l|x_i,\theta^g)$$

公式(3-18)中 $l\in 1,2,...,m$

$p(l|x_i,\theta^g)$ 表示在已知样本集 X 和模型参数 θ^g 情况下，某样本 x_i 属于混合模型中第 l 个模型的后验概率。由公式（3-19）计算，k 表示第 k 次循环：

$$p^{(k)}(l|x_i,\theta^g)=\frac{\alpha_l^{(k)}p_l(x_i|\theta_l^{(k)})}{\sum_{j=1}^{m}\alpha_j^{(k)}p_j(x_i|\theta_j^{(k)})} \qquad (3\text{-}19)$$

在 EM 算法的 M 步，需要找到在使似然函数最大化的模型参数。由上述似然函数 $\mathrm{Q}(\theta,\theta^g)$ 的展开式看到，该式由两项相加而得，第一项只跟 α_i 有关，后一项只跟 θ_l 有关。最大化前一项可得到新得混合系数估计；最大化后一项，得到新的均值和协方差矩阵估计。

新的参数为：

$$\alpha_l^{(k+1)}=\frac{\sum_{i=1}^{N}p^{(k)}(l|x_i,\theta^g)}{N} \qquad (3\text{-}20)$$

$$\mu_l^{(k+1)} = \frac{\sum_{i=1}^{N} x_i p^{(k)}(l \mid x_i, \theta^g)}{\sum_{i=1}^{N} p^{(k)}(l \mid x_i, \theta^g)} \tag{3-21}$$

$$\Sigma_l^{(k+1)} = \frac{\sum_{i=1}^{N} p^{(k)}(l \mid x_i, \theta^g)(x_i - \mu_l^{(k+1)})(x_i - \mu_l^{(k+1)})^T}{\sum_{i=1}^{N} p^{(k)}(l \mid x_i, \theta^g)} \tag{3-22}$$

可以看到上述公式(3-20，3-21，3-22)同时执行了求期望值（E 步）和最大化（M 步）。对上述公式迭代重复，就是对 EM 算法中 E 步 M 步的重复迭代。当找到似然函数得极大值使，停止迭代。对最大似然的估计训练样本得高斯混合模型参数来说，EM 算法是一个极佳得算法。

EM 算法的迭代由设定模型的初始参数开始，将设定的初始参数代入上述公式(3-20，3-21，3-22)中，把推导得到的新的模型参数用作下一次迭代过程中模型参数的初值，迭代直到满足收敛条件后结束。

由于 EM 算法搜索得到的是局部最大值，所以，参数初始值的取值至关重要。高斯混合模型初始参数设定是对混合系数、均值向量、协方差矩阵、高斯模型混合数的设定。其中，高斯模型混合数应该由作者先验的设定。即根据经验或一定依据来指定，而其余的参数初值，则可以取随机数作初值。另外，为了更好的确定初值，可用聚类等方法大致估计样本的分布，并据此设定初值。例如，确定均值矢量的初值。对协方差矩阵Σ简单的设为对角阵等。如果设为对角阵，则假设样本矢量各维之间不相关，这样可以有效减小计算量。

3.2 本章小节

在本文的说话人语音转换技术中，语音频谱特征矢量空间用高斯混合模型（GMM）来表示。为了得到语音转换函数，本章重点讨论了高斯混合模型以及用 EM（Expectation—Maximization 最大期望）算法如何求出 GMM 参数。

（1）本文中，语音频谱特征矢量空间用高斯混合模型（GMM）来表示，以便找到一个映射函数，通过它把源说话人语音特征矢量转换成目标说话人语音特征矢量。

（2）EM 算法能够最大似然地估计非完整数据集的概率分布模型的参数。GMM 的参数估计采用 EM 算法。

第四章 基于分段的说话人语音转换方法

在现有的说话人语音转换中适合跨语种说话人语音转换的方法是基于分段的转换方法，因为这种方法不需要源说话人和目标说话人说的训练语句是相同的，训练语句不同是实现跨语种说话人语音转换中最基本的要求，本章提出了一种基于分段的说话人语音转换方法。

4.1 语音编码器

要进行说话人语音转换，使得源说话人说的语音经过转换后能象是目标说话人说的，首先就要把预转换的源说话人的语音信号进行编码，求得一系列的语音特征参数，然后修改这些特征参数，最后根据这些修改后的特征参数合成出新的语音（转换后象是目标说话人说的语音）。具体过程如图 4-1。

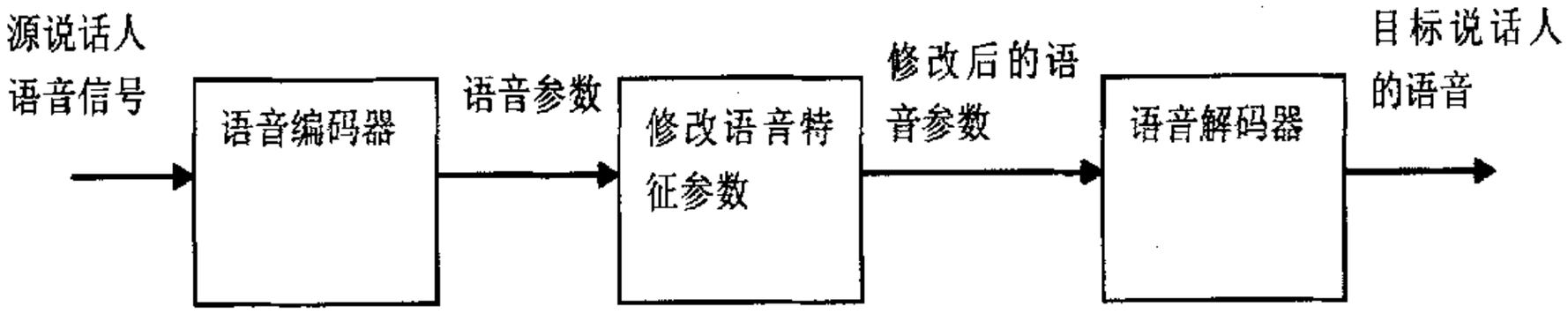


图 4-1 说话人语音转换过程

Fig. 4-1 voice conversion

这就要求找到一个合适的语音编码器，求得相应的特征参数，这些特征参数可以表现一个人的语音特征，而且要求通过修改这些参数应该可以容易的达到改变语音说话人特征的目的。研究表明基音频率和频谱特征是表现了一个人的语音特征的主要参数。

本文选用了“pitch＋mel 倒谱＋MLSA 滤波器”的语音编码器[27,28,29]。这个语音编码器比较简单，语音编码器的主要参数是基音周期（pitch）和 mel 倒谱（mel-cepstral），而且基音周期正好是基音频率的倒数，而 mel 倒谱可以很好的表现一个人的频谱特征，只要修改基音周期和 mel 倒谱就可以改变语音个人特征，达到说话人语音转换的目的。

语音信号是一种典型的时变信号，然而如果把观察时间缩短到 10 毫秒至几十毫秒，则可以得到一系列近似稳定的信号，所以语音是一种短时时变信号[2]。通过语言学的研究，知道发音器官的机理模型的方框图如图 4-2。

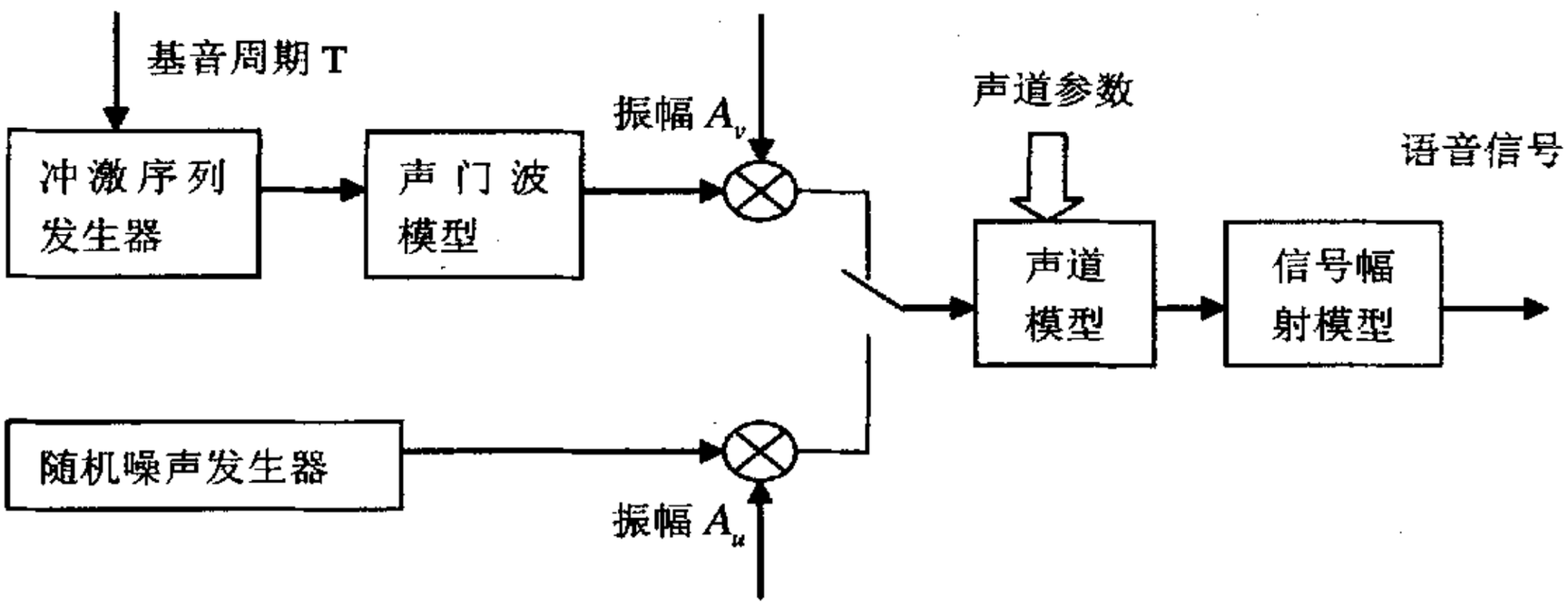


图 4-2 产生语音信号的方框图[2]

Fig.4-2 generation of speech[2]

长期研究证实，发不同性质的声音时，其激励的情况不同的[2]。大致上分为两大类：发浊音情况和发清音情况。

（1）发浊音情况时，此时的气流在通过绷紧的声带时，冲激声带产生振动，使声门处形成准周期性的脉冲串，并用它取激励声道。声带绷紧的程度不同，振动频率也不同。这个频率就是音调频率，它的倒数就是音调周期（基音周期）。男子、女子；老人、小孩的音调周期是不同的；男子低、女子高；老人低、小孩高。

（2）发清音情况时，声带松弛而不振动，气流通过声门直接进入声道。

为了使得计算机可以模拟发音过程，推导出了信号产生模型（图 4-3）。这里把图 4-2 中的辐射、声道以及声门激励的全部谱效应简化为一个时变的数字滤波器来表示，这个数字滤波器为合成滤波器。而且，对于浊音语音，这个系统受冲激序列的激励，各冲激之间的间隔为基音周期；对于清音语音，则受白噪声序列激励，它可以简单地由一个随机数发生器完成。

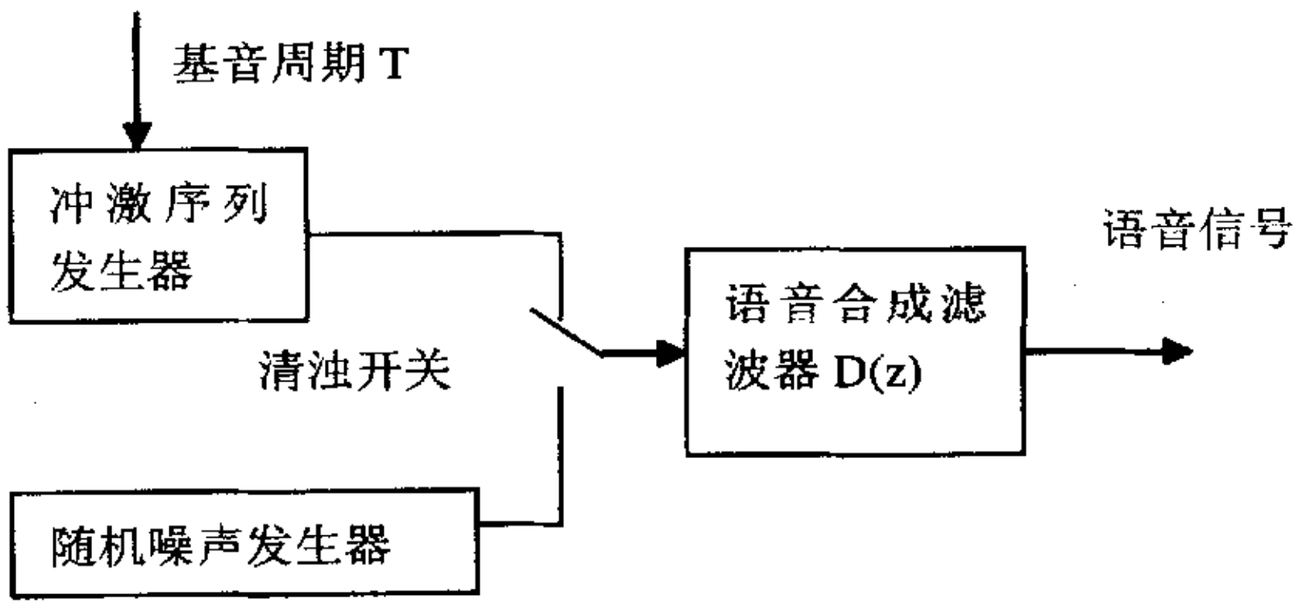


图 4-3 语音信号产生模型[2]

Fig.4-3 the model of speech generation[2]

在本文中，图 4-3 的语音合成器滤波器 $D(z)$ 是 MLSA(mel log spectrum approximation)滤波器。

$$D(z)=\exp\sum_{m=0}^{M}c(m)\tilde{z}^{-m},\quad \tilde{z}^{-1}=\left.\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right|_{z=e^{-j\omega}}=e^{-j\tilde{\omega}},\quad |\alpha|<1 \tag{4-1}$$

c 为 mel 倒谱系数，$c=\arg\max_{c} p(x|c)$；

$$x=[x(0),x(1),...,x(N-1)]^T$$
$$c=[c(0),x(1),...,x(M)]^T$$

令 $D(z)=\exp F(z)$，$F(z)=\sum_{m=0}^{M}c(m)\tilde{z}^{-m}$

$F(z)$的结构图如图 4-4：

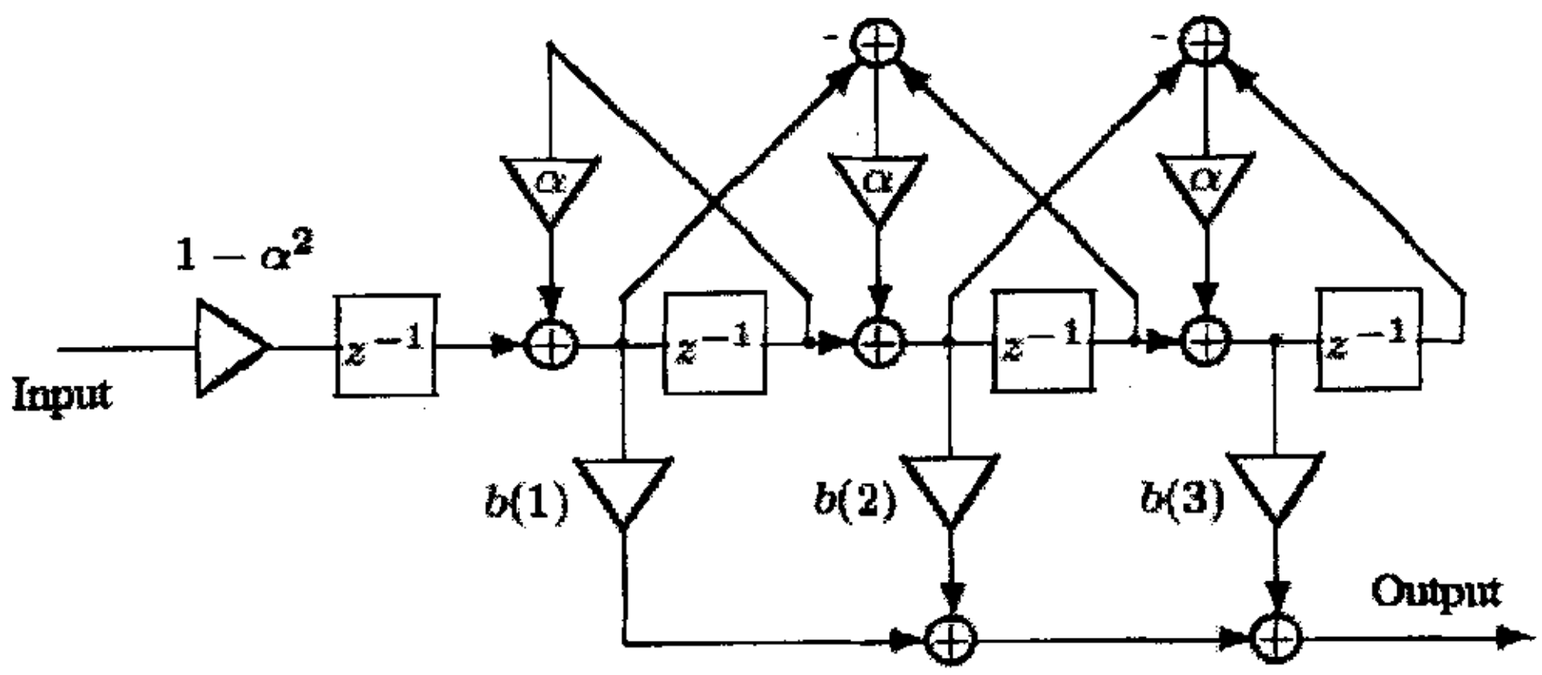


图 4-4 $F(z)$结构图

Fig.4-4 the structure of $F(z)$

经过转换：

$$D(z)=\exp F(z)\cong\frac{1+\sum_{l=0}^{L}A_{L,l}\{F(z)\}^l}{1+\sum_{l=0}^{L}A_{L,l}\{-F(z)\}^l} \tag{4-2}$$

MLSA 滤波器结构图为图 4-5：

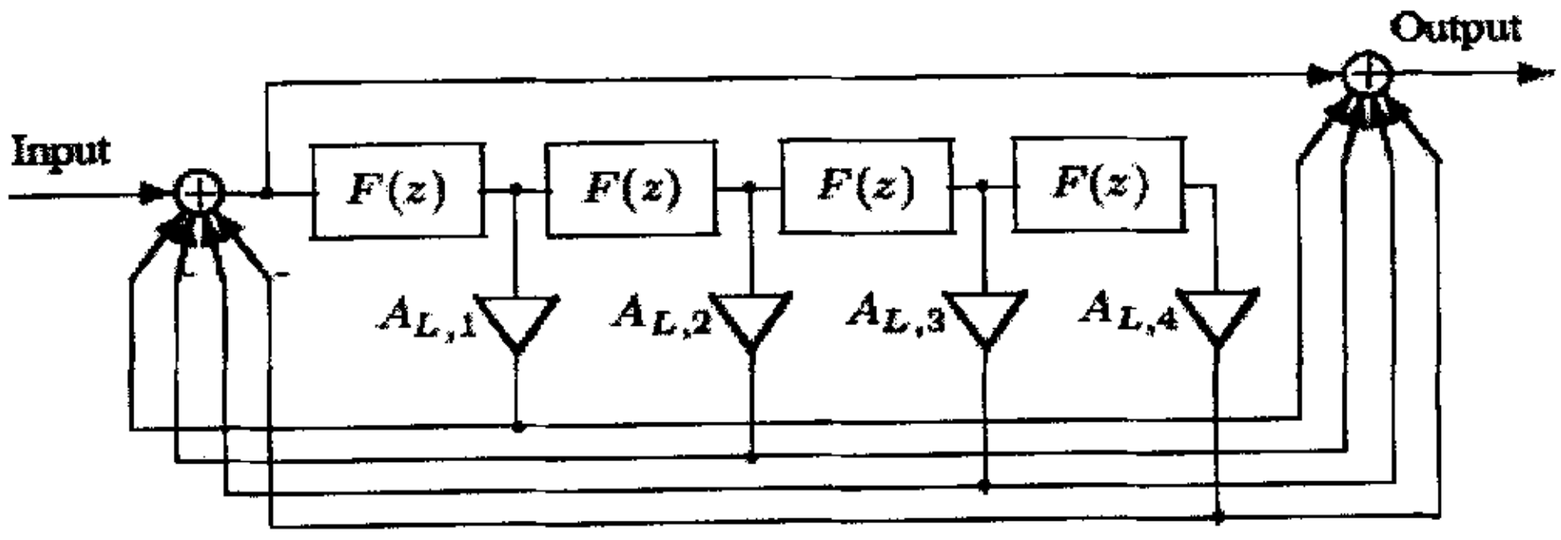


图 4-5 MLSA 滤波器结构图

Fig.4-5 the structure of MLSA filter

MLSA 滤波器具有以下的优点：

（1）精确度高：最大频谱误差为 0.24db。

（2）稳定性高。

4.2 频谱的转换方法

从上节的讨论中，可以知道本文的语音特征参数主要包括基音周期和频谱参数，进行说话人语音转换时就要分别对基音周期（或基音频率）和频谱参数（Mel 倒谱系数）进行转换。在本文中，频谱参数是 Mel 倒谱系数。为了对频谱进行修改，本文用 GMM 来统计源说话人和目标说话人语音的 Mel 倒谱系数的概率空间统计，然后求出转换函数进行修改。

在基于 GMM 的转换过程中，要求用来训练的源说话人语音特征参数和目标说话人语音特征参数要先对齐再进行训练，所以根据训练语音库的特点，频谱参数的转换分为两种方法：（1）不用分段的转换方法[5]。如果源说话人和目标说话人的训练语音库的内容是相同的，就不用对这些训练语句进行分段标注，直接对两个语音库进行动态时间规整（DTW）就可以了。（2）分段的转换方法[10]。如果源说话人和目标说话人的训练语音库的语音不同，就要先把这些语音进行分段标注，然后让源说话人和目标说话人的语音按语音段归类，然后再把两个说话人相同的语音段的类对应起来。第 1 种方法的优点是可以省去对训练语句进行分段标注的工作，缺点是要求训练语句一定要相同。第 2 种方法的优点是不要求训练语句相同，这对于跨语种说话人语音转换是很有利的。但是这种分段方法的缺点是训练语句必须进行分段标注，而且分段标注的好坏对转换的效果有一定的影响。

4.2.1 不用分段的转换函数

当源说话人和目标说话人的训练语句相同时，就可以分别求出两个训练语句的频谱参数，并对他们进行动态时间规整（DTW）[2]，得到一对频谱参数矢量序列。源说话人频谱参数矢量序列为 $\{X_t, t=1,2,...,n\}$，目标说话人频谱参数矢量序列为 $\{Y_t, t=1,2,...,n\}$。转换函数 $F(X_t)$ 的训练就是要找到一个转换函数 $F(X_t)$，使 $\{X_t, t=1,2,...,n\}$ 可以经过转换后得到 $\{Y_t, t=1,2,...,n\}$，即让 $F(X_t)$ 等于 Y_t。

先根据 $\{X_t, t=1,2,...,n\}$ 求出源说话人语音的高斯混合模型（GMM）参数 $(\alpha_i, \mu_i, \Sigma_i, i=1,2,....,m)$，特征矢量 X 属于第 i 类（C_i）的概率可以表示为：

$$p(C_i \mid X) = \frac{\alpha_i N(X;\mu_i,\Sigma_i)}{\sum_{j=1}^{m} \alpha_j N(X;\mu_j,\Sigma_j)} \tag{4-3}$$

下面求转换函数 $F(X_t)$，使

$$F(X_t)=\sum_{i=1}^{m} p(C_i \mid X_t)[\nu_i+\Gamma_i\Sigma_i^{-1}(X_t-\mu_i)]，\ i=1,2,...,m \tag{4-4}$$

$F(X_t)$ 由 p 维向量 ν_i 和 p×p 矩阵 Γ_i 确定。要求出 $F(X_t)$，就要求出 ν_i 和 Γ_i。

$$\varepsilon=\sum_{t=1}^{n}\| Y_t-F(X_t)\|^2 \tag{4-5}$$

要求出 $F(X_t)$，就要找出最佳 ν_i 和 Γ_i 使得最小平方差 ε 最小。

根据公式（4-4），要解决最小平方差问题，就是要解出下面的方程：

$$Y_t=\sum_{i=1}^{m} p_t(i)[\nu_i+\Gamma_i\Sigma_i^{-1}(X_t-\mu_i)] \ ，\ \mathrm{t}=1,2,...,\mathrm{n} \tag{4-6}$$

把上式转换成矩阵方程：

$$Y=\mathrm{P}\cdot\nu+\Delta\cdot\Gamma=\begin{bmatrix}\mathrm{P} & \vdots & \Delta\end{bmatrix}\cdot\begin{bmatrix}\nu\\ \cdots\\ \Gamma\end{bmatrix} \tag{4-7}$$

Y 是 n×p 的目标说话人频谱参数矢量矩阵：

$Y=[Y_1 \ \vdots \ \cdots \ \vdots \ Y_n]^T$。

P 是 n×m 的条件概率矩阵：

$$\mathrm{P}=\begin{bmatrix} p_1(1) & p_1(2) & \cdots & p_1(m)\\ p_2(1) & p_2(2) & \cdots & p_2(m)\\ \vdots & \vdots & \cdots & \vdots\\ p_n(1) & p_n(2) & \cdots & p_n(m)\end{bmatrix} \tag{4-8}$$

Δ是一个 n×pm 的矩阵，它与条件概率、源特征矢量、以及混合高斯模型 GMM 的参数有关：

$$\Delta=\begin{bmatrix} p_1(1)(X_1-\mu_1)^T\Sigma_1^{-1T} & p_1(2)(X_1-\mu_2)^T\Sigma_2^{-1T} & \cdots & p_1(m)(X_1-\mu_m)^T\Sigma_m^{-1T}\\ p_2(1)(X_2-\mu_1)^T\Sigma_1^{-1T} & p_2(2)(X_2-\mu_2)^T\Sigma_2^{-1T} & \cdots & p_2(m)(X_2-\mu_m)^T\Sigma_m^{-1T}\\ \vdots & \vdots & & \vdots\\ p_n(1)(X_n-\mu_1)^T\Sigma_1^{-1T} & p_n(2)(X_n-\mu_2)^T\Sigma_2^{-1T} & \cdots & p_n(m)(X_n-\mu_m)^T\Sigma_m^{-1T}\end{bmatrix} \tag{4-9}$$

而且还有两个未知矩阵 ν 和 Γ：

$$\nu=[\nu_1 \ \vdots \ \nu_2 \ \vdots \ \cdots \ \vdots \ \nu_m]^T_{(m\times p)}$$

$$\Gamma=[\Gamma_1 \ \vdots \ \Gamma_2 \ \vdots \ \cdots \ \vdots \ \Gamma_m]^T_{(m\times p)}$$

方程 4-7 变型为[30]：

$$\left(\begin{bmatrix}\mathbf{P}^{\mathrm{T}}\\ \cdots\\ \Delta^{\mathrm{T}}\end{bmatrix}\cdot[\mathbf{P}\quad\cdots\quad\Delta]\right)\cdot\begin{bmatrix}\nu\\ \cdots\\ \Gamma\end{bmatrix}=\begin{bmatrix}\mathbf{P}^{\mathrm{T}}\\ \cdots\\ \Delta^{\mathrm{T}}\end{bmatrix}\cdot Y \tag{4-10}$$

$$\text{或是}\begin{bmatrix}\mathbf{P}^{\mathrm{T}}\mathbf{P} & \vdots & \mathbf{P}^{\mathrm{T}}\Delta\\ \cdots & \vdots & \cdots\\ \Delta^{\mathrm{T}}\mathbf{P} & \vdots & \Delta^{\mathrm{T}}\Delta\end{bmatrix}\cdot\begin{bmatrix}\nu\\ \cdots\\ \Gamma\end{bmatrix}=\begin{bmatrix}\mathbf{P}^{\mathrm{T}}Y\\ \cdots\\ \Delta^{T}Y\end{bmatrix} \tag{4-11}$$

显而易见，求解这样的方程计算量非常庞大。实际上，如果设源说话人频谱参数矢量 X_t 服从高斯分布 $N(X;\mu,\Sigma)$，源说话人频谱参数矢量和目标说话人频谱参数矢量是联合高斯分布，则在 X_t 的前提下，Y 的期望值为[30]：

$$\mathrm{E}[Y\mid X=X_t]=\nu+\Gamma\Sigma^{-1}(X_t-\mu) \tag{4-12}$$

ν 表示目标说话人频谱参数矢量 Y 的均值，Γ 表示源说话人频谱参数矢量和目标说话人频谱参数矢量互相关矩阵。

$$\nu=E[Y] \tag{4-13}$$

$$\Gamma=E[(Y-\nu)(Y-\nu)^{T}] \tag{4-14}$$

在联合高斯分布，转换函数是简单的线性转换。把联合高斯分布推广到高斯混合模型，见公式(4-4)。把公式(4-13)和公式(4-14)代入公式（4-4），就可以求出转换函数 $F(X_t)$。

4.2.2 分段的转换函数

当源说话人和目标说话人的训练语句不相同时，就只能采用分段标注的转换方法。本文以基本音素作为语音段，比如英语中有 41 个基本音素，就表示英语语音有 41 个基本语音段，所有的英语语音都是用这些基本音素组成的。设语音的基本音素集为 $P=\{P_1,...,P_N\}$，N 为基本音素的个数。首先把训练语句根据基本音素集进行切分标注，把属于第 i 个音素 P_i(i=1,...,N)的频谱参数矢量集合在一起。然后属于第 i 个音素 P_i(i=1,...,N)的频谱参数矢量集合用一个高斯混合模型表示：

$$p(X)=\sum_{i=1}^{M}\alpha_i N(X;\mu_i,\Sigma_i) \tag{4-15}$$

其中：M 是高斯模型混合的个数，

α_i，i=1, 2, …, m 是高斯分布混合时的权重值，

$N(X;\mu_i,\Sigma_i)$ 是第 i 个单高斯分布概率密度函数。

混合权重值满足：

$$\sum_{i=1}^{M}\alpha_i=1 \text{ 而且 } \alpha_i\geq 0$$

在公式(4-15)中，$N(X;\mu,\Sigma)$ 表示 p 维正态分布概率密度函数，μ 是均值

矢量， Σ表示协方差矩阵，即

$$N(X;\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}}|\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right] \quad (4\text{-}16)$$

特征矢量 X 属于第 i 个基本音素（P_i）的概率可以表示为[10]：

$$p(P_i\mid X)=\frac{\beta_i\sum_{j=1}^{M}\alpha_{ij}N(X;\mu_{ij},\Sigma_{ij})}{\sum_{i=1}^{N}\beta_i\sum_{j=1}^{M}\alpha_{ij}N(X;\mu_{ij},\Sigma_{ij})} \quad (4\text{-}17)$$

β_i表示基本音素 P_i的概率，从统计语音库所含各个音素的数量得到。N 表示基本音素的个数。

基本音素 P_i的的 GMM 参数 (α,μ,Σ)和基本音素 P_i的概率 β_i就是基本音素 P_i的参数。

和不用分段的转换函数一样，设源说话人频谱参数矢量和目标说话人频谱参数矢量是联合高斯分布，求得基于音素的转换函数[10]：

$$Y=F(X)=\sum_{i=1}^{K}P(P_i\mid X)\phi_i \quad (4\text{-}18)$$

$$\phi_i=\sum_{j=1}^{M}\alpha_{ij}[\mu_{ij}^{Y}+\Sigma_{ij}^{Y}\Sigma_{ij}^{X^{-1}}(X-\mu_{ij}^{X})] \quad (4\text{-}19)$$

其中，X 为源说话人频谱参数矢量，Y 为目标说话人频谱参数矢量。μ_{ij}^{X}，Σ_{ij}^{X}分别为源说话人频谱参数矢量的第 i 个基本音素的高斯混合模型中第 j 个高斯分布的均值和协方差矩阵。μ_{ij}^{Y}，Σ_{ij}^{Y}分别为目标说话人频谱参数矢量的第 i 个基本音素的高斯混合模型中第 j 个高斯分布的的均值和协方差矩阵。M 为一个基本音素的 GMM 的高斯模型混合个数，N 为基本音素的个数。通常，$K\le N$，取 $P(P_i|X)$最大的前 K 个音素来计算。

4.3 基音频率的转换

基音频率是语音的一个重要特征，一般来说，每个人的基音频率都不同，通常来说，男子、女子；老人、小孩的基音频率是不同的；男子低、女子高；老人低、小孩高[2]。在说话人语音转换中，修改基音频率是必不可少的。

对于一个人来说，无论他说什么语言，他的基音频率的统计特性应该是不变的。基音频率的修改涉及到基音数值和基音的范围的修改，因此采用一个转换公式 $f_{0B}=af_{0A}+b$来进行基音频率 f_0的转换[8]。这个公式可以同时修改基音数值和基

音的范围。设定 a=$\sqrt{\frac{\delta_B^2}{\delta_A^2}}$，一旦 a 已经确定，可以设定$b=\mu_B-\sqrt{\frac{\delta_B^2}{\delta_A^2}}\mu_A$。$\mu_A$表示源说话人 A 的$f_0$的均值，$\delta_A^2$表示源说话人 A 的$f_0$的方差，$\mu_B$表示目标说话人 B 的$f_0$的均值，$\delta_B^2$表示目标说话人 B 的$f_0$的方差。均值和方差分别由公式（4-20）和公式(4-21)估计得。

$$\mu=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{4-20}$$

$$\delta^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{4-21}$$

求出均值和方差，f_0就可以通过公式(4-22)转换。

$$f_{0B}=\sqrt{\frac{\delta_B^2}{\delta_A^2}}(f_{0A}-\mu_A)+\mu_B \tag{4-22}$$

公式(4-22)中，f_{0B}表示转换后的基频，f_{0A}表示转换前的基频（源说话人的基音频率）。

通过公式(4-22)，可以对源说话人的基音数值和基音的范围进行修改，使得基音频率数值和基音频率的范围符合目标说话人的基音频率特征。

4.4 基于分段的说话人语音转换方法

通过前面对语音编码器、频谱的转换方法、以及基音频率转换方法的分析说明，现在对于基于分段的说话人语音转换方法的主要技术已经很清楚了。下面给出本文提出的基于分段的说话人语音转换的实现过程。本文提出的基于分段的说话人语音转换方法主要分为训练和转换两部分。训练部分又分成频谱转换函数的训练和基音频谱的训练。

4.4.1 频谱转换函数的训练过程

频谱转换函数的训练过程实际上就是求 mel 倒谱的转换函数的过程。先分别把源说话人和目标说话人的训练语音库进行下面的操作：

1.分段切分标注。把说话人的训练语句进行标注分段。

2.mel 倒谱分析。把说话人训练语音信号进行 mel 倒谱分析，得到 mel 倒谱系数。

3.对每个基本语音段训练一个高斯混合模型。

完成以上步骤后，再根据前面分析的频谱转换函数求出它的各个参数，从而

得到 mel 倒谱的转换函数。注意的是，基本语音段 P_i 的概率 β_i 是根据语音库统计出来的。

频谱转换函数的训练过程如图 4-6，（P1 表示第 1 个基本语音段，PN 表示第 N 个基本语音段）。

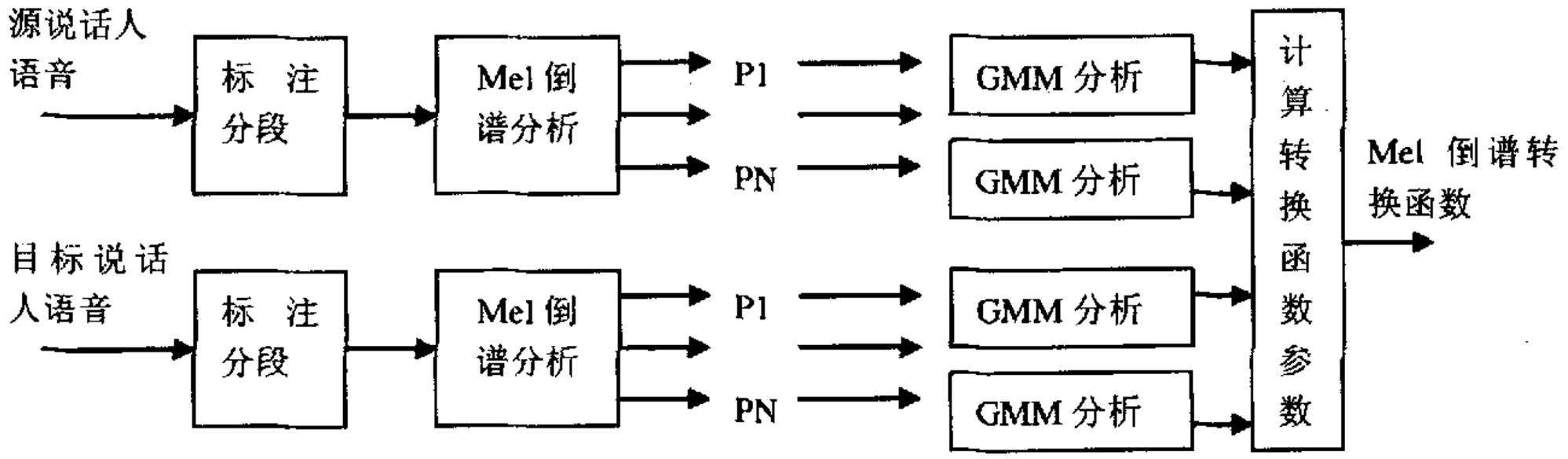


图 4-6 频谱转换函数的训练过程

Fig.4-6 Training the function for mel-cepstral

4.4.2 基音频率转换函数的训练过程

从上面讨论知道，基音频率的转换函数为 $f_{0B}=\sqrt{\frac{\delta_B^2}{\delta_A^2}}(f_{0A}-\mu_A)+\mu_B$，其中 μ_A 表示源说话人 A 的 f_0 的均值，δ_A^2 表示源说话人 A 的 f_0 的方差，μ_B 表示目标说话人 B 的 f_0 的均值，δ_B^2 表示目标说话人 B 的 f_0 的方差。那么训练实际上就是求 μ_A、δ_A^2、μ_B、δ_B^2 这四个参数。

首先，分别选取源说话人和目标说话人的一段语音，再分别求出这两段语音的基音频率序列，最后分别统计算出 μ_A、δ_A^2、μ_B、δ_B^2 的值。把这些参数代入转换公式中，就可以得到一个全局范围的转换函数。

4.4.3 转换的过程

在基于分段的说话人语音转换过程中，先把源说话人的语音分别进行基音频率分析和 mel 倒谱分析得到基音频率和 mel 倒谱系数，然后运用训练中得到的基音频率的转换函数和 mel 倒谱系数的转换函数对这些参数进行转换，最后通过 MLSA 滤波器合成出新的语音。这个新语音就是目标说话人说的语音。转换过程如图 4-7。

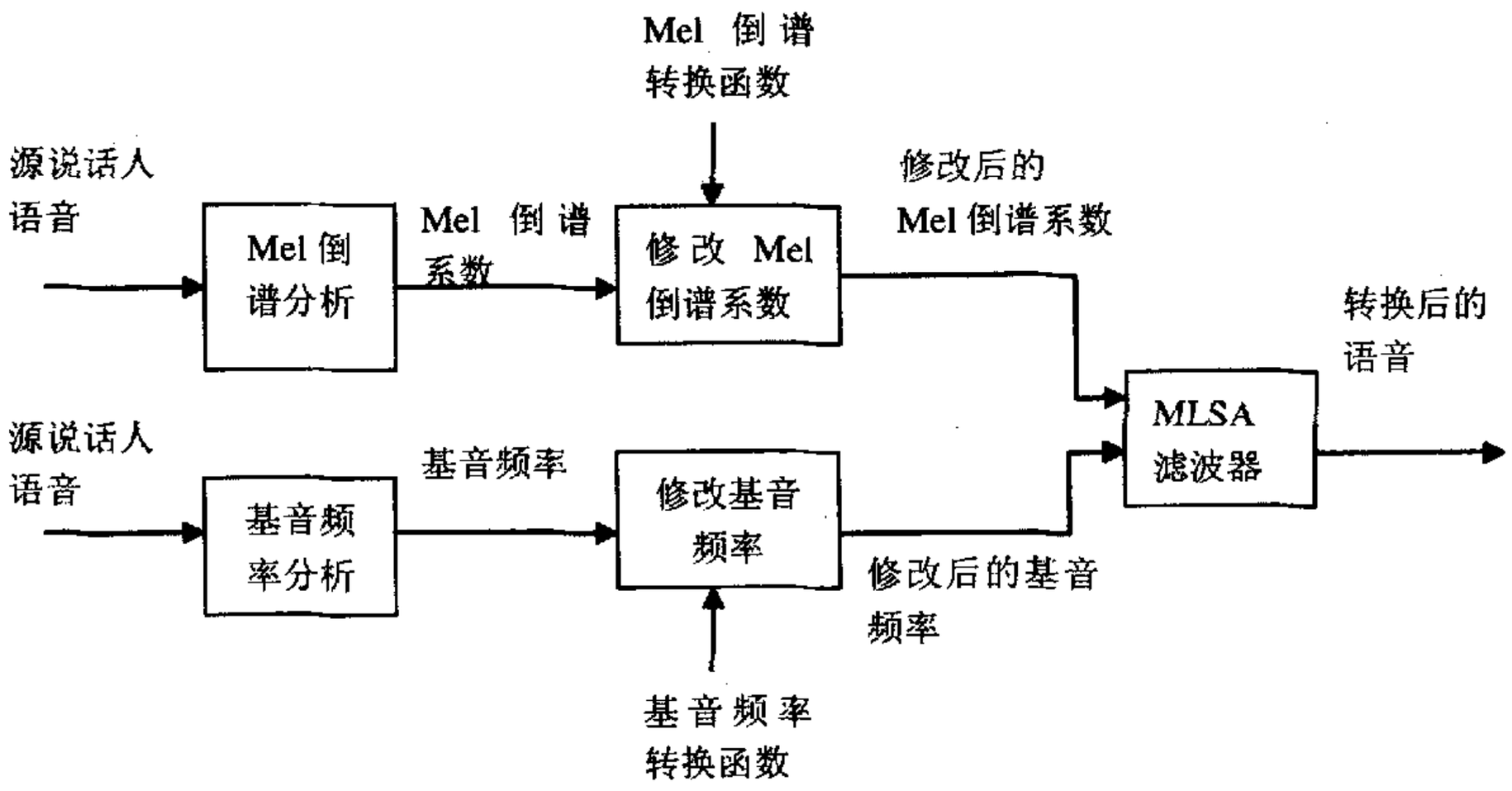


图 4-7 说话人语音转换过程

Fig.4-7 procession of voice conversion

4.5 本章小节

本章提出了一种基于分段的说话人语音转换方法。本章介绍了在试验中用到的语音编码器，以及分别论述了这种基于分段的说话人语音转换方法的频谱和基音频率的转换方法，最后给出基于分段的说话人语音转换方法的训练过程和转换过程。

（1） 本文提出了一种基于分段的说话人语音转换方法，因为这种方法不需要源说话人和目标说话人说的训练语句是相同的，适用于单语种和跨语种的说话人语音转换。

（2）研究表明基音频率和频谱特征是表现一个人的语音特征的主要参数。本文选用“pitch＋mel 倒谱＋MLSA 滤波器”的语音编码器，特征参数是基音周期（pitch）（或基音频率）和 mel 倒谱（mel-cepstral），通过修改这两个参数达到转换的目的。

（3）在对于 mel 倒谱参数的转换中，本文为每个基本语音段建立一个 GMM，采用基于分段的转换函数对 mel 倒谱参数进行修改。

（4）本文采用一个全局的转换函数 $f_{0B}=\sqrt{\frac{\delta_B^2}{\delta_A^2}}(f_{0A}-\mu_A)+\mu_B$ 对基音进行修改，这样可以达到一起修改基音数值和基音的范围的目的。

第五章 单语种（英语）说话人语音转换的研究

在第四章中，本文已经详细论述了本文提出的基于分段的说话人语音转换方法及相关技术。本章将给出基于分段的单语种（英语）说话人语音转换的实现方法。

5.1 英语的单音素集

基于分段的说话人语音转换方法要求找到一个基本语音段集作为说话人语音的基本构成单元集，使所有语音可以由这个基本语音段集中的元素构成。这个基本语音段集可以是单音素集、双音素集等，也可以是自己定义的非音素的基本语音段集。本文采用单音素(monophone)集作为基本语音段集。

英语中共有 40 个单音素，如果加上一个静音（pau），共 41 个单音素。英语音素分为元音和辅音两类，元音音素有 16 个(表 5-1)，辅音音素有 24 个(表 5-2)。其中的鼻音 m 和 n 既可做声母，又可做韵母。

5.2 语音数据库及其标注

本实验采用CMU(Carnegie Mellon University)的bdl英语语音数据库和jmk英语语音数据库分别作为源说话人语音库和目标说话人语音库。这里，bdl的母语为美语，其对应的英语语音库是标准的美音。而jmk是加拿大人，其说的英语具有很浓的加拿大口音。设bdl为源说话人，jmk为目标说话人，希望可以把bdl说的美式英语进行转换，使其听起来象是jmk的声音。

bdl 和 jmk 语音库分别含有 1132 句英语语句。在训练转换函数之前先对这些语句进行切分标注。在本实验中，因为 bdl 和 jmk 语音库已经带有各语句对应的基于单音素（monophone）的标注信息文件，所以在本实验中直接用到语音库带有的标注信息。如果语音库没有附有标注信息文件，就要采用英语语音切分系统对语音库里的语句进行标注。

标注文件里包含有 3 个元素：起始时间、中止时间、单音素符号。图 5-1 是 dbl 库中语句 arctic_a0001.wav 的标注文件 arctic_a0001.lab。

图 5-1 中第 1 列的数据表示单音素起始时间，第 2 列的数据表示单音素中止时间，第 3 列是单音素的符号。起始时间和中止时间是以 100 微秒作为单位。

表 5-1　英语元音音素表

Table 5-1 English vowel

	国际音标	音素字母	举例词
单元音	[i:]	iy	bee
	[i]	ih	it
	[e]	eh	red
	[æ]	ae	map
	[ʌ]	ah	but
	[ɑ:]	aa	part
	[ɔ:]	ao	pork
	[u]	uh	push
	[u:]	uw	true
	[ə]	ax	paper
	[ə:]	er	her
双元音	[ei]	ey	Israeli
	[ai]	ay	die
	[ɔi]	oy	oil
	[əu]	ow	hoe
	[au]	aw	out

表 5-2 英语辅音音素表

Table 5-2 English consonant

	国际音标	音素字母	举例词
阻塞音	[p]	p	**p**en
	[t]	t	**t**ake
	[k]	k	**k**eep
	[b]	b	**b**ike
	[d]	d	**d**eep
	[g]	g	**g**ate
磨擦音	[f]	f	**f**at
	[ʃ]	sh	**sh**op
	[s]	s	**s**eem
	[θ]	th	**th**in
	[ʒ]	zh	a**z**ure
	[z]	z	**z**one
	[v]	v	**v**ote
	[ð]	dh	**th**is
塞擦音	[tʃ]	ch	**ch**art
	[dʒ]	jh	**j**ob
鼻音	[m]	m	**m**eet
	[m]	m	bottom
	[n]	n	**n**ame
	[n]	n	butto**n**
	[ŋ]	ng	si**ng**
滑音	[w]	w	**w**eek
	[h]	hh	**h**ome
	[r]	r	**r**ich
	[l]	l	**l**ate
半元音	[j]	y	**y**es

```
0 400000 pau
400000 2000000 ao
2000000 3000000 th
3000000 4100000 er
4100000 4500000 ah
4500000 5100000 v
5100000 5700000 dh
5700000 6100000 ax
6100000 7200000 d
7200000 8400000 ey
8400000 9100000 n
9100000 10000000 jh
10000000 10600000 er
10600000 11900000 t
11900000 12900000 r
12900000 14100000 ey
14100000 15500000 l
15500000 17200000 f
17200000 17700000 ih
17700000 18500000 l
18500000 19000000 ax
19000000 19900000 p
19900000 21100000 s
21100000 21700000 t
21700000 23100000 iy
23100000 24700000 l
24700000 25800000 z
25800000 26300000 eh
26300000 27000000 t
27000000 28000000 s
28000000 28900000 eh
28900000 29300000 t
29300000 30600000 er
30600000 31800000 ax
31800000 32200000 pau
```

图 5-1 dbl 库中语句 arctic_a0001.wav 的标注文件 arctic_a0001.lab

Fig. 5-1 the label file of arctic_a0001.wav in dbl speech database

5.3 转换函数的训练

5.3.1 mel 倒谱转换函数的训练

本实验采用单音素(monophone)集作为基本语音段集。加上一个静音（pau），则英语单音素集由 41 个英语单音素组成：{'aa','ae','ah','ao','aw','ax','ay','b','ch','d','dh','eh','er','ey','f','g','hh','ih','iy','jh','k','l','m','n','ng','ow','oy','p','pau','r','s','sh','t','th','uh','uw','v','w','y','z','zh'}。

为了求得源说话人 bdl 和目标说话人 jmk 的每一个单音素的高斯混合模型参数，先根据标注文件把 bdl 和 jmk 的语音库中的语音都切成一个个单音素，而且把同一单音素的语音归为一个语音集，比如发'aa'音的语音段都集合在单音素语音库 Caa 中，这样 bdl 和 jmk 就分别有 41 个单音素语音库：
{'Caa','Cae','Cah','Cao','Caw','Cax','Cay','Cb','Cch','Cd','Cdh','Ceh','Cer','Cey','Cf','Cg','Chh','Cih','Ciy','Cjh','Ck','Cl','Cm','Cn','Cng','Cow','Coy','Cp','Cpau','Cr','Cs','Csh','Ct','Cth','Cuh','Cuw','Cv','Cw','Cy','Cz','Czh'}。

然后把每一个单音素语音库作为对应的单音素的训练语音集，来求出该单音素的 mel 倒谱对应的高斯混合模型参数。为了求得更好的高斯混合模型参数，在训练之前，要对每一个单音素语音库的语音进行了人工筛选，删去有明显错误和时间过短的语音。

在对声音文件进行 mel 倒谱分析时，采用的参数设置为：mel 倒谱的阶数为 24（包括 0 阶，共 25 阶 mel 倒谱系数），帧长是 25ms，位移为 5ms，α 取 0.42。这样，本实验的频谱参数为 25 维的 mel 倒谱系数。对每个单音素语音库的语音分别进行 mel 倒谱分析后得到单音素 mel 倒谱矢量集，然后运用 EM 算法，求出该单音素的高斯混合模型参数：$\lambda=\{\alpha_i,\mu_i,\Sigma_i\},i=1,2,...,m$。本实验中设高斯模型混合的个数 m 为 1，所以 α_i 为 1。

为了求得单音素 P_i,(i=1...,41)的概率 β_i，就要对 bdl 语音库进行统计，得到各个单音素的 β。统计的结果如下表 5-3 和表 5-4。

求出 bdl 和 jmk 每一个单音素的高斯混合模型参数 $\lambda_j, j=1,...,41$ 和每个音素的概率 β_i 后，就可以求得 mel 倒谱的转换函数：

$$Y=F(X)=\sum_{i=1}^{K}P(P_i\mid X)\phi_i \tag{5-1}$$

$$\phi_i=\sum_{j=1}^{M}\alpha_{ij}[\mu_{ij}{}^{Y}+\Sigma_{ij}^{Y}\Sigma_{ij}^{X^{-1}}(X-\mu_{ij}{}^{X})] \tag{5-2}$$

5.3.2 基音频率的转换函数的训练

要得到基音频率的转换函数，就是统计出求 μ_{bdl}、δ^2_{bdl}、μ_{jmk}、δ^2_{jmk} 这四个参数。在本实验中，在 bdl 和 jmk 语音库中选出 500 句（大概 30 分钟）语音作为训练语音。分别对 bdl 和 jmk 的语句求基音频率 f_0，然后用估计公式分别求出 μ_{bdl}、δ^2_{bdl}、μ_{jmk}、δ^2_{jmk}。

$$\mu=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{5-3}$$

表 5-3 辅音音素的 β

Table 5-3 β of the consonants

	国际音标	音素字母	统计个数 n	概率 β (%)
阻塞音	[p]	p	645	1.7
	[t]	t	2359	6.2
	[k]	k	940	2.5
	[b]	b	625	1.6
	[d]	d	1789	4.7
	[g]	g	406	1.1
磨擦音	[f]	f	726	1.9
	[ʃ]	sh	324	0.8
	[s]	s	1613	4.2
	[θ]	th	224	0.6
	[ʒ]	zh	36	0.1
	[z]	z	1016	2.7
	[v]	v	638	1.7
	[ð]	dh	1068	2.8
塞擦音	[tʃ]	ch	218	0.6
	[dʒ]	jh	228	0.6
鼻音	[m]	m	1074	2.8
	[n]	n	2431	6.4
	[ŋ]	ng	396	1.0
滑音	[w]	w	908	2.4
	[h]	hh	894	2.3
	[r]	r	1548	4.1
	[l]	l	1454	3.8
半元音	[j]	y	293	0.8
静音		pau	2300	6.0

表 5-4 元音音素的 β

Table 5-4 β of the vowel

	国际音标	音素字母	统计个数 n	概率 β (%)
单元音	[i:]	iy	1230	3.2
	[i]	ih	2162	5.7
	[e]	eh	1003	2.6
	[æ]	ae	1229	3.2
	[ʌ]	ah	777	2.0
	[ɑ:]	aa	810	2.1
	[ɔ:]	ao	606	1.6
	[u]	uh	185	0.5
	[u:]	uw	472	1.2
	[ə]	ax	2419	6.3
	[ə:]	er	943	2.5
双元音	[ei]	ey	618	1.6
	[ai]	ay	739	1.9
	[ɔi]	oy	92	0.2
	[əu]	ow	514	1.3
	[au]	aw	259	0.7

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{5-4}$$

求得的参数如表 5-5。

表 5-5 基音频率转换函数参数

Table 5-5 Parameters of the conversion function for f0

	μ_{bdl}	δ^2_{bdl}	μ_{jmk}	δ^2_{jmk}
500 句	127.995	1393.90	110.223	388.17

求出这些参数后，就根据公式 $f_{0B} = \sqrt{\frac{\delta_B^2}{\delta_A^2}}(f_{0A} - \mu_A) + \mu_B$ 进行基音频率的修改。

5.4 单语种说话人语音转换的过程

前面已经训练出 mel 倒谱和基音频率的转换函数参数，现在进行说话人语音转换（转换过程框图见第四章图 4-7）。在转换过程中，要先对源说话人的语音进行编码，从而得到基音频率和 mel 倒谱系数，然后把一帧帧的基音频率和 mel 倒谱系数分别进行转换，再通过 MLSA 滤波器得出新语音。

基音频率的转换只要直接通过转换函数就可以求出新的基音频率。本文把由 500 句（大概 30 分钟）语音求得的 μ_{bdl}、δ^2_{bdl}、μ_{jmk}、δ^2_{jmk} 代入转换公式 $f_{0B}=\sqrt{\frac{\delta_B^2}{\delta_A^2}}(f_{0A}-\mu_A)+\mu_B$，得：

$$f_{0B}=\sqrt{\frac{\delta_B^2}{\delta_A^2}}(f_{0A}-\mu_A)+\mu_B=0.5277(f_{0A}-127.995)+110.223 \qquad (5\text{-}5)$$

从 mel 倒谱转换函数 $Y=F(X)=\sum_{i=1}^{K}P(P_i\mid X)\phi_i$ 得出，转换的时候要求出特征矢量 X 属于第 i 个基本音素（P_i）的概率 $P(P_i\mid X)$，而且 $K\le N$，取的是 $P(P_i\mid X)$ 最大的前 K 个音素来计算。

在本文中，为了可以更好的进行说话人语音转换，转换之前已经对预转换的语音进行切分标注，即已知每帧的特征矢量是属于哪个音素的，也就是说本文选取 K=1，并已知特征矢量 X 属于某个确定基本音素（P_i）的概率 $P(P_i\mid X)$为 1。转换公式变为：

$$Y=F(X)=\phi_i=\sum_{j=1}^{M}\alpha_{ij}[\mu_{ij}^{\ Y}+\Sigma_{ij}^{Y}\Sigma_{ij}^{X^{-1}}(X-\mu_{ij}^{\ X})]$$

本文的 GMM 的混合个数 M 设为 1，转换公式可以简化为：

$$Y=F(X)=\phi_i=\sum_{j=1}^{M}\alpha_{ij}[\mu_{ij}^{\ Y}+\Sigma_{ij}^{Y}\Sigma_{ij}^{X^{-1}}(X-\mu_{ij}^{\ X})]=\mu_i^{\ Y}+\Sigma_i^{Y}\Sigma_i^{X^{-1}}(X-\mu_i^{\ X}) \quad (5\text{-}6)$$

5.5 实验结果与讨论

从语音学上知道，更能反应一个人说话特点的是浊音语音，大多数人发的清音是差不多的，所以在说话人语音转换中，重点进行了浊音的转换，清音部分没有进行转换。而浊音又以元音为主，因此下面重点分析一下元音（以“aa”为例）转换前后的语音特性，并对转换结果进行评价和讨论。

下边给出元音“aa”的转换结果（以发单个“aa”音素的语句作为语音转换）。其中 aa_212_bdl.raw 为源说话人说的语音，aa_212_jmk.raw 是目标说话人说的语

音，aa_212_bdl2jmk.raw 为由 aa_212_bdl.raw 转换成 aa_212_jmk.raw 的语音。

客观评价：下面给出各语音的 FFT 频谱图（图 5-2，图 5-3，图 5-4），FFT 频谱参数设定如 aa_212_bdl.raw 的 FFT 频谱图（图 5-2）所示。

通过观察 3 幅 FFT 频谱图，发现转换后的语音 aa_212_bdl2jmk.raw 的 FFT 频谱图中共振峰已经改变，与目标说话人语音 aa_212_jmk.raw 的共振峰更相近，而且转换后的语音 aa_212_bdl2jmk.raw 与目标说话人语音 aa_212_jmk.raw 的频谱距离小于源说话人语音 aa_212_bdl.raw 和目标说话人语音 aa_212_jmk.raw 的频谱距离。

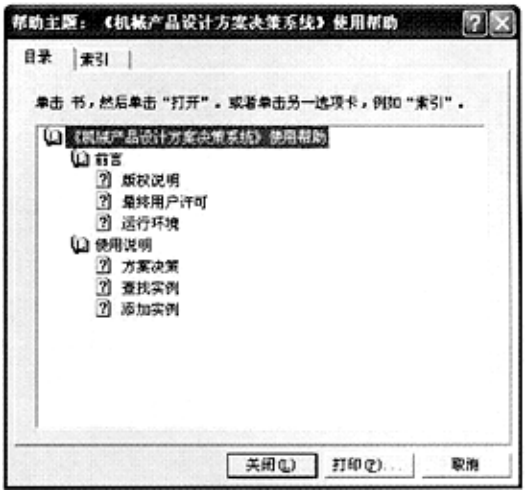


图 5-2　源说话人语音 aa_212_bdl.raw 的 FFT 频谱图

Fig. 5-2　FFT Spectrum of source speech aa_212_bdl.raw

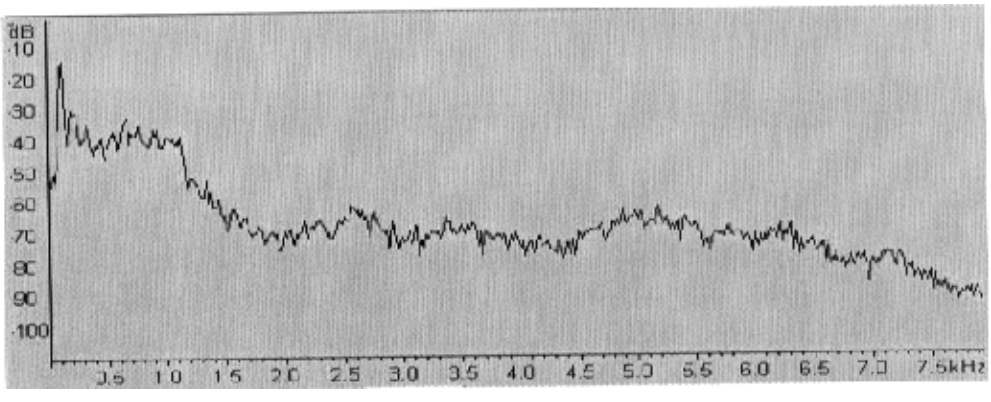


图 5-3　目标说话人语音 aa_212_jmk.raw 的 FFT 频谱图

Fig. 5-3　FFT Spectrum of target speech aa_212_jmk.raw

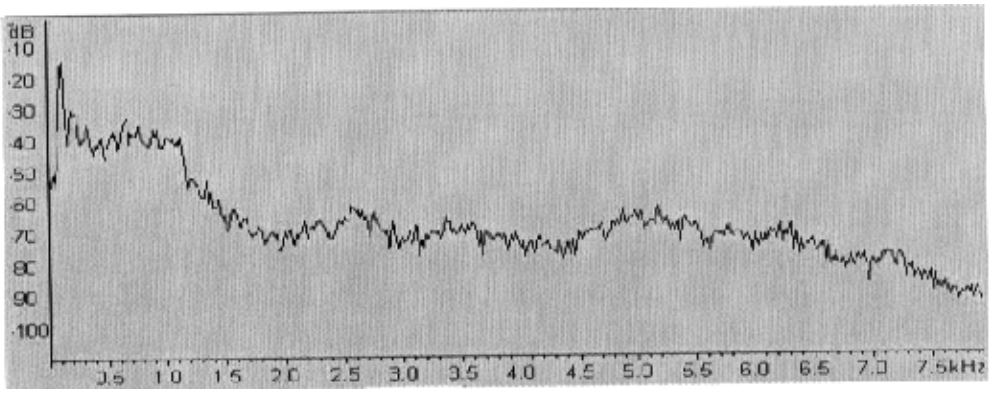


图 5-4 转换后的语音 aa_212_bdl2jmk.raw 的 FFT 频谱图

Fig. 5-4 FFT Spectrum of modified speech aa_212_bdl2jmk.raw

主观评价：对源说话人 bdl 说的 star_bdl.raw 语音(单词“star”)中的“aa”音素进行转换，其他音素不转换，得到 star_bdl2jmk.raw。判断转换后的语音 star_bdl2jmk.raw 更接近于 bdl 的源语音 star_bdl.raw，还是更接近于 jmk 的目标语音 star_jmk.raw，找 10 个人进行主观测试，10 个人都认为 aa_212_bdl2jmk.raw 更象是 jmk 说的。

对其他 15 个元音的转换结果进行客观和主观分析，分析结果和上面对音素“aa”的分析结果一样。结果表明本文提出的基于分段的说话人语音转换方法可以实现单语种的说话人语音转换。

5.6 本章小节

本章实现了基于分段的单语种（英语）说话人语音转换，并给出了实验的结果，最后进行分析。

（1）单语种（英语）说话人语音转换采用的基本音素是单音素（monophone），其中包括元音音素 16 个，辅音音素 24 个。说话人语音转换时主要是对元音音素的转换。

（2）进行 mel 倒谱转换时，先对转换的语音进行切分标注，即转换前已知特征矢量 X 属于某个确定基本音素（P_i）的概率 $P(P_i|X)$为 1，这样可以提高转换的效果。

（3）通过实验结果分析知道，转换后的语音的频谱包络和共振峰接近于目标语音，而且通过主观评价，也表示语音听起来更接近于目标语音，因此说明基于分段的单语种说话人语音转换方法是可行的。

第六章 跨语种(中英)说话人语音转换的研究

在第五章中已经对基于分段的单语种（英语）说话人语音转换进行了研究，本章重点讨论如何将这种基于分段的说话人语音转换应用于跨语种（中英）说话人语音转换，并根据跨语种说话人语音转换的特点，提出二叉树的转换方法。在本文中，设英文为源说话人的母语，中文为目标说话人的母语。

6.1 中英文比较

要进行跨语种（中英）说话人语音转换，首先就要了解这两种语言的异同点，而语音最基本的构造单元就是音素。第五章已经对英语的基本音素库进行了分析，下面分析一下中文基本音素库，并比较中英文音素的异同点。

6.1.1 汉语音素

汉语音素包括韵母和声母两部分，韵母 38 个（表 6-1），声母 21 个（表 6-2）。

6.1.2 中英语音音素比较

通过观察英语和汉语的音素表和听觉的判断，发现英语和汉语有一部分的音素是基本相同的，而一部分没有。

（1） 英语汉语基本相同的音素 （英--汉）（29 个，占 40 个英语音素的 72.5%）

aa－a (啊)； ao－o（哦）； aw－ao (熬)； ay－ai (哀)； er－e（鹅）； ey－ei (累)；
iy－i(衣)； ng－ing（ong、eng）； ow－ou (欧)； uw—u (乌)；
b－b (玻)； ch－q(去)； d－d (得) ； f－f (佛)； g－g (哥) ； hh－h (喝) ；
jh－j (决)； k－k(科)； l－l (勒)； m－m (摸)； n－n (讷)； p－p (坡)； r－r（日）；
s－s (思)； sh－x（续）； t－t (特)； w—w（我）； y—y (应)； zh－r（入）

（2）英语中有汉语中没有的音素（英--汉）（11 个,占 40 个英语音素的 27.5%）

ae； ah； ax； eh； ih； oy； uh； dh； th； v； z

其中 ah 是 aa 的短音，ax 是 er 的短音，uh 是 uw 的短音，ih 是 iy 的短音，这些可以通过切短相对应的长音得到。(4 个，占 40 个音素的 10%)。剩下的(ae；eh； oy； dh； th； v； z)（7 个，占 40 个音素的 17.5%）没有对应的中文音素。

表 6-1　汉语韵母音素表[2]

Table 6-1 Chinese vowel[2]

单韵母		**i**	**u**	**ü**
	a	ia	ua	
	o		uo	
	e	ie		üe
复韵母	**ai**		uai	
	ei		uei	
	ao	iao		
	ou	iou		
鼻韵母	**an**	ian	uan	üan
	en	in	uen	ün
	ang	iang	uang	
	eng	ing	ueng	
	ong	iong		
特殊韵母	**-i**			
	er			

注：特殊韵母（1）-i 有两种发音，在 z,c,s 后发资韵，在 zh,ch,sh,r 后发知韵；（2）er 是儿化音

表 6-2　汉语声母音素表[2]

Table 6-1 Chinese consonant[2]

		双唇音	唇齿音	舌尖前音	舌尖中音	舌尖后音	舌面前音	舌根音
塞音（清）	送气	p	f		t			k
	不送气	b			d			g
擦音	清			s		sh	x	h
	浊					r		
塞擦音（清）	送气			c		ch	q	
	不送气			z		zh	j	
鼻音（浊）		m			n			
边音（浊）					l			

6.2 建立二叉树的转换方法

从上面中英文音素的比较发现，有大部分的英文音素可以找出对应的中文音素，但是仍然有一部分的英文音素不能找到对应的中文音素。在基于分段说话人语音转换过程中，首先要对源说话人和目标说话人的语音的每一个基本音素（英语音素）都要建立一个 GMM，然后求出每个音素对应的转换函数，但是现在不可能做到对目标说话人语音的每一个基本音素（英语音素）建立一个 GMM。为了解决这个问题，本文提出了对英语基本音素建立二叉树的转换方法。在文献[31、32]中，MLLR 算法为了克服样本不足的问题，采用了 Regression Class Tree 进行聚类，本文就是从中得到启迪，提出对英语基本音素建立二叉树的转换方法，解决了中英文不对应的音素的转换问题。

在建立二叉树之前，要先对源说话人的的每个基本音素的 mel 倒谱都建立一个 GMM，得到每一个音素的均值。

建立一个二叉树的具体步骤为：

1. 把所有的音素都归为一类，作为根节点 1，求出根节点 1 的均值 mean1。

2. 选取一个适当的值 a，令 mean2=mean1-a，mean3=mean1+a 分别作为节点 2 和节点 3 的均值。

3. 对于每一个模型，分别求它与节点 2 和节点 3 的距离 d2、d3。如 d2>d3，这个模型归于节点 2，否则归于节点 3。

4. 重新求出节点 2 和节点 3 的均值。

5. 检查分类是否满足条件，如果不满足，则重复以上的步骤进行。直到满足聚类的条件。

假设已经建好一棵二叉树如图 6-1，图中有 4 个音素，分别位于节点 4，节点 5，节点 6 和节点 7。

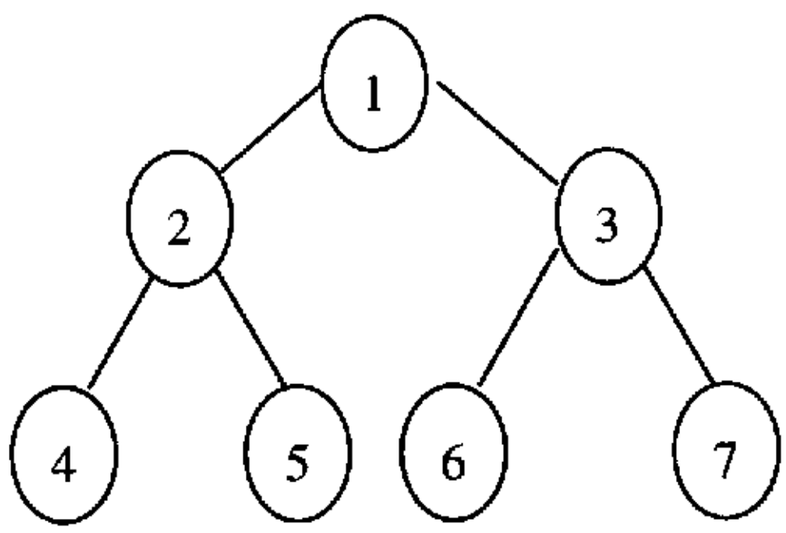


图 6-1 音素的二叉树

Fig. 6-1 the bitree of phones

假设节点 7 样本不够，或者是没有样本时，而节点 6 样本足够，可以先求出把节点 6 和节点 7 的音素归为一类的节点（节点 3）的转换函数，把这个节点 3

的转换函数作为节点 7 的转换函数，或者是直接采用节点 6 的转换函数作为节点 7 的转换函数。这样就可以使得没有足够训练数据的音素的 mel 倒谱系数也可以通过相应的转换函数进行转换。

6.3 实验和讨论

在实验中，采用 CMU(Carnegie Mellon University)的 bdl 英语语音数据库和摩托罗拉中国研究院提供的一个女声（用 female_ch 表示）中文数据库分别作为源说话人语音库和目标说话人语音库。

6.3.1 训练基音频率的转换函数

在 bdl 和 female_ch 语音库中各选出 500 句（大概 30 分钟）语音作为训练语音，求得的相关参数如表 6-3。

表 6-3 基音频率转换函数参数

Table 6-3 The parameters of conversion function for f0

	μ_{bdl}	δ^2_{bdl}	μ_{ch}	δ^2_{ch}
500 句	127.995	1393.9076	187.647	1423.9569

6.3.2 中英文相对应的音素的转换

对于中英文相对应的音素的转换方法和单语种的转换方法一样，直接求出该音素的转换函数，然后一帧帧进行转换。在这里，也是先对转换语音进行切分标注，已知特征矢量 X 属于某个确定的基本音素（P_i），即 $P(P_i|X)$为 1。下面以发单个“aa”音素的语句作为语音转换，并分析转换前后的频谱变化。其中 aa_212_bdl.raw 为源说话人说的语音，aa_ch.raw 是目标说话人说的语音，aa_212_bdl2ch.raw 为由 aa_212_bdl.raw 转换成 aa_212_ch.raw 的语音。

客观评价：下面给出各语音的 FFT 频谱图（FFT 频谱参数设定如源说话人语音 aa_212_bdl.raw 的 FFT 频谱图所示）。

通过观察 3 幅 FFT 频谱图（图 6-2、图 6-3、图 6-4），发现转换后的语音 aa_bdl2ch.raw 的 FFT 频谱图中共振峰已经改变，与目标说话人语音 aa_ch.raw 的共振峰更相近，而且转换后的语音 aa_212_bdl2ch.raw 与目标说话人 aa_ch.raw 的频谱距离小于源说话人语音 aa_212_bdl.raw 和目标说话人语音 aa_ch.raw 的频谱距离。

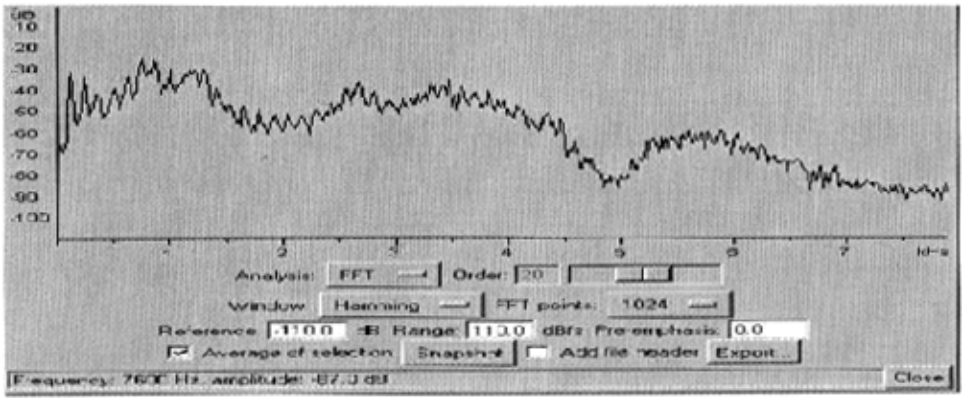


图 6-2 源说话人语音 aa_212_bdl.raw 的 FFT 频谱图

Fig.6-2 FFT spectrum of source speech aa_212_bdl.raw

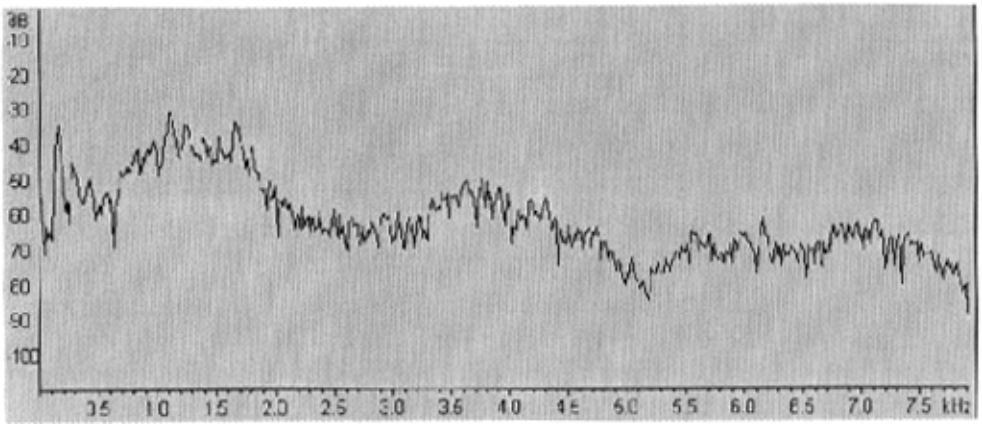


图 6-3 目标说话人语音 aa_ch.raw 的 FFT 频谱图

Fig.6-3 FFT spectrum of target speech aa_ch.raw

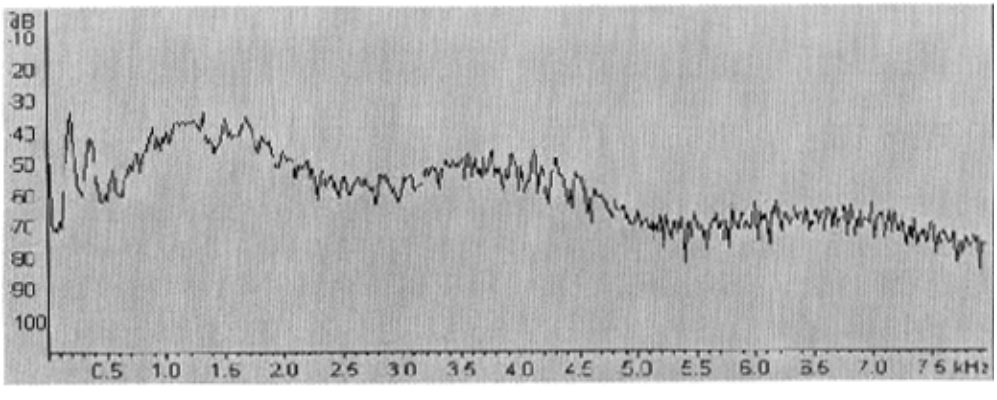


图 6-4 转换后的语音 aa_bdl2ch.raw 的 FFT 频谱图

Fig.6-4 FFT spectrum of modified speech aa_bdl2ch.raw

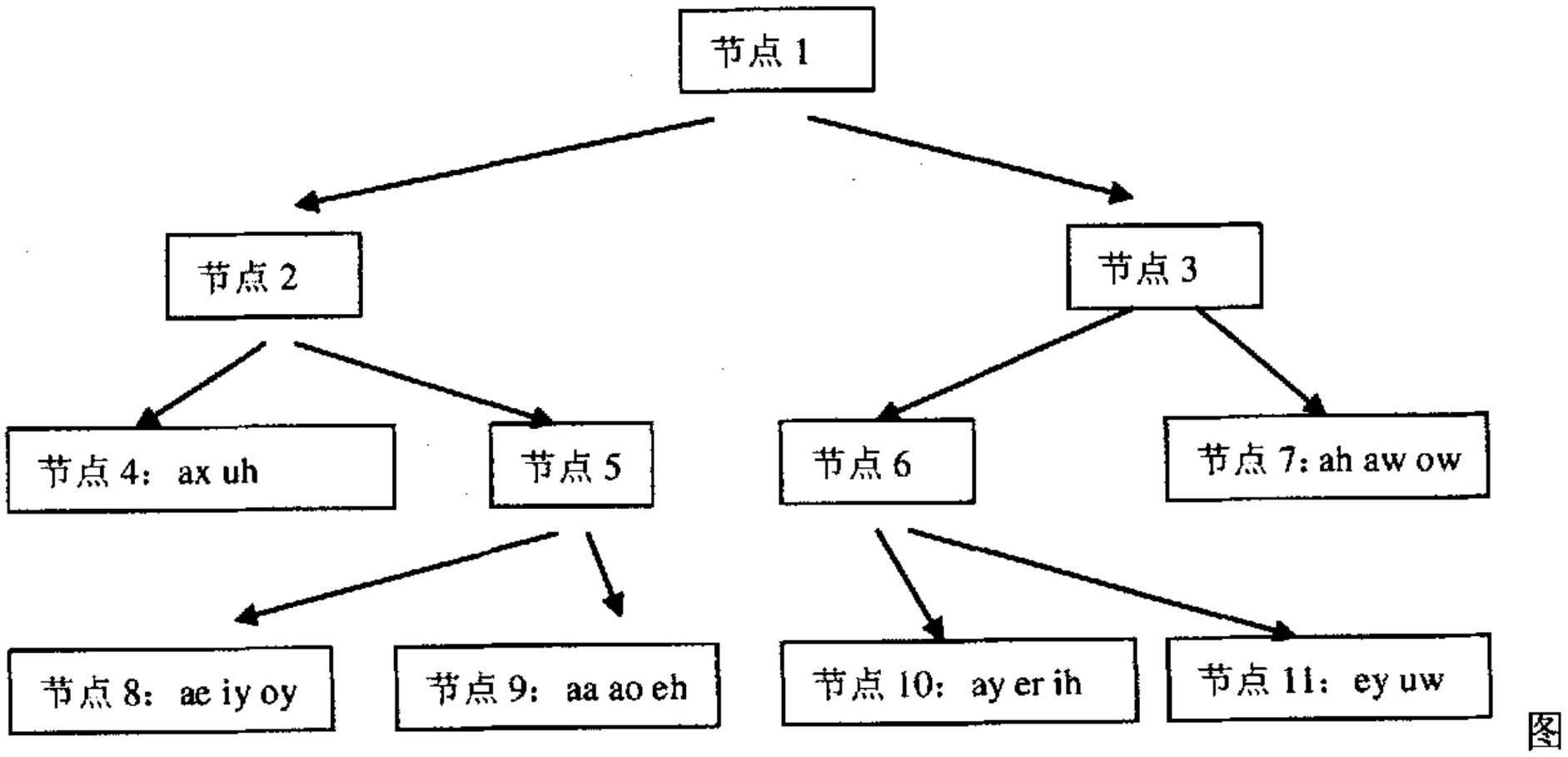
主观评价：对源说话人 bdl 说的 star_bdl.raw 语音（单词 star）中的“aa”音素进行转换，其他音素不转换，得到转换后的语音 star_bdl2ch.raw。判断转换后的语音 star_bdl2ch.raw 更接近于源说话人 bdl 说的，还是更接近于目标说话人 female_ch 说的。找 10 个人进行主观测试，10 个人都认为转换后的语音 aa_bdl2ch.raw 更象是 female_ch 说的。

对于其他 12 个英汉相对应的元音音素的转换结果进行客观和主观分析，分析结果和上面对音素“aa”的分析结果一样。结果证明基于分段的跨语种（中英）说话人语音转换中，汉英对应的元音音素可以得到转换。

6.3.3 建立二叉树

把英语音素集分成两类：元音和辅音。转换的时候，本文只对元音进行转换，转换之前先对元音建立二叉树。图 6-4 给出用 bdl 训练出来的 GMM 来求得的元音二叉树。

元音包括了 16 个音素{'aa','ae','ah','ao','aw','ax','ay','eh','er','ey','ih','iy','ow','oy','uh','uw'}。其中'ae','eh','oy'在中文中找不到对应的音素。



图

图 6-4 元音的二叉树

Fig. 6-4 The bitree of vowel

在转换过程中，当出现'ae','eh','oy'时，就采用该音素所在节点的转换函数对该音素进行转换。

6.3.4 根据二叉树进行转换

为了检验根据二叉树进行转换的结果，下面分析音素 aa 用音素 ao 的转换函数进行转换的结果。

从图 6-4 可以知道{aa、ao、eh}在同一节点 9，本实验假设 aa 没有训练样本，而 ao 有足够的样本，当对 aa 进行转换时，采用 ao 的转换公式对 aa 进行转换。aa_bdl.raw 为源说话人的语音，aa_jmk.raw 为目标说话人的语音，aa_bdl2jmkuseao.raw 为用 ao 转换函数对 aa 进行转换后的语音。

客观评价：通过观察 3 幅 FFT 频谱图(图 6-5、图 6-6、图 6-7)，发现用 ao 转换函数转换后的语音 aa_bdl2jmkuseao.raw 的 FFT 频谱图中共振峰已经改变，与目标说话人的语音 aa_jmk.raw 的共振峰更相近，而且用 ao 转换函数转换后的语音 aa_bdl2jmkusao.raw 与目标说话人语音 aa_jmk.raw 的频谱距离小于源说话人语音 aa_bdl.raw 和目标说话人语音 aa_jmk.raw 的频谱距离。

主观评价：对源说话人 bdl 说的 star_bdl.raw 语音中的“aa”音素用“ao”的转换函数进行转换，其他音素不转换，得到 star_bdl2jmkuseao.raw。判断转换后的语音 star_bdl2jmkuseao.raw 更象源说话人 bdl 说的，还是更接近于目标说话人 jmk 说的。找 10 个人进行主观测试，10 个人都认为转换后的语音 star_bdl2jmkuseao.raw 更象是目标说话人 jmk 说的。

根据这种方法对'ae','eh','oy'进行转换，通过主观评价，转换后的'ae','eh','oy'更象是目标说话人说的。结果表明，二叉树的转换方法可以对中英文不对应的元音音素进行转换，因此本文提出的基于分段的转换方法适用于跨语种（中英）说话人语音转换。

6.5 本章小节

本章重点研究基于分段的跨语种（中英）说话人语音转换，并对转换结果进行了分析。比较了中英文的基本音素库，以及提出了解决中英文不对应的音素的转换方法，最后给出跨语种说话人转换的一些实验，并对实验结果进行了分析。

（1）比较中英文的基本音素库，发现 72.5%的英语音素可以在中文音素中找到对应的，10%的英语音素可以把中文音素中的某些音素进行切短得到对应的音素，剩下的英语音素没有对应的中文音素。

（2）对于中英文相对应的音素的 mel 倒谱的转换方法和单语种的转换方法一样，直接求出该音素的 mel 倒谱的转换函数，然后用音素的转换函数进行转换。经过对转换结果的分析，发现转换后的语音的频谱包络和共振峰接近于目标语音，

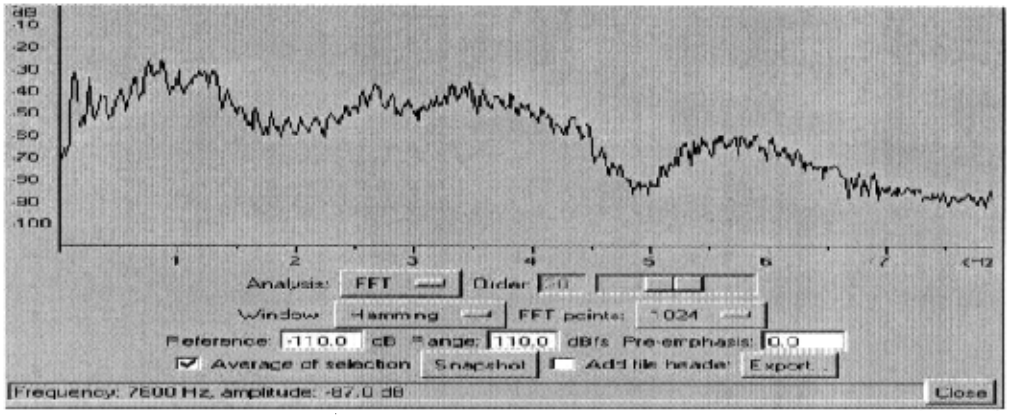


图 6-5 源说话人语音 aa_bdl.raw 的 FFT 频谱图

Fig.6-5 FFT spectrum of source speech aa_bdl.raw

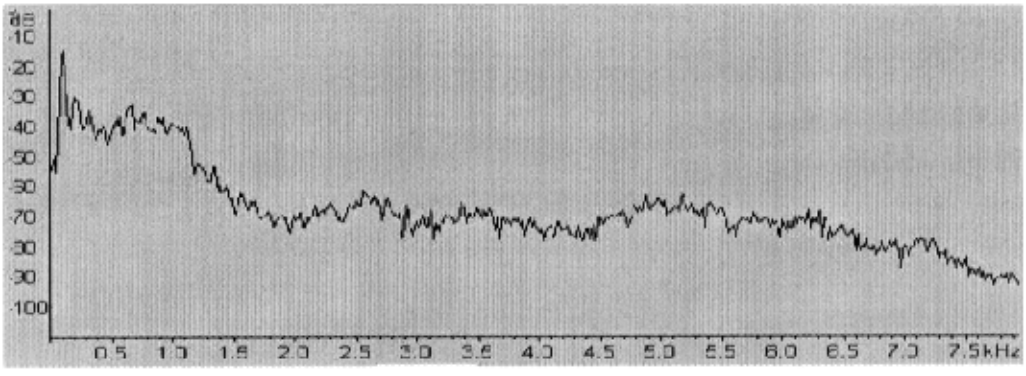


图 6-6 目标说话人语音 aa_jmk.raw 的 FFT 频谱图

Fig.6-6 FFT spectrum of target speech aa_jmk.raw

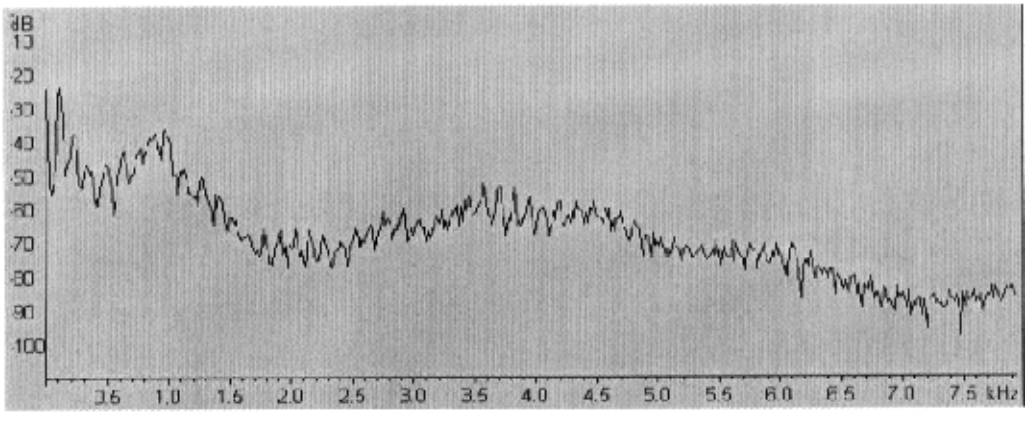


图 6-7 转换后的语音 aa_bdl2jmkuseao.raw 的 FFT 频谱图

Fig.6-7 FFT spectrum of modified speech aa_bdl2jmkuseao.raw

而且通过主观评价，也表示语音听起来更接近于目标语音。说明这种基于分段的说话人语音转换也适用于跨语种的说话人语音转换中汉英对应的音素的转换。

（3）为了对中英文不对应的音素进行转换，本文提出了二叉树的转换方法。经过对转换结果的分析，发现转换后的语音的频谱包络和共振峰接近于目标语音，而且通过主观评价，也表示语音听起来更接近于目标语音，因此表明这种二叉树的转换方法是可行的。因此本文提出的基于分段的转换方法适用于跨语种（中英）说话人语音转换。

结　论

本文主要研究说话人语音转换的方法，也就是希望找到一种方法，使得一个人（源说话人）说的语音经过转换使之听起来是另一个人（目标说话人）说的语音。近年来，说话人语音转换已经成为语音信号处理中的研究热点，而且国内外的一些研究机构也取得一定的成果。但是这些成果大多应用于单语种说话人语音，对训练样本有一定的限制性，一般都要求源说话人和目标说话人说一段相同的语音作为训练样本。可是，当源说话人和目标说话人不能说一段相同的语音时（比如跨语种说话人语音转换），这些对训练样本有限制性的方法就不可行了。针对这种情况，本文研究并提出了一种基于分段的说话人语音转换的方法，这种方法对训练样本不存在局限性，适用于单语种和跨语种的说话人语音转换。本文完成的主要工作包括：

1. 由于在基于分段的说话人语音转换中，需要对语音进行切分标注，因此本文研究了用HMM实现语音切分系统的方法，用HMM工具包（HTK）实现了用于把语音切分成音素串的语音切分系统。

2. 本文提出了一种基于分段的说话人语音转换方法，和以往的方法相比，这种基于分段的说话人语音转换不要求源说话人和目标说话人一定要说同样的训练语句，所以同时适用于单语种和跨语种的说话人语音转换。这种基于分段的说话人语音转换方法采用的是“pitch＋mel 倒谱＋MLSA 滤波器”语音编码器，修改的参数是基音周期 pitch（或基音频率）和 mel 倒谱。在对 mel 倒谱的转换中，先对源说话人和目标说话人的每段基本语音的 mel 倒谱参数训练 GMM 模型，求出一个转换函数，然后用这个转换函数对 mel 倒谱参数进行转换。而基音频率的转换则采用一个全局的转换公式，对基音频率的数值和范围进行修改。

3. 本文把基于分段的说话人语音转换方法应用于单语种（英语）说话人语音转换中，并对实验结果进行分析。说明这种基于分段的单语种说话人是可行的。

4. 研究中英文的语言特点，特别是两种语言的基本音素之间的异同点。通过比较，发现英文中有72.5%的英语音素可以在中文中找到相对应的音素；10%的英语音素可以通过对中文音素的切短，可以对应起来；17.5%的音素找不到中文对应的音素。

5. 为了实现基于分段的跨语种（中英）的说话人语音转换，解决中英文不对应的音素的转换问题，本文提出了一种二叉树的转换方法。实验表明，这种方法是可行的。

本文已经对基于分段的说话人语音转换方法的研究做了一定的工作，但是实

现一个质量很好的基于分段的单语种说话人语音转换系统或跨语种说话人语音转换系统，目前的工作还是远远不够的，以后应该进一步进行研究，主要方面是：

1．英语语音切分系统中用到的语法没有考虑到相邻音素之间的关系，所以目前切分的准确性较低。以后应该对语法进行改进，以达到较高的切分率。

2．语音转换中所采用的语音编解码器的效果直接影响到转换后的语音效果，本文采用的是“pitch＋mel倒谱＋MLSA滤波器”的语音编解码器，这个语音编码器相对比较简单，以后可以考虑采用其他较高质量的语音编解码器，比如CELP语音编码器。

3．本文用到的基本语音段是单音素（monophone），所以基本语音段的个数比较少，以后可以考虑采用不同的基本语音段，比如双音素（biphone）、三音素（triphone）等，这样基本语音段库就更丰富，这个语音段库更能很好的表示语音。这样对说话转换是很有利的。

让源说话人说的语音经过处理使其具有目标说话人语音特点是说话人语音转换的目的。说话人语音转换具有广泛的应用领域，比如文语转换（Text-to—Speech，TTS）、配音系统和翻译系统等。近年来说话人语音转换技术已经有了很大的发展，但是和真正做到音质较好地进行说话人语音转换还有一定距离。要想真正在说话人语音转换技术方面有进一步的突破，还有许多知识需要我们去学习，许多工作需要我们去研究、探索。

参考文献

[1] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara “ Voice Conversion Through Vector Quantisation”, proc. ICASSP-88, 1988

[2] 陈永彬，王仁华，“语言信号处理”，中国科学技术大学出版社，1990

[3] 陈立，“基于时域的男女声语音转换新途径的研究”，中国科学院声学研究所硕士学位论文，2000 年 1 月

[4] Kiyhiro Shikano, Satoshi Nakamura, Masanobu Abe, “Speaker Adaptation and Voice Coversion by Codebook Mapping”, IEEE, 1991

[5] Yannis Stylaniou , Oliver Cappe, “Continuous Probabilistic Transform for Voice Conversion”, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 6, NO. 2, MARCH 1998

[6] Yannis Stylaniou , Oliver Cappe. "A System for voice conversion based on probabilistic classification and a harmonic plus noise model." IEEE, 1998

[7] Masanobu Abe. "A segment-based approach to voice conversion." IEEE, 1991.

[8] LEvent M. Arslan, “Speaker Trainsformation Algorithm using Sepgmental Codebook(STASC)”, Speech Communication 28(1999)211-226, 1999

[9]Cheng-Yuan Lin and J. S Roger Jang, “New refinement schems for voice conversion”, IEEE, 2003

[10]Arun Kumer, Ashish Verma, “Using phone and diphone acoustic models for voice conversion: a step towards creation voice fonts”, IEEE, 2003

[11]Masanobu ABE, Kiyohiro SHIKANO, Hisao KUWABARA, “Cross-language voice conversion”, IEEE, 1990

[12] http://isw3.aist-nara.ac.jp/IS/Shikano-lab/e-home.html

[13] http://kt-lab.ics.nitech.ac.jp/~tomoki/demo_e.html

[14]BlackWell D. Koopmans L,“On the indentifiability problem for functions of finite Markov Chains”, Annals of Mathematical Statistics. 28 PP. 1-11-1015, 1957

[15]谢锦辉，“隐 Markov 模型（HMM）及其在语音处理中的应用”，华中理工大学，1995

[16] http://htk.eng.cam.ac.uk

[17] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, "The HTK Book (for HTK Version 3.2)", 2002

[18] 尉洪"基于说话人自适应的连接数字串语音识别",云南大学硕士学位论文, 2002年6月

[19] 雷鹏, "基于倒谱技术的说话人识别", 中国科学院声学研究所硕士学位论文, 2002

[20]M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996

[21]M. J. F. Gales & P. C. Woodland, " Mean and Variance Adaptation within the MLLR Framework," Computer Speech & Language, Vol. 10, 1996: pp. 249-264

[22]D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Processing, vol. 3, Jan. 1995: pp. 72 - 83

[23] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 1990: pp. 293 - 296,

[24] B. L. Tseng, F. K. Soong, and A. E. Rosenberg, "Continuous probabilistic acoustic map for speaker recognition," in *Proc.* IEEE Int. Conf. Acoustics, Speech, Signal Processing, 1992: pp. II-161 - II-164.

[25] 杨澄宇, "与文本无关说话人识别研究", 云南大学硕士学位论文, 2001年5月

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. B, vol. 39, 1977: pp. 1 - 22 and 22 - 38

[27] T. Fukada, K. Tokuda, T. kobayashi, S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech", Proc. ICASSP-92, 1992: pp. 137-140

[28]S. Imai, " Cepstral anaylysis synthesis on the mel frequency scale," ICASSP-83, 1983: pp. 93-96

[29]Keiichi Tokuda, Hidetoshi Matsumura, Takao Kobayashi, Satoshi Imai, "Speech coding based on adaptive mel-cepstral analysis", Proc. ICASSP-94, 1994

[30] S. M. Kay, "Fundamentals of Statistical Signal Processing: Estimation Theory", Englewood Cliffs, NJ: Prentice-Hall, 1993

[31]M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996

[32]M. J. F. Gales & P. C. Woodland, " Mean and Variance Adaptation within the MLLR Framework," Computer Speech & Language, Vol. 10, 1996: pp. 249-264

攻读学位期间发表的学术论文

在学期间已发表（包括已接受待发表）的论文，以及已投稿、或已成文打算投稿、或拟成文投稿的论文情况（只填写与学位论文内容相关的部分）：

序号	作者（全体作者，按顺序排列）	题　　目	发表或投稿刊物名称、级别	发表的卷期、年月、页码	相当于论学位的哪一部分文章（章、节）	被索引收录情况
1	洪穗　尹俊勋 刘睿　金连文	基于 HMM 语音合成的汉英跨语种说话人语音转换	2003 年度华南理工大学电子与信息学科研究生学术研讨会	2003 年 11 月 125 页－129 页	第四章 第六章	

致谢

首先，感谢我的导师金连文老师和尹俊勋老师。本论文的完成得到了两位老师悉心的指导、积极鼓励和无微不至的关怀。两位老师渊博的知识、严谨的治学态度和勇于开拓的创新精神使我受益非浅。从他们身上，我不仅学到了丰富的知识，更学到做人的道理。在此谨向两位老师致以最崇高的敬意和最真诚的谢意。

同时感谢电子与信息学院的领导和老师们对我的支持、帮助和关心。感谢谢胜利老师、贺前华老师、冯穗力老师、杜明辉老师、蔡汉添老师、马丽红老师对我的教育和指导。

在这里我还要感谢我们项目小组的成员吕声博士、刘睿同学、黄峰师弟和夏菁师妹，在完成本论文的过程中，他们给了我很大的帮助和支持。还要感谢摩托罗拉中国研究中心黄建成博士在科研上的指导和督促。

同时，我还要感谢实验室的徐睿、龙均宇、沈东升和各位师弟师妹们，感谢他们在这三年中对我的鼓励、支持和帮助。

最后，谨以此文献给一直默默关心和支持我的奶奶、爸爸妈妈、姐姐姐夫、弟弟。他们始终是我战胜各种困难的精神支柱和坚强后盾，我的每一步都凝聚着他们的心血和期望。借此机会，特向他们表示我最忠心的感谢和最深切的挚爱。