东南大学

硕士学位论文

说话人语音转换技术的研究

姓名：宋鹏

申请学位级别：硕士

专业：信号与信息处理

指导教师：赵力

20090112

摘 要

说话人语音转换技术的研究

学生姓名：宋鹏　导师姓名：赵力
东南大学 信息科学与工程学院

说话人语音转换是为了实现合成语音中的说话人的多样性而产生的语音信号处理的一个新兴的方向，说话人语音转换在很多的领域有着广泛的应用，如军事上的说话人伪装、机器翻译、医疗卫生等。目前说话人语音转换的方法有很多，高斯混合模型法是其中效果比较理想的一种方法，本文的主要工作是基于该算法提出改进的转换算法。

本文首先阐述了说话人语音转换的基本概念和现状、语音信号处理的相关基础知识，然后对常用的几种说话人语音转换方法进行了分析比较。提出了改进的基于高使混合模型的语音转换算法。针对高斯混合模型方法中转换语音的频谱过平滑问题，通过假定转换语音和目标语音有相同的协方差，在高斯混合模型法和线性多变量回归法之间设置阈值，很好的解决了这一问题；针对转换后的频谱参数帧与帧之间的不连续的现象，通过考虑帧与帧之间的动态特征参数很好地解决了这一问题。

通过客观和主观测试，证明了改进的算法在很大程度上解决了传统的转换算法存在的相关问题。本文最后对改进的说话人语音转换方法存在的缺陷和不足进行了讨论，并展望了今后下一步的主要工作。

关键词：语音转换；高斯混合模型；线性多变量回归；过平滑；不连续

Abstract

Voice conversion is a new branch of speech signal processing, which aims at realizing variety of synthesized speech. It can be used in many fields such as speaker disguise, machine translation and medical treatment. There are many methods on voice conversion, among which GMM method shows better performance than others'. So the main work is to realize the modified conversion method based on GMM.

Firstly, the concept and development of voice conversion and basic information of speech signal processing are analyzed. Then, several different conversion methods are discussed and compared. A modified voice conversion method based on GMM has been presented, in order to overcome the spectral over-smooth phenomena, the same covariance of the converted speech and target one are assumed, and a threshold between GMM method and LMR method is set. By considering the dynamic features of frame-by-frame, the discontinuity phenomenon between frames is also resolved efficiently.

By objective and perceptual tests, the problems are overcome to a large extent. Finally, the shortcomings of the modified conversion method are discussed, and next work in future is also prospected.

Keywords: Voice conversion; GMM; LMR; Over-smooth; Discontinuity

东南大学学位论文独创性声明

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

研究生签名：宋明鹏 日 期：2009.1.12

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学研究生院办理。

研究生签名：宋明鹏 导师签名：[illegible] 日 期：2009.1.12

第一章 绪论

1.1 说话人语音转换的概念

在日常生活当中，我们知道语音信号包含了很多信息，除了最为重要的语义信息外，还有说话人的个性特征（或者说身份信息）、情感特征、说话人的态度以及说话场景等信息。而语音中的说话人个性特征在现代通信及信息服务中扮演着越来越重要的作用，一个人的声音可以说就是他的“声音名片”，人们通常通过这张“声音名片”来辨别自己周围熟悉的人，人们常说到“未见其人，先闻其声”这些现象等成为说话人语音转换的研究的出发点。

说话人语音转换就是改变一个说话人的语音个性特征（我们称之为源说话人），使其听起来像是另一个说话人（我们称之为目标说话人）的语音个性特征。

说话人信息当中通常包括语义信息和个性信息，说话人语音转换就是保持语义信息不变，仅改变他的个性信息，使得一个说话人的语音听起来像是另外一个说话人说的一样。

1.2 说话人语音转换的意义

最初的语音转换技术来源于语音识别中的说话人自适应技术，作为说话人无关的语音识别系统的前端处理模块，用于降低说话人变化对语音识别系统的影响，提高系统的鲁棒性。有很多的说话人自适应技术仍被广泛的引用于说话人语音转换系统。但目前更多的说话人语音转换技术是应用于语音合成系统，它改变了合成语音个性特点单一的缺陷。不仅如此，说话人语音转换技术也越来越多的应用到其他场合，具有广阔的应用前景。

说话人语音转换作为语音信号处理领域的一个新兴的分支，研究说话人语音转换有着重要的研究价值和应用前景。通过对说话人语音转换的研究，可以进一步加强对语音的相关参数的研究，可以进一步探索人类的发音机制，掌握语音信号的个性特征参数由哪些因素所决定，因而人们可以通过控制这些参数来达到自己的目的。因此对说话人语音转换的研究可以推动语音信号的其它领域如语音识别、语音合成、说话人识别等的发展。具体来说包括以下几个方面[1]：

1. 在TTS（文语合成系统）中，目前高质量的语音合成系统或文语转换系统都是基于语音波形拼接的方法。做一个语音合成系统必须录制较大的音素语音库，在合成语音时，选择适当的音素语音，然后拼接成合成的语音。但是这种方法合成的语音个性特征一般是比较单一，缺乏相应的个性。如果要使合成的语音具有其他人的个性特征，必须重新录制一个语音库，而录制一个语音库是相当耗费时间和存储空间的事情。如果通过在语音合成系统中增加一个说话人语音转换系统，将合成的语音通过说话人语音转换系统或者将合成单元通过一个说话人语音转换系统再进行合成，将其转化为特定人的声音，使单调的语音具有更多的个性特征，满足不同人的应用需要。特别是在当今的多媒体应用时代，人们需求越来越多的多媒体业务。例如手机短信越来越被广大的用户接受，在市场生活中，人们经常利用短信来传递信息，将来可以发展成语音个性短信。在接收端如果短信用发送者的声音读出来，这样具有鲜明个性特征的语音短信势必会吸引到更多的用户，更好地拓展电信业务。

2. 说话人伪装身份通信。在不方便透露说话人身份的情况下，通信系统的前端安装说话人语音转换系统，则可以进行身份保密的语音通信。另外，在法庭上经常需要对控、辩双方提供的一些录音证据进行司法认证，如果说话人语音转换系统能对那些故意伪装了身份的录音恢复出原来的真实身份，这将为司法裁决提供很重要的判决依据，具有良好的社会效益。

3. 在医学领域，用于语音增强系统。对于声带等发音器官存在病变或者损伤的病人，其话音的质量严重受损，对方很难理解，严重影响了正常的交流。说话人语音转换可以用于帮助恢复受损语音，把受损语音变成一个清晰可懂的语音，这将极大地改变这些病人的生活。

4. 在电影配音中，特别是用另一种语音进行配音时，往往不是演员本身，配音演员和原演员的个性特征相差很大，配音效果不理想。对于观众来说，总是希望有该演员的说话效果，如果通过说话人语音转换系统，使之重新具有原演员的特征，增强了电影的效果。这种应用还体现在广播电台中。另外，电脑游戏越来越深入老百姓的生活，特别是网络游戏出现后，现在的游戏都是有声游戏，玩家

在游戏中扮演某个角色，如果该角色的声音能换成玩家自己的声音，则会使增加玩家身临其境的感觉，势必会吸引越来越多的游戏用户，使游戏提供商可以开拓更大的市场。

5. 在语音识别领域，一个非常重要的研究课题就是说话人自适应和说话人归一化的问题。当不同说话人在发同一个音时，由于生理结构的不同，其声学特征参数也是不一样的，如果以待识别的语音特征参数为源语音特征参数，以非特定人语音识别模板为目标特征参数，则采用说话人语音转换技术进行转换，就是说话人归一化，反之，如果以非特定的语音特征参数为源语音特征参数，以特定人语音识别模板为目标特征参数，则可以实现说话人自适应技术。在识别系统的前端，可以将说话人语音转换技术作为语音识别系统的前端模块，提高系统的鲁棒性。

6. 在通信领域，可以用来改变极低速率的编码的语音质量。相关研究表明，当语音的速率低于2.4Kbps 的情况下，解码出来的语音不再具有说话人的个性特征，这样可能使通信双方都感觉不舒服。如果我们将解码出来的语音经过一个说话人语音转换系统，恢复说话者的个性特征，将极大提高通信效果。

7. 用于机器语言翻译系统。机器语言翻译是国际上一个非常热点的研究课题，很多高校、科研机构都报道了他们开发的系统，但这些系统对任何使用者来说，其最后翻译合成出来的声音都没有了源说话人的个性特征信息，非常缺乏现场感，但如果对合成的语音进行转换，重新恢复出说话人的身份特征，具有良好的实际效果。

1.3 国内外研究现状

说话人语音转换是首先提取说话人身份相关的声学特征参数，然后再用改变后的声学特征参数合成出新的接近目标语音的语音。要完成一个说话人语音转换，一般包含两个阶段：训练阶段和转换阶段。

在训练阶段，我们首先提取源说话人和目标说话人的个性特征参数，然后根据某种匹配规则建立源说话人和目标说话人之间的匹配函数。

在转换阶段，利用训练阶段获得的匹配函数，对源说话人的个性特征参数进行转换，最后利用转换后的特征参数合成出接近目标说话人的语音。

说话人语音转换的核心问题就是找出源说话人和目标说话人之间的匹配函数。目前已经形成了很多比较经典的转换算法，尽管思路不同，但是一个完整的说话人语音转换系统一般会考虑以下几个因素：

1. 选择一个理想的分析合成模型，为了获得良好的语音转换效果，必须要建立一个有效的分析合成语音的数学模型。

2. 选择一种较为理想的转换算法，在源说话人和目标说话人的个性特征参数之间建立一个有效的匹配函数，这也是说话人语音转换的核心所在。

3. 选择一种有效的语音特征参数来表征说话人的个性特征。

在过去的近三十年的时间里，说话人语音转换逐渐引起人们的重视，国内外的语音工作者在这方面也做了大量的工作。总体来说，国外的研究比较深入、起步也比较早、取得的成果也比较多，而国内对于说话人语音转换算法的研究起步较晚。目前国内外对于说话人语音转换技术的研究已经取得了较为广泛的成果，对于说话人语音转换的算法研究主要集中在频谱特征参数的转换上。

1988 年 Abe 最早运用矢量量化的方法进行了说话人语音转换的研究[2]，其匹配函数是离散的，它的思想来源于语音识别中的说话人自适应技术，其码本映射的语音转换算法较为简单，容易实现，但是由于矢量量化的方法会引起特征空间的不连续性，使得转换后的语音的效果收到很大的影响；1992 年 Valbret 运用 LMR(线性多变量回归)和 DFW（动态频率调整）的方法进行说话人语音转换的研究[3]，LMR 方法用变换矩阵实现两个子空间之间的线性映射，考虑到人耳的听觉特性，可以在转换中使用感觉加性系数，提高转换后的语音质量，这种转换算法由于对特征矢量进行硬判决，在进行转换时只在一个单独的子空间中进行，丢失了和其它空间的相关信息，因而同样会引起转换后频谱的不连续性；基于 DFW 的说话人语音转换算法主要是将源说话人和目标说话人相对应语音帧的频谱曲线用某个函数进行表示，如果是线性函数则成为线性频率调整，如果是非线性频率调整，则

称为非线性频率调整。由于线性频率调整会丢失高频信息或者存在补零现象，转换的性能受到一定的影响，目前很少使用，现在主要运用非线性频率调整，而且有很多不同的非线性频率函数。频率调整法关键在于寻找一个合适的频率调整函数，从人耳的听觉感官来讲，不同频率的差别很大，因此，在寻找频率映射函数时，应当考虑使用分段、非线性的函数；1995 年 Kuwabara 引入了模糊矢量量化的方法用于说话人语音转换[4]，在一定程度上提高了说话人语音转换的质量；为了解决矢量量化引起的不连续性，Stylianou 引入了 GMM(高斯混合模型)的算法[5]，通过加权求平均的方法在很大程度上解决了特征空间的不连续性的问题，大大提高了转换后的语音质量，这也是目前公认的较为理想的用于说话人语音转换的方法，但是由于转换后的语音参数是加权求平均的结果，所以又会引起 over-smooth（过平滑）问题，同时由于参数的转换是基于每一帧单独进行转换，没有考虑相邻帧之间的关系，使得转换后的语音有一定的不连续性；针对这两个问题，很多人基于 GMM 转换模型进行了改进，2001 年 Toda 运用 GMM 和 DFW 加权的方式进行了说话人语音转换的研究[6]，使得转换后的语音相比传统的基于 GMM 模型的方法有一定的提高。

国内对于说话人语音转换的研究较晚，初敏等人利用 TD-PSOLA(时域基音同步叠加)进行男女声的语音转换研究[7]，基音周期的变换采用 TD-PSOLA 方法，声道响应的特性通过重采样的方法来实现；王聪修对噪音源的特性进行了研究[8]，基于噪音源进行韵律的变换，谱包络的转换通过线性和非线性的频谱搬移方法实现，以此来实现男女声语音之间的转换；2003 年微软亚洲研究院的陈一宁[9]提出了一种新的基于 GMM+MAP 的方法，仅仅将源说话人的特征参数的概率分布转移到目标说话人的特征参数的概率分布上，这样可以避免 over-smooth 的问题，为了解决帧之间的不连续性，通过设置后置的低通滤波器和中值滤波器很好地解决了这一问题；我国学者左国玉[10]提出了一种基于径向基函数神经网络的算法，该算法采用 LSP（线谱对）特征参数和遗传算法，试验结果表明，该算法的性能有所改善，提高了系统的稳定性；2005 年台湾学者 Chung-Hsien Wu[11]提出了将 Bi-HMM（隐马尔科夫）模型用于说话人语音转换，用 HMM 中的状态持续时间来刻画音素的时长信息，并利用 Gamma 函数分布来描述状态持续时间变量。实验结果表明该方法有效地改善了语音韵律特性，特别有利于语音情感信息的控制和转换。

对于说话人语音转换的影响参数，除了反映声道信息的频谱包络以外，语音信号的韵律特征也包含了丰富的反映说话人身份特征的参数。韵律特征参数主要包括基音周期、音素时长等参数。对于韵律的转换方面，由于语音信号的韵律信息具有较大的不稳定性，很难对其建立有效的数学模型进行控制和转换。虽然很多学者在这方面做了很多的工作，但是在在目前的说话人语音转换研究中，往往对基音周期和音素时长进行统计匹配，然后实现语音的平均基音周期和平均音素时长的转换。

对基音周期的转换主要分为在时域和频域的转换算法，主要是首先对语音进行 LPC（线性预测分析）分析，然后对残差信号进行转换。TD-PSOLA 是最常见也最容易实现的一种算法，学者 K.S.Rao[12]提出了一种基于残差信号中强激励脉冲的韵律改变算法，在提取强激励脉冲序列的同时，按一定的转换因子实现基音周期的转换；R.Muralishankar[13]提出了利用 DCT（离散余弦变换）在语音的残差信号中进行转换，它是通过 DCT 对参数信号进行变换，在变换时根据所需要的基音修改因子对 DCT 变换系数进行截取或者补零。然后再对修改后的参数进行 IDCT（离散余弦逆变换）。这样就可以获得基音周期变换后的残差信号；K.S.Lee[14]也是利用残差信号进行基音周期的转换。他是运用 FFT（快速傅立叶变换）对残差信号变换到频域，再在频谱中按照转换因子进行频率轴的拉伸或者压缩，最后通过 IFFT（快速傅立叶逆变换）变换回来；Stylianou[5]通过 HNM（谐波加噪声模型）对语音信号进行建模，然后叠接相加形成新的基音周期。另外还有一些算法并不是直接来修改语音的基音周期、音素时长等韵律信息；Kain[15]用转换后的特征参数来预测激励信号，从而来合成语音；H.Ye[16]在训练阶段保存目标语音的残差信号，语音谱特征参数转换，寻找与其谱距离最小的目标语音，而该目标语音所对应的残差信号用来合成所需要的语音。这些算法在一定程度上对基音周期的转换有所提高，但要使转换后的基音轮廓和目标语音的基音轮廓相接近还有很大的差距。对于能量、音素时长、发音速率等的改变，目前的转换算法都是较为粗糙的转换，都是通过在训练时对它们的平均值进行求取，求出相应的比例因子，然后在合成语音时按比例地增加或者减少来实现。

1.4 课题的研究目标和所要解决的问题

总的来说，说话人语音转换取得了一定的成果，无论是主观测试还是客观测试，转换语音相比源语音都更接近于目标语音，但是说话人语音转换还是一项很不成熟的技术，还有很多的问题需要解决，说话人语音转换的效果还不是很令人满意。主要表现在：说话人语音转换的精确度还不够高，转换后的语音和目标语音还有较大差距；语音经过分析、转换、合成后，语音质量有着较为明显的下降，有的还下降地很严重。

针对这些问题，本论文所要做的主要研究工作包括以下几个方面：

1. 对频谱参数转换方法的研究：通过比较常用的几种转换方法，从复杂度和有效性方面考虑提出一种比较有效的频谱转换算法；

2. 对基音周期转换算法的研究：目前的说话人语音转换算法都是对语音的基音周期的均值进行转换，通过对相关算法的研究，提出一种更加有效的基音周期转换算法；

3. 对频谱包络参数的研究：通过对相关语音特征参数的比较研究，分析在不同情况下各种特征参数的效果的组合，选择一种或者几种特征参数的组合作为频谱包络的特征表示；

4. 对语音分析合成模型的研究：通过对目前常用的语音分析合成模型的研究，找出一种性能更加优良的语音模型。

1.5 论文的组织结构

全文分为六章，论文的章节内容安排如下：

第一章：绪论部分

着重阐述了说话人语音转换的基本概念，以及说话人语音转换的研究意义和国内外研究现状。在本章的最后介绍了论文所要做的主要工作和需要解决的主要问题。

第二章：语音信号处理基本原理

着重阐述与说话人语音转换有关的语音信号处理的基本知识，对语音的发音机理、语音的数学模型、语音的时域和频域特性、语音的相关声学特征参数进行了分析，并对基音频率和共振峰这两种重要的语音特征参数分别进行了讨论，最后重点讨论了语音的两种常用的分析合成模型：共振峰模型和 LPC 模型。

第三章：语音转换方法和语料库的选择

本章第一部分对常用的几种频谱转换方法进行了分析，比较了基于矢量量化的方法、基于 LMR 的方法、基于神经网络的方法、基于 DFW 的方法等基本实现原理和优缺点，然后分析了常用的几种对于基音周期转换的方法，主要包括平均基音周期转换法、高斯模型法、句子码书法等。对基于韵律的转换方法也进行了相应的讨论。

本章的第二部分阐述了用于语音转换的语音库的设计要求和设计方案，分析了语音合成的基本概念，重点讨论了基于 HMM（隐马尔科夫模型）的语音合成系统，详细讨论了 HMM 的相关原理，对 HTK 工具箱也进行了详细的研究，最后给出了基于 HMM 的语音合成方法的详细流程和具体的步骤。

第四章：基于 GMM 模型的语音转换

本章阐述了一个完整的基于 GMM 的语音转换系统的实现，给出了语音转换系统的详细流程和各个部分的具体实现，对 GMM 的训练和转换方法进行了详细的讨论。

第五章：基于改进的 GMM 模型的语音转换

本章给出了论文所提出的基于改进的 GMM 的语音转换方法，针对高斯混合模型中转换语音的频谱过平滑问题，通过假定转换语音和目标语音有相同的协方差，在高斯混合模型和线性多变量回归法之间设置阈值，很好的解决了这一问题；通过考虑帧与帧之间存在的动态特征参数来解决转换后的语音帧之间的不连续问题。

最后总结了改进的基于 GMM 的说话人语音转换系统的实现过程、实验条件和具体步骤。并对改进的说话人语音转换系统从主观和客观方面分别进行了评价。

第六章：总结和展望

在本章中，对于提出的改进的基于 GMM 的转换方法的优点进行了总结，并对存在的不足进行了详细的讨论。最后对说话人语音转换方法未来可能需要解决的一些问题进行了展望。

1.6 本章小结

本章主要阐述了说话人语音转换的基本概念、应用前景，对国内外的研究现状进行了分析，最后对文章的主要章节的内容进行了介绍。

第二章　语音信号处理基本原理

2.1 语音的产生机理和数学模型

2.1.1 语音的产生机理

声音是一种波，能被人耳听到，它的振动频率在20～2000Hz之间。自然界中包括各种各样的声音，如风声、雷声、雨声、机械发出的声音、乐器发出的声音等。而语音是声音的一种，它是由人的发音器官发出的，具有一定的语法和意义的声音[17]。

人类的语音是由人体发音器官在大脑控制下的生理运动产生的。人的发音器官包括：肺、气管、喉（包括声带）、咽、鼻腔和口腔，如图所示，这些器官共同构成一条形状复杂的管道。在发音器官中，肺和器官是整个系统的能源，喉是主要的声音生成机构，而声道则对生成的声音进行调制。

一般把声门以上，经咽喉、口腔的这一管道叫做主声道；经舌和鼻的这一管道成为鼻道；经肺、气管和支气管的管道成为次声门系统。肺部的压力产生气流，流经气管和喉部，从口腔和鼻腔流出，这四种气流产生四种类型的原始声波：送气噪音、摩擦噪音、爆破音和浊音。这些类型的原始声波再经由不同形状的声道改变，不同的声道形状对应着不同的共振峰频率，称为共振峰。这可用来产生不同的声音，成为音素。音素根据发音方式，可以分为元音、鼻音、爆破音、摩擦音、无擦通音等。

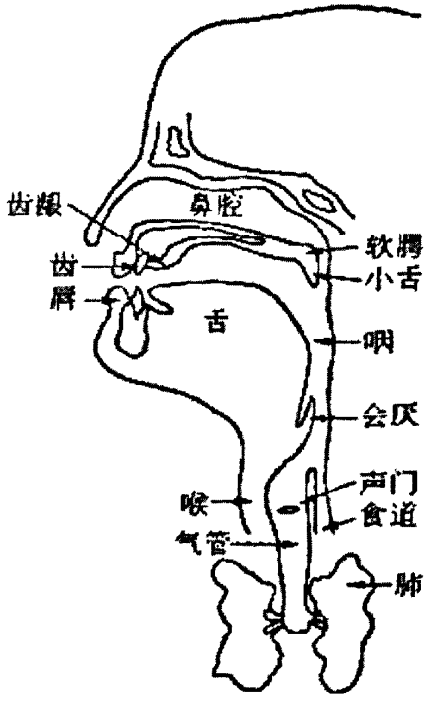


图 2-1 人的发音器官简图

2.1.2 语音的数学模型

通过对发音器官与语音产生机理的分析，可以将语音分成三个部分：在声门（声带）以下，称为“声门子系统”。它负责产生激励振动，是“激励系统”；从声门到嘴唇的呼气通道是声道，是“声道系统”；语音从嘴唇辐射出去，所以嘴唇以外是“辐射系统”。如图所示：

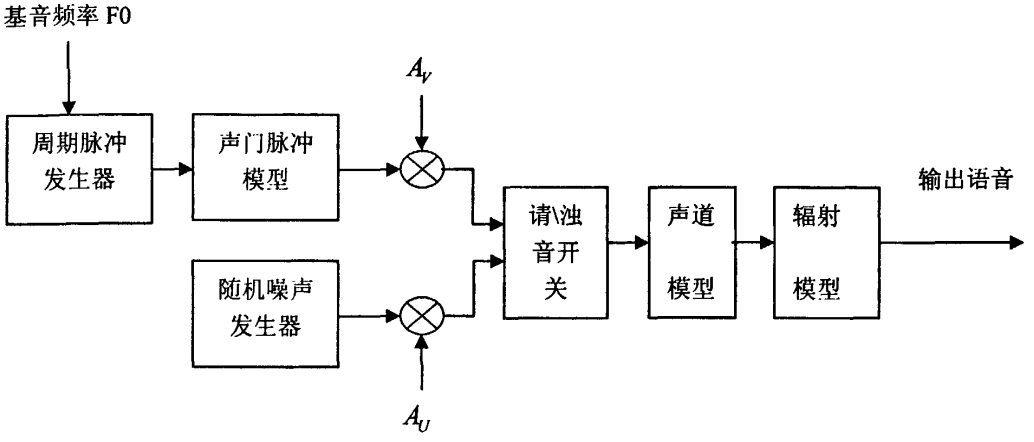


图 2-2 语音的数学模型

传输函数[18]可以用下式来表示：

$$H(z) = AU(z)V(z)R(z) \tag{2.1}$$

其中$U(z)$是激励信号，浊音时$U(z)$为声门脉冲即三角形脉冲序列的z变换；在清音的情况下，$U(z)$是一个随机噪声的z变换。$V(z)$是声道传输函数，即可用声管模式也可用共振峰模型等来描述。A是加权系数。如图 2-2 所示$A_v + A_u = 1$。

2.2 语音信号特性

2.2.1 语音信号的统计特性

语音信号的统计特性可以用它的波形振幅概率密度函数和一些统计量如均值和自相关函数等来描述。通过对语音信号统计特性的研究表明，其幅度分布的概率有两种表达方式：一种是修正的 gamma 分布概率密度函数：

$$f(x) = \frac{\sqrt{k}e^{-k|x|}}{2\sqrt{\pi\sqrt{|x|}}} \tag{2.2}$$

其中 k 为常数，它可由标准差求取

$$k = \frac{\sqrt{3}}{2\sigma_x} \tag{2.3}$$

另外一种稍微差一点的表达方式是拉普拉斯分布函数：

$$f(x) = 0.5\alpha e^{-\alpha|x|} \tag{2.4}$$

其中α为常数，由另一个标准差决定。

$$\alpha = \frac{\sqrt{2}}{\sigma_x} \tag{2.5}$$

另外还有一种表示方式就是用高斯分布函数来近似，但是相比其他两个函数效果要差一些。同时语音信号的振幅通常集中在低电平范围内，同时还应注意到语音信号的强度要经过压缩，而振幅的概率分布不仅反映从一个瞬时到另一个瞬时的取样值的分布，而且还反映语音强度总的变化。

2.2.2 语音信号的时域特性

贯穿于语音分析全过程的是“短时分析技术”。因为语音信号从整体来看，其特性及表征其本质特征的参数均是随时间时间而变化的，所以它是一个非平稳过程。不能用处理平稳信号的数字信号处理技术对其进行分析处理。但是，由于不同的语音是由人的肌肉运动构成声道某种形状而产生的响应。而这种口腔肌肉运动相对于语音频率来说是非常缓慢的，所以从另一方面看，虽然语音信号具有时变特性，但是在某一段较短的时间内（一般 20～30ms 的短时间内），其特性基本保持不变即相对稳定，因而可以将其看作是一个准稳态过程，即语音信号具有短时平稳性。

所以，任何语音信号的分析和处理都必须建立在“短时”的基础之上，即进行“短时分析”，将语音信号分成一段一段来分析其特征参数，其中每一段成为一“帧”，帧长一般为 10~30ms。这样对于整体的语音信号来讲，分析出的是由每一帧特征参数组成的特征参数序列。

虽然短时分析是语音处理的基本方法，但是对于某些要求较高的研究领域或应用场合（如语音识别），应该考虑语音信号是时变的或非平稳的，此时可以采用“隐马尔科夫模型”进行分析。

2.3 语音信号的个性特征参数

语音信号的特征参数根据性质的不同，分为时域和频域特征参数，由于语音信号最重要的感知特性反映在功率谱当中，而相位信号只是起着非常小的作用，所以相比时域特征参数，频域特征参数更重要一些。

语音信号包括多种类型的信息，大致分为三类：第一类为语义信息，即说话的内容；第二类为说话人信息，即是谁说的话；第三类为语音的环境信息。说话人语音转换描述的是说话人信息的转换，与语义信息和环境信息无关。

由于生理结构、生活方式、情感、知识层次的不同，语音会有所差别，人们据此可以区分不同的说话人，表征说话人特征的参数有很多，一般分为音段特征、超音段特征和语言特征[1]：

1. 音段特征信息：音段特征主要包括共振峰频率、共振峰带宽、频谱倾斜、基音周期等。音段特征主要取决于发音器官的生理物理特性，具有一定的稳定性，不宜在短时间内发生改变。

2. 超音段特征信息：超音段信息反映说话人的韵律特征，和人的发音风格紧密联系。例如音素时长和基音轨迹等。这些特征信息主要受到社会和自身心理的影响，具有不稳定性，不容易进行精确的数学建模。

3. 语言特征信息：主要包括用词造句、方言口语等。这受成长环境、教育程度、社会背景等多方面的影响。目前的说话人语音转换系统还没有对高层次的语言模型进行建模，没有考虑说话人的语言习惯。

在过去的几十年里，对于说话人语音转换的研究，主要是对音段特征信息进行转换，对于基音周期、音素时长等超音段信息一般是通过平均值转换以和目标的平均值相匹配，之所以没有对超音段信息进行详细的建模、匹配和转换，主要是由于在现有的语音技术水平下，很难对高层的语音特征进行提取和操作，对于语言特征的转换，在说话人语音转换上几乎没有研究。

对于各种声学特征参数的研究，不同的研究人员得出的结果不尽相同，Matsumoto 研究了男声元音的基音频率、共振峰频率、谱包络等其它一些声学特征参数对个性信息贡献的大小[18]，得出的结论是基音对个性信息的贡献最大，其次是共振峰；Furui 研究报告指出，由倒谱系数平滑得到的平均频谱的贡献最大，特别是 2.5～3.5KHz 这个频段的频谱，其次才是基频[19]；Nakatsui 等研究了三个说话人的元音部分，通过交换他们的激励信号和声道模型，得出的结论是基音对语音的个性特征影响最大；而学者 Itoh 和 Saito 研究后宣布，频谱包络的贡献最大，其次才是基音周期[20]。从这些结果不难看出，决定语音个性特征的参数主要是基频和频谱包络。虽然我们很难确定哪一种特征参数对说话人个性特征的影响最大，但是可以肯定的是，语音的个性特征参数是由多种声学特征参数决定的，不是某个特定参数就可以决定的。

2.3.1 基音频率

基音频率作为重要的说话人个性特征参数，对其进行研究有着重要的意义。对于基音频率进行

描述的声学特征参数主要有：

1. 平均基频。平均基频是大多数基音频率研究的主要的研究对象。语音的基音频率跟说话人的种族、年龄、性别等密切相关。对于基音频率的研究已经取得了很多的成果。一般情况下，随着年龄的增加而降低。男性的基音频率比女性的基音频率要低。Klatt 的研究结果表明：成年女性的基音频率大约是成年男性的 1.7 倍[21]；Childers 的研究结果表明：成年女性的基音频率是男性 1.7-1.8 倍；吕士楠的研究结果表明：女性的基音频率是男性的 1.61 倍[22]。

2. 基音的分布。基音的分布反映了说话人的基音频率的覆盖范围。男性的基音频率的范围大约为 60-200Hz。女性和儿童的基音频率大约为 200-450Hz。

3. 基音的轮廓。汉语是一种一种有调语音，其声调和语调是由基音曲线决定的。

4. 基音的颤动。基音的颤动是指基音的某些不规则变化，在自然语音中广泛存在这一现象。

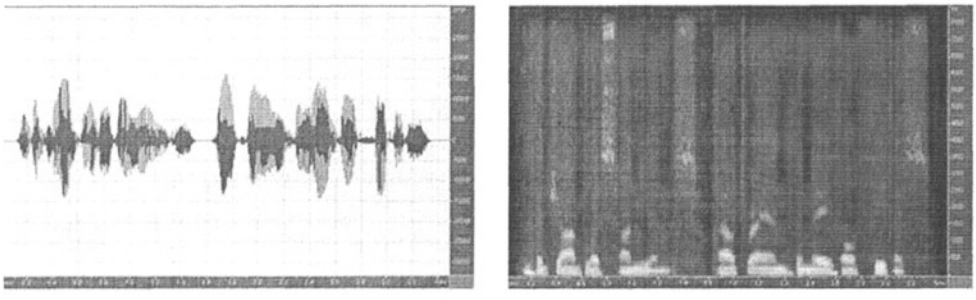
2.3.2 共振峰参数

共振峰是决定元音音色的主要因素，共振峰的位置、带宽、频率等对元音的音色起了决定的作用。一切元音都是以较低层次的两到三个共振峰来代表其主要特性。对于某一个或者某一类的特定的说话人来说，决定共振峰音色的是共振峰的绝对频率而不是相对频率。但是不同的说话人在说同一个元音时，有相同或者相近的共振峰模式，这些共振峰之间有着一定的相对关系。

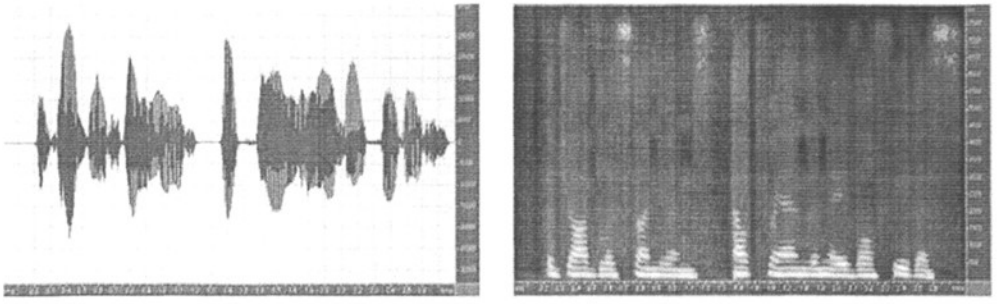
1. 女性的共振峰频率比男性的共振峰频率要高一些，不同的元音的不同共振峰，高出的比例不同。

2. 对于绝大多数元音来说，男性的共振峰带宽比女性的共振峰带宽要窄。

不同年龄、不同性别的说话人，其语音的共振峰位置、带宽、幅度是不一样的，但是不同的说话人发同一个音时，相应的共振峰之间有着一定的关系。在说话人语音转换中，就是要找出这种关系，利用其对源语音进行转换，得到近似目标语音的转换语音。图 2-3 为一男性和女性说相同语句得到的语音波形图和语谱图，我们可以看到语音的基音周期，波形幅度有着明显的不同，共振峰的位置、带宽、幅度等也存在着较大的不同。



(a)



(b)

图 2-3 一男声和女声的相同语音的波形图和语谱图

(a) 男声的语音波形图和语谱图；

(b) 女声的语音波形图和语谱图

2.4 语音模型

语音信号模型是语音信号的数学模型，选择适当的模型可以更加有效地控制语音的特征参数，从而生成高质量的语音，在说话人语音转换中我们用的一般都是源—滤波器模型。

源—滤波器语音合成是基于这样一种声学理论，这种理论认为声音由激励和响应的滤波器形成。其中激励主要分为两种：一种是类似噪声的激励，主要形成非浊音语音信号；另外一种是周期性的激励，主要产生浊音信号。这两种激励也会共同使用，如产生某些浊辅音信号。在这种方式里，语音库中预先存放各种语音合成单元的声道参数，这些参数根据控制规则的要求进行修正，以合成出各种语言环境下的语音。

在源—滤波器的参数合成中，合成器的工作流程主要分为三步：

1. 首先根据待合成音节的声调特性构造出响应的声门波激励源；

2. 然后再根据协同发音、速度变换（时长参数）等音变信息在原始声道的基础上构造出新的声道参数模型。

3. 最后将声门波激励源送入新的声道模型中，其输出就是符合给定韵律特征的合成语音。

共振峰合成和 LPC 合成是上述源—滤波器模型结构的参数合成器中最常用的两种方法。它们实现原理基本上类似，只是所用声道模型不同。同时针对声道模型的特性，在源的选取上略有差别。下面我们将对这两种参数合成器分析。

2.4.1 共振峰合成器模型

共振峰合成器模型将声道视为一个谐振腔，利用腔体的谐振特性如共振峰频率及带宽，以此为参数构成一个共振峰滤波器，因为音色各异的语音有不同的共振峰模式，以每个共振峰频率及带宽为参数，可以构成一个共振峰滤波器。将多个这种滤波器组合起来模拟声道的传输特性，对激励源产生的信号进行调制，经过辐射便可得到合成语音。

由语音产生的模型可知，语音信号谱的谐振特征（对应声道传输函数的极点），完全由声道的形状决定，与激励源的位置无关。语音谱中的反谐振特征（对应声道传输函数的零点）出现在下面两种情况：一是当激励源位置不在喉部（如发摩擦音时），二是发鼻音时。所以对于一般元音，传输函数可以采用全极点模型，对于鼻音和大多数辅音，声道模型应采用极零点模型。

对于全极点传输函数 $V(z)=\dfrac{G}{1-\sum_{i=1}^{p}a_i z^{-i}}$，可以将 $V(z)$ 分解成多个二阶极点的网络的串联，即：

$$V(z)=\prod_{i=1}^{M}\frac{a_i}{1-b_i z^{-1}-c_i z^{-2}} \tag{2.6}$$

由于二阶谐振器的传输函数的参数与其共振峰之间有着简单明确的对应关系，而谐振器串联时各部分的共振峰会保留，所以这种方法可以很方便地模拟全极点滤波器的共振峰特性。而对于零级点模型，则可以用串并联共振峰模型来实现。它们可以模拟谐振和反谐振特征。因而可以被用来合成辅音和鼻音。

实际上共振峰滤波器的个数和组合形式是固定的，只是是共振峰滤波器的参数，随着每一帧输入的语音参数而改变，以此表征音色各异的语音的不同的共振峰模型。共振峰的系统模型如下所示：

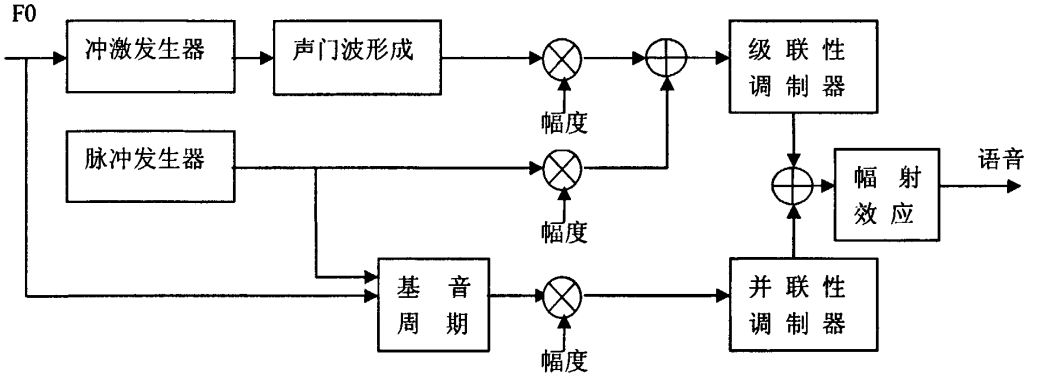


图 2-4 共振峰合成系统模型

上图是共振峰合成的系统模型，我们可以看出激励声源发生的信号，经过模拟声道传输特性的共振峰滤波器的调制，再经过辐射传输效应后即可得到合成的语音输出。

从图中我们可以看出合成共振峰激励源有三种类型：合成浊音语音时用周期激励信号、合成清音信号时用伪随机噪声、合成浊擦音时用周期激励调制的噪声。激励源对合成语音的质量有明显的影响。发浊音时，最简单的是三角波脉冲，但这种模型不够准确，对于高质量的语音合成，激励源的脉冲形状选择是十分重要的，可以采用其他更为精确的形式，如多项式波、滤波成形波等。

对于声道模型，在上图中共使用了两种类型的声道模型：一种是将其模型化为二阶数字谐波器的级联，另一种将其模型化为并联形式。级联性结构可以模拟声道谐振特性，可以很好地逼近元音的频谱特性。这种形式结构简单，每个谐振器代表了一个共振峰特性，只需要用一个参数来控制共振峰的幅度。采用二阶数字滤波器的原因是因为它对单个共振峰特性提供了良好的物理基础，同时在相同的频谱精度上，低阶的数字滤波器量化位数小，在计算上也十分有效。并联型结构能模拟谐振和反谐振特性，可以用来合成辅音。

对于辐射模型，一般取一阶的高通滤波器。由于除了激励脉冲串之外，斜三角波模型是二阶低通，而辐射模型是一阶高通。所以在实际信号分析时，常用所谓“预加重技术”。即在取样之后，插入一个一阶的高通滤波器，这样只剩下声道部分，便于声道参数的分析，在分析转换结束后进行语音合成时，再进行去预加重处理，就可以很好地恢复原来的语音。

2.4.2 LPC模型

LPC 方法是日前较为实用和简单的一种语音合成模型，它以极低数据率、低复杂度、低成本受到特别的重视，LPC 方法可以有效地估计基本语音参数，如基音、谱、声道面积函数等，可以对语音的基本模型给出精确的控制。LPC 语音合成器利用 LPC 语音分析方法，通过分析自然语音样本，计算出 LPC 系数，就可以建立语音信号的产生模型，从而合成出语音。

LPC 模型是一种源—滤波器模型，由白噪声序列和周期脉冲序列构成的激励信号，经过选通、放大并通过时变数字滤波器（由语音参数控制的声道模型），就可以再获得原语音信号。模型如下图所示：

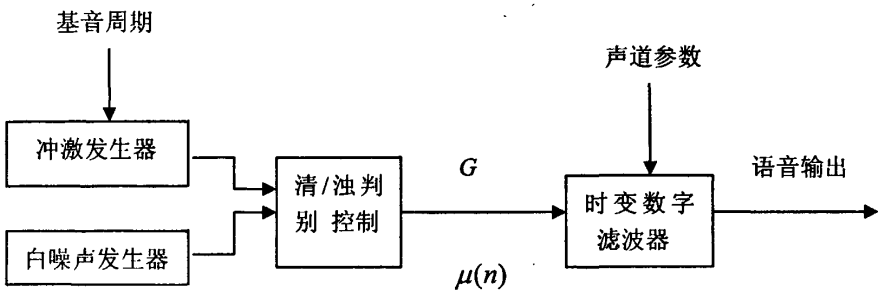


图 2-5 线性预测系统合成模型

如上图所示的线性预测合成的形式有两种：一种是直接用线性预测系数构成的递归合成滤波器，用这种方法定期地改变激励参数和预测其参数，就能合成出语音，这种方式简单直观，为了合成一个语音样本，需要 p 次乘法和 p 次加法；另外一种是采用反射系数构成的格性滤波器，只要知道反射系数、激励位置和模型增益，就可由反向误差序列迭代地计算出合成语音,。需要 $(2p-1)$ 次乘法和 $(2p-1)$ 次加法。它与直接型相比具有一系列的优点：如滤波器稳定、对有限字长的量化效应灵敏度较低等等。下面我们简要介绍一下直接型合成滤波器的实现。

由于语音信号在时域上的相关性，可以考虑用信号 $x(n)$ 过去的 p 个样本来预测当前值 $x(n)$：

$$\hat{x}(n)=\sum_{i=1}^{p}a_i x(n-i) \tag{2.7}$$

对应的线性误差为：

$$e(n)=x(n)-\hat{x}(n) \tag{2.8}$$

然后在最小均方误差准则下，即最小化 $E[e(n)^2]$，即可计算出响应的 LPC 系数 α_i，$i=1,2...p$。由最小均方误差准则，我们可以得到 LPC 系数满足如下公式：

$$\frac{d(E[e(n)^2])}{da_i}=-2E[e(n)x(n-i)],\quad i=1,2...p \tag{2.9}$$

将（2.8）式代入上式可得：

$$\phi(i,0)-\sum_{j=1}^{p}a_j\phi(i,j)=0,\qquad i=1,2,...p \tag{2.10}$$

其中

$$\phi(i,j)=E[x(n-i)x(n-j)] \tag{2.11}$$

通过求解上式，我们就可以得到所有的预测系数 a_i。需要注意的是，由于语音信号的短时平稳性，在实际中一般采用加窗分帧的方式处理。因此求解 $\phi(i,j)$ 分为自相关法和协方差法。除了直接求解外，也可以采用列文森—杜宾法进行求解，与传统方法相比，这种方式在实际中的效率和精确度更高。

反之，如果我们已知$e(n)$，则可以通过传递函数为$1/A(z)$的滤波器，就可以在最小均方误差意义下把$x(n)$恢复出来，实际中的LPC语音合成器，正是利用下式构造声道模型：

$$H(z)=\frac{G}{A(z)}=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{2.12}$$

G为模型增益，只要我们输入一个单位方差的白噪声$e(n)$，就可以恢复出原始语音信号。在实际合成系统中，激励源要根据实际语音的清浊不同来生成，而不是简单的单位方差的白噪声序列。由于激励源在绝大部分时间很小，可以采用均方误差最小准则使$e(n)$逼近实际的激励源。因而从原理上仍是相洽的。

2.5 本章小结

本章阐述了语音信号处理的相关基础知识，分析了语音的发音机理、数学模型、语音的时频特性和相关声学特征参数。对语音信号的两种重要的特征参数：基音频率和共振峰进行了讨论，最后重点分析了语音信号处理的两种重要的分析合成模型：共振峰模型和LPC模型。

第三章　语音转换方法和语料库的选择

对于韵律信息的转换和声音频谱参数的转换是说话人语音转换的最基本的内容，在说话人语音转换方法的选择上，国内外的研究主要集中在频谱参数的转换方法上，本章中我们提到的说话人语音转换方法主要是指基于频谱参数的转换方法，本章将详细阐述这些转换方法，同时，为了保证语音转换算法的效果，设计一个优良的语音库是至关重要的，语音库设计的好坏往往对实验结果有着重大的影响，在本章中将重点讨论语料库的设计原则和方法。

3.1 频谱参数的转换

3.1.1 矢量量化法

Abe[2]在上世纪 80 年代最早用矢量量化的方法进行了不同说话人之间的说话人语音转换的研究，取得了较为理想的效果。主要分为训练阶段和转换阶段两个过程。

训练阶段的过程如下图所示:

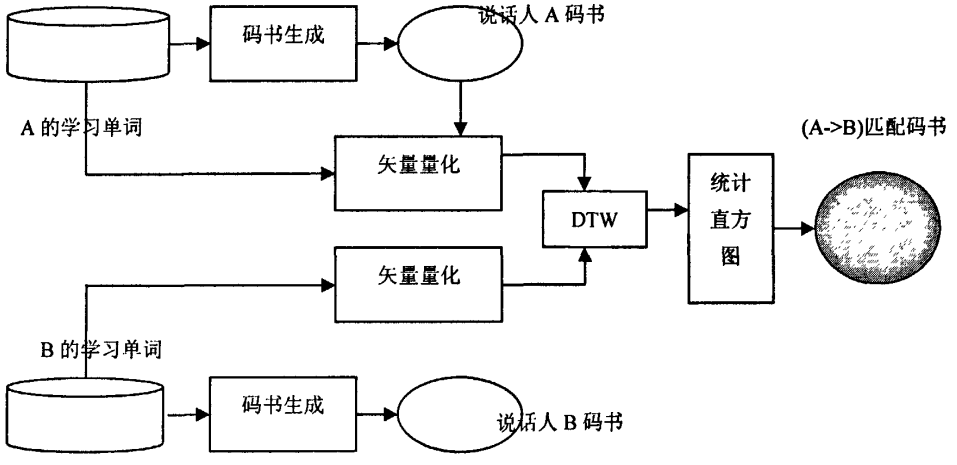


图 3-1 匹配码书的生成

1. 对源语音和目标语音的频谱特征参数空间进行量化，得到具有相同码字数目 M 的码本分别为 V、U；

2. 由源说话人和目标说话人分别产生学习集，然后对所有的单词逐帧进行矢量量化；

3. 运用 DTW（动态时间调整）对两说话人的相同的单词进行对齐；

4. 两说话人之间的矢量量化对应关系累积成柱状图，将柱状图作为加权系数，映射码书即为目标语音矢量的线性合成时的加权系数。

转换阶段的过程如下所示：

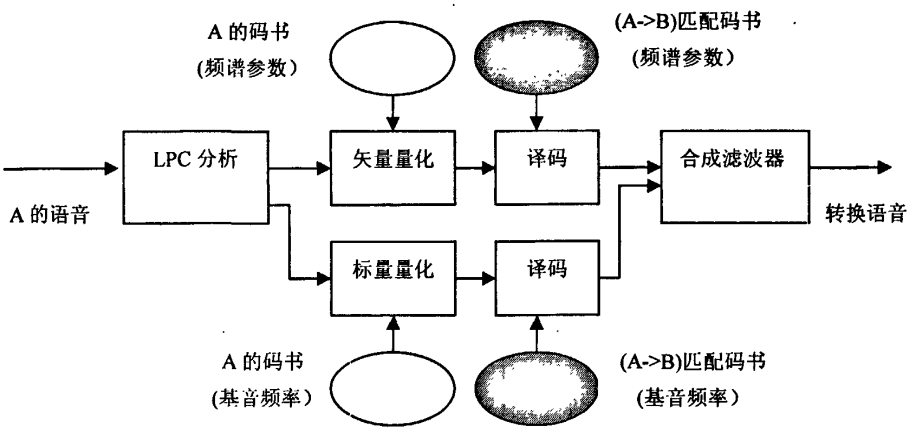


图 3-2 应用匹配码书的语音转换

在转换阶段，先将语音特征矢量进行矢量量化，假设量化成第 l 个码字,则转换后的特征向量为：

$$y_n = \sum_{k=1}^{M} h_{l,k}\mu_l \tag{3.1}$$

其中 h_{lk} 是映射码书 H 中的元素，满足 $\sum_{k=1}^{K} h_{lk} = 1$，$u_k$ 是目标码书 U 的第 k 个码字。

矢量量化的方法在一定程度上实现了不同说话人语音之间的转换，但是由于矢量量化的方法是在每一个特征子空间上进行转换，忽略了相关各个子空间的联系性，会引起转换频谱帧之间的不连续性，使得转换后的语音的效果不是很理想。

3.1.2 线性多变量回归法

90 年代初，Valbret 提出了 LMR（线性多变量回归）的方法[3]，训练时首先对源特征参数和目标特征参数进行归一化，用 DTW 方法将源语音和目标语音的频谱包络特征参数进行对齐，然后应用非监督的分类技术将源说话人和目标说话人的声学空间分成非叠加的子空间，通过在每一个子空间中运用 LMR 算法对源特征参数和目标特征参数建立一个简单的线性关系，可以更好地进行频谱特征的转换。

在训练阶段，转换方程可以用下式表示：

$$\hat{y}_i = A_i x_i \tag{3.2}$$

A_i 的估计可以通过最小平方误差的方法

$$\| \hat{y}_i - y_i \|^2 \tag{3.3}$$

进行求取。

其中 A_i 代表转移矩阵，x_i 代表归一化的源特征矢量，$\hat{y}_i$ 代表归一化的转换后的特征矢量，y_i 代表归一化的目标特征矢量，i 代表第 i 个子空间。

在转换阶段，首先对源特征矢量进行归一化处理，然后对其进行量化归类，确定所用的转移矩阵，再将归一化之后的特征矢量乘以转移矩阵，再对得到的矢量进行解归一化，即得转换后的频谱

特征参数。

3.1.3 神经网络法

学者 Narendranath[23]提出了一种使用神经网络实现说话人语音转换的算法。神经网络共分为四层结构：两个隐层、三个输入单元、三个输出单元。他提取源说话人的前三个共振峰用作输入，其对应的目标说话人的前三个共振峰作为输出，采用含有 8 个神经元的两个中间隐含层，运用 BP 方法进行训练。在转换后合成时，将转换后的共振峰频率和平均基音频率进行合成来得到最终的语音。

学者 Baukoin[24]也采用 BP 神经网络进行类似的实验，他采用两种类型的神经网络：一种神经网络包含两个隐含层，每个隐含层含有 15 个神经元；另一种神经网络包含三个隐含层，每个隐含层含有 12 个神经元，采用的特征参数是倒谱参数。主要步骤如下：

训练阶段：将源语音的谱参数用均值和协方差进行归一化处理，然后进行分类，对于源特征参数和目标特征参数进行动态时间调整，将其分别作为神经网络的输入和输出。训练阶段的优化原则是使转换的倒谱矢量和目标矢量的平均距离最小。

转换阶段：先对源特征矢量进行归一化处理，将归一化后的特征矢量进行归类，再用对应类的神经网络进行转换，再用均值和协方差进行解归一化处理。

我国学者左国玉[10]提出了一种基于径向基函数神经网络的说话人语音转换算法，该算法输入采用的是 LSP 特征参数，神经网络参数的训练采用的是遗传算法，该算法使性能有所改善，增强了系统的稳定性。

3.1.4 多说话人插值法

多说话人插值法是通过对预先存储的多个说话人的频谱包络进行插值得到目标的频谱包络，频谱包络通过慢变化的插值率来进行平滑的转换。在进行插值之前，首先对多个说话人的语音频谱参数序列进行时间对齐，然后再进行下面的转换：

$$\hat{y}_n = \sum_{k=1}^{K} a_k x_k^n + b \tag{3.4}$$

其中 x_k^n 是第 k 个说话人的第 n 帧频谱参数，K 是说话人的个数，a_k 是第 k 个说话人的加权系数，b 是偏移向量，$\hat{y}_n$ 是经插值转换后得到的第 n 帧语音的频谱参数，a_k 和 b 可以通过 LMR 方法或者神经网络的方法计算得到。

多说话人插值法在说话人数量较少时，可以获得比其他方法好的效果，但是当说话人数量足够多时，这种方法就不可取。

3.1.5 高斯混合模型法

Stylianou、Kain 等引入了 GMM(高斯混合模型)的算法。GMM 方法是现在使用最为频繁的说话人语音转换算法，相比矢量量化法会使特征参数离散的特点，GMM 方法通过加权求平均的方法在根本上解决了这一问题。

Stylianou 和 Kain 的方法基本上是相同的，不同点在于，Stylianou 是对源特征参数和目标特征参数分别进行建模，而 Kain 是将源特征参数和目标特征参数进行联合概率密度建模，实验表明[6]这两种方法取得的效果基本上一致。

对于未知矢量的求解，我们便可以建立起源特征参数到目标特征参数之间的匹配关系。Stylianou 采用了最小二乘法来进行估计，而 Kain 采用了更加简单、更易实现地联合概率估计算法。具体的步骤如下：

1. 提取源语音和目标语音的基音周期和频谱特征参数，分别形成两个特征参数空间；

2. 对于源语音频谱特征参数和目标特征参数运用 DTW 的方法进行动态调整，形成一一对应的特征矢量对；

3. 进行 GMM 模型的训练。运用高斯模型对特征参数空间进行建模，在 GMM 模型的每一部分里建立源语音和目标语音的线性关系；

4. 在转换阶段，对于源语音的每一帧特征矢量，计算在 GMM 模型每一部分的转换矩阵并加权求和。

虽然 GMM 方法在很大程度上克服了 VQ 和 LMR 算法的使得转换后的频谱离散的缺陷，但是，GMM 方法仍然存在则较大的缺陷：

1. 由于 GMM 模型是对特征矢量加权求平均进行求取，必然会引起参数的过平滑。

2. 同时由于参数的转换是基于对每一帧单独进行转换，没有考虑相邻帧之间的关系，使得转换后的语音有一定的不连续性。

3.1.6 频率弯折法

频率弯折实际上是一种非线性的频谱搬移，在非特定人语音识别中用来对声道长度进行归一化。这种频谱搬移的方法在不同年龄段的非特定人语音识别的研究中，对降低语音识别错误率起了重要的作用。

从以前的研究中我们可以发现：双线性变换可以很好地模拟人类语音的非线性特征，因此在倒谱域中利用双线性变换进行频率弯折。双线性变换是复平面中的一种映射，将单位圆映射到它本身上。令 z 为原始的复变量，s 和 μ 分别为一次和两次线性变换之后的复变量。 它们之间的关系为

$$s^{-1}=\frac{z^{-1}-a_1}{1-a_1z^{-1}} \tag{3.11}$$

$$z^{-1}=\frac{s^{-1}-a_2}{1-a_2s^{-1}} \tag{3.12}$$

由上述两式可得

$$\mu^{-1}=\frac{\dfrac{z^{-1}-a_1}{1-a_1z^{-1}}-a_2}{1-a_2\dfrac{z^{-1}-a_1}{1-a_1z^{-1}}}=\frac{z^{-1}(1+a_1a_2)-(a_1+a_2)}{(1+a_1a_2)-(a_1+a_2)z^{-1}} \tag{3.13}$$

在双线性变换的两个阶段应用频谱弯折系数 a_1 和 a_2 相当于在双线性变换的一个阶段应用。令 $a=\dfrac{a_1+a_2}{1+a_1a_2}$。由此可得双线性方程的转换公式

$$\hat{z}^{-1}=\frac{z^{-1}-a}{1-az^{-1}},-1<a<1 \tag{3.14}$$

其中 a 为频率弯折系数，通过式中的 $\hat{z}=e^{j\hat{w}}$ 来代替 $z=e^{j\omega}$ 进行频率的转换

$$\hat{\omega}=\theta(\omega)=\arctan[\frac{(1-a^2)\sin\omega}{(1+a^2)\cos\omega-2a}]$$
$$=\omega+\arctan[\frac{a\sin\omega}{1-a\cos\omega}] \tag{3.15}$$

在说话人语音转换中，可以运用上式进行频谱的搬移，达到改变单位声道冲激响应频谱的目的，这种 z 变换具有如下特点：单位圆内的零极点转换后仍落在单位圆内；单位圆外的零极点仍落在单位圆外；单位圆上的零极点仍落在单位圆上。

频谱弯折的特点决定了经过频谱搬移后的稳定性，只要频谱搬移前系统是稳定的，则经过频谱弯折后的系统仍然是稳定的。不同的弯折系数对频谱的弯折作用不同，由式子取弯折系数 a 的不同值所得到的结果如下图所示：

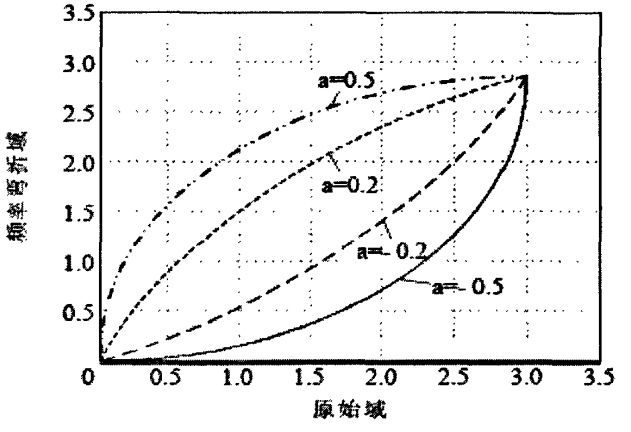


图 3-3 不同 a 值的频谱弯折函数

根据共振峰的变化规律，在语音转换过程中采用频谱搬移的方法，通过对参数的调整控制各个共振峰的搬移幅度和压扩程度，使声道模型得到了转换。

同样，Valbret[1]提出了运用频率规整法进行说话人语音转换，频率规整法是将源说话人和目标说话人相对应语音帧的频谱曲线用某个函数来进行匹配，分为线性函数规整和非线性函数规整。然而线性频率调整存在补零或者丢失高频信息的现象，使得系统的转换性能受到一定的影响，目前已经很少使用，现在用的较多的是非线性频率调整。频率规整算法的步骤如下所示：

1. 用 DTW 算法对源语音和目标语音的特征参数进行时间规整，将相同音素的语音帧一一对齐；
2. 使用预加重函数去除语音帧的频谱斜率；
3. 用循环迭代算法训练最佳的频率规整函数，使得转换后的语音频谱和目标频谱的距离最少；
4. 利用求得的频率规整函数对源语音特征参数进行规整匹配，得到所需的转换语音特征参数，再合成语音。

频率规整法的关键是寻找一个合适的频率映射函数，从人耳的听觉感性来说，不同频段的作用相差很多。因此在寻找频率映射函数时，应考虑使用分段地、非线性地匹配函数。

目前对于说话人语音转换的研究，主要是基于以上几种方法，当然还有其他的转换方法，如基于 HMM（隐马尔科夫模型）的方法的研究、PCA（概率主成份分析）法、基于 Fuzzy VQ（模糊矢量量化）的转换方法的研究等也都取得了不错的转换效果。

3.2 基音周期的转换

基音周期是很重要的说话人特征参数，在说话人语音转换中需要对其进行有效的建模和转换，以便使转换后的语音的基音周期尽可能的接近目标语音的基音周期。下面介绍几种常用的基音周期

的建模和转换方法。

3.2.1 平均基音周期转换法

对基音周期进行转换时，常用的是方法是分别提取源说话人和目标说话人的平均基音周期[25]，分别记为$\overline{p}_s$、$\overline{p}_t$。则平均基音周期转换率α等于目标说话人的平均基音周期除以源说话人的平均基音周期

$$\alpha = \frac{\overline{p}_t}{\overline{p}_s} \tag{3.16}$$

在转换阶段即用源语音的基音周期p_S乘以α即得转换语音的基音周期p_c。

$$p_C = \alpha p_S \tag{3.17}$$

3.2.2 高斯模型转换法

在这种方法中，我们假定源说话人的基音周期和目标说话人的基音周期都服从高斯分布[27],[28]。我们首先获得源说话人和目标说话人基音周期的均值和方差，分别记为$(\mu_s,\sigma_s),(\mu_T,\sigma_T)$。假定转换后语音的基音周期的均值、方差与目标语音的相同，并且转换后语音的基音周期和源说话人的基音周期服从相同的高斯分布。可得：

$$p_c = Ap_s + B \tag{3.18}$$

$$A = \frac{\sigma_t}{\sigma_s},\quad B = \mu_t - A\mu_s \tag{3.19}$$

3.2.3 句子码书模型转换法

Chappell 提出采用建立句子级别的基音周期轮廓码书的方法[29]，这种方法可以直接运用目标语音的基音轮廓。但是由于基音的随意性很大，这种方法必须包含大量的基音轮廓的码书，合成出所有类型的基音轮廓是不可能的，而且对于基音周期轮廓的选择也是非常复杂的。这种方法对于有限词汇量和某些特定的应用效果是十分明显的，因为这时所需的基音周期的轮廓数量有限。

3.3 韵律信息的转换

在表征说话人信息的特征参数中，除了表示声道信息的特征参数外，还包括说话人的韵律信息，它同样能丰富的反映说话人的个人信息，韵律信息包括：说话人的说话时长、能量、基音频率等等，语音的韵律信息具有很大的不稳定性，很难对其进行有效的建模。虽然在这方面做了大量的工作，但是目前的研究中，主要是对基音周期和音素时长进行统计匹配，按照它们的平均值求出相应的比例因子，然后在合成语音时按比例地增加或者减少帧间叠加的样本点数目，或者通过复制或者删除一定的残差信号，实现基音周期平均值和音素时长平均值的转换。

也有一些算法并不是直接修改语音的基音周期和音素时长，而是利用语音库中目标说话人的残差信号来确定转换语音的激励信号，H.Ye[16] 利用训练阶段保存目标说话人的残差信号，语音谱特征参数转换，寻找与其谱距离最小的目标语音，而该目标语音对应的残差信号就用来合成所需要的语音，Kain[15] 利用转换后的谱特征参数来预测激励信号，从而来合成语音。

3.4 语音库的设计

3.4.1 语音库的设计原则

语音库是指以语音波形文件和相应的参数文件组成的数据库。语音库为实验收集必需的语音素材，一个语音库设计的好坏往往对实验结果有着重大的影响。不同的研究领域往往对所需的实验语音素材有着不同的要求。在设计语音库时我们主要从以下四个方面考虑：

1. 语音库的大小

语音库的大小是指语音库中每个说话人的语音数据。在进行录音时，为每个发音者设计多大的文本素材。从目前的说话人语音转换来看，不同的转换算法所需要的语音数据量是不同的，有些需要较多的语音材料，而有些只需要少数几个语句就可以实现较为理想的转换效果。

2. 音素覆盖范围

即文本内容，描述了语音库覆盖整个语音空间的程度，如音素或者双音素，从一个音素到另一个音素的过渡等。对于汉语来说，由于汉字的发音是声母—韵母结构，并且语音的能量和信息主要集中在韵母身上，一个汉字就是一个音节，是一种单音节的有调语音。在设计语音库时，要尽可能地考虑大部分的汉语音节，至少有大部分的韵母，并且还要考虑到同一语音的不同声调。另外，为了能更好地实现韵律特性的转换，语音库中还应该有单个的词、词组和句子等，即既要有单个词的语音，也要有连续的语音。

3. 说话人数量

语音库中的说话人只是整个说话人中的很小的一部分，测试大量源—目标说话人的说话人语音转换系统更能表明说话人语音转换的一般性能，在现在报道的说话人语音转换系统中，所用的语音库包含的说话人数量最多六个，最少两个，包含的说话人越多，越有利于评估语音转换系统。

4. 时间对齐

在说话人语音转换研究中，为了训练一个恰当的特征参数匹配函数，往往在训练阶段需要对源语音特征参数序列和目标语音特征参数序列进行时间对齐。为了减少在训练阶段的匹配误差，在进行语音库的录制时，每个说话人应当尽量以相同的发音速率、节奏、语气、情绪等方式发音，这样可以提高说话人语音转换系统的性能。

为了进行有效地说话人语音转换，训练的语料库一般比较复杂，语料的选择也比较严谨，下面简要回顾一下相关文献报道中的用于说话人语音转换的语料库：

Kain[30]的训练语料是从 TIMIT 和 Harvard Psychoacoustic Sentence 中的数据库中选出 1170 个句子，用 greedy search 算法进行音素标注，选择标准是使稀少的音素出现次数尽可能多，而且包含尽可能多的双音素。最后选出 50 个句子，通过最后的选择，每个音素出现的平均次数为 41 次，最少为 17 次。

Stylianou[31]指出，满足全矩阵和对角形矩阵类型的转换函数的性能，至少需要 3.3 分钟（约 22000 个频谱特征矢量）的对齐的训练数据。

Aslan[32]的训练数据包含三个说话人的语音，每个说话人包括 10 分钟左右的训练语音和 5 分钟左右的测试语音。

Narendranath[23]的说话人语音转换系统中，训练语料来自五男五女组成的说话人的录音，每个录音内容为五个元音/a/、/i/、/u/、/e/、/o/。

另外在其他文献中，Kain[33]和 Stylianou[34]用于说话人语音转换的训练语料都是从 TTS 数据库中训练得到的。

从各方面考虑，在实验室中录制一个标准的语料库比较困难，鉴于此，本次论文选取的是基于文本合成之后的语音作为实验的语料库。

3.4.2 基于HMM的语音合成

由人工制作出语音称为语音合成，语音合成是人机语音通信的一个重要组成部分，语音合成赋予机器“人工嘴巴”的功能，它解决的是如何让机器像人那样说话的问题。语音合成的目的是制造

一种会说话的机器，使一些以其他方式表示或存储的信息能转换为语音，使人们能够通过听觉方便地获取这些信息。

语音合成的研究已有多年的历史，现在研究出的语音合成方法的分类，从技术方式讲可以分为波形合成法、参数合成法和规则合成方法，从合成策略上讲可以分为频谱逼近和波形逼近。

无论是波形编码合成还是参数合成，其原理都等同于语音通信中的波形编码器和声码器中的接收端的工作过程。区别只是在于其是以分析或者变换得到的存储在语音库中的参数或码序列作为合成数据来合成语音。

相关论文与实验表明，基于 HMM 的语音合成系统能够合成较为理想的语音。因此本文对于语料库的选择运用的是基于 HMM 的语音合成方法。HMM 的应用需要一套有效的工具，80 年代末剑桥大学最早开发出 HMM 应用的开发包，不仅功能强大，而且简单易用，使得 HMM 逐渐得到广泛的应用。在下面的讨论中，我们将对基于 HMM 的语音合成系统进行详细的介绍

基于HMM的语音合成系统是由日本名古屋理工大学的Tokuda等人提出的[35]，在这个系统当中，语音的频谱包络（声道）、基音周期（声源）和时长（韵律）通过 HMM 的方法来进行建模，语音波形通过最大似然准则由 HMM 模型自身来产生。基于 HMM 的语音合成系统的系统原理图如下所示，包含训练和合成两个过程：

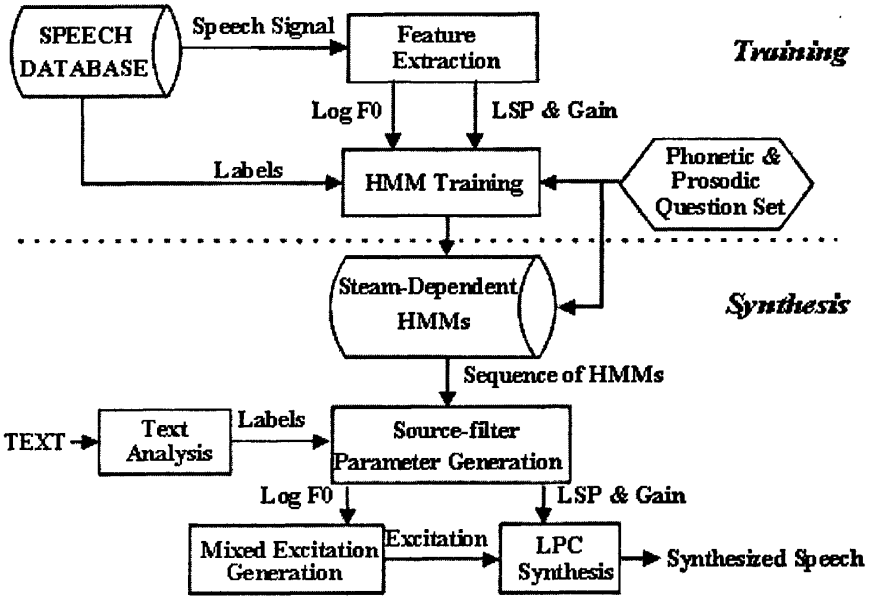


图 3-4 基于 HMM 的语音合成

在训练阶段，语音信号通过特征提取模块来获取特征矢量序列，特征矢量序列通过 HMM 模型进行建模，特征矢量包括激励参数和频谱参数，频谱参数由线谱对（LSP）参数和对数增益组成，而激励参数由对数基音频率组成。它们被分成不同的参数流进行建模，LSP 参数通过连续 HMM 进行建模，而对数基音频率由多空间似然分布 HMM（MSD-HMM）进行建模。通过 MSD 的方法可以对基音频率进行有效的建模，不需要任何的假设和插补。上下文相关的音素模型被用来解决音素间的协同发音现象。根据决策树和最小描述长度准则获取的状态被用来解决训练时的数据稀疏问题，参数流独立的模型用来将频谱、韵律和时长特征进行聚类生成不同的决策树。

在合成阶段，输入文本首先通过文本分析转换成一系列带有上下文属性的标签，通过决策树获得相应的带有上下文属性的 HMM 模型，通过基于最大似然准则的参数生成算法来获取带有动态特征参数和全局协方差信息的 LSP、增益和基频轨迹，最终语音波形通过 LPC 合成所产生的频谱参数和激励参数合成得到。

基于 HMM 的语音合成是在 HTK（HMM based toolkit）工具的基础上开发得到的，下面一节我

们将对 HMM 和 HTK 工具做一详细的说明。

大概 90 年前，人们就已经知道马尔科夫链了，有关 HMM 的基本理论[36],[37]早在 20 世纪 60 年代末被提出并加以研究，它在语音中的应用和实现的研究工作，在 70 年代中期开展开来，然而对它的理论的广泛和深入的了解，它在语音处理中的成功应用，还是最近一二十年的事。作为语音信号处理领域的一项重要进展，它在语音识别和合成领域已经取得了很大的进展。

3.4.2.1 HMM模型

为了说明隐马尔科夫模型的基本概念和原理，我们从离散时域有限状态机开始分析[38]，离散时域有限状态机是一个简单的马尔科夫模型。在每一个离散时刻 t，它处于状态 s_n 中的一种，记为 q_t。设此自动机开始运行的时间起点为 t=1，那么在以后每一时刻 t 它所处的状态以概率方式取决于初始状态概率矢量 π 和状态转移矩阵 A。π 的每一个分量 π_n 表示 $q_1 = s_n$ 的概率。即

$$\pi_n = P(q_1 = s_n), n = 1 \sim N \tag{3.20}$$

矩阵 A 的每一个元素 $a_{i,j}$ 表示已知相邻两个时刻中前一个时刻的状态为 s_i 的条件下，后一个时刻为 s_j 的概率。表示如下：

$$a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i), t \geq 1 \tag{3.21}$$

可以看到，对于任何时间 t，自动机的状态 q_t 取 $s_1 \sim s_N$ 中哪一种的概率只取决于前一个时刻 t-1 所处的状态，而与更前的任何时刻所取的状态无关。由此产生的状态序列 $q_1, q_2, ..., q_T$ 是一条一阶马尔科夫链。对于任何一个特定的 $Q = q_1, q_2, ..., q_T$，其出现的概率为：

$$P(X) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} ... a_{q_{T-1} q_T} \tag{3.22}$$

HMM 类似于一阶马尔科夫过程，不同之处在于 HMM 是一个双内嵌式随机过程，HMM 是由两个随机过程组成：一个是状态转移序列，它对应着一个单纯的马尔科夫过程；另外一个是状态转移时的符号组成的符号序列。这两个随机过程其中一个过程是不可观测的，只能通过另外一个随机过程的输出观察序列观测，设状态转移序列为 $S = S_1, S_2, ..., S_T$，输出的符号序列为 $O = O_1, O_2, ..., O_T$。则在单纯马尔科夫过程和相邻符号之间是不相关的情况(即 S_{i-1} 和 S_i 之间转移时的输出观察值 O_i 和其他转移之间无关)，则下式成立：

$$P(S) = \prod_i P(S_i \mid S_1^{i-1}) = \prod_i P(S_i \mid S_{i-1}) \tag{3.23}$$

$$P(O \mid S) = \prod_i P(O_i \mid S_1^i) = \prod_i P(O_i \mid S_{i-1}, S_i) \tag{3.24}$$

因为是 HMM，把所有可能的状态转移序列都考虑进去，则有：

$$P(O) = \sum_S P(O \mid S) P(S) = \sum_S \prod_i P(S_i \mid S_{i-1}) P(O_i \mid S_{i-1}, S_i) \tag{3.25}$$

为了更好地理解隐马尔科夫的含义，我们来看一个著名的说明 HMM 概念的例子—球和缸(Bal and Um)的实验[39]。

设有 N 个缸，每个缸中装有很多彩色的球，在同一个缸中不同颜色球的多少由一组概率分布来描述。实验是这样进行的，根据初始概率分布，随机的选择 N 个缸中的一个缸，假设第 i 个缸。再根据这个缸中颜色的概率分布，随机的选择一个球，记下球的颜色，记为 O_1，再把球放回缸中。又根据描述缸的转移的概率分布，选择下一个缸。假设第 j 个缸。再从缸中随机选择一个球，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。如果缸中只装有一种彩色的球，则根据球的颜色的序列 $O_1,O_2,...,$ 就可以知道缸的排列。但球的颜色跟缸之间不是一一对应的，所以缸之间的转移以及每次选择的缸被隐藏起来了，并不能直接观察到。而且从每个缸中选择什么颜色的球是由彩球颜色概率分布随机决定的。因此，每次选择哪个缸是由一组转移概率所决定。

由以上分析可知，HMM 用于语音信号建模时，是对语音信号的时间序列结构建立统计模型，它是数学上的双重随机过程：一个是具有有限状态数的马尔科夫链来模拟语音信号统计特性变化的隐含的随机过程，另一个是与马尔科夫链的每一状态相关联的观测序列的过程。

前者通过后者表现出来，但前者的具体参数（如状态序列）是不可测的，人的言语过程实际上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，是由大脑根据语法结构和言语需要（不可观测的状态）发出的音素的参数流。可见，HMM 合理地模仿了这一过程，很好的描述了语音的整体非平稳性和局部平稳性，是一种较为理想的语音信号模型。

欲使所建立的 HMM 模型对于实际应用有效，有三个问题需要解决[40]：

1. 对于一个给定的序列 O 和模型 λ，模型可能建立的序列的概率是什么，可以通过前向—后向算法来进行求解。

2. 对于一个给定的序列 O 和模型 λ，什么样的最可能的状态序列可以创建输出序列，Veterbi 算法可以解决此问题。

3. 给定一个输出序列和拓扑结构，怎样调整模型参数，包括状态转移和输出的概率分布，使模型创建的输出序列具有最大概率。在实际 HMM 训练中，EM（期望最大）算法和 Baum-Welch 算法（最大似然准则）是最常用的。

3.4.2.2 HTK开发包

80年代HMM技术已经非常成熟，在语音信号处理中也有了较为成熟的应用，但是由于缺乏一套有效的工具，限制了其进一步的应用。1989年，剑桥大学的Steve Young[41]等人开发出一套HMM的应用开发包——HTK 1.0，它由一系列的C函数和应用模块构成。最初被专门用作语音识别的研究开发。但自它发布以后，很快就得到广泛的重视，并且变成了一个通用的开发工具包。它也被广泛的应用于其他方面，例如语音合成，字符识别和DNA排序等。

从 2000 年开始，HTK 工具的源代码可以从 CUED 的网站上免费获取，此举是为了进一步发展语音识别和完善 HTK。HTK 工具历经十几年的发展，目前已经推出了 HTK 3.4 版本，随着 HTK 源代码和工具的免费发布，HTK 工具得到了不断的完善和提高。虽然微软保留了对 HTK 源代码的版权，但是他鼓励每一个 HTK 的使用者去修改和完善源代码。共同致力于 HTK 工具包和语音识别的发展。

HTK 的许多功能被建立和封装为一系列的函数模块，这些模块可以通过相同的接口方式和外界进行交互。下图说明了 HTK 的软件结构和输入输出接口。

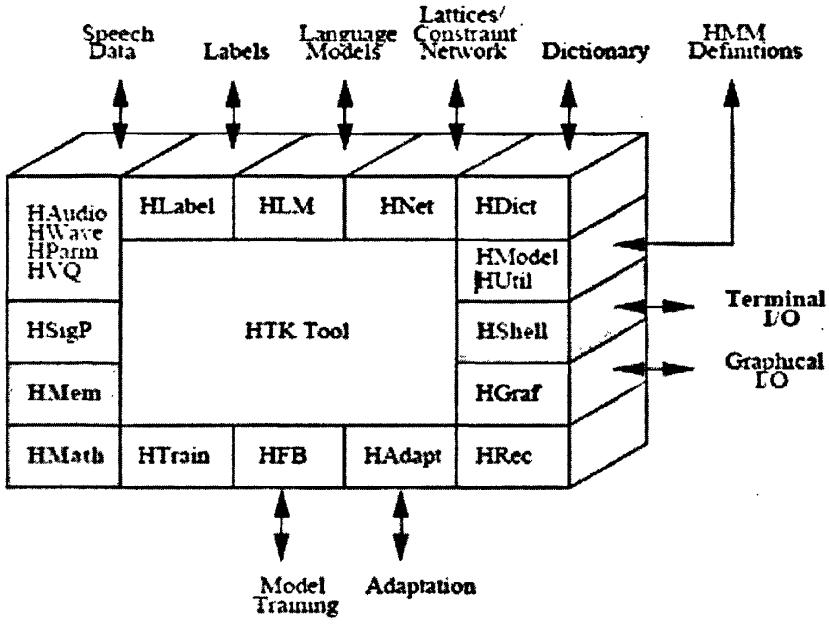


图 3-5 HTK 工具箱

上图各主要函数模块的功能如下所示：

HShell 控制用户的输入输出和与操作系统的接口。

HMem 负责内存的管理。

Hmath 用于数学函数的支持。

HSigP 用于进行语音分析所需的处理操作。

HLabel 用于提供标签文件的接口。

HLM 用于语言模型的建立。

HNet 用于语法网络的建立。

HDict 用于词汇的发音词典的建立。

HVQ 用于矢量量化码本的建立。

HModel 负责 HMM 模型的定义和建立。

HWave 针对所有语音数据的输入输出，提供各种波形数据的格式支持。

HParm 提供几种特征参数文件的格式支持。

HUtil 提供许多有关 HMM 模型的应用例程。

HTrain 和 HFB 提供对多种 HTK 训练工具的支持。

HAadpt 为 HTK 的自适应提供支持。

HRec 包含了 HTK 中主要的识别处理函数。

从孤立词识别来看，HTK 工具的实现过程主要包括四个阶段：数据准备、数据训练、数据测试和数据分析。如下图所示：

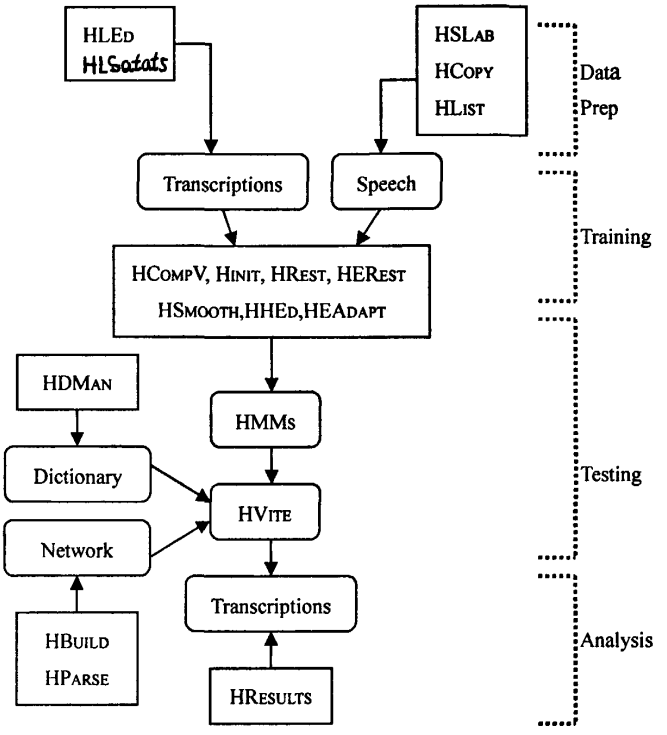


图 3-6 HTK 处理过程

1. 数据准备工具

为了建立一套 HMM 系统，需要准备一组语音数据文件和相关的标签文件。同时语音数据文件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转换为 HTK 支持的文件格式。相应的工具是 HCOPY 和 HLED。HCOPY 用来将语音波形数据转化成我们需要的特征参数，HLED 用来建立实验所需的各种语音标签文件，HLSATATS 用来获取并显示标签文件的一系列统计特性，HQUANT 可以为我们建立一个离散的 HMM 系统。

2. 训练工具

语音识别系统的第二步是为每一个 HMM 模型建立拓扑机构，并利用已有的数据对 HMM 模型进行训练。有两种训练方式：当训练用的语音数据中的每个发音单元（词或音素）的边界确定时，可以用 HINIT 和 HREST 进行训练，即孤立词方式的训练。如果训练用的语音数据没有任何的边界信息，我们采用 HEREST 提供的嵌入式的整体训练方法。

3. 识别工具

HTK 提供唯一的识别工具 HVITE，它采用符号传递算法来完成 Viterbi 搜索的语音识别。HVITE 的使用需要提供两个输入参数：一个是允许的发音词的语法网络，一个是为所由此建立的一本发音词典。发音词的语法网络可以通过 HBUILT 或 HPARSE 来建立，发音字典由 HDMAN 来完成。

4. 分析工具

HMM 的识别系统建立以后，下一步的任务就是如何对系统的性能进行评价。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、词的错误删除率等。

基于 HTS 2.1 语音合成系统，我们对一男声语音数据和一女声语音数据分别进行建模，选取的训练用的数据库大小为 10000 句的平均长度约为 5ms 的英文语音和对应的文本信息。这里我们选取的特征参数为 41 阶的 LSP，通过主观测试，我们可以观察到通过 HTS 系统合成出的语音具有很高的自然度，在很大程度上接近于原始自然语音。

3.5 本章小结

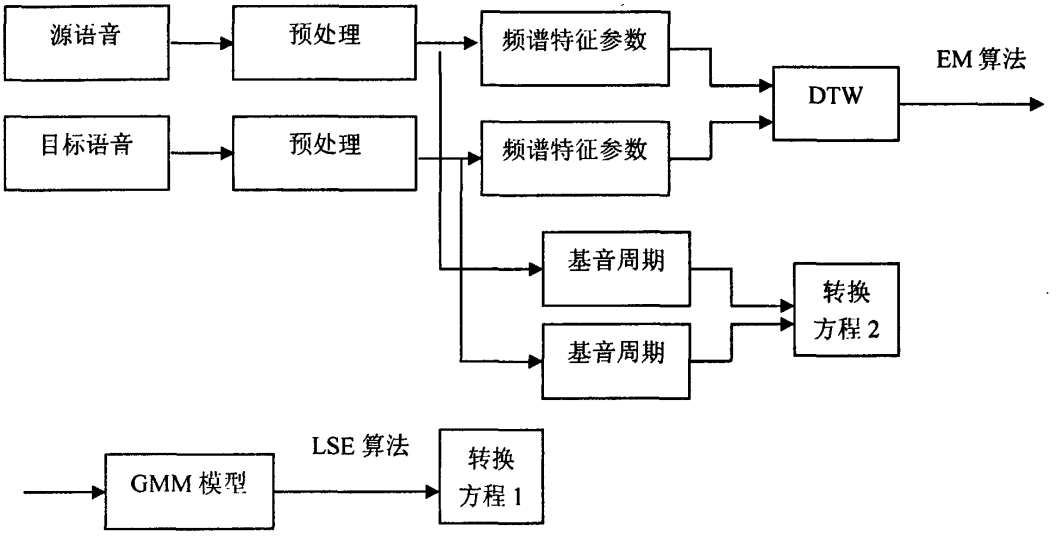
本章首先分析比较了语音转换的几种常用的方法，对常用的频谱转换方法进行了讨论，对基音周期和韵律的转换方法也进行了阐述。接着重点讨论了说话人语音转换的语音库的设计，对 HMM 和 HTK 工具箱进行了详细的研究，最后对基于 HMM 的语音合成方法进行了详细的阐述。

第四章 基于GMM模型的语音转换

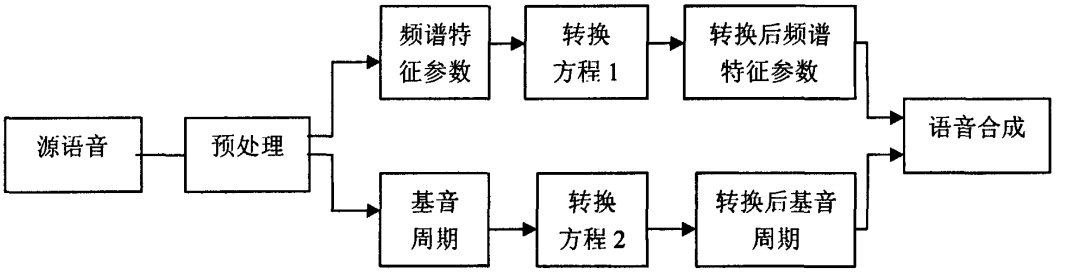
4.1 说话人语音转换系统

说话人语音转换系统是通过改变说话人的声学特征参数来改变说话人的个性特征。一般分为两个阶段：训练阶段和转换阶段。训练阶段通过获取说话人的特征参数，对特征参数空间进行训练，得到两个特征参数之间的匹配关系；在转换阶段利用训练得到的匹配规则，将源说话人的特征参数映射成目标说话人的特征参数，再利用这些特征参数来合成近似于目标说话人个性特征的语音。

说话人语音转换系统的实现分为训练和转换两个阶段，流程图如下所示：



(a)



(b)

图 4-1 语音转换的流程图

(a) 训练过程; (b) 转换过程

训练阶段：

1. 进行特征参数的提取，包括基音周期和频谱特征参数的提取。

2. 对源说话人和目标说话人的混合频谱特征参数运用 DTW 方法进行对齐。

3. 对经过时间对齐后的混合特征参数进行 GMM 模型的建模，运用 LBG 方法和 EM 方法训练出一个有效的 GMM 模型。分别获得源说话人和目标说话人特征参数的均值、方差和混合协方差，获得频谱转换函数。

4. 分别对源说话人和目标说话人的平均基音周期F0，求出平均基音周期转换率α。

转换阶段：

1. 运用训练获得的频谱转换函数进行频谱特征参数的转换，运用基音周期转换函数对基音周进进行转换。

2. 对转换后的频谱特征参数和基音周期运用语音分析合成模型进行合成。

4.2 说话人特征参数的选择

在说话人语音转换的研究中，对于说话人频谱特征参数的选择同样至关重要。一般来说，从语音信号中提取的说话人特征参数应满足以下准则：

1. 对局外变量（如说话人的健康状况和情绪、系统的传输特性等）不敏感。
2. 能够长期地保持稳定。
3. 可以经常表现出来。
4. 易于对之进行测量。
5. 与其他特征不相关。

下面我们详细分析一下在语音识别、语音合成中常用的几种特征参数。

4.2.1 LSP参数

LSP（线谱对）在语音合成和语音编码中有着广泛的应用[42]，根据LPC滤波器的$A(z)$，我们可以构造如下的两组方程。

$$P(z)=A(z)+z^{-(p+1)}A(z^{-1}) \tag{4.1}$$

$$Q(z)=A(z)-z^{-(p+1)}A(z^{-1}) \tag{4.2}$$

由上边两式可以直接得出

$$A(z)=\frac{1}{2}(P(z)+Q(z)) \tag{4.3}$$

如果知道了$P(z)=0$和$Q(z)=0$的值后，我们可以很方便的求取$A(z)$。不难看出，$P(z)$和$Q(z)$的零点都在单位圆上，并且沿着单位圆随ω的增加而交替出现。设$P(z)$和$Q(z)$的零点分别为$e^{j\omega_i}$，$e^{j\theta_i}$。则P(z)和Q(z)可写成下列因式的分解形式：

$$P(z)=(1+z^{-1})\prod_{i=1}^{p/2}(1-2\cos\omega_i z^{-1}+z^{-2}) \tag{4.4}$$

$$Q(z)=(1+z^{-1})\prod_{i=1}^{p/2}(1-2\cos\theta_i z^{-1}+z^{-2}) \tag{4.5}$$

并且ω_i和θ_i按下列关系排列：

$$0<\omega_1<\theta_1<\ldots<\omega_{\frac{p}{2}}<\theta_{\frac{p}{2}}<\pi$$

因为因式分解中的ω_i和θ_i都是成对出现的，反映了谱的特性，故称为“线谱对”。而且可以保

证 $P(z)$ 和 $Q(z)$ 的零点相互分离，是保证合成器稳定的充分必要条件之一。

LSP 得到如此广泛的应用，主要是由于它具有以下特性：

1. 敏感性，其一阶 LSP 参数的量化误差只会影响其对应频率附近的频谱，对较远处的频谱基本上没有影响。

2. 线性内插型。

3. 高效性，即 LSP 参数对应的频谱失真较小。

4. 稳定性，只要 LSP 随阶次增高而增大的次序不发生改变，其对应的 LPC 合成器的稳定性就能得到保证。

4.2.2 LPC复倒谱

LPC 系数是线性预测的基本参数，可以把这些系数转变为其他参数，以得到语音的其他替代方法。LPC 系数可以表示整个 LPC 系统冲激响应的复倒谱。

设通过线性预测分析得到的声道传输函数为：

$$H(z)=\frac{1}{1+\sum_{i=1}^{p}a_i z^{-i}} \tag{4.6}$$

其冲激响应为 $h(n)$，设 $\hat{h}(n)$ 表示 $h(n)$ 的复倒谱，则有：

$$\hat{H}(z)=\ln H(z)=\sum_{n=1}^{\infty}\hat{h}(n)z^{-n} \tag{4.7}$$

将(4.6)式带入并将其两边对 z^{-1} 求导，则有：

$$(1+\sum_{i=1}^{p}a_i z^{-i})\sum_{n=1}^{\infty}n\hat{h}(n)z^{-n+1}=-\sum_{n=1}^{p}ia_i z^{-i+1} \tag{4.8}$$

令上式左右两边的常数项和各次幂的系数 z^{-i} 分别相等，从而可由 a_i 求取 $\hat{h}(n)$：

$$\hat{h}(0)=0$$

$$\hat{h}(1)=-a_1$$

$$\hat{h}(n)=-a_n-\sum_{i=1}^{n-1}(1-k/n)a_k\hat{h}(n-k)\quad(1\le n\le p)$$

$$\hat{h}(n)=-\sum_{i=1}^{n-1}(1-k/n)a_k\hat{h}(n-k)\quad(n\ge p) \tag{4.9}$$

LPC 倒谱参数的优点：

1. 利用了线性预测传输函数 $H(z)$ 的最小相位特性，避免了相位卷绕问题。

2. LPC 倒谱的计算量小，它仅是 FFT 求复倒谱的计算量的一般。

3. 当 $p\to\infty$ 时，语音信号的复倒谱 $S(e^{j\omega})$ 满足 $|S(e^{j\omega})|=|H(e^{j\omega})|$，因而可以认为 $H(e^{j\omega})$ 包含了语音信号的频谱包络信息，可以近似把 $\hat{h}(n)$ 当作 $s(n)$ 的复倒谱 $\hat{s}(n)$，来分别估计出语音短时

谱包络和声门激励参数。

4.2.3 MFCC参数

对于语音参数的选择我们选取的是MFCC（Mel Frequency Cepstral Coefficient）参数，MFCC的分析更符合人耳的听觉特性。所谓Mel频率尺度，它的值大体上对应实际频率的对数关系。关系式可用下式表示：

$$Mel(f)=1127\ln(1+f/700) \tag{4.10}$$

在1000Hz以下，大致呈线性分布，带宽为100Hz左右，在1000Hz以上呈对数增长。类似于临近频带的划分，可以将语音频率划分成一系列三角形的滤波器序列，即Mel滤波器组。取每个三角形的滤波器频带内所有信号幅度加权和作为某个带通滤波器的输出，然后对所有滤波器输出做对数运算，再进一步做DCT变换即得到MFCC。

4.2.4 LPC美尔倒谱参数

在计算出复倒谱$\hat{h}(n)$后，由公式$C(n)=[\hat{h}(n)+\hat{h}(-n)]/2$即可求出倒谱$C(n)$，但是这个倒谱$C(n)$是实际频率尺度的倒谱系数，根据人耳的听觉特性可以把上述的倒谱系数进一步按符合人耳听觉特性的*Mel*尺度进行非线性变换，从而可以求出如下形式的LPC美尔倒谱系数（LPC-MCC）。

$$MC_k(n)=\begin{cases}\alpha+MC_0(n+1) & k=0\\ (1-\alpha^2)MC_0(n+1)+\alpha MC_1(n+1) & k=1\\ MC_{k-1}(n+1)+\alpha(MC_k(n+1)-MC_{k-1}(n)) & k>1\end{cases} \tag{4.11}$$

式中C_n表示倒谱系数，MC_k为美尔倒谱系数，n为迭代次数，k为倒谱阶数。取$n=k$，迭代次数从高到低，即n从大到0取值，当抽样频率分别为8KHz、16KHz时，取α=0.35，这样可以近似于美尔尺度。

4.3 频谱转换

通过对各种说话人语音转换方法的比较，我们可以发现GMM转换方法应该是当前最为有效的频谱转换方法。

按照Kain的观点，假设源特征矢量和目标特征矢量符合联合高斯概率分布，利用高斯混合模型对混合频谱特征参数进行进行建模，表示为：

$$p\begin{pmatrix}x\\y\end{pmatrix}=\sum_{m=1}^{M}\alpha_m N\left(\begin{pmatrix}x\\y\end{pmatrix},\mu_m,\sum_m\right) \tag{4.12}$$

$$\mu_m=\begin{pmatrix}\mu_m^x\\\mu_m^y\end{pmatrix},\sum_m=\begin{pmatrix}\sum_m^{xx} & \sum_m^{xy}\\\sum_m^{yx} & \sum_m^{yy}\end{pmatrix} \tag{4.13}$$

其中$N\left(\begin{pmatrix}x\\y\end{pmatrix},\mu_m,\sum_m\right)$代表第$m$个高斯分量，$\mu_m^x$，$\mu_m^y$分别是第$m$个高斯分量对应的源特征矢量与目标特征矢量的均值向量，$\sum_m^{xx}$，$\sum_m^{yy}$分别是第m个高斯分量对应的源特征矢量与目标特征矢量的自相关矩阵，$\sum_m^{xy}$，$\sum_m^{yx}$是第m个高斯分量对应的源特征矢量与目标特征矢量的协方差矩阵，α_m

是第 m 个高斯分量的概率比重，

利用 EM（期望最大）算法对高斯概率分布的各个参数进行估计。在每一个高斯分量里对源特征矢量和目标特征矢量建立线性关系。

$$\hat{y}_t = \sum_{m=1}^{M} P(c_m \mid x_t)(A_m x_t + B_m) \tag{4.14}$$

其中 $P(c_m \mid x_t)$ 表示第 t 帧特征矢量在第 m 个高斯分量中的后置概率，A_m 表示第 m 个高斯分布的转换矩阵，B_m 表示偏移向量。运用最小平法误差准则如式（3.3）所示，对 A_m，B_m 进行求解。

$$A_m = \sum_m^{yx} \sum_m^{xx-1}, B_m = \mu_m^y - A_m \mu_m^x \tag{4.15}$$

$$P(c_m \mid x_t) = \frac{\alpha_m N(x_t, \mu_m^x, \sum_m^{xx})}{\sum_{m=1}^{M} \alpha_m N(x_t, \mu_m^x, \sum_m^{xx})} \tag{4.16}$$

其中 $\mu_m, \sum_m$ 分别表示 GMM 模型的第 m 个分量的均值和协方差，α_m 表示第 m 个分量的概率，x_t 表示第 t 帧源特征参数矢量。

4.3.1 动态时间调整

目前的说话人语音转换系统是基于对称的语音库（即源说话人和目标说话人的训练语句相同）进行训练的，不能简单地将源说话人和目标说话人的特征参数进行直接匹配，因为语音信号具有相当大的随机性，即使同一个人在不同时刻所讲的同一句话、发的同一个音，也不可能具有完全相同的时间长度，在进行模板匹配时，这些时间长度的变化会造成转换参数精度的下降，从而影响转换效果。

日本学者 Itakura 将 DP（动态规划）算法用于解决孤立词识别时的说话人不均匀的难题，提出了著名的 DTW（动态时间调整）算法[45]。DTW 是把时间调整和距离测度结合起来的一种非线性规整技术，假定共有 I 帧矢量，而参考模板有 J。且 $I \neq J$，动态时间调整就是寻找一个动态时间函数 $j = \omega(i)$，它将测试矢量的时间轴 i 线性地映射到模板的时间轴 j 上，并使该函数 ω 满足下式：

$$D = \min_{\omega(i)} \sum_{i=1}^{D} d[T(i), R(\omega(i))] \tag{4.17}$$

式中 $d[T(i), R(\omega(i))]$ 是第 i 帧测试矢量 $T(i)$ 和第 j 帧特征矢量 $R(j)$ 之间的距离测度，d 则是出于最优时间调整情况下两矢量之间的距离。

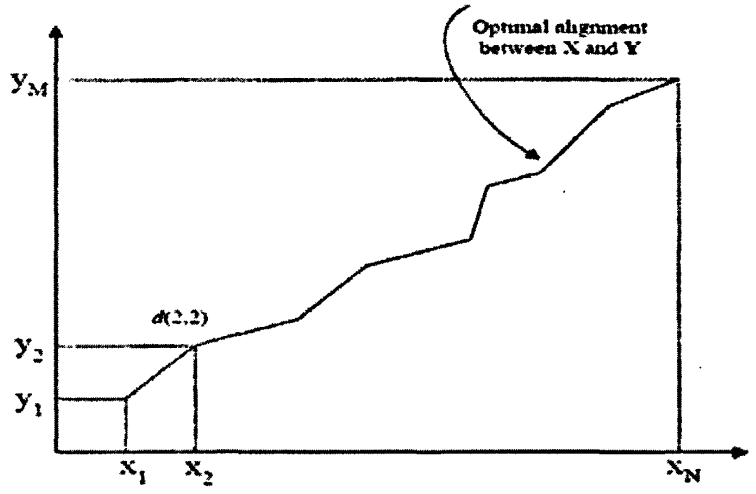


图 4-2 两个语音模板 $X = x_1, x_2, ..., x_N$， $Y = y_1, y_2, ..., y_M$ 之间的直接匹配

由于 DTW 不断地计算两矢量的距离以寻找最优的匹配路径，所以得到的两矢量匹配是累积距离最小的规整函数，这就保证了他们之间存在最大的特征相似性。

实际中，DTW 是采用 DP 技术来具体实现的，动态规划是一种最优化算法。在 GMM 转换算法的实现中，我们把规整函数 $\omega(i)$ 被限制在一个平行四边形当中，它的一条边的斜率为 2，另一条边的斜率为 1/2。规整函数的起始点为（1,1），终止点为 (I, J)，$\omega(i)$ 的斜率为 0、1、2，否则为 1、2。我们的目的是在平行四边形内，找到一个规整函数，使得（1,1）到 (I, J) 具有最小代价函数，由于对路径进行了限制，计算量会相应的减少，此时代价函数的计算公式如下所示：

$$D(c(k)) = d[c(k)] + \min D[c(k)] \tag{4.18}$$

上式中，$D[c(k)]$ 匹配 $c(k)$ 本身的代价，$\min D[c(k)]$ 是在 $c(k)$ 以前由路径限制的所有允许值中最小的一个。

通常动态规划算法是从过程的最后阶段开始，即最优决策是逆序的决策过程。进行动态调整时，对于每一个 i 值要考虑沿纵向可以到达当前值的所有可能的点，由路径限制减少这些可能的点，对于每一个新的可能按上式寻找最佳先前点，得到此点的代价。随着过程的进行，路径不断分叉，并且分叉的可能性不断变大，不断重复这一过程，得到从 (I, J) 到 $(1,1)$ 的最佳路径。

4.3.2 GMM模型的初始化

对特征参数进行 DTW 调整后，下一步要做的工作就是对调整后的特征参数进行 GMM 模型训练建模的工作。

GMM 模型[43]是只有一个状态的模型，在这个状态里有多个高斯混合分布函数。一个 M 阶的高斯混合模型的概率密度函数时由 M 个高斯概率密度函数加权求和得到的，形式如下所示：

$$P(X \mid \lambda) = \sum_{i=1}^{M} \omega_i b_i(X) \tag{4.19}$$

其中 X 是一个 D 维德随机向量，$b_i(X)$ 是子分布，ω_i 是混合权重。每个子分布是一个 D 维的联合高斯概率密度分布。可以表示为：

$$b_i(X)=\frac{1}{(2\pi)^{D/2}}e^{[-\frac{1}{2}(X-\mu_i)'\sum_i^{-1}(X-\mu_i)]} \tag{4.20}$$

其中 μ_i 是均值向量，$\sum_i$ 是协方差矩阵。混合权重满足：

$$\sum_{i=1}^{M}\omega_i=1 \tag{4.21}$$

完整的高斯混合模型由参数均值向量、协方差矩阵和混合权重组成

$$\lambda=\{\omega_i,\mu_i,\sum_i\},i=1,.2,...,M \tag{4.22}$$

利用给定的时间序列 $X=\{x_t\},t=1,2,...t,...T$，利用高斯模型求得的对数似然度

$$L(X\mid\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log P(X_t\mid\lambda) \tag{4.23}$$

对于GMM模型的初始化，也即矢量量化器的最佳设计问题，即从大量样本中训练出好的码本。对于码本设计的一种非常有效的算法就是LBG算法[44]，整个算法就是最近邻准则和最小畸变准则的反复迭代过程，即从初始码书寻找最佳码本的迭代过程。它由对初始码本进行迭代优化开始，一直到系统性能满足要求或不再有明显的改进为止。对于LBG算法的具体的实现步骤如下所示：

1. 设定初始码书和训练参数：设全部输入训练矢量 X 的集合为 X；设置码本的尺寸为 J；设置迭代算法的最大迭代次数为 L；设置畸变的改进阈值为 S。

2. 设定初始值：设 J 个码字的初始值为 $Y_1^{(m)},Y_2^{(m)},.....Y_J^{(m)}$；设置畸变的阈值为 $D^{(0)}=\infty$；设置迭代次数初值为 $m=1$。

3. 假设经过 m 次迭代，根据最近邻准则将 S 分成了 $S_1^{(0)},S_2^{(0)},.....S_J^{(0)}$，当 $X\in S_l^{(m)}$ 时，下式应成立：$d(X,Y_1^{(m-1)})\le d(X,Y_i^{(m-1)}),\forall i,\ i\neq l$。

4. 计算总的畸变 $D^{(m)}$：$D^{(m)}=\sum_{l=1}^{J}\sum_{X\in S_l^{(m)}}d(X,Y_i^{(m-1)})$。

5. 计算畸变改进量 $\Delta D^{(m)}$ 的相对值 $\delta^{(m)}$：$\delta^{(m)}=\frac{\Delta D^{(m)}}{D^{(m)}}=\frac{|D^{(m-1)}-D^{(m)}|}{D^{(m)}}$。

6. 计算新码本的码字：$Y_1^{(m)},Y_2^{(m)},.....Y_J^{(m)}$：$Y_l^{(m)}=\frac{1}{N_l}\sum_{X\in S_l^{(m)}}X$。

7. 阈值判断1：判断 $\delta^{(m)}$ 是否小于 δ，若是，则转入9执行；否则，则转入8执行。

8. 阈值判断2：判断 m 是否小于 L，若是，则令 $m=m+1$，转入3执行；若否，则转入9执

行。

9. 迭代终止：输出 $Y_1^{(m)}, Y_2^{(m)}, \ldots\ldots Y_J^{(m)}$ 作为训练生成码本的码字，并输出总的畸变 $D^{(m)}$。

4.3.3 GMM模型的训练

GMM 模型的训练就是给定一组数据，依据某种准则确定给定的参数。通常用的估计方法是 ML（最大似然）法。但是对于位置参数很难直接求出最大似然度，一般采用 EM 算法来估计参数 λ[46]。GMM 模型的参数估计是从参数 λ 的初始值开始，运用 EM 算法得到一个新的参数 $\hat{\lambda}$。使得在新的模型参数下的似然度 $p(x|\hat{\lambda}) > p(x|\lambda)$，新的参数 $\hat{\lambda}$ 再作为当前参数反复迭代，直到模型收敛。

混合权重的重估公式：

$$\omega_i = \frac{1}{T}\sum_{t=1}^{T}\log P(i|X_t,\lambda) \tag{4.24}$$

均值的重估公式：

$$\mu_i = \frac{\sum_{t=1}^{T}\log P(I|X_t,\lambda)X_t}{\sum_{t=1}^{T}\log P(I|X_t,\lambda)} \tag{4.25}$$

方差的重估公式：

$$\sigma_i = \frac{\sum_{t=1}^{T}\log P(I|X_t,\lambda)(X_t-\mu_i)^2}{\sum_{t=1}^{T}\log P(I|X_t,\lambda)} \tag{4.26}$$

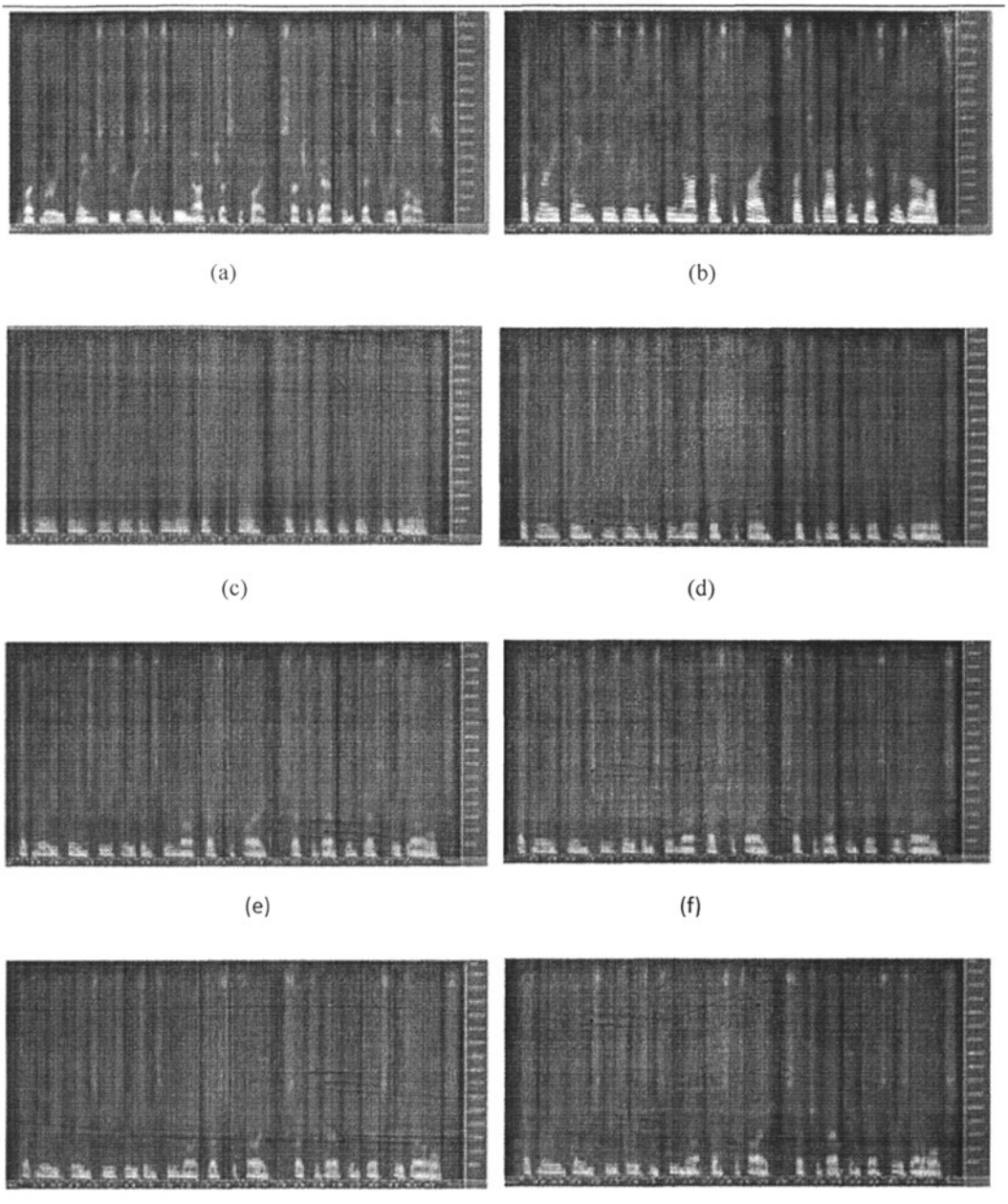
后置概率密度公式：

$$P(i|X_t,\lambda) = \frac{\omega_i b_i(X_t,\lambda)}{\sum_{k=1}^{M}\omega_k b_k(X_t,\lambda)} \tag{4.27}$$

在实际应用中，往往会碰到训练数据不充分的情况，结果会使协方差的一些分量很小，这些很小的分量对模型参数的似然度影响很大，会严重影响系统的性能。为了避免这种情况的发生，在 EM 算法的迭代计算中，对协方差的值设定一个阈值，在训练过程中，若协方差的值低于设定的阈值，则用响应的阈值代替。

GMM 模型参数的个数大小 M 对于转换效果的影响也很大，若训练模型的个数太小，则训练出来的模型不能有效的刻画说话人的特征，从而使整个系统的性能下降；若 M 值取值过大，则模型参数会很多，从有效的训练数据中得不到收敛的模型参数，同时得到的模型参数误差会很大，而且太多的模型参数要求更多的参数存储空间，而且训练和识别的复杂度也大大增加。对于 M 值的大小，很难从理论上推导出来，我们可以通过实验获取，一般 M 值取 2 的 n 次方，选择不同数量的 GMM 模型：

分别选取分量个数为 1、2、4，8、16、32、64、128 的 GMM 模型。对源语音和目标语音进行训练，得到的转换语音的语谱图如下所示：



(a) (b)

(c) (d)

(e) (f)

(g) (h)

图 4-3 源语音和目标语音不同 GMM 分量的转换语音的语谱图

(a), (b) 分别为源语音和目标语音的语谱图；

(c), (d) M 为 1、4 的转换语音的语谱图；

(e), (f) M 为 16、32 的转换语音的语谱图；

(g), (h) M 为 64、128 的转换语音的语谱图

我们发现当 M 比较小时，转换语音的语谱图与目标语音的差距较大，当 M 达到 64 后，语谱图的效果差别很小，通过主观侦听转换语音，同样证明了这一点。通过对计算量和转换效果进行折中，最终选取 GMM 分量大小 M 为 64。

4.4 基音周期的转换

对基音周期进行转换中，基音周期需要保持短时包络特征以及源语音的时长信息不被改变。如下图所示：

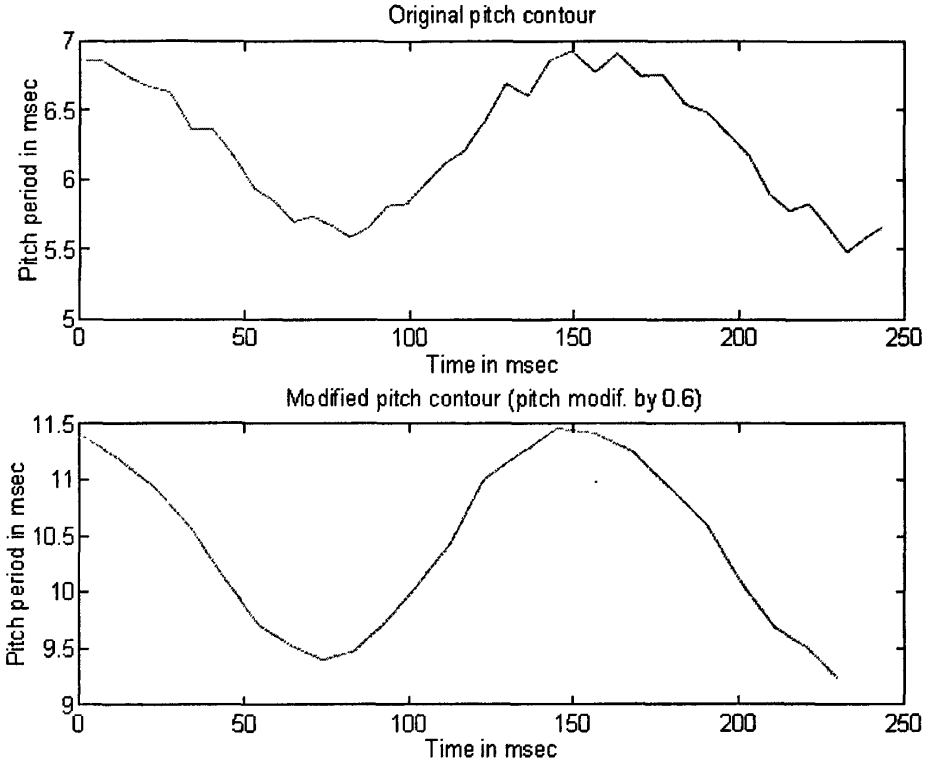


图 4-4 基音周期变换 0.6 时的频谱包络

对于基音周期的转换，目前的做法基本上是取出训练语句基音周期的中值，求出平均转换率进行求解。

$$\alpha = \frac{\overline{P}_{Target}}{\overline{P}_{Source}} \tag{4.28}$$

$$P_{Target} = \alpha P_{Source} \tag{4.29}$$

假定源语音的基音周期是 $P_S(t)$,语音修正系数为 $\alpha(t)$。并且

$$t_a^{i+1} = t_a^i + p(t_a^i) \tag{4.30}$$

将源语音的时间坐标与新的时间坐标进行匹配：$t_a^i \to t_s^i$。

$$t_s^{i+1} - t_s^i = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s}^{t_s^{i+1}} \frac{p(t)}{\alpha(t)} dt \tag{4.31}$$

下图所示为转换系数 $\alpha(t)$ 为 1.5 时对基音周期进行转换的过程。

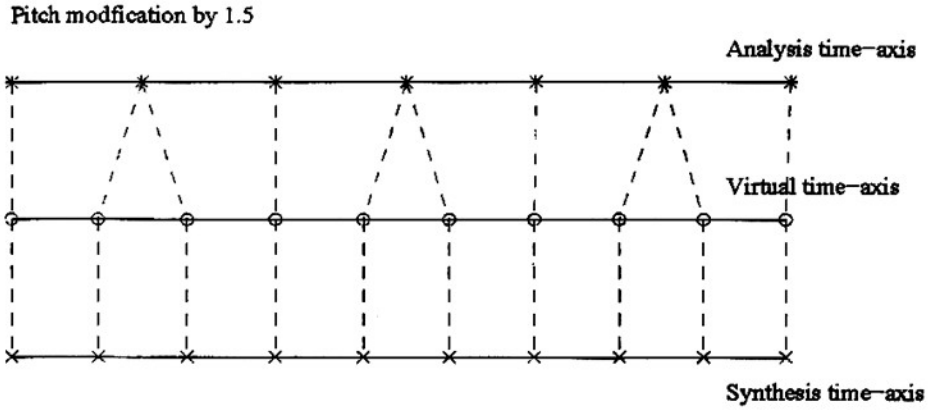
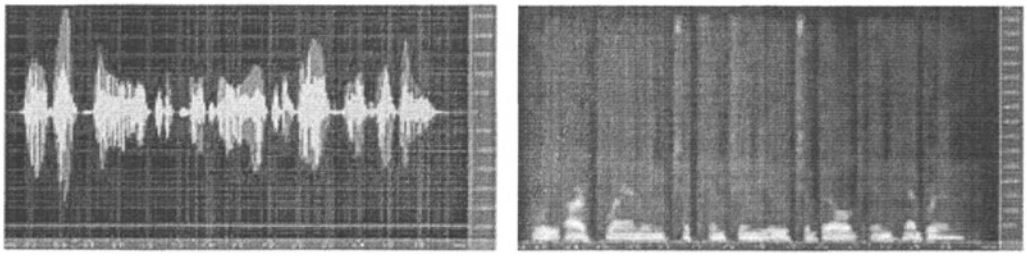
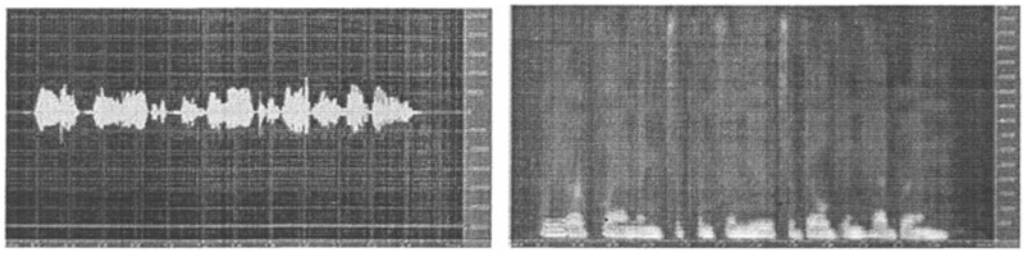


图 4-5 基音周期变换 1.5 倍时的示意图

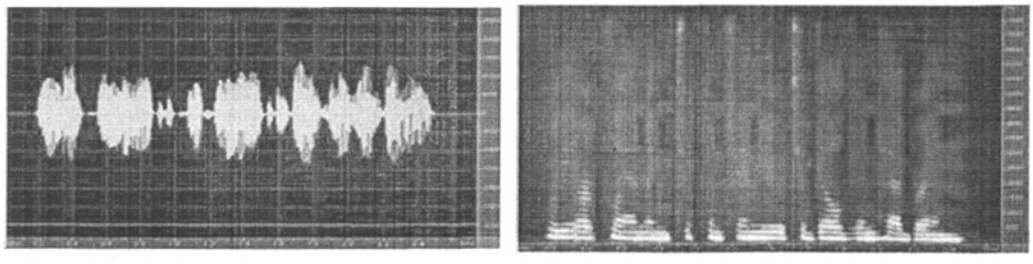
我们对源说话人的基音周期进行了转换，效果如下图所示：



(a)



(b)



(c)

图 4-6 语音的波形和语谱图

(a) 源语音的波形和语谱图；

(b) 目标语音的波形和语谱图；

(c) 改变基音周期后的转换语音的波形和频谱图

4.5 本章小结

本章实现了一个完整的基于 GMM 模型的语音转换系统，并对 GMM 模型的训练和转换方法进行了说明，对语音转换的流程和具体实现进行了详细的讨论，

第五章 基于改进的GMM模型的语音转换

5.1 改进的语音模型

在改进的 GMM 模型算法中，我们采用了性能更加优良的 STRAIGHT 模型进行语音参数的分析和合成。

STRAIGHT（Speech transformation and representation and interpolation using weighted spectrogram）即自适应加权谱内插，是由 Kuwahara 教授等人于九十年代末提出的[47]，主要是针对对于语音参数的修改和恢复而提出的一种非常优秀的语音模型。

以前的合成器（如共振峰合成器、倒谱合成器、LPC 合成器等）虽然其原始语音参数分析合成后的语音效果还不错，但是对于调整后的参数，其合成语音的效果都比较差。因此，STRAIGHT 模型的提出，在一定程度上弥补了这种不足。下面我们对 STRAIGHT 分析合成算法的核心思想和实现进行详细的讨论。

STRAIGHT 是一种针对语音信号的分析合成算法，它的核心思想也是一种源—滤波器的思想，强调将频谱中的激励的影响完全去除，最终将语音分解为相互独立的频谱参数和一系列脉冲的卷积，优点是通过对语音短时谱时频域的自适应内插平滑来提取精确的谱包络，它利用提取的语音参数能够恢复出高质量的语音，并能对时长、基频和谱参数进行高灵活度的调整，主要由以下几个部分组成[48]：

1. 去除周期影响的谱估计

a) 去除时间轴上的周期性：采用基音同步并叠加补偿窗的方法来计算频谱，并在时域上平滑。

b) 去除频率轴上的周期性：通过对线谱卷积三角窗，并进行频率轴上的平滑，得到最终的谱包络。

在分析语音信号的频谱时，我们获取的语音信号频谱需要进行加窗处理，计算得到的短时谱在频率轴和时间轴上都表现出和基音周期相关的周期性，STRAIGHT 模型的核心内容就是去除时间轴和频率轴上的周期性，STRAIGHT 采用二维三角窗的方法来去除二维空间上时域轴和频域轴的周期性。

$$h_t(\lambda,\tau)=\frac{1}{4}(1-|\frac{\lambda}{\omega_0(t)}|)(1-|\frac{\tau}{\tau_0(t)}|) \tag{5.1}$$

$$s(w,t)=[g^{-1}(\int h(\lambda,\tau)g(|F(w-\lambda,t-\tau)|^2)d\lambda d\tau]^{\frac{1}{2}} \tag{5.2}$$

$s(w,t)$ 表示平滑后得到的谱包络，$F(\omega,\tau)$ 表示计算得到的短时谱，函数 $g(\ )$ 定义平滑时保留谱参数的何种特性。

2. 平滑可靠的基频轨迹的提取

使用小波分析方法提取语音信号中的基频，根据语音信号中的基频成分，计算出瞬时基频，通过在频谱上进行谐波分析，并进行频率轴上的平滑，得到最终的基频轨迹。

3. 合成器的实现

合成语音时需要输入的数据是语音的基频曲线数值，经过时间轴和频率轴平滑后的语音的二维的谱包络。在合成时需要使用基音同步叠加和最小相位冲激响应的方法，在合成时可以实现对时长、基频和谱特征参数的调整。

$$y(t)=\sum_{t_i}\frac{1}{\sqrt{G(f_0(t_i))}}u_{t_i}(t-T(t_i)) \tag{5.3}$$

$$u_{t_i}(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}V(\omega,t)\phi(\omega)e^{j\omega(t)}d\omega \tag{5.4}$$

$\phi(\omega)$ 为具有附加的控制相位的激励，用来改善听觉，$V(\omega,t)$ 表示最小相位冲激响应的傅立叶变换。

式(5.3)是基音同步叠加的过程，$y(t)$ 表示恢复后的语音信号。Q 表示用于语音合成的基音同步标记的集合，函数 $G(\ )$ 表示基频的调整函数，它可以是任意形式的映射关系。

$$T(t_i)=\sum_{t_k\in Q,k<i}\frac{1}{G(f_0(t_k))} \tag{5.5}$$

上式表示基音同步标记的确定过程。基音同步标记的位置定义为：在语音信号的浊音部分，标注在基音脉冲的短时能量的尖峰时刻上，这样可以使得相邻标注位置的变化比较平滑；在语音信号的清音部分，标注的位置都是相对于临近的浊音来说是比较平滑的地方。在语音的浊音部分，它的设置与基音周期同步，而清音部分没有基音周期，为了保持算法的一致性，可以设置为等距离分布。时域基音周期同步叠加中对基频和时长韵律参数的控制，是根据基音同步标志对短时信号的复制和删除来实现的。在合成中，基音同步标记在很大程度上影响着情感语音的合成效果，所以要合成出比较好的语音，不但要精确提取出基音周期，还要准确的记录最大基音周期脉冲的位置，两者都很重要，偏差过大的基音同步标注，将会给合成语音带来一定的噪声。

一帧语音对应的冲激响应的求取过程如下式所示：

$$v_u(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}V(\omega,t)\phi(\omega)e^{j\omega}d\omega \tag{5.6}$$

$\phi(\omega)$ 为具有附加的控制相位的激励，用来干预听感。$V(\omega,t)$ 用来表示最小冲激响应的激励，可以从实现得到的平滑谱 $s(w,t)$ 计算得到。即将 一般相位的谱转化为最小相位，采用的是基于倒谱的转换的方法：

$$V(\omega,t)=\exp(\frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}h_t(q)e^{j\omega q}dq) \tag{5.7}$$

$$h_t(q)=\begin{cases}0 & q<0\\ c_t(0) & q=0\\ c_t(q) & q>0\end{cases} \tag{5.8}$$

$$c_t(q)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-j\omega q}\log A(S(u(\omega),r(t)),u(\omega),r(t))d\omega \tag{5.9}$$

其中 $A(\ )$，$u(\omega)$，$r(t)$ 分别表示对平滑谱 $s(w,t)$ 在幅度、频率和时间轴上的调整。STRAIGHT 分析合成模型如下图所示：

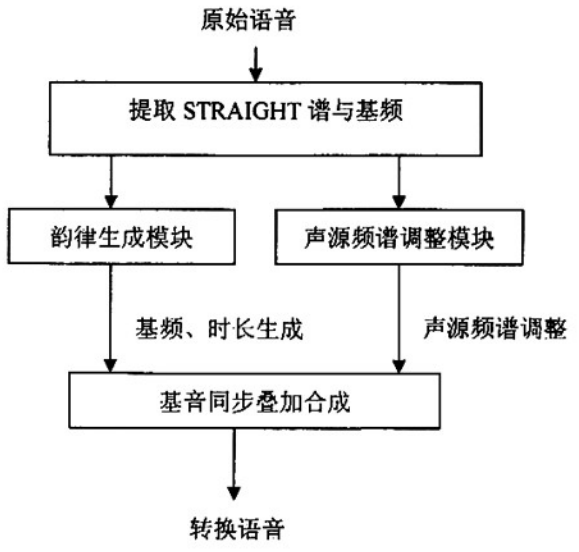


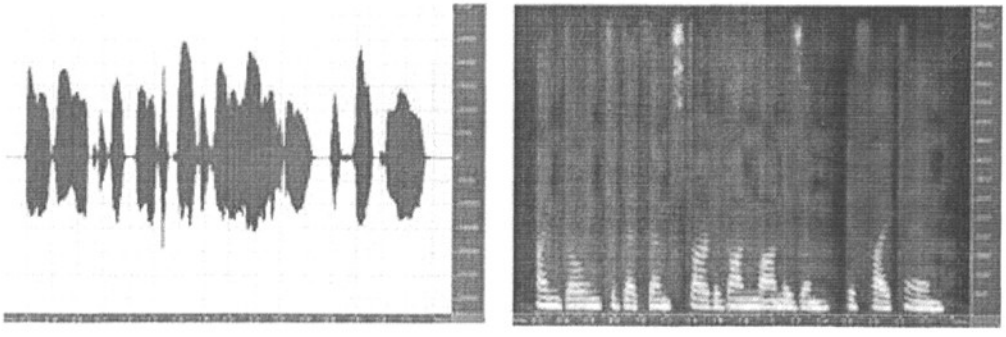
图 5-1 STRAIGHT 分析合成过程

当然 STRAIGHT 模型也有一些缺点，主要包含以下几个方面：

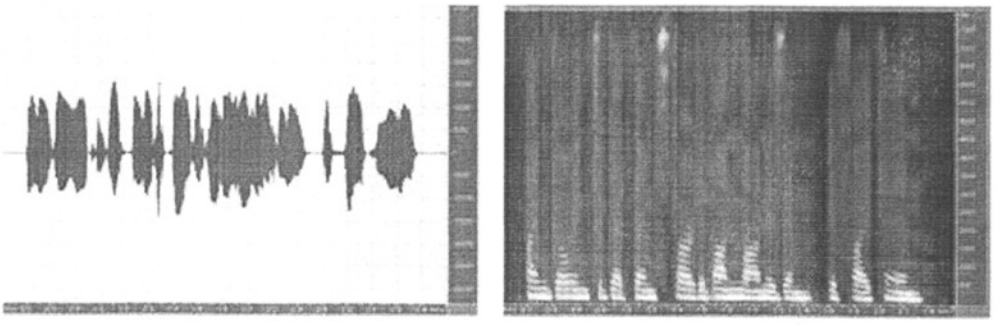
1. 基频和频谱参数需要同时进行调整。
2. 只对有音段进行处理，对于无音段没有考虑。
3. 信噪比较低时效果较差。

基于 STRAIGHT 模型提取的谱包络并不能直接的用来进行参数的建模，需要对谱包络进行特征参数化，然后再对参数进行建模。选择一个合适的谱包络参数化方法非常重要，直接影响 GMM 建模和模型的训练，进而影响最终合成的语音。

基于 STRAIGHT 模型的分析合成效果如下图所示：



(a)



(b)

图 5-2 STRAIGHT 模型分析合成的语音的波形图和语谱图

(a) 源说话人语音的波形和语谱图;

(b) 经 STRAIGHT 模型分析合成后的语音的波形和语谱图

通过对比，源语音信号和经过 STRAIGHT 模型分析合成的语音信号的波形图和语谱图，可以发现通过 STRAIGHT 模型分析合成后的信号可以很好地逼近原始语音。同时通过主观侦听，我们发现通过 STRAIGHT 模型恢复后的语音信号具有较高的自然度。所以，我们最终选择 STRAIGHT 模型作为我们的说话人语音转换中的语音分析模型。

5.2 改进的频谱特征参数

在实际应用过程中，同时满足优良的特征参数的五个条件的语音特征参数几乎不可能，在大多数情况下我们是对这几条准则进行折中。LPC系数和倒谱系数可以很好的表示语音的频谱包络。但人耳对于语音频率的感知是非线性的，对低频的感知比高频要灵敏。Mel尺度的频率弯折可以很好的描述人耳对于频率感知的非线性特征。在语音识别中，运用MFCC取得了很好的识别效果，同样在说话人语音转换中精确地表示低频信号也是很重要的。

对于基于Mel尺度的特征参数的获取有很多种方法：在语音识别中，MFCC是通过采用对数能量谱用Mel频率尺度的三角滤波来进行求解，这种方法获得的谱包络不是很理想，在说话人语音转换中不宜采用。Tokuda[48]采用由LPC系数求取MFCC，这种方法的精确度受LPC系数提取的准确度的影响；Stylianou[5]先对频率进行Bark尺度的弯折，采用解矩阵求解方程法来获取到谱，以此Bark尺度获得的频谱包络来进行频谱包络的转换。但随着阶数的增加，计算量会显著增加，而且这种方法是基于谐波处的谱幅度来进行求解，当谐波特性不好时，性能就会受到影响。

因此，我们提出了一种简单有效的Mel倒谱系数，首先获取语音信号的频谱，经过Mel尺度的弯折，再来获取基于Mel尺度的倒谱系数，具体步骤如下所示：

1. 求取语音信号的短时幅度谱。在语音信号处理中我们采用的是2K点的DFT变换。首先对语音信号进行加窗，对加窗后的信号进行补零，使其长度为2K。进行离散傅立叶变换，得到幅度谱$S(k)$。

2. 对幅度谱进行美尔尺度的调整。得到基于美尔尺度的幅度谱。

3. 由对数幅度谱计算Mel倒谱系数，如下所示：

$$Mel\_Cep(n)=\frac{1}{K}\sum_{k=0}^{K-1}\ln|S(k)|\cos(\frac{2\pi k}{2K}),n=0,1,...N-1 \tag{5.10}$$

基于Mel尺度的倒谱系数的计算量和复杂度大大减少。因此，论文实验中我们用到的是20阶的Mel倒谱系数，

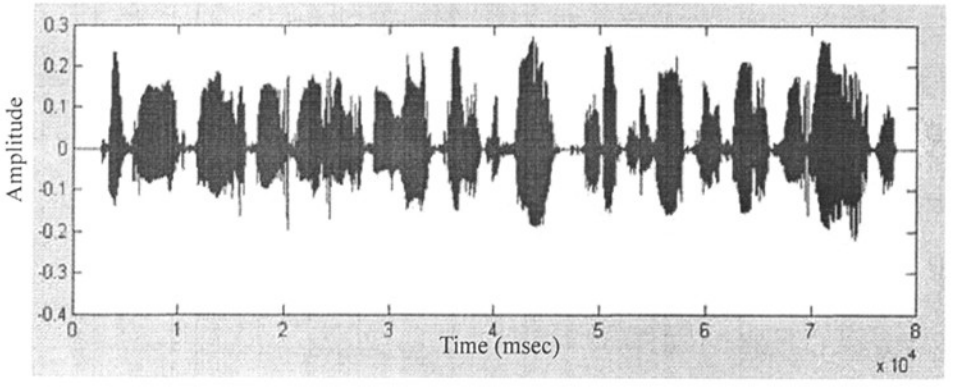
5.3 改进的频谱转换方法

通过对各种语音转换方法的比较，我们可以发现 GMM 方法应该是当前最为有效的频谱转换方法，因此在本次论文中，我们提出了改进的 GMM 转换算法。

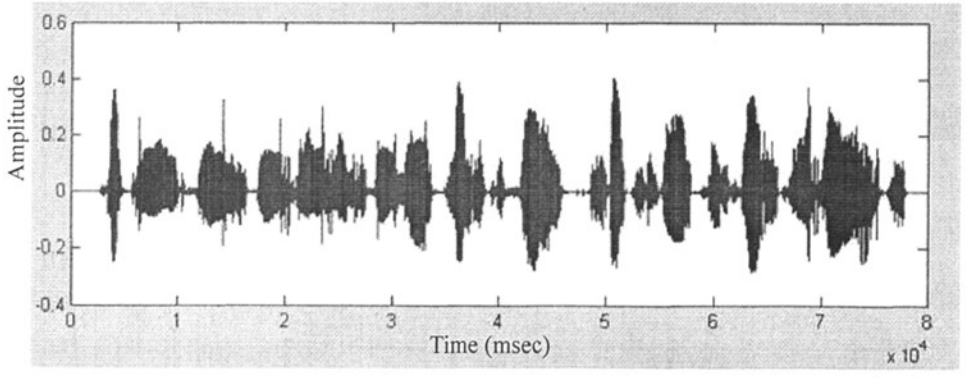
一方面，传统的 GMM 转换方法会引起转换后的特征参数的过平滑现象，对于这一问题，我们一共采取了两种方式进行消除，首先是考虑转换语音的协方差，为了防止过平滑现象的出现，在转换特征参数中我们需要保持足够大的协方差，但是很难将源说话人的协方差和目标语音的协方差进行匹配，因此，我们可以假定转换语音和目标语音有着相同的协方差。则 GMM 方法的转换公式(4.14)变成如下所示：

$$\hat{y}_t = x_t + \sum_{m=1}^{M} p(c_m \mid x_t)(\mu_m^y - \mu_m^x) \tag{5.11}$$

通过下图我们可以发现：通过考虑协方差，可以显著地减少语音的过平滑的缺点，通过主观侦听也证实了这一点，考虑了转换语音协方差后得到的转换语音相比传统的 GMM 转换方法的语音质量有了一定的提高。



(a)



(b)

图5-3 不同的GMM方法得到的转换语音波形图

(a) 传统GMM方法得到的语音波形图；

(b) 考虑动态参数的GMM方法的语音波形图

接着，对于过平滑的缺陷进行处理通过设置阈值在 GMM 方法和 LMR 方法之间进行折中，经过实验统计 20000 帧的频谱特征参数，结果如下图所示，不存在后置概率密度低于 20%的参数帧，只

有5%的参数帧的后置概率密度在50%以上，超过65%的帧的后置概率密度在90%以上。

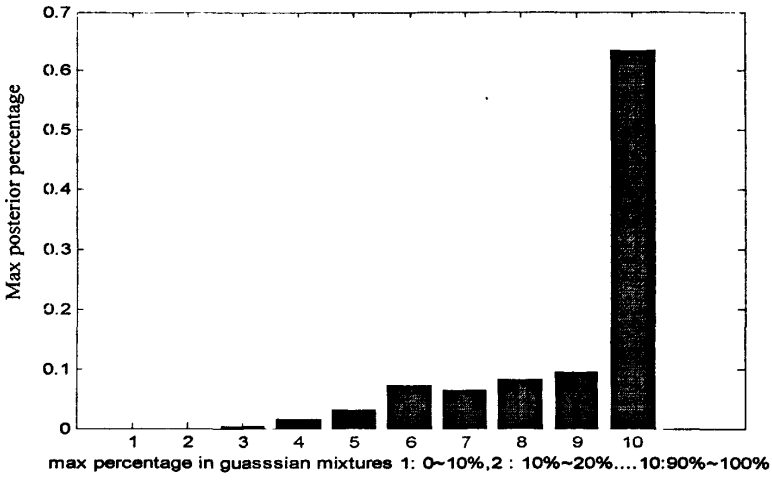


图 5-4 GMM 模型中最大后置概率密度

我们通过给后置概率密度设置一个阈值 β，当概率低于 β 时，我们采用 GMM 的方法，反之我们采用 LMR 的方法，即仅在一个单独的特征空间进行转换。通过这种方法可以在过平滑和离散之间进行一个权衡。通过循环迭代的方法得到 β 的最优值为 0.915。

另一方面，传统的 GMM 转换方法对于语音的转换是对每个语音帧单独进行转换的，忽略了相邻帧之间的联系性，造成了转换语音的不连续性，为了克服这一问题，我们考虑了相邻语音帧之间的 Δ,Δ^2 参数。首先获取源说话人特征量 x 和目标说话人特征矢量的特征参数 y。然后再乘以相应的转移矩阵 W [50]求取包含 Δ,Δ^2 信息的源混合特征参数 X 和目标混合特征参数 Y。如下所示：

$$X = Wx, Y = Wy \tag{5.12}$$

$$X = [x_1, x_2, ...x_t, ...x_T]^T, Y = [y_1, y_2, ...y_t, ...y_T]^T$$

$$x = [x_1, x_2, ...x_t, ...x_T], y = [y_1, y_2, ...y_t ...y_T]$$

$$W = [w_1, w_2, ..., w_t, ...w_T]^T \tag{5.13}$$

$$W_t = [w_t^0, w_t^1, w_t^2]$$

$$w_t^n = \begin{pmatrix} 0_{M*M}, ..., & 0_{M*M} & w^n(-L_-^n)I_{M*M}, \\ & & {\scriptstyle (t-L_-^n)-th} \\ ..., w^n(0)I_{M*M}, & ..., & w^n(L_+^n)I_{M*M} \\ {\scriptstyle t-th} & & {\scriptstyle (t+L_+^n)-th} \\ 0_{M*M}, & ..., & 0_{M*M} \\ & & {\scriptstyle T-th} \end{pmatrix}^T, \quad n = 0,1,2$$

我们选取的Δ窗是[0.5,0,0.5]，Δ^2 窗是[1,−2,1]。

如下图所示，我们可以看到考虑了动态特征参数的 Mel 倒谱失真[51]随着高斯分量的增加，倒谱失真有了显著的降低。

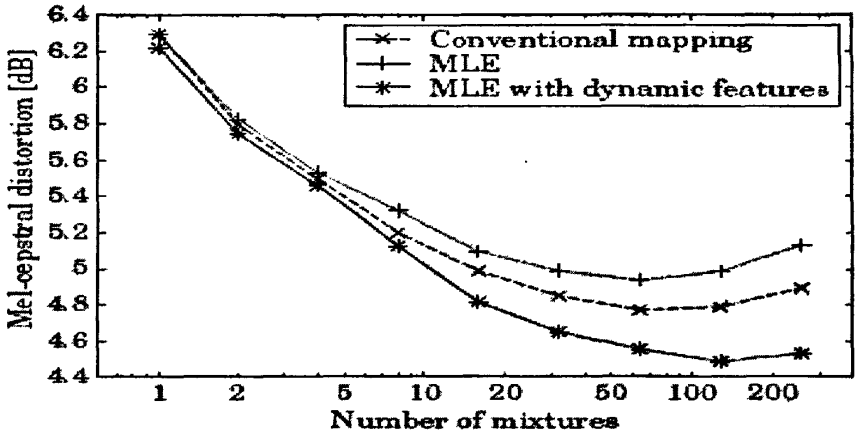


图 5-5 考虑动态特征参数的 GMM 转换方法和传统的 GMM 方法的频谱失真距离

转换后的混征参数如下式所示：

$$\hat{Y}_t = \sum_{m=1}^{M} P(C_m \mid X_t)(X_t + \mu_Y - \mu_X)$$

$$W\hat{y}_t = \sum_{m=1}^{M} P(C_m \mid X_t)(X_t + \mu_Y - \mu_X)$$

$$\hat{y}_t = \sum_{m=1}^{M} P(C_m \mid X_t)(W^T W)^{-1} W^T (X_t + \mu_Y - \mu_X) \tag{5.14}$$

其中 μ_X, μ_Y 分别代表源说话人混合特征参数和目标说话人混合特征参数的均值，C_m 代表第 m 个高斯混合分量，X_t 表示第 t 帧源说话人的混合特征参数。

通过主观侦听测试，发现考虑了语音动态特征参数以后，语音帧与帧连接处的“吱吱”声有了较为明显的减少，语音质量有了较为显著的提高。

5.4 改进的基音周期转换方法

对于基音周期的转换，相关报道中的做法基本上是取出训练语句基音周期的中值，求出平均转换率进行求解。我们的做法则考虑了基音周期的 Δ, Δ^2 信息。

$$\begin{aligned} &P_S = [P_{Source}, \Delta P_{Source}, \Delta^2 P_{Source}], P_T = [P_{Target}, \Delta P_{Target}, \Delta^2 P_{Target}], \\ &P_C = [P_C, \Delta P_C, \Delta^2 P_C] \\ &P_C = AP_S + B \end{aligned} \tag{5.15}$$

$$A=\frac{\sum(P_T,P_S)}{\sum(P_T,P_S)},B=E(P_T)-AE(P_S) \tag{5.16}$$

$$P_{Convert}=(W^TW)^{-1}W^TP_C \tag{5.17}$$

其中 P_{Source}、P_S 分别表示源语音的基音周期和混合基音周期，P_{Target}、P_T 分别表示目标语音的基音周期和混合基音周期，$P_{Convert}$、P_C 分别表示转换语音的基音周期和混合基音周期。

5.5 时长的转换

从国内外相关论文来看，很少有对语音的时长进行转换，在改进的语音转换算法中，我们考虑了时长的转换以使转换语音更好的接近目标语音。对时长转换，我们所要达到的目标是对语音的发音速率进行改变而不影响说话人的发音质量。如下图所示，我们对一段语音进行了时长变换系数为0.6的时长变换。而变换语音仍然能够保持较高的质量。

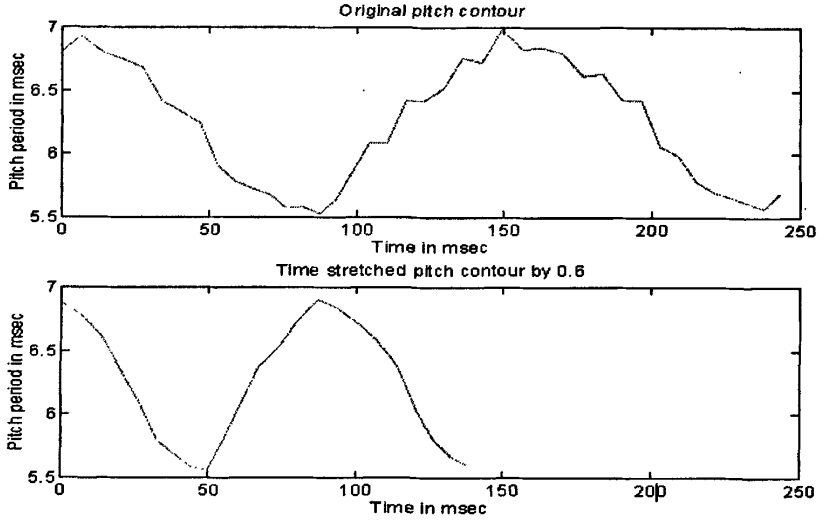


图 5-6 时长改变 0.6 倍时的语音波形

假定时域修正函数为 $\beta(t)$，论文中我们是采用平均时长转换的方法对时长进行转换的。所以 $\beta(t)$=常数。

时域弯折函数

$$D(t)=\int_0^t\beta(t)dt,\ t_a^{i+1}=t_a^i+p(t_a^i) \tag{5.18}$$

将语音的时间坐标与新的时间坐标进行匹配：$t_a^i\rightarrow t_s^i$。

$$t_s^{i+1}-t_s^i=\frac{1}{t_u^{i+1}-t_u^i}\int_{t_u}^{t_u^{i+1}}\frac{p(t)}{\alpha(t)}dt \tag{5.19}$$

其中 t_u^i 为虚拟时间常量定义如下：

$$t_s^i = D(t_u^i) \tag{5.20}$$

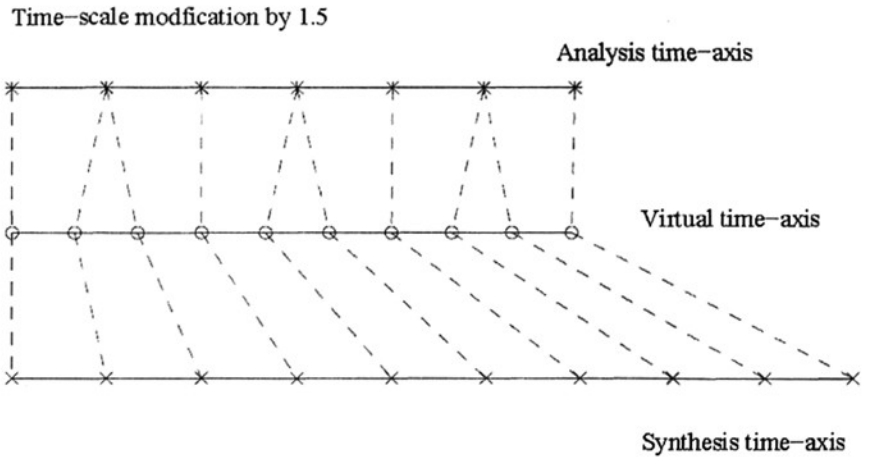


图 5-7 时长改变 1.5 倍时的示意图

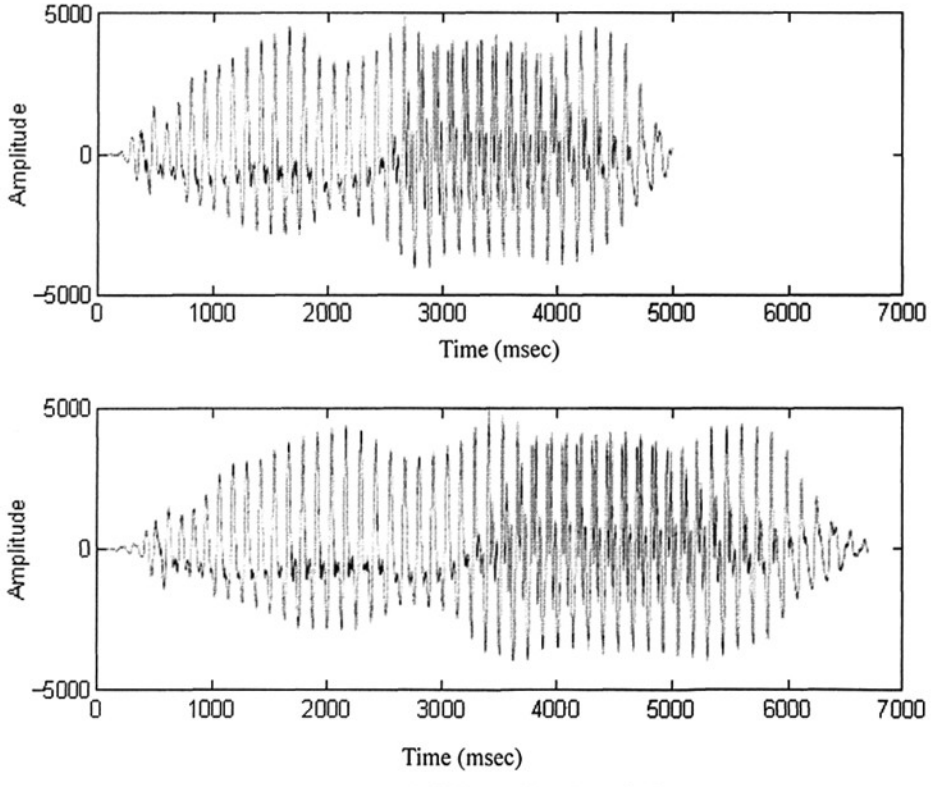


图 5-8 时长变换 1.5 倍时的语音波形

5.6 说话人语音转换系统的评价标准

说话人语音转换性能评估构成了说话人语音转换技术的重要组成部分，设计一个好的效果评价方案对于提高系统的性能有着重要的作用。合成语音可以根据清晰度、自然度和使用场合的实用性来比较评估。对于转换语音，适用性指标是指说话人的可识别性，即目标说话人特征倾向性。语音质量是一个多维性术语，还没有一种评估体系能够完全适合或自动评测合成语音或转换语音的质量。测试系统性能是一个复杂的实验环节，采用不同的评价标准，则有不同的测试方法。语音转换系统

的性能评价既有客观测试，也有主观测试。

5.6.1 客观测试

语音的客观评价一般建立在对语音幅度谱计算的基础上，主要有以下几种方案：

1. 频谱失真度[52]

转换语音与目标语音之间的频谱失真距离为一项客观评测内容被广泛采用，采用恰当的频谱失真是一种有效的客观测试方法，其基本形式可表示如下：

$$D=\frac{1}{T}\sum_{t=1}^{T}d(\hat{y}_t,y_t) \tag{5.21}$$

其谱失真测度可用上式表示，$\hat{y}_t$ 表示转换后的特征参数，y_t 表示目标特征参数，$d(\hat{y}_t,y_t)$ 表示某种距离测度。而目前广泛使用的是一种相对距离测度，如下式所示：

$$D=\frac{\frac{1}{T}\sum_{t=1}^{T}d(\hat{y}_t,x_t)}{\frac{1}{T}\sum_{t=1}^{T}d(\hat{y}_t,y_t)} \tag{5.22}$$

上式表示转换语音和源语音的谱距离与转换语音和目标语音的谱距离的比值。这个比值越大，则转换效果越好，转换后的频谱越接近于目标频谱。

2. 说话人辨识[53]

其主要思想是，将转换语音作为说话人识别系统的输入，以确定目标说话人辨识的似然性。在这里，目标说话人与源说话人的对数似然比被用于说话人决策的置信度测量。其测试的模型可以通过下式表示：

$$Ratio=\log\frac{\sum_{n=1}^{N}P(x_n|\theta_T)}{\sum_{n=1}^{N}P(x_n|\theta_S)} \tag{5.23}$$

上式中，θ_S 表示源说话人的模型，θ_T 表示目标说话人的模型，这个模型可以是 GMM，也可以是 HMM。当似然度变大时说明转换效果越好，即转换后的频谱越接近目标频谱，反之，转换效果越差。

5.6.2 主观测试

生成的语音信号最终要通过人的主观测试来判别转换效果的好坏，因此，主观测试是语音系统评价不可缺少的手段之一。主观测试主要有以下几种：

1. ABX测试

在主观测试中，说话人的ABX测试方法被广泛采用。X表示转换语音，A和B分别代表源说话人的语音和目标说话人的语音，话语内容相同但与X不同。每一个三元组就是这样三句话语的组合。参与测试的听者要求对A和B中哪一个和X声音最接近做出选择[54]。尽管可能有100%的选择正确率，但是转换语音也不能被认为是目标说话人说出来的，还存在着一定程度的差距。

2. 倾向性测试

倾向性测试（Preference Test）评价[34]由两种不同算法（如码本映射和混合高斯模型）实现的语音中哪一种最接近于期望的说话人声音特征．测试也是用ABX方法实现。其中X代表自然语音，A和B代表两种不同算法实现的转换语音对。话语内容相同但与X不同。听者被要求在转换语音对中做出倾向性选择．倾向性测试是对语音转换系统性能的一种横向评估方法。

3. MOS意见法

MOS意见法即观点测试是为了评估转换算法的整体性能而设计的。语音对包含所有的源、目标说话人的语音和由不同算法实现的转换语音，不同的句子用来组成这些语音对。要求听者按照不同的等级对每一对语音的相似性打分，不同的等级用来表示语音对中的语音相似性。

5.7 改进的语音转换系统

本节将给出我们所设计的改进的基于 GMM 模型的语音转换系统，转换的训练过程和转换过程如图 4-1 所示。

1. 语料库的选择：基于 HMM 方法的 TTS 系统合成的英文语音。一共 3 人，2 女 1 男。训练语句和测试语句时长分别为 5 分钟左右。

2. 频谱参数的选择：改进的 Mel 倒谱参数。

3. 基音周期的变换：运用平均基音周期转换法，这里基音周期的转换考虑了基音周期的动态参数即 Δ,Δ^2 信息。

4. 时长的变换：运用了平均时长转换法。

5. 频谱参数的转换：采用基于改进的 GMM 转换算法。相比传统的 GMM 转换方法，采用假定转换后的特征参数的协方差和目标特征参数的协方差相同，以及通过在 LMR 方法和 GMM 方法设定阈值 β，来解决频谱参数的过平滑问题。我们考虑了特征参数的动态特征信息以减少转换后语音帧的不连续的现象，

6. 语音合成模型的选择：选用 STRAGHT 模型进行语音的分析合成。

5.8 效果评价

本节我们将测试提出的改进的说话人语音转换算法的性能。转换方法分成两类：第一类为女声到男声之间的转换，第二类为女声到女声之间的转换。实验条件如下表所示：

表 5.1 实验条件

实验人数	2 女 1 男
实验用的语音采样频率	16KHz
语音的帧长	10ms
语音的帧移	5ms
GMM 模型的大小	64
训练语句的数量	100 句平均时长为 3s 的语音
测试语句的数量	100 句平均时长为 3s 的语音

5.8.1 客观评价

为了能够正确地对转换语音的效果进行客观评价，相比说话人辨识测试法，频谱失真度测试被更为广泛的应用于转换语音效果的评价上。本文中我们选取频谱失真度公式(5.22)对转换后语音的质量进行评价。

测试语句的数量为 100，不包含训练语料库中的语句。阈值 β 被设定为 0.915。一共进行了三组实验，训练语句数量分别为 10、50、100。三种方法被比较：GMM 算法、GMM+LMR 算法、考虑动态特征参数的 GMM+LMR 算法（改进的 GMM 算法）。

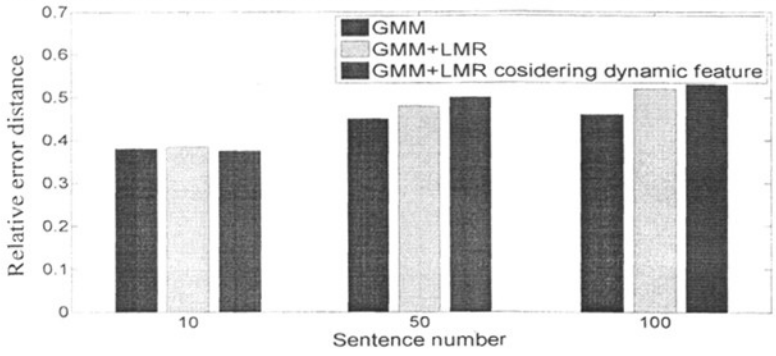


图 5-9 训练语句分别为 10、50、100 时的女声到男声的频谱失真图

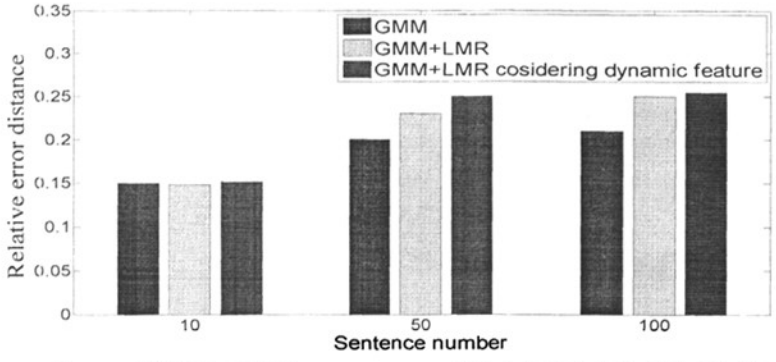


图 5-10 训练语句分别为 10、50、100 时的女声到女声的频谱失真图

正如上述两图所示，当训练语句较少时，三种方法得到的频谱失真度相差不大。但随着训练语句的增加。我们可以发现，考虑动态特征参数的 GMM+LMR 方法（改进的 GMM 算法）较其他两种算法在性能上有了很大的提高。

5.8.2 主观评价

对于语音的转换效果的评价，仅通过客观测试进行评价是不全面的，我们通过主观测试法来侦听语音转换的效果和语音的转换质量。测试人为实验室的 20 名同学，共 10 男 10 女。

第一类测试：ABX 测试法，判断转换后的语音 X，更接近说话人 A 和说话人 B。对测试结果进行平均。

表 5.2 ABX 方法

转换类别	女—男（GMM 算法）	女—女（GMM 算法）	女—男（改进的 GMM 算法）	女—女（改进的 GMM 算法）
转换语音测试结果	100%	85%	100%	90%

第二类测试：倾向性测试，将转换语音和目标语音的相似度分成 10 个等级，用 0～9 个来表示分别表示相似度由小到大，0 表示完全不同，9 表示完全相同。

表 5.3 倾向性测试

转换类别	女—男（GMM 算法）	女—女（GMM 算法）	女—男（改进的 GMM 算法）	女—女（改进的 GMM 算法）
源语音和目标语音	1.0	5.5	1.0	5.5
转换语音和目标语音	8.0	7.0	8.2	7.5

第三类测试：MOS 意见法，将语音质量分成五个等级，分别为“很差”，“差”，“一般”，“好”，“很好”。分别用 1～5 来表示。

表 5.4 MOS 意见法

转换类别	女—男（GMM 算法）	女—女（GMM 算法）	女—男（改进的 GMM 算法）	女—女（改进的 GMM 算法）
源语音	4.0	4.0	4.0	4.0
转换语音	3.5	3.6	3.7	3.8

通过主观测试，我们可以看到经过改进的 GMM 转换算法可以很好地对语音进行转换，尤其女声到男声的转换最为明显，相比传统的 GMM 转换算法有了很大的提高。通过上述表格可以发现，经过转换后语音的质量都有所下降，女声到男声之间的转换幅度最大，语音质量下降的也较大，而女声到女声之间的转换幅度较小，语音质量下降的也较小。

5.9 本章小结

本章给出了我们所提出的基于改进的 GMM 模型的语音转换方法，即考虑动态特征参数的 GMM+LMR 方法，针对传统的 GMM 转换方法造成的频谱过平滑问题，通过假定转换语音和目标语音有相同的协方差，在 GMM 和 LMR 方法之间设定阈值，很好的加以解决；通过考虑帧与帧之间的动态特征参数，很好地解决了语音帧之间的频谱不连续问题。通过主客观评价分析，改进的 GMM 算法在很大程度上提高了说话人语音转换的性能。

第六章 总结与展望

6.1 论文的主要工作和贡献

本文主要是基于说话人语音转换的方法的研究，目的是通过对目前的算法进行研究与优化，能够找到一种更加有效的转换方法，最终使转换语音听起来更像是目标说话人的语音。而且转换后的语音质量能够达到一个较为理想的状态。最近十几年说话人语音转换逐渐成为国内外高校和各个研究机构的研究热点，在一定程度上已经取得了一定的成果，但这些成果在应用上还存在着很大的局限性，转换的语音质量与理想语音还有着很大的差距。本文的主要工作是在分析比较了几种经典的说话人语音转换方法的基础之上，对基于 GMM 模型的说话人语音转换方法进行了改进，提出了一种新的改进的语音转换算法，实验结果表明，相比传统的说话人语音转换方法，我们提出的算法更加有效。

在本次论文中，作者的主要工作是：

1. 首先为了能够更好的进行研究，设计一个优良的语音库是基础，为了能够训练一个有效的说话人的语料库，我们采用了基于 HMM 的语音合成后的语料库，利用该方法得到比较理想的对称语音。

2. 针对传统的 GMM 转换方法后语音的频谱的过平滑现象，我们通过在 LMR 方法和 GMM 方法之间设置后置概率阈值，并且假定源说话人和目标说话人具有相同的协方差很好的解决了这一问题，通过主客观测试，转换后的语音的频谱过平滑现象得到了很好的解决，同时语音质量也有了显著的提高。

3. 传统 GMM 方法得到的语音是基于频谱参数的每一帧单独进行转换的，忽略了相邻帧之间的关系，因而得到的转换语音会存在较为明显的不连续性，我们通过考虑相邻帧之间的特征参数的动态特征，加强了相邻帧之间的联系性。使得转换后的语音质量有了明显的提高。

4. 传统的语音合成模型如 LPC 模型、共振峰合成模型等在基音频率、时长和频谱参数有了较大的改变后，得到的语音质量会有明显下降，针对这个问题，我们采用了 STRAIGHT 模型很好的加以解决。STRAIGHT 模型在说话人的基频、时长和频谱参数等有较大的改动的情况下仍然能够很好地恢复语音。

5. 传统的说话人语音转换方法对于基音周期的改变是基于平均基音周期进行转换的，在本次论文中我们对于基音周期的改变同样考虑了相邻帧之间的动态基音周期，取得了很好的效果。

6.2 后继的主要工作

本文主要是基于 GMM 模型的说话人语音转换方法的提出了新的算法，但是对于实现一个优良的说话人语音转换系统，目前的工作还远远不够，以后需要进一步加强对相关算法的研究，目前的研究的缺陷主要包括以下几点：

1. 目前的说话人语音转换对于频谱和基音周期的转换是单独进行分析的，而没有考虑激励和声道之间的相互作用。然而研究表明，在实际语音的产生过程中，声源的振动对于声道中传播的声波有着不可忽略的作用，我们需要进一步研究激励和声道之间的关系，对它们进行同时转换。可以做为提高转换语音音质的一个重要途径。

2. 目前用于说话人语音转换的语料库都是对称的，在实际生活中更多的是非对称的语料库，我们目前的基于 GMM 的转换方法是不可行的，因此我们需要加强基于非对称语料库进行训练的方法的研究。

3. 目前的说话人语音转换方法都是针对同一语种进行转换的，在实际的应用中我们需要加强对不同语种说话人语音转换的研究。当然这也必将推动机器翻译的发展。

4. 当前的说话人语音转换算法仍然停留在理论研究阶段，距离实际开发应用还有很长的路要走，我们需要进一步加强对相关算法的研究，以期进一步减少算法复杂度和运算量，到达语音的实时转换，在复杂度和实用性方面达到一个很好的折中。

5. 实际的语音信号存在很大的噪声，这些噪声会对说话人语音转换得到的语音质量造成很大的影响，我们需要对相关语音信号进行去噪处理。

相信随着语音技术的发展，说话人语音转换技术会越来越广泛的应用于社会生活的各个领域。让转换后的语音具有目标说话人的语音特点是说话人语音转换的目的，但是目前的转换后的语音质量和目标语音还有着较大的差距，要想在说话人语音转换研究领域获得进一步的突破，需要我们不断的去研究和探索。

6.3 本章小结

本章首先对提出的改进的语音转换算法进行了总结，对算法的优缺点进行了详细的讨论，最后对语音转换未来需要解决的一些问题进行了展望。

致 谢

两年半的研究生生活即将结束，很荣幸能够来到东南大学信息处理与工程应用研究中心学习，这是我人生中非常重要的一段时光，中心良好的工作和学习氛围给了我很大的帮助。

首先要感谢我的导师赵力教授，本次论文是在赵老师的精心指导下完成的，从选题、实验到论文写作都是与赵老师的悉心指导分不开的。赵老师做事严谨，对学生在学习上严格要求，在生活上又关心备至。谨此致以诚挚的敬意和衷心的感谢。

同时我还要感谢微软亚洲研究院的陈一宁研究员，陈老师学识渊博，严谨治学。作为我在研究院实习期间的导师，陈老师无论从生活上，学习上还是工作上都对我关心备至，一幕幕令我终身难忘。

同时感谢同宿舍的舍友吴广、张光亮、方伟，你们为我创造了一个轻松惬意的生活环境，为我的两年半的硕士生活留下了一个美好的回忆。

感谢实验室的同学为课题研究提供的讨论和学习氛围，谢谢你们！

最后还要特别感谢我的父母，正是你们一如既往的关心、支持和鼓励，给了我战胜困难的勇气，使我能够最终完成学业！

参考文献

[1] 李波. 语音转换的关键技术研究[D]：［博士学位论文］. 长沙：国防科技大学，2005

[2] M.Abe, S.Nakanura, K.Shikano and H.Kuwabara. Voice conversion through vector quantization[J]. Proc. ICASSP, 1988: 655-658

[3] H.Valbret, E.Mulines and J.Tubach. Voice transformation using PSOLA techniques, Speech Communication[J]. Vol. 11. No. 2-3, 1992: 175-187

[4] H.Kuwabara and Y.Sagisaka. Acoustic characteristics of speaker individuality: control and conversion[J]. Speech Communication, Vol. 16. No .2, 1995: 165-171

[5] Y.Stylianou, O.Cappe and E.Moulines. Statistical methods for voice quality transformation[J]. Proc. EUROSPEECH, 1995: 447-450

[6] T.Toda, H.Saruwatari and K.ShikaNo. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum[J]. Proc. ICASSP, 2001: 841-844

[7] Chu Min. Voice conversion between female and male in a TD-PSOLA based Chinese TTS system[J]. ISLSLP. Singapore, 1998: 113-117

[8] 王重修. 语音转换及相关问题的研究[D]：［博士学位论文］. 北京：中国科学院声学所，2001

[9] Y.Chen, M.Chu, E.Chang, J.Liu and R.Liu. Voice conversion with smoothed GMM and MAP adaptation[J]. Proc. EUROSPEECH, 2003: 2413-2416.

[10] G.Y.Zuo, W.J.Liu and X.G.Ruan. Genetic algorithm based RBF neural network for voice conversion[J]. IEEE Transactions on Speech and Audio processing, Vol. 6, No. 2, Mar. 1998: 131-142

[11] C.H.Wu, C.C.Hsia and T.H.Liu. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis[J]. ICASSP, Vol. 14, No. 4, 2006: 1109-1116

[12] K.S.Rao, B.Yegnanarayana. Determination of instants of significant excitation in speech using group delay function[J]. ICASSP, Vol. 3, No. 5, 1995: 325-333

[13] R.Muralishankar, A.G.Ramakrishnan and P.Prathibha. Modification of pitch using DCT in the source domain[J]. Speech Communication, Vol. 42, No. 2, 2004: 143-154

[14] K.S.Lee. Statical approach for voice personality transformation[J]. ICASSP, Vol. 15, No. 2, 2007: 654-651

[15] A.Kain, M.W.Macon. Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction[J]. ICASSP, Salt Lake City, USA, 2001: 813-816

[16] H.Ye, S.Young. Quality-enhanced voice morphing using maximum likelihood transformation. ICASSP, Vol. 14, No. 4, 2006: 1301-1312

[17] 胡航. 语音信号处理[M]. 哈尔滨工业大学出版社. 哈尔滨，2004

[18] 赵力. 语音信号处理[M]. 机械工业出版社. 北京，2005

[19] H.Matsumoto, S.Hiki, S.Sone and N.Tadamoto. Multidimensional representation of personal quality of vowels and its acoustical correlates[J]. IEEE Trans., 1973

[20] T.Galas, X.Rodet. An improved cepstral method for deconvolution of source-filter system with discrete spectral: Application to musical sounds[J]. Proc. of International Computer Music Conference, Glasgow, 1990: 82-84

[21] K.Itoh, A.Saito. Effects of acoustical feature parameters of speech on perceptual identification of speaker[J]. IECE Trans., 1982, J65-A: 101-108

[22] D.H.Mart, L.C.Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers[J]. J. Acoust.Soc.Am, 1990, 87 (2): 820-857

[23] M.Nadrendranath, H.A.Murthy, S.Rajendran, B.Yegnanarayana. Transformation of formants for voice conversion using artifical neural networks[J]. Speech Communication, 1995.16(2): 207- 216

[24] G.Baudoin,Y.Stylianou. On the transformation of the speech spectrum for voice conversion[J]. In

Proceedings of ICSLP96, 2: 1405-1408

[25] 吕士楠,周同春. 用KLATT合成器合成汉语的初步研究. 第二届全国人机语音通讯学术会议论文集[C]. 1992: 281-286

[26] K.S.Lee, W.Doh and D.H.Youn. Voice conversion using low dimensional vector mapping[J]. IEICE Transaction Information&System, 2002

[27] L.M.Arslan, D.Talkin. Speaker transformation using sentence HMM based alignments and detailed prosody modification. ICASSP, Seattle, WA, May. 1998

[28] O.Turk. New methods for voice conversion[D]: [Master degree thesis]. Bogaziqi University, 2003

[29] D.T.Chappell, J.H.L.Hansen. Speaker-specific pitch contour modeling and modification[J]. ICASSP, Seattle, Washington, 1998: 885-888

[30] A.Kain. High resolution voice transformation[D]: [Ph.D. thesis]. OGI school of science and Engineering at Oregon Health and Science University, Oct. 2001

[31] Y.Stylianou. Harmonic plus noise model for speech combined with statistical methods for speech and speaker Modification[D]: [Ph.D. thesis]. Ecole Nationale Suqerieure des Tele communications, Paris, France, Jan. 1996

[32] L.M.Aslan. Speaker transformation algorithm using segmental codebooks(STASC)[J]. Speech communication, 1999, 28(3): 211-226

[33] A.Kain, M.Macon. Voice conversion for text-to-speech speech synthesis[J]. Proceedings of ICASSP, Seattle, Washington, May. 1998: 285-288

[34] Y.Stylianou, O.cuppe and E.Moulines. Stastical methods for voice quality transformation[J]. In Proceedings of Eurospeech, Madrid, Spain, Sep. 1995: 447-450

[35] T.Masuko, K.Tokuda, T.Kobayashi and S.Imai. Speech synthesis from HMMs using dynamic features[J]. Proc of ICASSP, 2000: 937-940

[36] L.E.Baum and T.Petriee. Statistical inference for probabilistic functions of finite state Markov chains[J]. Ann. Math. Star., Vol. 37, 1966: 1554-1563

[37] L.E.Baum, T.Petriee, G.Soules and N.Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains[J]. Ann. Math. Star., Vol. 41, 1970: 164-170

[38] 史忠植. 知识发现[M]. 北京：清华大学出版社，2001

[39] L.Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proc. IEEE, Vol. 77, 1989: 257-286

[40] L.Rabiner and B.H.Juang. Fundamentals of speech recognition[M]. Prentice Hall PTR, 1993

[41] Steve Young, D.Kershaw, J.Odell, D.Ollason, V.Valtchev and P.Woodland. The HTK Book[M]. Entropic Ltd., 1999

[42] F.Itakura. Linear spectral representation of linear predictive coefficients[J]. Journal of Acoustic Society of America, 87(4), 1990: 1738-1990

[43] H.Sakoe, S.Chiba. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Trans. on Acoustics, Speech and Signal Processing, 1978, 26(1): 43-49

[44] 杨行峻, 迟惠生. 语音信号数字处理[M]. 北京: 电子工业出版社, 1995

[45] Y.Linde, A.Buzo and R.M. Gray. An algorithm for vector quantizer design[J]. IEEE Trans. on Communication, 1980, COM-28(1): 84-95.

[46] C.K.Chow. An optimum character recognition system using decision functions[J]. IRE Trans., 1957: 247-254.

[47] H.Kuwahara, I.Masuda-Katsuse and A.deCheveigne. Restructuring speech representation using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds[J]. Speech Communication, Vol. 27, No. 3, 1999: 187-207

[49] 吴义坚. 基于隐马尔科夫模型的语音合成技术研究[D]: [博士学位论文]. 合肥: 中国科学技术大学，2006

[49] K.Tokuda, T.Kobayashi and S.Imai. Recursive calculation of mel-cepstrum form LP coefficients [EB/OL]. http://kt-lab.ics.nitech.ac.jp/tips/mgceptr_sa2.pdf

[50] K.Tokuda, T.kobayashi and S.Imai. Speech parameter generation from HMM using dynamic features[J]. Proc. ICASSP, 1995: 660-663

[51] T.Toda, A.Black and K.Tokuda. Spectral conversion based on maximum likelihood Estimation considering global variance of converted parameter[J]. Proc. ICASSP, 2005: 9-12

[52] A.Mouchtaris, J.V.derSpiegel and P.Mueller. Non-parallel training for voice conversion based on a parameter adaptation approach[J]. IEEE Transactions on Speech and Audio Processing, Vol. 14, No. 3, May. 2006: 952-963

[53] L.M.Arslan. Speaker transformation algorithm using segmental codebooks[J]. Speech Communication, Vol. 28, 1999: 211-226

[54] Y.Stylianou, O.Cappe and E.Moulines. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 2, Mar. 1998: 131-142

作者简介

1. 基本情况：

宋鹏，男，25 岁，籍贯为山东省烟台市。

2. 学习经历：

2006.9—至今： 东南大学信息学院信号与信息处理专业攻读硕士学位。

2002.9—2006.7：山东科技大学信电学院电子信息工程专业取得了学士学位。

3. 研究生阶段的学习成绩：

课程包括 8 门学位课、4 门非学位课，平均成绩为 84 分。

4. 研究生阶段的主要课题：

2007.6—2008.10：从事说话人语音转换算法的研究，对国内外相关说话人语音转换算法的论文进行比较实验，最终提出了改进的基于 GMM 模型的说话人语音转换方法。

2007.3—2007.4： 参与了说话人识别系统的实现，主要负责 GMM 识别算法的改进与实现、相关代码的修改与编写工作。

5. 研究生阶段发表的论文：

Song Peng, Zhao Li. Improving the Performance of GMM Based Voice Conversion Method, IEEE PACIIA, 2008.12 (EI 收录)

宋鹏，赵力. 基于频谱包络的语音转换方法的研究，第 22 届南京地区研究生通信年会，2007.12

6. 研究生阶段的实践活动：

2007.12.4—2008.5.31：在 Microsoft Research Asia（微软亚洲研究院）语音组实习，从事语音信号处理的相关研究，负责语音识别、语音合成、说话人自适应等相关算法的实现、优化和改进等工作。