声　明

本人郑重声明：此处所提交的硕士学位论文《　　》，是本人在华北电力大学攻读硕士学位期间，在导师指导下进行的研究工作和取得的研究成果。据本人所知，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得华北电力大学或其他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。

学位论文作者签名：童波　　日　期：2010.3.14

关于学位论文使用授权的说明

本人完全了解华北电力大学有关保留、使用学位论文的规定，即：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的,复制赠送和交换学位论文；⑤同意学校可以用不同方式在不同媒体上发表、传播学位论文的全部或部分内容。

(涉密的学位论文在解密后遵守此规定)

作者签名：童波　　导师签名：

日　　期：2010.3.14　　日　　期：2010.3.14

摘　　要

特定对象的语音转换系统目的在于在不改变语义的前提下，改变源说话人的语音特征，使其声音听起来更像目标说话人的声音。近年来，对语音转换系统的研究已经成为了语音信号处理中一个非常关键的领域，由于它涉及到很多其他的信号处理技术，如语音识别，语音合成等等，所以对它的研究势必会推动这些领域的发展。本文从汉语语音的音频特征入手，在预处理阶段，对于声道频谱包络特性和基音频率特性提出了两种不同的混合分帧策略，消除了传统定长分帧对汉语语音特点的掩蔽；在训练阶段，采用了以音素为单位对高斯混合模型（GMM）的进行训练的方法，提高了语音信号建模的精度；在转换阶段，采用基于 GMM 和码本映射的混合算法，有效解决了 GMM 转换频谱包络时，过于平滑的问题，提高了转换后语音的质量。

关键词：汉语，语音转换，高斯混合模型，码本映射

ABSTRACT

A Voice Conversion System can change the speech of source speaker into the target speaker's without changing the meanings. Recent years, the research of Voice Conversion System becomes a key area of the speech signal process, because it will improve and push the development of other area of SSP, like speech recognition, speech synthesis, etc. Based on the special feathers of Mandarin, This paper will give a new combined enframing algorithm during pre-processing period, a new GMM training strategy based on Phoneme during training period, and in transform period, a combined method of GMM and codebook mapping algorithm will be used, which will highly improve the quality of the transformed voice.

****** (Signal and Information Process)

Directed by prof. *******

KEYWORDS: mandarin, voice conversion, GMM, code mapping

目　录

第1章 绪论

1.1 语音转换系统的基本概念

语音转换技术是语音信号处理领域中的一个重要分支。对语音转换系统的研究几乎延伸到了语音信号处理的各个领域。在本节中，我们首先介绍有关语音转换系统的基本概念，包括语音转换系统的定义和分类，以及语音转换的基本步骤。

1.1.1 语音转换系统的定义

语音转换（Voice Conversion）系统是指，在不改变语义的前提下，改变源说话人（Source Speaker）的语音个性特征，使其听起来更像目标说话人（Target Speaker）的人的声音[1]。语音信号除语义以外，还包括很多声学个性化信息，例如说话的语音，语调，情绪，韵律等等。语音转换系统主要集中于如何把这些信息模型化，并且实现它们之间的相互转化。研究表明反映语音信号个人特性的因素一般分为两个方面，一个是声学参数，如共振峰位置及其带宽，基音频率等等，这主要是由不同说话人的发声器官差异所决定的。另一方面是韵律学参数，如不同说话人说话的快慢、节奏、口音等等。这和人们所处的社会环境和心理环境有关。

1.1.2 语音转换系统的分类

语音转换系统大致可以分为以下三类：

1，特定对象的语音转换系统，就如语音转换的定义一样，实现语义内容不变而只改变语音的说话人个性特征，使源说话人的语音经语音转换以后听起来像是目标说话人的语音，本文主要讨论这一类语音转换系统。

2，群体对象的语音转换系统，如男声与女声之间的转换，儿童、成年人、老年人语音之间的相互转换等，这种转换是研究语音个性特征的群体特性，并加以转换。

3，广义语音转换系统定义为改变源说话人语音个性特征，使之听起来像是另外一个人的语音。此处只是使语音的说话人个性特征发生改变，并没有确定转换的目标语音，例如像变声器，声音伪装器等。应该说较前两种语音转换，这种语音转换系统有着宽松的语音转换要求。实现难度上也比前两种简单。

1.1.3 实现语音转换的步骤

语音转换系统一般包括以下几个阶段：

1，信号预处理阶段 语音信号是非平稳性信号，及其分布特性随时间的变化和变化，所以对语音信号的处理都是建立在短时的基础上的，使用一定长度的窗函数，可以认为在窗函数内的语音信号是平稳的，所以对语音信号的分帧加窗处理是必不可少的。

2，参数提取阶段 收集源说话人和目的说话人的语音特征，组成特征参数向量，一般情况，特征参数包括两方面信息：声门信息和声道信息。例如，基音周期，共振峰等等。

3，训练阶段 对源说话人和目标说话人的语音特征进行对齐，调整，建立 GMM 模型，计算相关参数和转换规则。

4，转换阶段 根据转换规则和源说话人的语音特征参数，得到新的转换后语音信号的特征参数

5，合成阶段 根据转换后得到的特征参数，合成出新的语音信号。

1.2 语音转换系统的研究意义

语音转换系统有着重要的理论价值和实用价值。对语音转换系统的研究涉及到说话人识别，文语转换，语音编码，语音识别，等各个领域，对语音转换系统的研究势必会影响对这些领域的研究。

1，在说话人识别中的应用：语音转换系统的研究必然要对说话人的语音特性进行建模，分析其语音特征，这些特征也正是构成每个人说话差异的原因。所以语音转换系统中对说话人语音特性的研究和建模势必为说话人识别领域的研究，提供了重要的理论和实验依据。

2，在文语转换（TTS）系统中的应用：在文语转换系统中，根据文本合成的语音特性往往是单一的，缺乏自然度。可以把 TTS 合成的语音通过一个语音转换系统，将其转换成特定人的语音，这样就减少了文语转换系统中的大量繁重的音频录制工作。

3，在语音编码中的应用，通过语音转换系统，可以实现极低速率的语音编码。前提是文本的内容已知，把说话人的语音特性和文本内容，先通过一个 TTS 系统进行语音合成，然后再通过一个语音转换系统，传换成原来说话人的语音。

4，在语音识别中的应用：在语音转换系统中，对于谱包络的参数的提取和转换是非常重要的一项工作，而在语音转换系统中，也需要把一个类似的自适应过程。

在语音识别的过程中，可以先通过一个语音转换系统，达到语音信号自适应的要求，在通过语音识别的系统，可以达到良好的识别效果。

5，在其他领域的应用：语音转换系统不仅在上述领域，在军事，医学，通信，航海，外语学习等领域中都有应用。

1.3 语音转换系统的研究背景及国内外动态

1.3.1 语音转换系统的研究背景

随着国内外科学技术的发展，人们在于计算机交互的过程中，发现需要一种更快能直观的方式来获取信息。而语音和图像是人类获取信息最自然，也是最有效的工具。因此对于音信信号的研究越来越引起人们的重视。语音转换系统正是其中之一。

对于语音转换系统，一些国内外学者都做了大量的工作，但是总体来说，由于国外对语言转换系统研究的实际较长，而国内时间较短，所以国内的研究状况还是落后于国外。

1.3.2 语音转换系统研究的国内外动态

1988 年 Abe 等人采用矢量量化的方法进行语音转换，但是不可避免的出现了转换语音不连续、细节丢失严重等问题。[1]1991 年 H.Kuwabara 采用分析-合成方法来进行语音转换，研究影响语音个人特征和话音质量的声学参数，他通过线性预测解卷积算法将语音信号分解为噪音源信号和声道传输函数，通过求解声道传输函数的零点得到共振峰的位置信息，改变共振峰的位置和宽度，采用线性预测的残差信号作为激励来合成语音 [2] 。同年 Childers 等人采用基于固定长度帧的非基音同步的方法和与信号相关的方法来进行分析，通过线性频谱搬移和分析合成的方法，进行了男女音之间的语音转换，取得了较好的效果 [3] 。1997 年 Arslan 等人采用对线谱频率和激励频谱的码本映射的方法，对语音信号进行转换 [4] 。1998 年，Chapell 等人提出了一种对基于说话人特点对其基音周期轨迹进行建模的算法 [5] 。1998 年，Kain 等人采用 GMM 法来对谱包络参数进行建模转换方法 [6] 。Stylianou 等人在研究语音转换系统时使用了一种连续概率模型对语音信号进行建模 [7] 。Shikano 在 2001 年提出了基于具有动态频率规整的 GMM 算法进行语音转换，还提出了基于 GMM 的转换频谱与频率规整频谱之间的差异，这可以避免说话人个性上转换准确性的下降 [8] 。2006 年 Yamagishi 等人在对语音合成的研究中提出了自适应参数的

HMM 模型，提高了合成语音的准确度和真实感[9] 。国内有关语音转换的研究起步比较晚一些，主要的研究也都集中在时域。1996 年初敏等人采用 TD-PSOLA 方法进行男女语音转换，频谱包络特性的转换是通过重采样的方法来实现的 [10] 。2000 年刘立采用矢量量化的方法，进行了男女语音转换的研究，其基音频率的变换通过把一个基音周期内部的语音信号幅度最小的一部分截去或添加来实现，频谱包络特性的变换是通过 DTW 技术从矢量库里寻找一定宽度的半波波形替换原语音信号中的半波波形来实现 [11] 。2006 年，康永国等人提出一种高斯混合模型和码本映射相结合的语音转换算法。很好的解决了 GMM 转换后，频谱过平滑的问题 [12] 。

1.4 目前语音转换系统的研究中存在的问题

尽管在过去的 20 年中，国内外很多学者在语音转换方面做了大量的工作，但是语音转换质量和效果依然不能令人满意，主要有几方面原因：

1，现今对于语音转换系统的研究，主要集中于语音的音段特征，因为很难对超音段特征建立数学模型，所以一般的转换系统对于超音段特征，都采取平均值转换，而超音段特征对于语音来说往往又是非常重要的语音特征。所以转换出来的声音往往缺乏真实感。

2，在语音转换阶段，对语音进行韵律调整，也会导致语音质量的下降，引起失真。例如对基音周期的调整，当基音周期改动较大时，会使语音信号的其他特征也跟着改变，这就导致转换后的语音缺乏真实感。

3，在声道相应的转换算法中，矢量量化法会引起不连续现象，GMM 法能较好的克服这种不连续现象但却又引起共振峰的平滑现象。这也导致了语音质量的下降。

在以后进一步的研究中，需要解决的一些关键性的问题如下：

（1）对于超音段特征模型的研究。所谓超音段特征是语音信号的动态特征，这些特征在说话人每次说时都有不同的表现，无法通过训练进行量化或比较。例如时长的变化、能量的变化，基音周期的变化、以及谱包络的变化等等。而这些超音段特征也从很大程度上反映了说话人人的个性特征。

（2）对于中文语音特征的研究，中文语音和英语不同，由于中文发音有字正腔圆的特点，所以在语音转换系统中，可以通过对中文语音特点的学习，在特定情况下，对中文语音指定高效的，准确的转换和合成方案。

（3）对于提高合成语音的质量的研究，目前的语音转换系统合成的语音都有不自然的缺点，如果提高合成语音的质量是语音转换系统中，需要解决的一个重要问题。

1.5 本文的主要工作

本文的主要工作是实现一个特定对象的语音转换系统，并对于传统语音转换系统中的算法都做了相应的研究和改进。主要内容如下：

1，在语音信号的预处理阶段，对语音信号的传统分帧算法做了比较和讨论，针对汉语语音的特点，提出根据音节、音素及基音周期的变化，采用不同的非定长动态分帧策略，提高了基音周期的检出概率。在音节内部采用基于 MFCC 短时相关性分析的方法，识别并记录音节类型，实现了汉字和其音频特征的对应动态识别和存储。

2，在模型训练阶段，提出基于汉语音素的训练模式，减少了 DTW 对齐的误差和噪声对 GMM 协方差矩阵的干扰，在对基音频率进行建模时，只用元音进行训练。提高了模型训练的正确度和充分度。

3，在语音的转换阶段，使用了 GMM 和码本映射的混合转换算法，由于 GMM 模型对声道频谱包络的描述不够细腻，所以在其基础上加入了可调整的码本映射的算法，提高了转换的精度。

1.6 本章小结

在本章中，我们简要地介绍了语音转换系统的分类、定义、相关国内外研究背景及其研究意义。初步的对语音转换系统有了大致的了解。并且确定了本文的研究目标和主要工作。下一步我们要了解实现语音转换系统所需要的理论知识，包括语音信号发声的机理和它的数学模型、汉语语音的特点、及其相关的处理分析方法和工具等等。

第2章 语音转换系统的基础理论

从本章开始，我们将从三个方面对语音转换系统的理论基础作详细的介绍：语音信号的基础知识，语音信号的处理方法，语音转换系统的基本模型。

语音转换系统的前提和基础是对语音信号进行处理，只有分析出可表示语音信号本质特征的参数，才有可能利用这些参数进行高效，准确的识别，转换，合成。转换的音质好坏，语音识别率的高低，也都取决于对语音信号处理的准确性和精确性。

2.1 语音信号处理基础

这一节中，我们将从二个方面介绍语音信号处理的基础理论，一方面是物理和声学特性，包括发声机理，语音特征，数学模型等。另一方面是语音信号处理中常用到的数学工具。包括时域分析，倒谱分析，线性预测等。

2.1.1 语音信号的定义及其发声机理

语音信号（Speech）是人们讲话时发出的声音，它区别于声音信号（Sound）和乐音信号（Music）的主要特点是语音信号带有语义，是人们可以理解的。所以说语音信号是声音和语言的结合体。

语音信号产生过程如图 2-1 所示，可以具体描述成以下几个过程：当一个人想要说时，在大脑里首先会出现想要说的词汇，语句，神经系统会产生电信号刺激发声系统。气流首先从肺里喷出，通过声带的震动，变成有一定变化频率的气流，并加载一定的噪音，再通过口腔，鼻腔等腔体的共振，使其周围的空气发生震动而变成我们所听到的声音。因此可把人体的发声器官概括为两个部分：声门系统和声道系统。声门主要指喉部，由于气流通过喉部时不断地周期开合，才产生语音信号的激励；声道又分成主声道，鼻声道和此声道，可以把声道理解成一个时变的滤波器，激励源信号通过滤波器产生我们听到的语音。

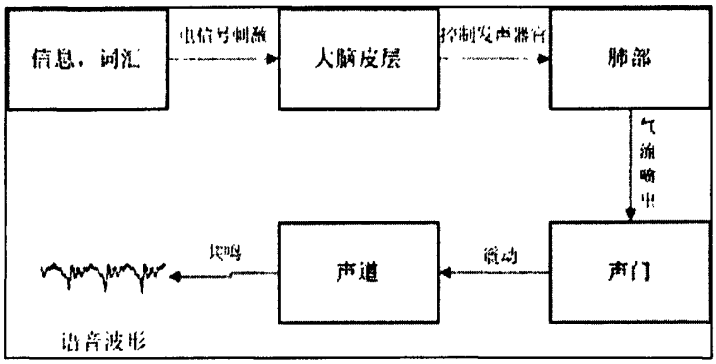


图 2-1 语音信号的发声过程

从人体发声的机理来看，使用基于声源-滤波器的语音模型来描述语音信号的产生是非常有效的，在这种模型下，声音源及激励波通过时变的滤波器及声道腔体产生语音信号。它能解释大多数的语音现象。从而使对语音信号的研究分别转换为对声源激励和声道时变滤波器两个部分的研究。

2.1.1.1 语音信号的基本声学特征

语音（speech）是语言（word）和声音（acoustic）的结合体[13]。语言代表语义，声音是语言的载体。语音信号是人类的发声器官周期性振动，而发出的声波。人们说出的语音，是由于肺部气流推动，声带振动，和口腔，鼻腔等腔体共鸣等一系列复杂的生理活动，相互作用而形成的。它具有一定的音色，音调，音强和音长。如图 2-3 所示，语音是由句子组成，当说一个句子时，可以明显被感觉到的语音片段被称为音节（Syllable），每个音节又是由一个或多个音素（Phoneme）组成的。音素是发音的最小单位，音素分为元音（Vowel）和辅音(Consonant)。其中元音是顺畅的气流通过声带振动而发出的。辅音则是由于气流受到阻碍而发出的。辅音又分为清辅音和浊辅音，其中清辅音发声时，声带不震动；浊辅音发音时声带振动。

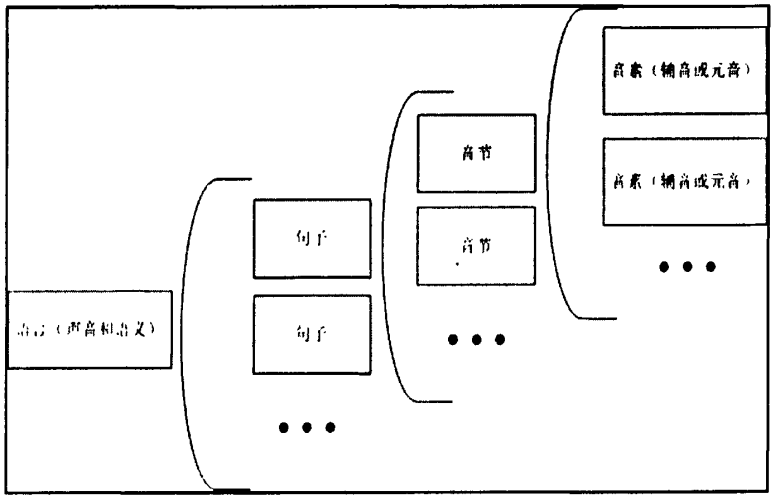


图 2-2 语音信号的基本组成结构

从时域上来说，语音信号具有很强的时变特性。这表明语音信号的频谱特性随时间的变化也会很大。图 2-3 显示了一段已经归一化的语音信号时域波形，其采样频率为 10kHz，16bit 量化：

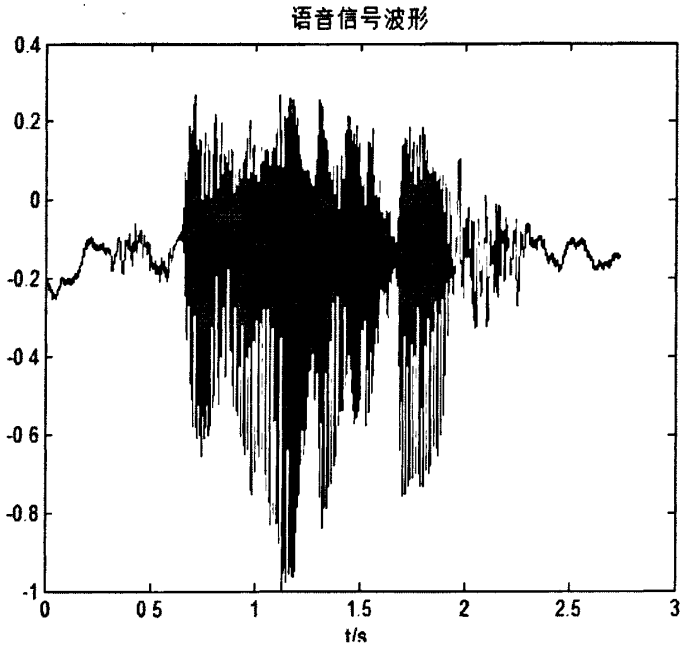


图 2-3 语音信号“语音文学”的波形图

语音信号的另一个重要特征就是它的非平稳性，但是不同语音是由人的口腔肌肉运动构成的声道的某种形状而产生的响应，而这种肌肉运动的频率相对于语音频率来说是相对缓慢的，因而再一个短时间范围内，其特性是基本稳定的。正是由于这个原因，“短时分析技术”贯穿于语音信号分析的全过程。图 2-4 显示了上述语音信号的 512 点快速傅里叶变换波形图：

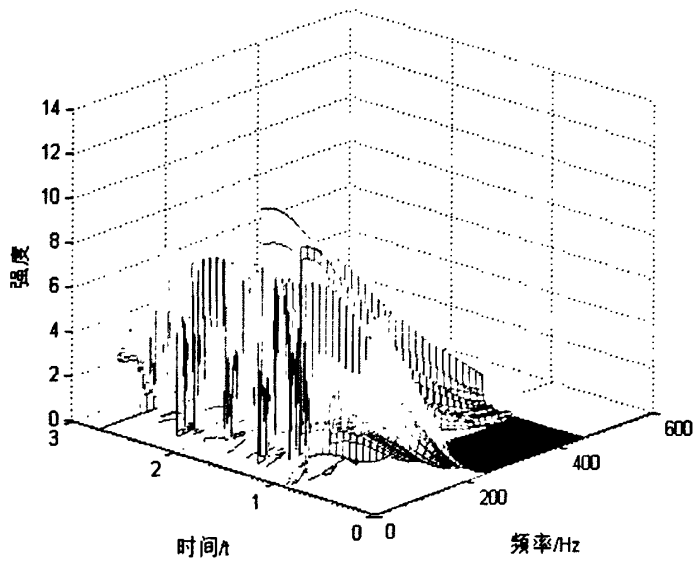


图 2-4 语音信号“语音文学”的快速傅里叶变换波形图

2.1.1.2 语音信号的个性声学特征

语音信号的个性声学特征，是我们区别不同说话人的重要语音特征[14]。语音转换系统主要对个性声学特征进行转换。他主要分为以下几类：

1，音段特征 音段特征主要描述了语音信号的音色特征。主要包括共振峰的位置、共振峰的带宽、基音频率、能量等。音段特征主要与发音器官的生理学和物理学特征有关。其中共振峰（formant）是元音的重要特征，它表示了声音能量比较集中的位置。它表征了语音信号在通过腔体共鸣时，在这些共振峰频率附近信号会被增强。所以每个元音都对应一组不同的共振峰位置及其频带宽度。基音频率是元音和浊辅音共有的一个语音特征参数，它表示声带振动的基本频率。清辅音因为发音时声带不震动，所以不存在基音频率。

2，超音段特征 描述的是语音的韵律特征。主要包括音素的时长、基音频率的变化规律等，这些特征受社会的和心理的环境影响。

3，语言特征 主要包括习惯用语、方言、口音等。

2.1.1.3 语音信号的数学模型

一个完整的语音信号模型是由三部分组成的：激励模型，声道模型，辐射模型。目前大部分的语音模型都是属于声源-滤波模型。一个语音信号的传递函数可以表示成：

$$H(z) = AU(z)V(z)R(z) \qquad \text{公式（2-1）}$$

其中 A 表示增益系数，U（z）表示激励信号，激励信号的表示在发清音和浊音的情况下是不同的，发清音时，U（z）表示一个随机白噪声，发浊音时，由于声带的不断张开和关闭，U（z）可模拟成一系列声门脉冲。如果假设脉冲波类似于斜三角波，那么 U（z）可以表示为：

$$U(z) = G(z)E(z) \qquad \text{公式（2-2）}$$

其中 $E(z) = \frac{A}{1-z^{-1}}$ 为脉冲的激励的传递函数，$G(z) = \frac{1}{(1-e^{-cT}z^{-1})^2}$ 为单个斜三角波的传递函数。

V（z）表示声道传递函数，一般用级联方式表示，在 LPC 模型中，假设声道是由一系列的管子连接组成的。使用共振峰模型来描述，每个管子为一个一阶极点模型，则整个声道就可以表示成一个全极点模型。它的传递函数为：

$$V(z)=\frac{G}{1-\sum_{k=1}^{N-1}a_k z^{-k}} \qquad \text{公式（2-3）}$$

其中 G 为振幅因子，a 为极点的常系数。

R（z）表示辐射模型，辐射模型的建模基于假设口唇张开的面积远小于头部的表面积，则可以近似成平板开槽的辐射情况，那么可以推到出辐射阻抗的公式如下：

$$Z_L(\Omega)=\frac{j\Omega L_r R_r}{R_r+j\Omega L_r} \qquad \text{公式（2-4）}$$

然后使用数字滤波器设计的双线性变换方法将上式转换成 z 变换形式：

$$R(z)=R_0\frac{1-z^{-1}}{1-R_1 z^{-1}} \qquad \text{公式（2-5）}$$

由公式(2-1)可知，语音信号数学模型可以表示为激励模型，声道模型，辐射模型的级联。它的整体数学模型如图 2-5 所示。

需要说明的是，上文中描述的语音模型只是一个简单的声源-滤波语音模型，真正语音产生的机理要复杂的多。首先来说气流通过声门是不肯能是匀速模型，并且只用开合的频率来描述声门信息是远远不够的，声道包括口腔，鼻腔等多个腔体，且去形状各异，发音时不断变换，只用一系列极点模型表示显然是不足的[15]。

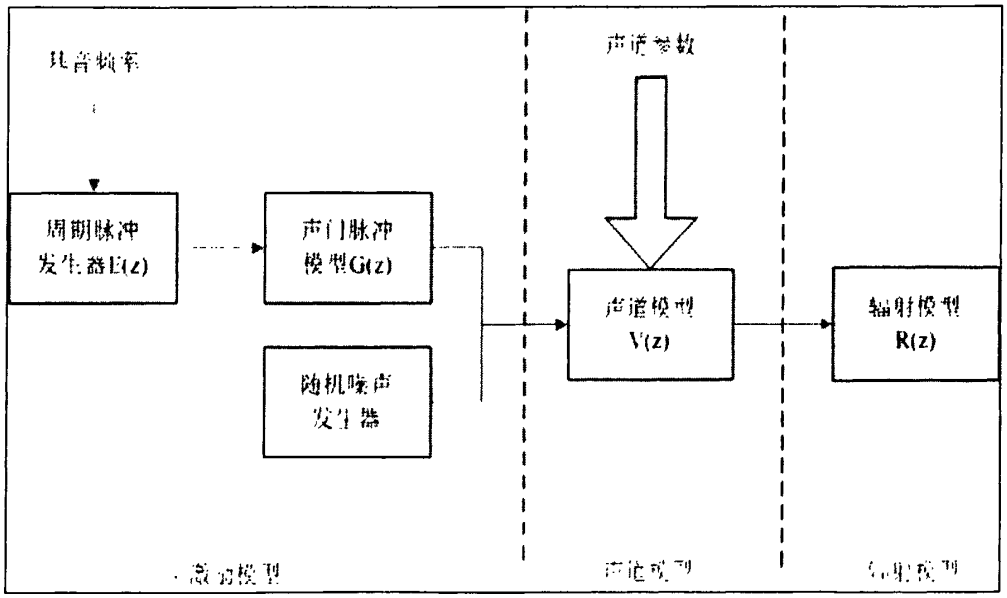


图 2-5 语音信号产生的数学模型

2.1.2 语音信号分析的常用方法

2.1.2.1 语音信号时域分析

语音信号的时域分析方法包括短时能量分析、短时过零率分析，短时相关性分析等等。语音信号的时域的特点是：信号比较直观，计算量小。以下介绍几种常用的时域分析方法。

1，语音信号的短时能量是指一个语音帧的能量大小，短时能量分析，主要用于区分清音浊音，或者区分明显的语音边界。在一个语音帧内，它的短时能量可表示为：

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \qquad \text{公式（2-6）}$$

2，语音信号的短时过零率，表示一个语音帧内，信号符号变换的次数。也常用于区分清音浊音，以及有声阶段和寂静阶段，因为清音有着较高的频率，和浊音的频率较低。所以清音的过零率高，浊音的过零率低。

3，短时相关性分析，一个语音信号的短时自相关函数可以表示为：

$$\sum_{m=0}^{N-1-k} x_n(m)x_n(m+k) \qquad \text{公式（2-7）}$$

它可以用于检测语音信号的基音周期。自相关函数有着非常好的抗噪性。

2.1.2.2 语音信号的倒谱域分析

倒谱分析是语音信号中最常用的分析方法，倒谱实际上是一个信号功率谱对数形式的逆傅里叶变换。计算过程如图 2-6 所示。

倒谱所表示的是一个信号，在不同频带上变化速率的信息，最早应用于地震回声和对爆炸声的分析。通常应用上，使用倒谱的自相关系数，因为它比倒谱更能显示信号的特征[16]。

倒谱的一个更典型的应用时用在同态信号的分析上，对于语音信号的声源-滤波器模型中，为了分别提取声源和滤波器的特性，利用倒谱解卷积，把声源特性和滤波器特性映射到线性叠加的倒谱域中。

正是因为倒谱的以上特性，所以它还可以用来检查语音信号的基音周期，如图 2-7 所示，低频部分的波峰突击表示是基音周期的位置，可以很明显地看出女声的女基音周期要小于男声的基音周期。

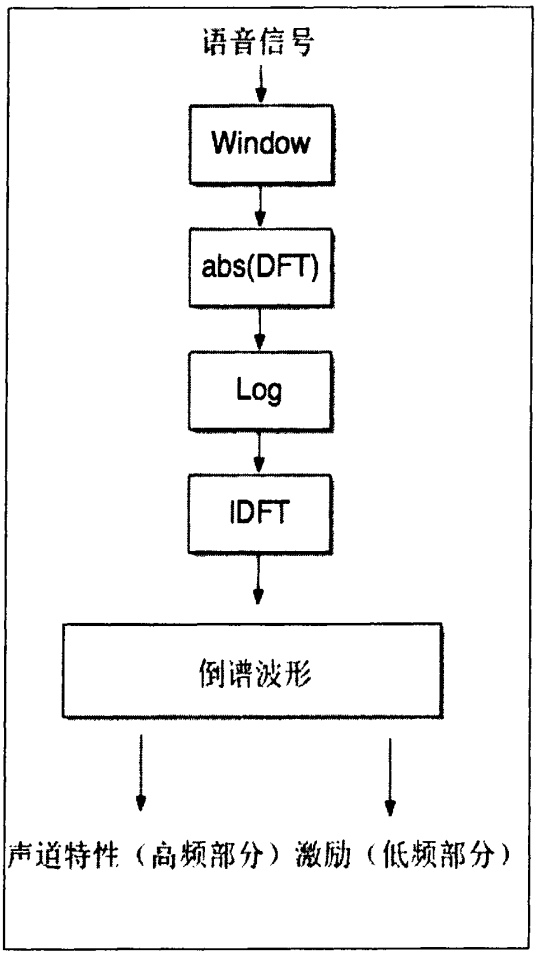


图 2-6 倒谱的计算流程图

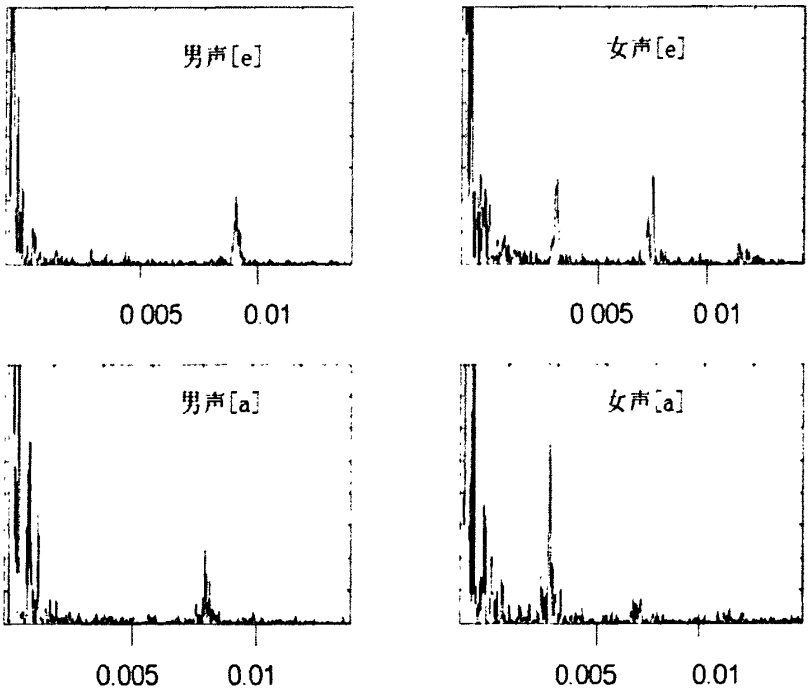


图 2-7 男女声[e][a]的倒谱基音检查

Mel 倒谱系数编码（MFCC）是一种常用的语音信号编码形式，他是把信号的功率倒谱映射到 mel 域，因为 mel 域低频细分，高频粗略的特点，充分考虑了人的听觉特性，且语音信号的特征主要集中在低频部分。所以常用于语音识别，语音编码及其特征提取等等。图 2-8 显示一个语音片段的 12 维 MFCC 的波形。频域到 mel 域的转换公式如下：

$$Mel(f) = 2595\lg(1 + f/700) \qquad \text{公式(2-8)}$$

要得到 MFCC 参数首先要在语音帧的 Mel 频域上划分出一系列的三角形滤波器序列，再对每个滤波器再与其频带内的信号幅度加权和作对数运算，最后一步通过离散余弦变换（DCT），得到 MFCC。计算公式如下：

$$c_{mfcc}(i) = \sqrt{\frac{2}{n}}\sum_{l=1}^{L}\log m(l)\cos\{(l-\frac{1}{2})\frac{i\pi}{L}\} \qquad \text{公式(2-9)}$$

其中 m(l)是信号幅度和滤波器的加权和：

$$m(l) = \sum_{n=s(l)}^{e(l)} W_l(k)\left|X_n(k)\right| \qquad \text{公式(2-10)}$$

其中 s(l)、e(l) 分别表示三角形滤波器的起点和终点。

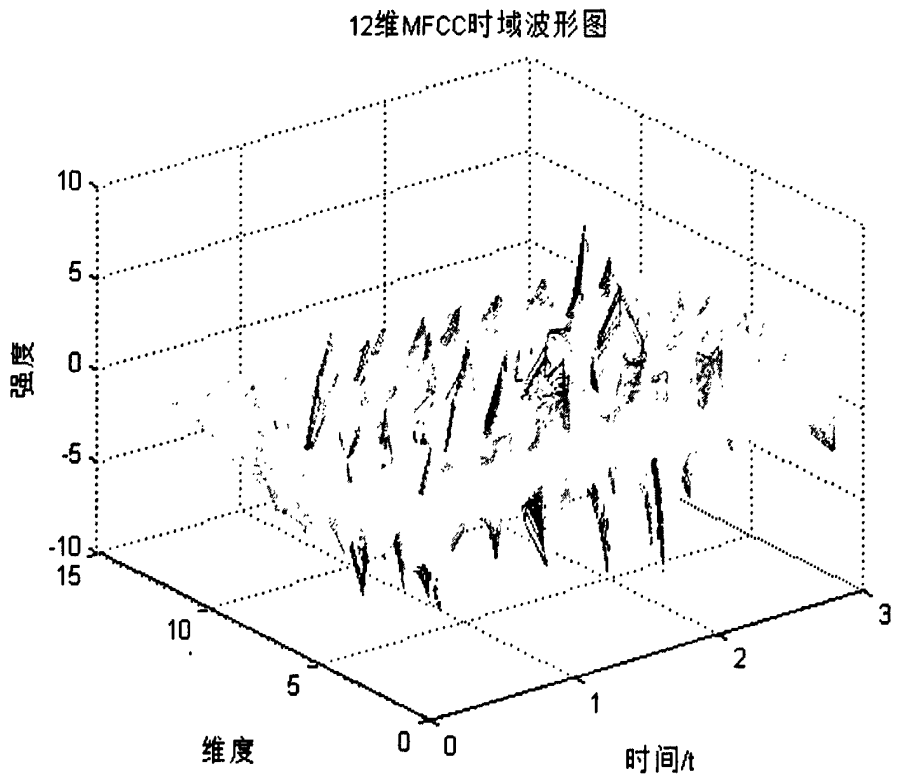
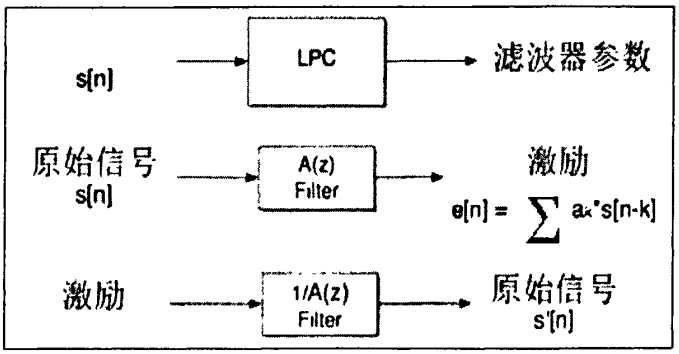


图 2-8 语音“语言文学”的 12 维 MFCC 波形图

2.1.2.3 语音信号的线性预测

语音信号的线性预测编码是对语音信号产生系统的一种近似模拟。在声源-滤波器模型中，声门产生的激励信号可以用响度和频率表示，声道则使用它的共鸣特性来表示，比如对于共振峰频带上信号的增强，还有一些齿擦音和破擦音是由舌头，嘴唇等的活动引起的。

LPC 语音信号分析的整个过程如图 2-9 所示，实际上就是通过对原始信号共振峰的位置（滤波器参数）的估计，并把共振峰对信号的影响，从信号中除去，并估计剩余信号的强度和频率，这个过程被称作逆滤波，剩余的信号减去滤波器模型的信号，被称作残差。利用这些残差就可以建立一个激励模型，利用估计的共振峰参数可以建立一个滤波器模型，这就形成了对语音信号的估计。



图片 2-9 语音信号线性预测过程

2.2 汉语语音的特点

2.2.1 汉语语音的基本特征

汉语语音的标准发音是北京语音，又称普通话。本文所提到的汉语语音转换系统主要指普通话语音之间的转换。汉语是世界上最大的语种，使用人口达到几十亿。汉语的音节结构有很强的规律性，主要特点有以下几点：

1，汉语以汉字为单位，一个句子由多个字组成，每个字又对应一个音节。中国传统上把一个音节分为声母、韵母和声调 3 部分。音节与音节之间的边界清晰明确。汉语的音系相对简单，大约有 60 个音素，400 多个音节 5 种音调。

2，每个音节都以元音为主，汉语讲究字正腔圆。一个字的发音中必然会有一个元音。且一个音节内部最多可以连续出现 3 个元音。以元音为中心，会有辅音和鼻音与之组合。

3，汉语是有调语音，每种语调对应一种基音周期变化规律。分为“阴、阳、上、去、清”五种声调。

4，汉语有着鲜明的轻重感和儿化音。使得语气活泼，语义明确，感情流露。

2.2.2汉语语音的拼音分类

汉语以字为单位，每个字是一个音节，一个音节一般上有2到3个音素组成，一个音素可能是元音或是辅音。汉语是用拼音来标注发音的。按照辅音和元音，分成声母和韵母。其中声母根据发音方式的不同又可以分为6大类：摩擦音，塞音，塞擦音，边音，鼻音，零声母。总计22个。韵母可以分为3个大类：单元音，复合元音，复合鼻元音。总计38个。声母和韵母的表示分别如下表所示：

表2-1 汉语声母韵母表

声母	摩擦音	塞音	塞擦音	边音	鼻音	零声母
	[f]、[h]、[s]、[sh]、[x]、[h]	[b]、[d]、[g]、[p]、[t]、[b]	[zh]、[z]、[j]、[ch]、[c]、[q]	[l]	[n]、m]	[i]、[u]、[y]
韵母	单元音		复合元音		复合鼻元音	
	[a]、[o]、[e]、[i]、[u]、[ü]		[ai]、[ei]、[ao]、[ou]、[ia]、[ie]、[iao]、[iou]、[ua]、[uo]、[uai]、[uei]、[üe]		[an]、[en]、[ang]、[eng]、[ong]、[ian]、[in]、[iang]、[ing]、[iong]、[uan]、[uen]、[uang]、[ueng]、[üan]、[ün]	

2.2.3汉语的音节结构

根于汉字发音方式的不同，汉字的音节结构种类大致可以分为以下4种：

1，元音形式(V) 汉语中一个元音就可以组成一个音节，例如：啊，鹅。

2，辅音+元音形式(CV) 汉语中最常用也是最多的音节形式，例如：他，你。

3，元音+鼻音形式(VG) 例如：昂。

4，辅音+元音+鼻音形式(CVG) 例如：梦，零。

我们把汉语的音节结构归类的目的在于，根据不同的音节结构，采取的不同的参数提取策略。例如，清音不存在基音周期，则在对音节提取基音周期轨迹时，就可以忽略掉清音部分。

表2-2 汉字拼音结构

模式	子模式	意义	举例
V		元音	[a]、[e]
CV	C1V	清辅音+元音	[shi]、[cha]
	C2V	浊辅音+元音	[pa]、[bo]

CVG	C1VG	清辅音+元音+鼻音	[shang]、[cheng]
	C2VG	浊辅音+元音+鼻音	[peng]、[meng]

2.2.4 汉语语音的音调

汉语是一种有调语音，相同的声母和韵母，组合不同的声调，可以表示不同的音节。声调的变化实际上就是浊音中基音周期的变化，各个韵母段中，基音周期随时间变化的曲线，称为声调曲线。不同的声调对应不用的声调曲线，单独说一个音节时的声调曲线，如图 2-10 所示：

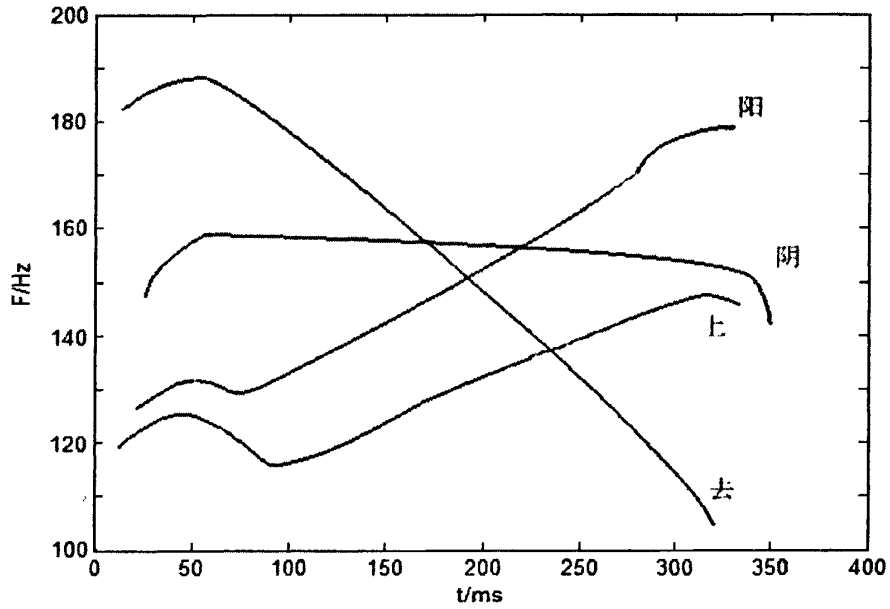


图 2-10 汉语中 4 种语调的基音频率轨迹（男声）

2.3 语音转换系统的建模

本文中语音转换系统的模型建立是基于特定对象的。至于非特定对象的语音转换系统的建模只是再训练阶段上较之简单。我们将从以下两个方面来介绍语音转换系统的模型建立。

2.3.1 语音转换系统的基本模型

特定对象的语音转换系统如图 2-11 所示，一般分为两个阶段，训练阶段和转换阶段。在训练阶段，主要是分别提取源说话人和目标说话人的语音特征参数，建立并训练 GMM 模型，在两组参数中建立一套匹配规则。在转换阶段，则提取源说话

人语音的特征参数，根据匹配规则，转换语音参数，最后合成出目标说话人的语音。

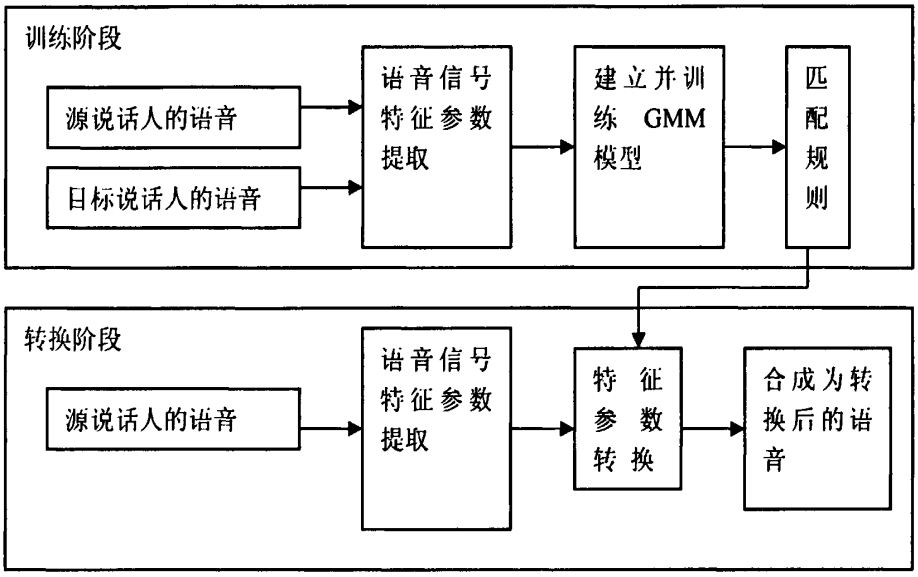


图 2-11 特定对象的语音转换系统模型

2.3.2 语音转换系统工作流程

语音转换系统的基本工作流程如下：

1，模型参数提取阶段：使用声源-滤波（source-filter）模型。将语音信号分解成声源激励和声道滤波两个部分。再通过解卷积的方法，对他们进行分别的处理。常用的语音信号模型有倒谱包络模型和线性预测（LPC）模型。LPC 模型是应用较多的语音参数模型，它也可以有效地把语音信号分解成激励部分和谱包络部分。谱包络部分由 LPC 系数表示，激励部分由 LPC 参差表示。对于谱包络部分，由 LPC 系数得到的推演参数 LSF 可以与频谱包络的共振峰很好的对应。对于激励部分，可以对 LPC 参差进行韵律的转换，以达到提高转换质量的目的。

2，训练阶段：先对提取出来的源说话人和目标说话人的语音特征参数在相同字词上进行对齐，常用的方法有，动态时间规整（DTW），再对高斯混合模型进行训练，建立一套参数对应转换的规则 [17] 。并计算相对应的转换函数。

3 转换阶段：动态提取源说话人的语音特征参数，根据第二部得到的转换函数，把相关特征向量带入到转换函数中，得到新的语音的特征参数的概率模型。

4，合成阶段：通过训练阶段得到的转换规测，转换得到一系列语音特征参数，再根据这些参数合成出最终的语音信号。

2.4 本章小结

本章中，我们介绍了在语音转换系统中经常用到的语音信号几种基本处理方式，常用的语音信号处理方式包括时域处理，倒谱处理，线性预测等。并且对汉语语音的发音特点和音节的结构组成，也做了相关的介绍。最后对语音转换系统的建模做了简要的介绍。下一章中，我们将汉语语音的特有语音特征用于我们的语音转换系统中，并详细介绍语音转换系统的各部分所用的算法和实现原理。

第3章 语音转换系统的实现原理

3.1 语音信号预处理

对语音信号进行预处理的目的在于，一方面语音信号是非平稳的信号，任何对语音信号的处理都建立在短时的基础上。所以首先要选择适当的窗函数对语音信号进行分帧处理。另一方面，由于采集语音信号的过程中必然会加入噪声，所以对语音信号进行预处理，也包括语音信号预增强、滤波、去噪等技术。本文中讨论的语音信号预处理技术主要指语音信号的分帧算法。

3.1.1 语音信号动态分帧算法

对语音信号进行动态分帧包括两方面含义，如图 3-1 所示，一方面是基于汉语语音的特征，按照字（音节），音素的二分策略进行分帧。这样做的目的，是为了更好的让音素进行对齐，提高训练 GMM 时的正确度。另一方面是指在基音频率提取是使用一种动态策略来选取帧长，以提高基音检出的概率，同时也提高了语音信号韵律转换的精度。

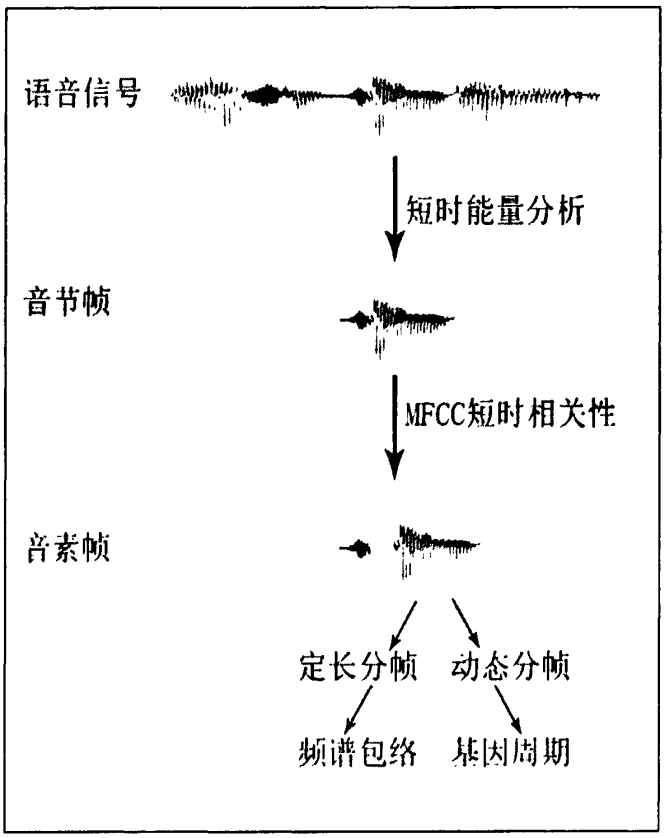


图 3-1 语音信号动态分帧含义及过程

3.1.1.1 动态分帧算法总述

语音信号分帧是语音信号预处理的基础步骤。在分析语音信号时，先要对语音进行加窗分帧处理，一般取帧长为 10ms～30ms 之间。传统的分帧算法都采取固定帧长。固定帧长的优点在于便于计算和存储，但是其缺点也显而易见，在语音信号的加窗操作中，窗函数长度的选取及帧长的选取是其中至关重要的问题，若帧长 N 取的太大，波形的高频部分被阻止，短时能量随时间变化慢，不能真实的反映语音信号的幅度变化；若 N 取得太小，则信号变化太快，不能得到平滑得能量函数 [18] 。窗长得选择更重要得是要考虑语音信号的基音周期，只有合适的帧长，才能提取出正确可靠的基音周期，而基音周期又是语音信号中相当重要的特征参数。通常在一个语音帧内包含 3－7 个基音周期即可，然而不同人的基音周期变化很大，即使是同一个人在不同的情态下基音周期也不一样，加之基音周期还受到单词发音音调的影响。所以使得 N 的选取十分的困难 [19] 。

不仅如此，对于汉语语音来说定长分帧，显然浪费了很多汉语语音特有的特征参数。利用汉语音节边界明显，且音节格式固定等特点，本文提出了一种混合型的动态分帧算法，在谱包络和基音频率的转换上采用不同的分帧算法，步骤如下：

1，在谱包络的转换上，首先采用短时能量分析法，对整体语音片段进行端点检测，找出语句中音节与音节间的端点。再通过分析他们谱包络的相关性识别出其汉语结构模型，并记录他们各自的长度。

2，在基音频率的转换上，基于上一步的转换，采用一种基于动态规划的平滑滤波迭代分帧算法，只对每一个音节的浊辅音，元音和鼻音计算基音周期轨迹，提高了基音周期的检出率。

3.1.1.2 基于短时能量分析的语音信号端点检测

本文中我们采用短时能量分析来检测音节的端点，由于汉语语音是以汉字单位的，所以汉语的音节之间有着比较明显的边界，且在边界处语音信号的能量很小，利用汉语语音的这个特点，我们采用短时能量分析，并选取一个能量阈值，就可以很快的得到音节的边界参数。该算法在实际应用中也取得了较好的效果。它的优点是计算量较小，算法速度快。我们把这样得到的语音帧称为音节帧。

3.1.1.3 基于 MFCC 短时相关性的汉语语音类型识别

该算法也是基于汉语语音的以下特点：在一个音节内，基于因素的谱包络变化有一定规律。两个相邻的不同音素，从元音跳变到辅音或不同的两个元音跳变时，谱包络有较大变化；在一个音节中，浊音与元音结合时，谱包络变换不大；轻音与

元音结合时，谱包络从轻音转到元音时，变换明显；普通元音和鼻音结合是也有明显的谱包络变化。本文把汉语音节分为可以识别的四种：C1V 型，C1VG 型、C2V 型和 C2VG 型。其中 C1 代表清辅音，C2 代表浊辅音，V 代表元音，G 代表鼻音。特别需要注意的是 C1 是不必须要出现的。及 V 型和 VG 型包含在 C1V 和 C1VG 型之中。

通过对一个音节中谱包络激变端点的位置和次数来建立类型识别模型。模型的具体描述如下：

1，一个语音帧内没有激变端点，则认为是 C1V 型。不在对音节进行再分。

2，一个语音帧内只有 1 个激变点，如果这个激变点在语音帧的前半段，则认为是 C2V 型，如果这个激变点在语音帧的后半段则认为是 C1VG 型。这样把音节再分成 2 帧。

3，一个语音帧内有 2 个激变点，那则认为是 C2VG 型。这样把音节再分成 3 帧。

传统的端点检测方法，例如短时能量谱和短时平均过零率分析，他们的优点是算法效率较高，但是在单音节内部对端点的识别率很差，抗噪性也不好，只有在信噪比较高且语速较慢的情况下，才能取得比较好的效果。而 MFCC 能很好的反映了谱包络的特性，计算也比较方便，而且互相关特性的抗噪性较好。选取二者的结合作为音节内端点检测的工具。其中互相关系数定义为：

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{DX}\sqrt{DY}} \qquad \text{公式(3-1)}$$

它的物理意义在于相关性系数越小，则说明相邻两帧之间线性相关程度越小，及变化越大，就可能是我们要检测的语音端点。

算法具体步骤如下：

1. 首先滤除语音信号直流和高频部分，对音节定长分帧加窗，窗函数选取汉明窗。为了能够更好的看清变化，帧与帧之间采用 0%堆叠。
2. 对于每一帧语音信号，我们选择 12 维的 mfcc 参数，24 个滤波器组计算其 mfcc 参数。
3. 对每一帧的 mfcc 参数数据，对相邻帧做相关性分析，求其相关系数曲线。
4. 根据得到的互相关系数曲线，选取阈值，提取端点。

基于以上算法检出的帧，我们称之为音素帧。

3.1.1.4 基于基音周期的语音信号平滑滤波迭代分帧算法

基于上一节的算法，只对音素帧内存在基音周期的部分进行二次分帧。本文使用了一种基于基音周期变化的动态分帧算法。使得一个语音帧内基本上可以确定包含一定数目的基音周期。而其可以认为在一个存在基音周期的音素帧中，基本不存

在基音周期的跳变，一般可以认为帧与帧之间的基音周期变化是连续的。[19]根据以上特点，可以提出以下 3 种算法：

1，简单迭代算法

假定对于一段长度的语音片段 s(n)为了保证其平稳性，其长度 n 不应该超过(30ms)来说存在一个最佳的基音周期长度 T_p，利用 T_p 可以算出一个最佳帧长 $Np=\partial(T_{n-1})$，为了求出 Tp，我们可以先选取一个帧长初始值 N_0，根据基音周期检出的算法 DT（s_0），计算出基音周期 T_0，再根据 T_0 算出一个新的帧长 N_1，以此类推，直到 N_n 和 N_{n-1} 之间的差值小于某个限值 θ。则可认为 N_n 就是最佳帧长 N_p。迭代公式：

$$\begin{cases} N_0 \quad \text{初始条件} \\ T_{n-1} = DT(s(N_{n-1})) \\ N_n = \partial(T_{n-1}) \end{cases} \qquad \text{公式（3-2）}$$

具体算法流程如下：

首先选取一个帧长阈值为 N_{max} 和 N_{min}，分别表示帧长的最大值和最小值，T_n 表示第 n 个语音帧的基音周期，N_n 表示第 n 帧的帧长。参数 ∂ 为一个语音帧内基音周期的数目（N_{max} 、N_{min} 和 ∂ 的值可根据目标语音的特征选取不同的值，例如，处理男声时，∂ 的值可选的小一些，阈值 N_{max} 和 N_{min} 之间的宽度可以宽一些。处理女声时，则相反）。对于第 n 帧，先设帧长 N_{n1} 为 $\frac{N_{max}+N_{min}}{2}$，计算这一帧的基音周期 T_{n1}。

再根据计算出的基音周期重新确定这一帧的帧长，计算原则如下：

首先设定基音周期用 T_p 表示，则新的帧长：

$$N_{n2} = \begin{cases} N_{min} & N_{n1} < N_{min} \\ T_{n1} \times \partial & N_{min} < N_{n1} \le N_{max} \\ N_{max} & N_{n1} > N_{max} \end{cases} \qquad \text{公式（3-3）}$$

最后计算新帧的基音周期，并返回第二步继续确定新的帧长。直到满足条件：

$$|N_n - N_{n-1}| < \theta \qquad \text{公式（3-4）}$$

此算法的问题在于，因为基音周期的检测算法很复杂，且存在一定的随机性，所以迭代不一定收敛。算法时间复杂度较高，效率也比较低。

2，简单代替算法

考虑到基音周期在帧与帧之间的变化不会太大，所以完全可以用前一帧的基音周期代替后一帧的基音周期，算法流程：

首先选取一个帧长阈值为 N_{max} 和 N_{min}，先顺序取出第一帧，，可用算法(1)的方法，算出第一帧的帧长 N_1，并计算其基音周期 T_1。

然后计算第二帧的帧长 $N_2 = T_1 \times \partial$ 并计算第二帧的基音周期 T_2。

重复以上步骤，则第 n 帧的帧长 $N_{max} = T_{p\,n-1} \times \partial$ 。

较前一种算法，这种算法效率大大提高。且能得到较为平滑的基音周期轨迹，但是一旦算错一个帧长，便无法保证其后面每一帧长的正确性。所以算法的可靠性并不高。

3，平滑滤波算法

本文中采用这种方法作为基音周期提取的主要方法。该算法在前两种算法的基础上，把平滑滤波算法加入到语音信号的分帧算法中，具体算法流程如下：

先用算法（2）算出前 s 帧的帧长 N_1、N_2……N_s 及其基音周期 T_1、 T_2……T_s。

设计一个平滑滤波器，因为求第 s+1 帧时，参数 s 帧的信息的可能权重较大，所以滤波器参数可以选择一个从大到小的排列的数列，例如当 s=4 时，滤波器的参数可选择为[0.7，0.2，0.05，0.05]。

较前一种算法，此算法能够得到能为更为平滑的基音周期轨迹，降低了“野点”对基音周期轨迹的干扰，提高了算法的可靠性。图 3-2 为传统的平均分帧算法和本文的平滑滤波算法所提取的元音[a][o][e]的基音频率轨迹对比图。可以看出平滑滤波算法可以得到更加平滑且野点更好的基音频率轨迹。

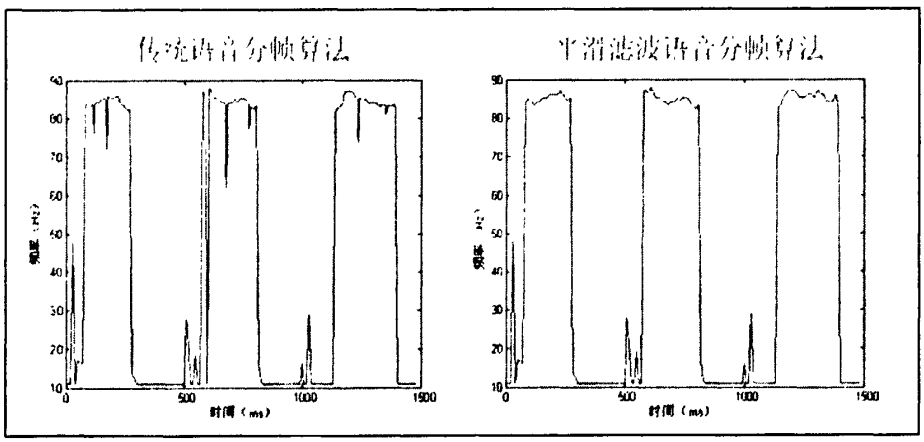


图 3-2 传统算法和平滑滤波算法提取[a][o][e]的基音轨迹

3.1.2 窗函数的选取

窗函数的选择包括窗函数类型的选择和窗长的选择，选取不同的窗函数对信号频谱的影响不一样，这主要是因为不同的窗函数，产生频率泄漏的大小不一样，频率分辨能力和时间分辨能力也就不一样。信号的截断产生了能量泄漏，而用 FFT 算法计算频谱又产生了栅栏效应，从原理上讲这两种误差都是不能消除的，但是可以通过选择不同的窗函数对它们的影响进行抑制。根据语音信号的特点，音节端点检测中选取矩形窗函数，它的时域和频域波形如图 3-3 所示，从图中可以看出矩形窗函数旁瓣较宽，所以频率分辨率较差，因为我们对语音信号进行端点检测时更注重

于它在时域上的特性，所以我们选取矩形窗作为音节检测的窗函数，它可以表示为：

$$w(n)=1,0\le n\le N \qquad \text{公式（3-5）}$$

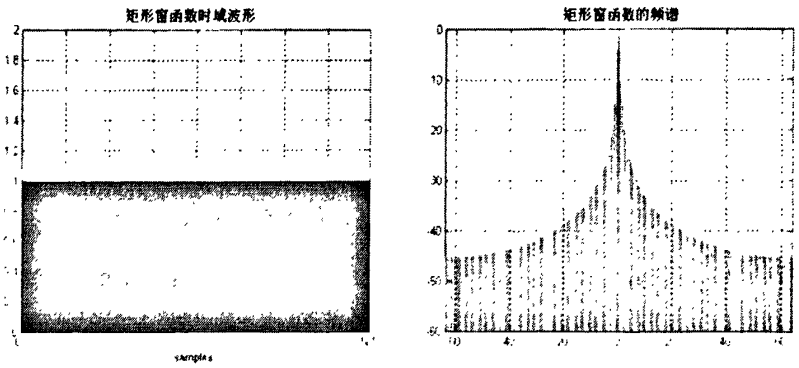


图 3-3 矩形窗函数的时域和频域波形

在对语音信号类型判别和基音检测时都选择频谱特性较好的汉明窗作窗函数。如图 3-4 所示，汉明窗又叫升余弦窗，因为汉明窗主瓣窄，旁瓣低的特点，因而有着较好的频率分辨性。汉明窗的表达式为：

$$w(n)=0.54-0.46\cos(\frac{2\pi n}{N-1}),0\le n\le N \qquad \text{公式（3-6）}$$

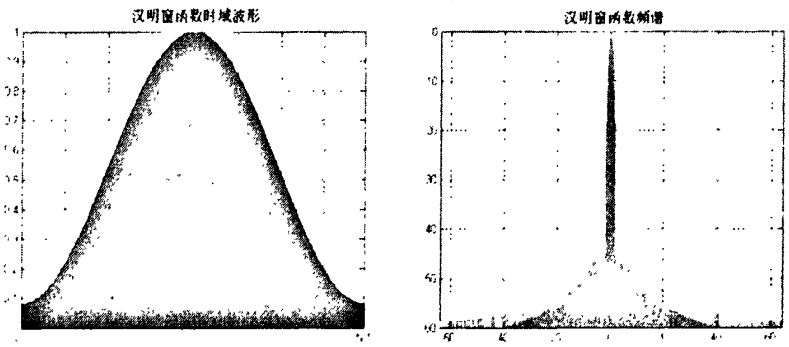


图 3-4 汉明窗函数时域和频域波形

3.2 语音信号特征提取

3.2.1LPC 参数的提取过程

语音信号的线性预测编码（LPC）是一种基于先前的语音样本，预测以后将会出现的语音样本的一种编码方法 [20] 。和倒谱分析方法相同的是，LPC 参数也把语音信号分为，声源激励和声道滤波两部分。线性预测一直是语音信号处理中的核心技术，它在语音识别、语音合成、语音编码、等方面都得到了成功的应用。作为 LP 系数的推演参数线谱频率（LSF）也具有良好的特性，非常有利于在语音转换中进行频谱包络的转换处理。LPC 参数的定义如下：

对一个语音片段 s，我们可以用前 p 个样本的加权和来预测第 n 个样本。其中 p 称为 LCP 的阶。其具体表示为：

$$\tilde{s}[n]=\sum_{k=1}^{p}a_k s[n-k] \qquad \text{公式(3-7)}$$

上述公式中 p 越大，线性预测越精确。但是考虑计算的复杂度，一般在实际应用中，p 的值选取 10~20 为宜。权值 a_k 的确定方法，是使得实际值与预测值之间的误差最小，也可以理解为去除了语音信号中滤波器的影响，那么剩下的残差及是声门的激烈信息。残差 e[n]表示为：

$$e[n]=s[n]-\tilde{s}[n]=s[n]-\sum_{k=1}^{p}a_k s[n-k] \qquad \text{公式（3-8）}$$

对以上公式两边取 Z 变换：

$$E(z)=S(z)-\sum_{k=1}^{p}a_k S(z)z^{-k}=S(z)[1-\sum_{k=1}^{p}a_k z^{-k}]=S(z)A(z) \qquad \text{公式（3-9）}$$

由上式可得：

$$S(z)=\frac{E(z)}{A(z)} \qquad \text{公式（3-10）}$$

其中 E(z)代表残差，S(z)代表原始语音信号，A(z)代表一个全极点的声道滤波器。每一个权值都是一个极点。且极点必须都位于单位圆以内，来确保滤波器的稳定性。这就达到了把声源激烈部分和声道滤波部分分开的目的。

3.2.2LSF 特征参数提取过程

线谱频率(line spectrum frequency，LSF)与 LPC 中的权值向量是相互对应的。LSF 参数完全是 LPC 参数的另外一种形式，它的计算是通过对 A(z)的变换得来的：

$$F_1(z)=A(z)+z^{-(p+1)}A(z^{-1}) \qquad \text{公式（3-11）}$$

$$F_2(z)=A(z)-z^{-(p+1)}A(z^{-1}) \qquad \text{公式（3-12）}$$

$$A(z)=\frac{1}{2}(F_1(z)+F_2(z)) \qquad \text{公式（3-13）}$$

LSF 参数是多项式 $F_1(z)$ 和 $F_2(z)$ 的复数根或者复数零点。它们的零点在单位圆上且交替出现。因此，A(z)可以表示为一系列的由 $F_1(z)$ 和 $F_2(z)$ 得到的交替出现的角频率。

本中采用多项式求根的方法来计算 LSF 参数。

LSF 参数叫 LPC 参数具有如下一些更有用的特性：

1，LSF 参数具有相对独立的性质，如果某个特定的 LSF 参数中只移动其中任意一个线谱频率的位置，那么它所对应的频谱只在附近频率与原始语音频谱有差异，而在其它 LSF 频率上则变化很小。

2，当改变 LSF 值时，可以保证滤波器稳定性和最小相位特性。

3，由于 LSF 是频率值，则可以很容易的根据人的听觉特性进行 Mel 尺度的规整。

4，LSF 参数能够反映声道幅度谱的特点，在幅度大的地方分布较密，反之较疏，这样就相当于反映出了幅度谱的共振峰特性。

3.3 语音转换的训练阶段

3.3.1 语音信号时间对齐

在对两个人说话的参数进行转换时，会发现即使是单个音节，单个音素，不同的人，所用的时长是不同的，所以在构造转换函数之前必须对语音信号进行时间参数的对齐。动态时间规整（DTW）是最常用的，也是最有效的对齐方法之一。DTW 的基本思想是，用源说话人的每一帧声音信号的特征矢量（例如：LSF）s[n]对齐目表说话人的特征矢量 t[m]，动态的复制或者删除 s[n]的中的值，满足下面两个条件：

$$Sizeof(s) = Sizeof(t) = \mathrm{m} \qquad \text{公式（3-14）}$$

$$\sum_{k=0}^{m-1} D(s[k], t[k]) \text{的值最小} \qquad \text{公式（3-15）}$$

其中 D（s[k],t[k]）表示为失真度函数，失真度函数的可以简单的定义为 s[k]和 t[k]的绝对值。

在对齐的过程中，对于删除的帧数和复制的帧数都一定的限制。原则上不许连续删除 2 个以上的帧，也不许连续赋值 2 个以上的帧。具体算法流程图 3-5 所示：

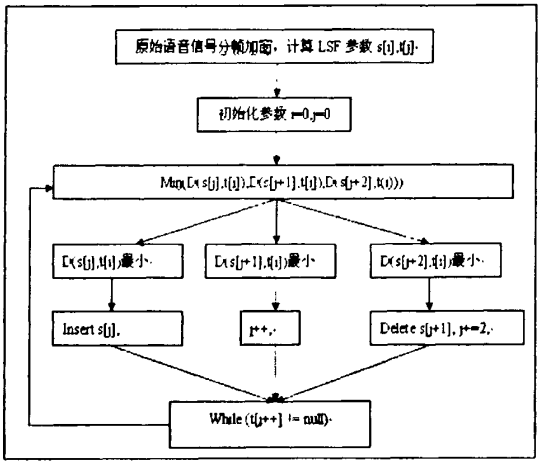


图 3-5 DTW 动态时间规整算法流程图

3.3.2基于音素的语音信号建模

3.3.2.1 基于高斯混合模型的谱包络转换模型的训练

基于上文音素对齐的基础上，本文将采用一种改进的 GMM 的训练方法，基于音素，对源说话人和目的说话人的语音进行训练。使用这种方法的原因在于由于音素、音节之间相互转换时，声道的变化会会加入非声道特定以及其他噪声的干扰，且基于整句对齐由于 DTW 是一种估计插值的算法，所以对齐队列太长会添加不必要的噪声，使得 GMM 训练的充分度下降。

高斯混合模型（GMM）作为一种分类工具，每一帧的特征向量可以用一个 GMM 模型描述，一个音素可以用几个 GMM 来描述，如图 3-6 所示，从数学上来描述，就是一组服从高斯分布的概率密度函数的带权线性组合。其定义如下：

$$p(X/\lambda)=\sum_{i=1}^{q}a_i N(X;\mu_i;\sum_i) \qquad \text{公式（3-16）}$$

$$\sum_{i=1}^{q}a_i=1, a_i\geq 0 \qquad \text{公式（3-17）}$$

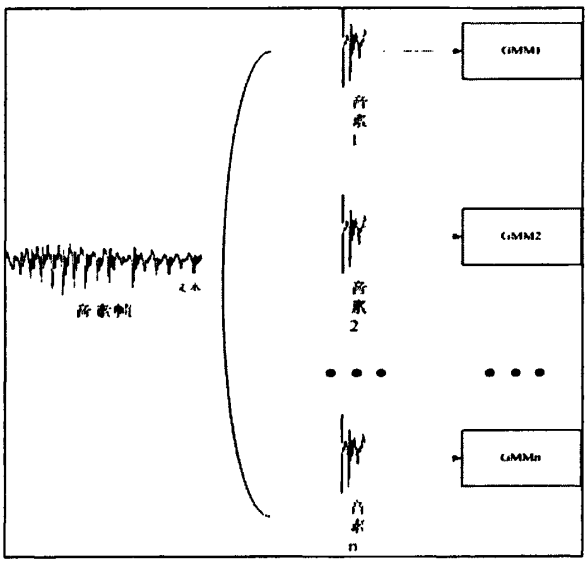


图 3-6 一个音素用多个 GMM 描述

其中 a_i 表示混合权重，q 表示混合高斯函数的维度，μ_i 表示向量均值，$\sum_i$ 表示协方差矩阵。从理论上来说，只要高斯函数的维度够大，GMM 可以描述任何概率密度函数模型。这就解决了在矢量量化过程中，导致语音信号不连续的问题。在一个语音转换系统中，可以认为每一个音素帧是由一组特征向量组成，这样的一组特征向量可以用若干个 GMM 模型划分到向量空间，所以可以得到一组基于 GMM 的不定长的模式向量来表征每一个音素。

要用 GMM 对语音信号进行建模，首先要对矢量空间进行划分。本文采用一种基于二分法和 k 均值的聚类方法划分矢量空间，具体步骤如下：

先对整体的训练数据进行聚类，算出其向量均值μ_0和协方差$\sum_0$。使用下面的值$\mu_0-\sqrt{\sum_0}$ $\mu_0+\sqrt{\sum_0}$作为下一次聚类的中心。

用新的聚类中心的均值μ_k和协方差$\sum_k$再计算下一次的聚类的均值$\mu_k-\sqrt{\sum_k}$和协方差$\mu_k+\sqrt{\sum_k}$，并重复上面的过程。

这样就可把特征矢量根据我们的训练数据，划分成任意 2 的指数个向量空间。避免了随机选点带来的算法不稳定的问题，提高了聚类的速度。

用以上方法，把向量空间分为 64 个空间，以每一个向量空间的中心，作为训练 GMM 参数的初始值。训练的过程实际上就是通过训练数据，求解模型$\lambda=(q,a,\mu,\sum)$中的参数，最常用的参数估计方法是期望最大(EM)算法。对于一个特征矢量 X，其 GMM 模型的似然度为：

$$\begin{aligned}p(X/\lambda)&=\sum_{i=1}^{q}a_iN(X_n;u_i;\sum_i)\\&=\sum_{i=1}^{q}a_i\prod_{n=0}^{M-1}N(X_n;u_i;\sum_i)\\&=\prod_{n=0}^{M-1}\sum_{i=1}^{q}a_iN(X_n;u_i;\sum_i)\end{aligned}$$

公式（3-19）

$\lambda_0=(q,a_0,\mu_0,\sum_0)$已经有上面的聚类方法求出，下面的目的就要找到一个$\lambda_1=(q,a_1,\mu_1,\sum_1)$满足下面的条件：$p(X/\lambda_1)>p(X/\lambda_0)$具体来说就是在$\lambda_{k+1}$的条件下，使得特征向量 X 的最大似然概率的对数为最大值：

$$E(\log[p(X/\lambda_{k+1})]=\sum_{i=1}^{q}P(X,i/\lambda_k)\log[p(X,i/\lambda_{k+1})])$$

公式（3-20）

利用最大似然概率求解的方法，对每个参数求导，使其结果为 0，并计算出λ_{k+1}中的参数。并重复上面的过程。直到模型基本收敛为止。各个参数的迭代公式如下：

$$a_i^{k+1}=\frac{1}{M}\sum_{n=0}^{M-1}p((i_n=i)/X_n,\lambda_k$$

公式（3-21）

$$u_i^{k+1}=\frac{\sum_{n=0}^{M-1}X_np((i_n=i)/X_n,\lambda_k}{\sum_{n=0}^{M-1}p((i_n=i)/X_n,\lambda_k}$$

公式（3-22）

$$\sum_i^{l+1} = \frac{\sum_{n=0}^{M-1} X_n X_n^T \; p((i_n = i)/X_n, \lambda_k}{\sum_{n=0}^{M-1} p((i_n = i)/X_n, \lambda_k} - u_i^{k+1}(u_i^{k+1})^T$$

公式（3-23）

$$p((i_n = i)/X_n, \lambda_k) = \frac{a_i^k N^k(X_n; \mu_i; \sum_i)}{\sum_{i=1}^{q} a_i^k N^k(X_n; \mu_i; \sum_i)}$$

公式（3-24）

最后一步判断一下模型是否收敛，我们规定一个范围域 e，这个范围在 1 左右，当得到的 u_i^{k+1} / u_i^k <e 时，就可以迭代停止，确定模型收敛。如图 3-7 所示，在训练量基本一致的情况下，右边的矩阵是使用基于音素的方法对 GMM 进行训练所得多的协方差矩阵。明显比左边使用基于整句对齐训练的协方差矩阵进过小值要少得多。

0.7	1.7	0.6	0.5	1.5	1.2	2.0	1.2	0.1	0.0	1.8	2.1
	1.7	0.0	0.0	1.4	1.6	2.0	0.0	0.0	0.0	1.4	0.1
		0.2	1.9	0.1	0.3	0.0	0.2	2.1	1.6	0.1	0.0
			0.2	0.6	0.4	0.1	1.1	0.0	0.9	1.7	1.0
				0.2	0.3	0.5	1.9	0.2	0.1	0.6	0.1
					0.3	1.9	1.6	0.3	1.6	0.4	1.6
						0.5	1.6	0.0	2.7	0.0	0.0
							0.2	0.0	1.9	0.6	1.6
								0.6	0.2	1.6	0.0
									1.2	0.0	2.7
										0.0	0.1
											0.9

0.9	2.1	0.6	0.7	1.6	1.1	2.4	1.5	0.1	0.4	1.8	2.7
	1.7	0.7	0.6	1.4	1.6	2.0	1.5	0.0	0.5	1.4	0.7
		0.2	1.9	0.1	0.3	0.0	1.0	2.1	1.6	0.1	0.2
			0.6	0.6	0.4	0.1	1.1	0.0	0.9	2.1	1.0
				0.6	0.3	0.5	1.9	0.6	0.1	0.6	0.1
					0.6	1.9	1.6	0.3	1.6	0.4	1.6
						0.5	1.6	0.0	2.7	0.4	0.0
							0.2	0.6	1.9	0.6	1.6
								0.6	0.2	1.6	0.3
									1.6	0.4	2.7
										0.0	0.1
											1.0

图 3-7 基于音素和基于整句提取元音[e]对 GMM 进行训练得到的协方差矩阵

3.3.2.2 基音频率的转换模型的建立

基音频率的转换是语音转换算法中非常重要的一个步骤，基音频率是语音信号的重要声学特征。由于汉语是有调语音，不同的声调，对应基音频率的变化是有一定规律。

和谱包络转换一样，本文同样采用高斯混合模型对基音频率轨迹，选定固定维度 n=24 的基音频率轨迹作为训练样本，针对汉语中只有五种不同的声调训练 GMM 模型，得到不同的模型参数：

$$\lambda_k^{tone} = N(q, a_k^{tone} \mu_k^{tone}, \sum_k^{tone}))$$

公式（3-25）

由于汉语语音的音调只有 5 种，并且为了更好的提前基音周期，我们只用元音对声调模型进行训练。所以模型的训练量比谱包络转换中模型的训练量要少得多。

3.4 语音信号的转换阶段

3.4.1基于码本混合映射的谱包络转换函数的建立

对于训练好的高斯混合模型（GMM）建立转换函数，一种方法是使用源矢量 X 与目标矢量 Y 的联合矢量：$Z_k = \begin{bmatrix} X_k \\ Y_k \end{bmatrix}$ 其中 k 是按时间对齐的。我来用 Z_k 做 GMM 训练，得到模型 $\lambda = N(q,a\ \mu,\sum)$ 其中期望和协方差分布可以表示为：

$$\sum = \begin{bmatrix} \sum^{XX} & \sum^{YX} \\ \sum^{XY} & \sum^{YY} \end{bmatrix} \quad \mu = \begin{bmatrix} u^X \\ \mu^Y \end{bmatrix} \qquad \text{公式（3-26）}$$

最后得到的条件期望转换函数为：

$$F(x) = \sum_{q=1}^{Q} p_q(x)[\mu^Y + \sum^{YX} (\sum^{XX})^{-1}(x-\mu^X)] \qquad \text{公式（3-27）}$$

GMM 语音转换的算法是目前最主流的算法之一，但是通过实现发现，由于在训练过程中很难得到准确的协方差矩阵，这样就会导致一些频谱细节的丢失，使转换以后的语音信号过于平滑 [21] 。

为了解决这个问题，本文采用混合码本映射的算法，在 GMM 的基础上，根据目标码本对协方差矩阵进行矫正，增加丢失的细节信息。具体步骤如下：

1，在得到 GMM 转换函数的基础上，对某一联合矢量求其偏移矢量：

$$X_{off} = \sum_{q=1}^{Q} p_q(x)(x-\mu_q^X) \qquad \text{公式（3-28）}$$

$$Y_{off} = \sum_{q=1}^{Q} p_q(x)(y-\mu_q^Y) \qquad \text{公式（3-29）}$$

2，使用基于音素绑定的码本映射算法来转换偏移矢量：

$$y_{off} = F_{off}(x_{off}) \qquad \text{公式（3-30）}$$

最终的特征可以表示在 GMM 模型和码本映射中间调节的转换函数：

$$y = (1-\lambda)y_{gmm} + \lambda y_{off}, \lambda \in [0,1] \qquad \text{公式（3-31）}$$

如图 3-8 所示，这是一段语音信号使用不同方法转换后的频谱包络和源语音信号的频谱包络的波形图。可以看出使用 GMM 和码本映射混合方法转换后的频谱包络比只使用 GMM 转换后的频率包络拟合度要好得多。很好的解决了转换频谱过于平滑的问题。

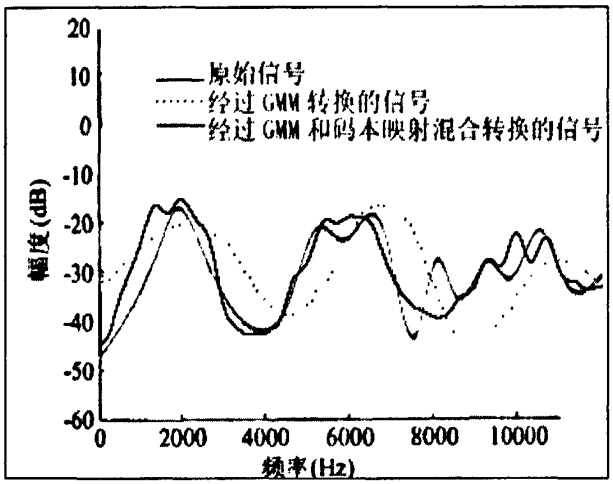


图 3-8 源信号的频率包络和使用两种不同方法得到的频谱包络

3.4.2 基音频率转换函数的建立

考虑如果每一种声调，都服从高斯分布，那么他们都可以用一组期望和标准差表示：

$p(f_k^{tone}) = N(\mu_k, \sum_k^{tone})$ 那么他们的转换关系就可以用下式表示：

$$f_t^{tone} = u_t^{tone} + \frac{\sum_t^{tone}}{\sum_s^{tone}}(f_s^{tone} - f_t^{tone}) \qquad \text{公式（3-32）}$$

前文中我们采用了高斯混合模型对音频进行建模，同理可知，相应的转换函数为：

$$f_t^{tone} = \sum_{i=1}^{Q} p_i(f_s^{tone})u_t^{tone} + \frac{\sum_t^{tone}}{\sum_s^{tone}}(f_s^{tone} - f_t^{tone}) \qquad \text{公式（3-33）}$$

其中：

$$p_i(f_s) = \frac{a_i N(f_s;\mu_i;\sum_i)}{\sum_{i=1}^{q} a_i N(f_s;\mu_i;\sum_i)} \qquad \text{公式（3-34）}$$

3.5 语音信号的合成阶段

语音信号的合成是相对于参数提取的一个逆过程。它的理论基础是，通过一个信号的短时傅里叶变换，可以通过其逆变换在时间域叠加得到原信号函数[22]。本文中使用时域基音同步叠加（TD-PSOLA）方式对语音信号的基音频率进行调整。调整过程分为以下三个步骤：

3.5.1基音同步分析

基音的同步分析是与合成单元浊音段的基音保持同步的一系列位置点，用他们来确切反映基音周期的起始位置，对语音合成单元进行标记设置是同步分析过程的主要工作。PSOLA 技术中，包括短时信号的截取和叠加，时长的选取等等。对于浊音段有基音频率，而清音段属于白噪声，在对浊音进行基音标注的同时，为保证算法的一致性，一般令清音的基音周期为一常数 [23] 。以语音合成单位的同步标记为中心，选择适当长度的窗函数对合成单元做加窗处理，窗长一般却基音周期的 2~4 倍。

3.5.2基音同步修改

基音的同步修改在合成规则的指导下，调整同步标记，产生新的基音同步标记，具体的说，就是通过对合成单元同步标记的插入，删除来改吧合成语音的时长，通过对合成语音单元标记间隔的增加和减少来改变合成的基音频率。这些短时合成信号序列在修改时与一套新的合成信号基音标记同步。

3.5.3基音同步合成

基音的同步合成是利用短时合成信号进行叠加合成，如果合成信号仅仅在时间长度上有变化，则相应减小，或增加合成语音的长度，如果是在基音频率上有变化，则首先将短时合成信号变成合成符合要求的短时信号，再进行合成。一般采用原始信号谱与合成谱差异最小的最小平方合成法，最终合成的信号为：

$$\bar{x}(n)=\sum a_q \bar{x}(n)\bar{h}_q(\bar{t}_q-n)/\sum \bar{h}_q^2(\bar{t}_q-n) \qquad \text{公式(3-15)}$$

其中分母是时变单位化因子，是窗之间时变叠加的能量补偿，$\bar{h}_q(n)$ 为合成窗序列。a_q 为一个相加的归一化因子。补偿音高修改时能量的损失 [24] 。

3.6 本章小结

本章中，我们给出了一个语音转换系统从预处理阶段，训练阶段，转换阶段，到合成阶段的所有理论算法，在预处理阶段我们采用二次动态分帧算法，最终的目的是为了检测音素的边界，让其在训练时更好的进行对齐，再训练阶段，我们基于音素对 GMM 模型进行训练，在训练次数尽量少的情况下，提高了训练的充分度。在转换阶段，我们采用 GMM 和码本映射的混合算法，提高了对谱包络的拟合度，弥补了 GMM 对频谱细节建模的不足。在转换阶段，我们采用 TD-PSOLA 三步合成技术对语音信号进行合成。提高了合成的效率。下一章中，我们将着重于如何实现一个语音转换系统，及其具体的算法步骤。

第4章 试验结果及相关讨论

4.1 系统实现及界面设计

本文中的语音转换系统使用 Visual studio 2008 开发 逻辑层使用 C#语音进行开发，底层算法由 matlab 7.4 实现，并通过 deploy tool 生成.net 组件，再通过 C#启动 matlab compiler runtime 对算法进行调用。软件界面如下：

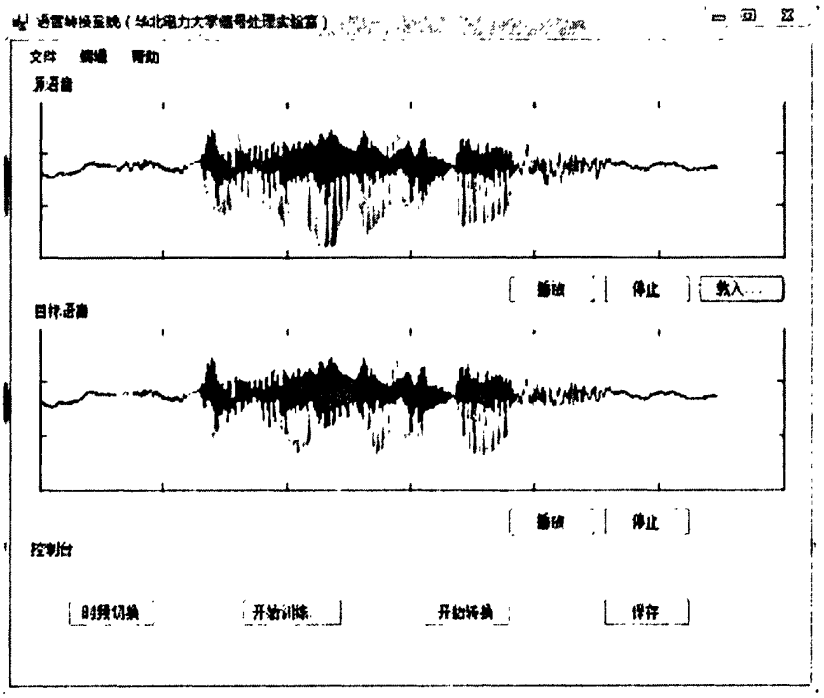


图 4-1 语音转换系统应用程序主界面

系统主要的功能有，语音信号的采集，播放，时间-频率视图切换，数据训练，数据转换，保持转换后语音等等。

4.2 实验具体步骤

4.2.1 语音信号采样阶段

我们尽量减少噪声对转换效果的影响，我们对语音信号进行采样在安静的室内环境中进行，采样的频率为 11025Hz，量化阶 16bit。并要求说话人的语速尽量平稳。字与字之间间隔 0.2 秒以上。且每个字长保证在 0.3~05 秒之间。

4.2.2模型训练阶段

以音素为单位训练 GMM 模型，对语音信号的频谱包络信息进行建模。首先对源说话人和目的说话人的音素语音进行采样，为了使音素库尽量小，我们只对 6 个单元音[a]、[o]、[e]、[i]、[u]、[ü]。8 个常用浊辅音[b]、[p]、[m]、[d]、[n]、[l]、[g]、[w]和 2 个鼻音[n]、[ng]进行采样，由于清音对基本频谱包络几乎没什么贡献，所以在对清辅音进行转换时，我们根据转换元音的平均能量，对清音作等比例的能量的补偿。

在训练过程中，每一人要求对每个音素重复 5 次，最后以每一帧的均值，作为训练用特征矢量。再对该音素帧采用 DTW 方法进行对齐，提取 LSF 参数，计算 GMM 模型的相关参数。我们用 64 个高斯混合模型（G[0]~G[63]）把特征向量所在的向量空间分成 64 份，且把截取定长音素，时长 0.3 秒，我们把音素平均分为 10 帧，每一帧为 30ms，加汉明窗，则每一个音素可以用一组高斯模型来表示，我们以元音[a]、[o]为例，得到的源说话人和目标说话人高斯混合模型序列如下表：

表 4-1 元音[a][o]的 GMM 序列

源说话人	元音	[a]	[o]
	GMM 序列	[34][36][34][30][21] [26][21][22][17][11]	[34][31][30][44][42] [39][21][22][15][17]
目标说话人	GMM 序列	[20][23][26][19][21] [17][17][19][16][9]	[20][18][19][18][18] [21][17][18][11][11]

通过对 GMM 模型的训练，我们得到一组基于概率最大的映射关系，并计算其相应的转换函数。到此我们就完成了利用 GMM 对频谱包络的模型建立。

下一步再对源说话人和目的说话人的音调进行建模，为了更好的提取基音周期，再建立 GMM 模型时，只使用 6 个单元音（[a]、[o]、[e]、[i]、[u]、[ü]）的不同音调进行训练，由于声调的情况较频谱包络简单，为了减少计算量，我们采用 16 个 GMM（$G_{tone}[0]$~ $G_{tone}[15]$）混合模型对基音频率的向量空间进行划分。同样采用固定 10 帧的 DWT 对齐，可得到一组基音频率 GMM 序列和映射关系，如表 4-2 所示。并计算其转换函数。

表 4-2 不同声调的 GMM 序列

源说话人	声调	阴（一声）	阳（二声）	上（三声）	去（四省）
	GMM 序列	[6][8][7][11] [13][12][13][10] [9][7]	[11][13][12][14] [14][15][14][12] [12][9]	[6][9][10][11] [15][14][12][12] [10][9]	[6][11][12][9] [10][11][9][9] [8][9]
目标说话人	GMM 序列	[7][10][8][11] [14][13][14][12] [11][9]	[11][14][13][15] [14][15][12][12] [11][11]	[7][11][12][11] [15][12][11][12] [12][11]	[7][12][12][11] [12][11][11][12] [10][10]

4.2.3 语音转换阶段

提取源说话人的一段语音，作为待转语音片段，我们这里在与训练时相同的环境，相同的采样频率和量化阶的条件下，提取一句由 5 个汉字组成的 “我是中国人”的语音片段，其波形图和语谱图分辨如图 4-2 和图 4-3 所示：

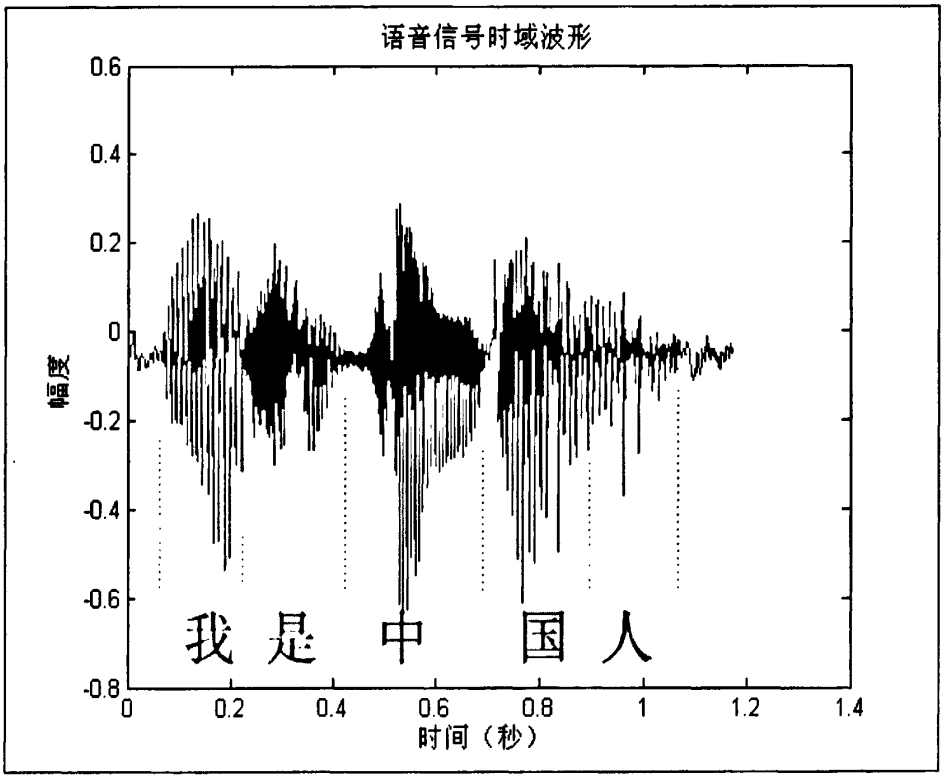


图 4-2 源说话人语音时域波形

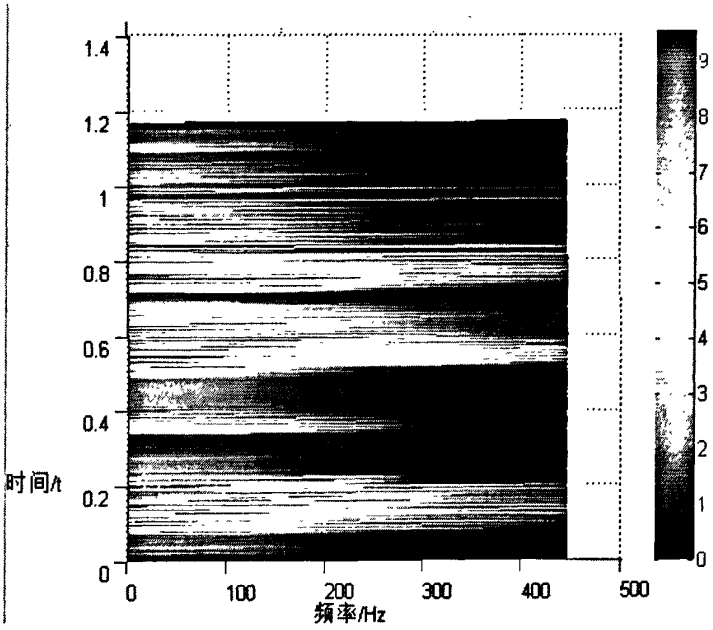


图 4-3 语音信号“我是中国人”的语谱图

对这段语音信号进行短时能能量分析如图 4-4 所示，计算其音节端点位置。得到的端点值如表 4-3：

表 4-3 源说话人语音中音节端点位置

端点编号	spos [0]	spos [1]	spos [2]	spos [3]	spos [4]	spos [5]
位置（t/ms）	63	218	417	691	905	1080

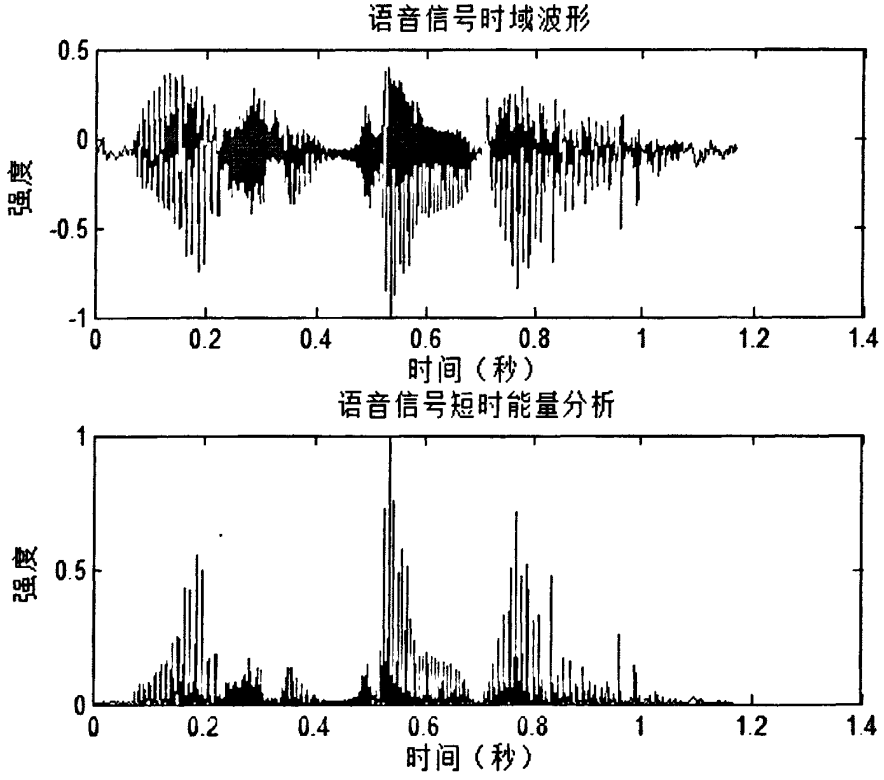


图 4-4 语音信号时域波形及其短时能量分析对比图

再对每个语音帧，进行 MFCC 相关性分析，对其发音模式进行识别。并用参数来表示相关音素帧的位置，如果没有参数，则表示这一个音素来表示这个音节。分析得到的曲线图如图 4-5 所示：

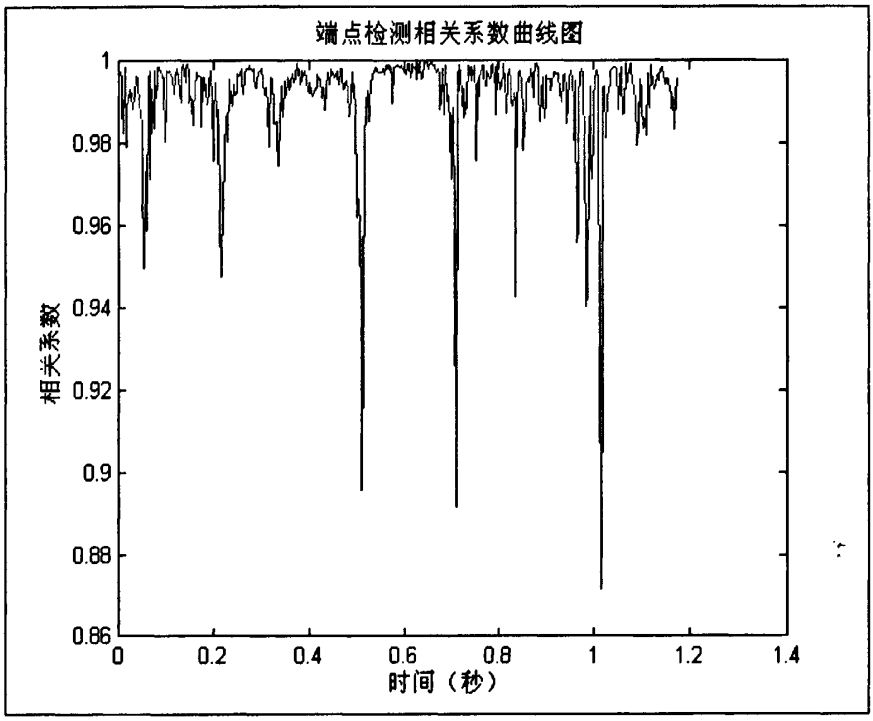


图 4-5 MFCC 短时相关性分析曲线

由上图可知相关性较差的点就是我们找的激变点。选定阈值 σ=0.96 发音类型的识别结果见过如表 4-4 所示：

表 4-4 源说话人语音发音类型

音节帧编号	发音类型	参数集合	spos_start	spos_end
0	C1V		63	218
1	C2V	310	218	417
2	C2VG	489,621	417	691
3	C1V		691	905
4	C1VG	962	905	1080

根据上文的类型识别，我们得到的音素帧长度和位置如表 4-5，并根据判别类型对是否是清音进行标记：

表 4-5 源说话人语音中音素位置

音标	[wo]	[sh]	[i]	[zh]	[o]	[ng]	[guo]	[re]	[n]
起点	63	218	310	417	489	621	691	905	962
终点	218	310	417	489	621	691	905	962	1080
清音	false	true	false	true	false	false	false	false	false

最后对音素帧平均分帧，对其谱包络进行 GMM 分类识别，根据转换规则，配合码本映射算法进行参数转换。对于清音，只作能量上的补偿。这样就得到了一组新的语音信号的谱包络特征向量集合。

在对基音轨迹的转换中，对音素帧采用基于平滑滤波的动态分帧算法，得到其基音频率轨迹，通过 GMM 进行转换。对清音不做处理，只在合成阶段做时长的调整，得到一组新的基音频率特征参数。

4.2.4 语音合成阶段

基于声源-滤波器模型理论，通过新的谱包络特征参数和基音频率轨迹特征参数相卷积的方法就可以得到要合成的新的语音信号。最后通过 TD-SOLA 算法对语音信号的基音频率进行微调，最后得到我们转换后的语音信号，如图 4-6 所示：

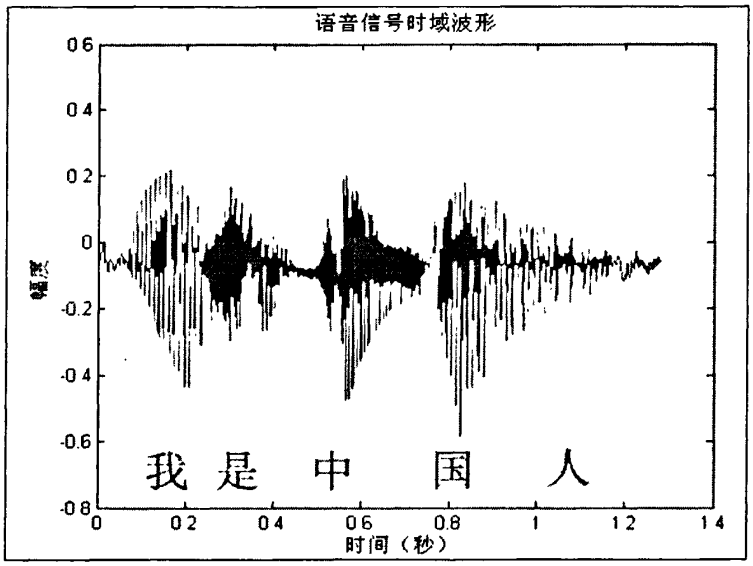


图 4-6 转换后的语言时域波形图

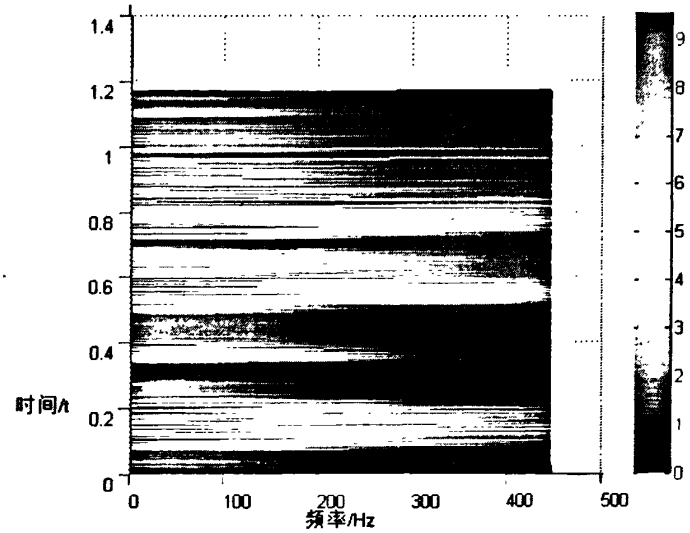


图 4-7 转换后语音的语谱图

4.3 实验结果分析

4.3.1 语音转换系统的主观评价系统

ABX 测试方法是用来检测说话人识别度的一种方法，在这里 A 和 B 分别指源说话人语音和目标说话人的语音，X 是指经语音转换系统后得到的语音。在测试中，测听者会听到语音 A，B 和 X，并判断所听到的语音的个性特征方面语音 A 还是 B 更接近于 X。这种测试方法的缺点是没有对转换质量进行量化，测听者常常认为自己听到的是第三方的语音。ABX 测试方法的一个改进是配对比较(pair--comparison)或者相似度测试(similarity test)。在这种类型的测试中，测听者首先测听一个语音对(不同的语音内容，例如不同的单词)，接着对说话人的相似度进行等级划分。应用多维的等级技术，结果可以映射成一个两维平面，表示两个语音的相关感受距离。从 0(完全相同)到 9(非常不同) 。本文根据此方法，找来 20 位测听者，其中 10 男，10 女，对语言转换质量进行评分。我们的测试句子为以下 10 组，每个句子的字数控制在 10~20 之内，所用的测试句子如表 4-6 所示：

表 4-6 ABX 测试用语音数据

S1	聪明出于勤奋，天才在于积累。
S2	人的每一步行动都在书写自己的历史。
S3	财富不应当是生命的目的，它只是生活的工具。
S4	贫穷的人往往富于仁慈。
S5	节约与勤勉是人类两个名医。
S6	决心就是力量，信心就是成功。
S7	一经打击就灰心泄气的人，永远是个失败者。
S8	最严重的浪费就是时间的浪费。
S9	劳动是一切知识的源泉。
S10	我们唯一不会改正的缺点是软弱。

4.3.2 不同性别间的转换评级

男声和女声之间的声学参数是有一定的差异的。主要表现在两个方面：

1， 基音频率参数的差异。成年女子的平均基音频率大约是成年男子的 1.7 倍，男音的基频范围大约为 60-200Hz，女音的基频范围大约为 200-450Hz。

2， 共振峰参数的差异。女子的语音的共振峰频率比男子的高一些，大约高出 1 到 1.3 倍左右，女子的共振峰的带宽比男子的相应的共振峰的带宽要稍宽一些。

作者发现，在转换过程中，高频向低频转换的效果要优于低频向高频转换，这可

能是在语音和成阶段 TD-PSOLA 算法，需要对频率低的部分进行截断，从而导致部分元音的细节信息的丢失造成的。所以总体上说女声转男声的效果要优于男声转女声。以下是 20 位测听者，对转换效果的平均评级如表 4-7 所示：

表 4-7 ABX 测试不同性别转换的评介结果

转换方法	男声转女声	女声转男声
测试结果	4.6	6.8

4.3.3 GMM 不同训练语料对转换结果的影响

我们采用句子而不是基于音素，对 GMM 进行训练，那么同等条件下，以下是 20 位测听者，对男声转男声和女声转女声的转换效果的平均评级如表 4-8 所示：

表 4-8 不同语料训练的 GMM,ABX 测试评价结果

转换方法	男声转男声	女声转女声
测试结果 1（使用句子训练）	6.7	6.4
测试结果 2（使用音素训练）	8.9	8.2

从评价结果上来看，基于音素训练 GMM 的转换方法，明显比基于句子训练 GMM 的方法转换效果好。

4.3.4 码本映射对 GMM 补偿对转换结果的影响

我们只使用 GMM 的概率模型进行语音转换，基于上述同条件下，20 为测听者，给出的评分如表 4-9 所示

表 4-8 是否使用码本映射,ABX 测试评价结果

转换方法	男声转男声	女声转女声
测试结果 1（不使用码本映射）	7.7	6.9
测试结果 2（使用码本映射）	8.8	8.2

4.4 本章小结

在本章中，我们根据前面几章所讨论的方法设计了一个特定对象的语音转换系统，并实现了源说话人和目标说话人直接之间的语音转换，并对转换结果从不同的角度，进行了评估。评估结果表明，同性别间的转换效果较好，不同性别的转换中，

女声转男声的效果要优于男声转女声；基于不同训练方式的 GMM 模型转换效果差别很大，基于音素训练的 GMM 模型的转换效果，要明显优于基于整句训练 GMM 模型；使用码本映射的方法对 GMM 进行补偿的方法也明显优于不使用码本映射的 GMM 方法。

第5章 结果与展望

通过上一章的实验结果，可以得到以下结论：

1，用单音素训练 GMM 模型确实比使用随机的句子训练 GMM 模型得到较好的效果。原因是在整句训练时，由于 DTW 对齐的序列较长，其误差也随着增大。且对齐时，并不能保证反应相应包络谱特性的部分对齐，而且对无声部分，清音部分对齐在训练 GMM 时，相当于加入了噪声。降低了协方差矩阵的训练充分度。

2，码本映射是很好的对 GMM 模型的补充，由于 GMM 边界过于平滑，使得一些明显的特征被忽略，从实验结果来看，码本映射和 GMM 相结合的谱包络转换算法取得了很好的效果。

3，LSF 参数确实能够很好的模拟声道共振峰的特性，在谱包络的转换中，也收到了很好的效果，但是其对声道特性的建模是基于声道级联模型的假设，这导致随着说话人音色差别变大，转换精度也随之下降。可以采用基于自适应加权谱内插的语音转换和重构（STRAIGHT 分析-合成）的方法来提高精度。

4，当说话人声调差别较大时，TD-PSOLA 算法在调整基音频率时，实际上是按基音周期调整了音素的时长，这是导致转换结果和源说话人语音时长不一样，且听上去语音不够连续。

要提高语音转换的精度，作者提出以下建议：

1，从生理语音学的角度，建立更准确的语音参数模型，提取更多更有效的语音特征参数。包括谱包络特征和基音特征。

2，加强对超音段特征的建模，加强对语音韵律的研究，在语音合成阶段，找到一种合适的方法，可以在基音周期改变的同时，不影响语音信号的其他参数。

参考文献

[1]. M. Abe, S. Nakamura, K. Shikano "Voice conversion through vector quantization," in ICASSP, New York, 1988: 565-568

[2]. Kuwabara H.,TAkigi T., Acoustic Parameters of voice individuality and voice quality control by analysis-synthesis method. Speech Communication, 1991, 10(5-6):491-495

[3]. D.G.Childers, Chieteuk Ahn, Modeling the glottal volume-velocity waveform for three voice types, J.Acoust. Soc. Am. 1995 97(1):505-519

[4]. Arslan L M, Talkin D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum Proceedings of European Conference on Speech Communication and Technology. Rhodes, Greece, 1997

[5]. D.T. Chapell and J.H. Hansen, "Speaker-specific pitch contour modeling and modification," in ICASSP, Seattle, 1998: 885-888

[6]. A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, 1998: 285–288

[7]. Y. Stylianou, 0. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. on Speech and Audio Processing, vol. 6(2),1998: 131-142

[8]. T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," Proc. ICASSP, 2001:841–844

[9]. J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average voice-based speech synthesis," in ICASSP, Toulouse,2006,1: 77-80.

[10]. 初敏，高清晰度高自然度汉语文语转换系统的研究，中科院声学所博士论文，1996

[11]. 刘立，基于时域的男女声语音转换新途径的研究，中科院声学所硕士学位文，2000

[12]. 康永国，双志伟，陶建华等，基于混合映射模型的语音转换算法研究，声学学报,2006,32(6): 555-562

[13]. Z. Inanoglu, "Transforming a pitch in a voice conversion framework," M.S.

thesis, University of Cambridge, 2003

[14]. A. Verma and A. Kumar, "Voice fonts for individuality representation and transformation," TSLP, vol. 2(1), 2005: 1-19

[15]. Turk and L.M. Arslan, "Robust processing techniques for voice conversion," Computer Speech and Language, vol. 4(20), 2006: 441-467

[16]. J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss, "A parametric approach for voice conversion," in TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, 2006:225-229

[17]. Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," Proc. Euro speech,2003: 2413–2416

[18]. A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," Proc. ICASSP, 2003: 720–723

[19]. H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including dynamic and phonetic information in voice conversion systems," Proc. ICSLP,2004:1193–1196

[20]. H. Ye and S. Young, "Quality-enhanced voice morphing using maximum like hood transformations," IEEE Trans. on Audio, Speech and lang. Proc.,2006: 1301–1312

[21]. K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through combining modified GMM and formant mapping for Mandarin," Proc. ICDT, 2007: 10

[22]. D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," Proc. Inter speech, 2007: 1965–1968

[23]. 赵力，语音信号处理，北京:机械工业出版社，2003

[24]. B. P. Nguyen and M. Akagi, "Control of spectral dynamics using temporal decomposition in voice conversion and concatenative speech synthesis," Proc. NCSP, 2008: 279–282

致　谢

首先，我要忠心感谢我的导师。在我研究生学习过程中，无论是在课程的选择、研究课题的确立，还是在课题的具体研究、论文的精心审阅等方面，我的导师都付出了辛勤的汗水。许教授严谨求实的治学态度、渊博的学识、严格的要求和耐心的教诲对我的硕士论文的顺利完成起着至关重要的作用，也将使我终生受益。

感谢学院的各位老师，他们不仅教给我知识，更重要的是教给我严谨的治学态度和科学的学习方法，使我在学习的道路上不断前进。感谢各位同学以及实验室的师弟师妹们，在我的学习和论文工作中给予我的帮助和支持。

最后衷心感谢我的家人在我的学习生涯中给予我的关心、鼓励和支持！

在学期间发表的学术论文和参加科研情况

[1]. "Robust Endpoint detection in Mandarin Based on MFCC and Short-time Correlation"ICICTA2009 Vol.2: 336-339