个性化文本语音转换系统（TTS）的设计与实现

作 者 姓 名：　潘鹤

指 导 教 师：　王义 教授

单 位 名 称：　信息科学与工程学院

专 业 名 称：　计算机科学与技术

东 北 大 学

2014 年 6 月

Design and Implementation of Individuation Text-Voice Conversion (TTS)System

By Pan He

Supervisor ： Professor Wang Yi

Northeastern University

June 2014

毕业设计（论文）任务书

毕业设计（论文）题目：

个性化文本语音转换系统（TTS）的设计与实现

设计(论文)的基本内容：

（1）基于 Microsoft 提供的 Speech SDK 语音工具包实现 TTS 的基本功能。

（2）实现个性化朗读角色的设定，建立语音转换模型，实现系统语音转换成个性化朗读者语音。

（3）声音情感特征的分析和提取，求取不同情感间的转换关系。

（4）设计说话人识别系统，对语音转换结果进行评价。

毕业设计（论文）专题部分：

题目： _____

设计或论文专题的基本内容：

学生接受毕业设计（论文）题目日期

第 1 周

指导教师签字：

2014 年 3 月 3 日

个性化文本语音转换系统（TTS）的设计与实现

摘 要

语音技术是近半个世纪以来崛起的一门新型科学技术，主要包括语音识别技术和语音合成技术。其中,语音合成是指将文本转化为人类可以理解的声音信号的相关技术。事实上，文字-语音的转换（简称 TTS）是近年来发展起来的一种应用非常广泛的技术，它可以将任意输入文本信息转换为语音信号。因此,在汽车导航、电信的呼叫服务、公交车到站站名自动播报中具有广阔的应用前景。

然而，当前的 TTS 大多仅能使用系统自带的朗读角色进行文本朗读，语速平淡不具有任何感情色彩，在用户体验上欠缺了一些个性化元素。因此，本文根据语音特征分析，建立了个性化语音转换函数，从而实现用户根据自己的个人喜好进行朗读角色的任意设定，为后续的语音信号个性化处理奠定了理论基础。

本文首先详尽地阐述了基于微软提供的 Speech SDK 语音工具包实现 TTS 的基本功能；然后，利用 Speech SDK 的 API 接口，将输入的文本内容准确朗读并且实现了中英文的混合朗读；接着，为了实现个性化语音处理，本文建立了一个语音转换模型。语音转换是针对于源说话人和目标说话人，即使一段源说话人的语音转换后具有目标说话人的声音特征。此外，本文对于声音的情感因素进行研究，提取分析不同情感语音的特征参数，获得了不同情感和中性语音间的转换关系，使朗读出来的语音具有了用户需求的情感色彩。

最后，本文使用 C++和 Matlab 混合编程设计和实现了一个个性化文本语音转换系统，既可使用系统自带的朗读角色也可根据个人喜好进行自定义，语音朗读的效果较为理想，验证了上述语音转化算法的有效性。

关键词：TTS，个性化，情感因素，说话人识别，GMM 模型

Design and Implementation of Individuation Text-Voice Conversion (TTS) System

Abstract

The voice technology is a new science and technology nearly half a century, mainly including speech recognition and speech synthesis technology. Among them, speech synthesis converts text into sound signals that human can understand. In fact, text-voice conversion (abbreviated TTS) has very broad applications and it can convert any text to voice signals. Thus, it has broad application prospects in car navigation, telecommunications call and the bus train station names automatically broadcast.

However, current TTS system only can speak text in system roles without any emotion, so there is a little lack in personalization elements for user experience. Therfore, after voice feature analysis, this article sets up personalization voice convertion functions. So users can set any speaking roles by themselves. It also establishes theoretical foundation for speech signal personalized processing.

Firstly, this article expounds TTS basic functions focused on Microsoft Speech SDK. Then, using the API interface of Speech SDK, we can let the computer read the text and achieve mixed reading in English and Chinese. Next, to get personalization, we use a voice conversion model. Voice Conversion is for the source speaker and the target speaker, namely converting a source speaker speech into a speech with target speaker acoustic features. In addition, this article talks about voice emotion factors. Extracting and analyzing different emotional speech feature parameters to get conversion relationship among different emotions and neutral speech. Finally, we can get a speech with emotional features meeting user need.

Finally, this article designs a personalization text-voice conversion system with C++ and Matlab mixed programming. It reads text in both system roles and customed roles. The speaking result is acceptable and the effectiveness of the voice conversion algorithm is demonstrated.

Key words: TTS, personalization, emotion factors, speaker recognition, GMM model

目 录

毕业设计（论文）任务书 I

摘 要 I

Abstract II

第 1 章 引言 ..... - 1 -

1.1TTS 概述 ..... - 1 -

1.2 源－目标说话人语音转换 ..... - 2 -

1.2.1 语音转换简介 ..... - 2 -

1.2.2 语音转换的意义 ..... - 3 -

1.2.3 语音转换技术现状 ..... - 4 -

1.3 论文研究的主要内容和目标 ..... - 5 -

1.4 论文的组织结构 ..... - 6 -

第 2 章 Microsoft Speech SDK 及 TTS 实现 ..... - 7 -

2.1Microsoft Speech SDK 简述 ..... - 7 -

2.1.1 Microsoft Speech SDK 结构 ..... - 7 -

2.1.2 Microsoft Speech SDK 的使用 ..... - 9 -

2.2Text To Speech 简介 ..... - 9 -

2.2.1 声学处理 ..... - 10 -

2.2.2 基于语音库的 TTS 系统基本框架 ..... - 10 -

2.3COM 技术相关介绍 ..... - 11 -

2.3.1COM 技术的主要优点 ..... - 11 -

2.3.2 COM 技术的主要接口类 ..... - 12 -

2.4TTS 实现 ..... - 12 -

第 3 章 语音转换理论基础 ..... - 15 -

3.1 语音信号的基本特征 ..... - 15 -

3.2 语音信号的预处理 ..... - 17 -

3.2.1 语音信号的采集与数字化 ..... - 17 -

3.2.2 语音信号的预加重和加窗 ..... - 17 -

3.2.3 语音信号的端点检测 - 18 -
3.3 特征参数提取 - 20 -
3.3.1 基音频率检测 - 20 -
3.3.2 基音检测后的平滑处理 - 22 -
3.3.3 频谱参数 - 23 -
3.4GMM 模型基本知识 - 25 -
3.5 语音转换的评价标准 - 29 -
3.5.1 客观评价标准 - 29 -
3.5.2 主观评价标准 - 29 -
3.6 说话人识别 - 30 -
第 4 章 语音转换和语音合成 - 35 -
4.1 基频目标模型 - 35 -
4.1.1 传统 Pitch Target 模型 - 35 -
4.1.2 改进的 Pitch Target 模型 - 36 -
4.2 基于基频目标模型的语音转换 - 38 -
4.2.1 基音频率转换 - 38 -
4.2.2 频谱转换 - 42 -
4.3 语音合成 STRAIGHT 模型 - 44 -
4.4 情感语音特征分析 - 46 -
4.4.1 语音的情感定义及分类 - 46 -
4.4.2 情感语音特征参数分析 - 47 -
第 5 章 实验验证与系统实现 - 50 -
5.1 语音转换 - 50 -
5.2 情感特征 - 52 -
5.3 系统运行环境 - 57 -
5.4MFC 界面设计 - 57 -
第 6 章 结论与展望 59
6.1 本文工作总结 - 59 -
6.2 进一步工作进展 - 59 -

参考文献.....- 61 -

致　谢.....- 63 -

第 1 章 引言

1.1TTS 概述

文本转换成语音仍是当前国内外语音方面研究的一个热点，目前很多领域都用到了文本转换成语音系统，已为大众生活带来了很大的便利，如手机播报来电显示号码，小霸王点读机等，其核心主要是语音库的建立与搜索引擎的实现与优化问题。因语音在生活中时刻都存在着，我们经常与语音直接打交道，也是生活交流的主要方式，文本转换成语音系统就是根据语音与文本之间存在的规律而开发的、一种为人民提供便利的、改善大众生活的新型应用软件系统。

把文本转换成语音，这是一种把人的视觉改变成听觉去交换感受的新理念，同时也为一些在视觉上存在问题的人们提供了大大的方便性，解决了残疾人在视觉存在的问题，又如，当人们工作对视觉产生疲劳时，希望把文档内容以声音形式展现出来以方便与改变人们的工作方式，达到一种轻松愉悦感。

目前世界上已研究出多种语言的 TTS 系统，如英、汉、日、法、德等，还有汉语几种语种，如普通话、粤语、四川话等，但这些都是基于微软的语音搜索引擎进行改变而设计的。在世界上，Bell 实验室、ATR 和 Siemens 公司已研制出多语种 TTS 系统，法国 CNET 实现的多语种 TTS 已用于电话网中的公共话音服务。语音合成所追求的目标，是让计算机输出的“合成语音”应该是具有可懂性、清晰性、自然性、具有较强的表现力。20 世纪 60 年代，TTS 英语系统首先被研制成功。80 年代，我国开始介入汉语语音合成领域的研究。中科院声学研究所首先开始汉语合成的研究。之后，社科院语言所、清华大学、中国科技大学、北方交通大学等单位陆续开展了对汉语 TTS 的研究。近些年来，在国家的大力支持下，汉语 TTS 技术有了长足的进步。清华大学、中国科技大学、中科院声学研究所等单位在这一领域取得了很好的成绩，有些研究成果已经转化为产品得到了实际的应用。如各种的有声阅读软件、手机语音助手、公交车报站等，可以说其应用非常之广泛。虽然目前语音合成技术已走向实用，但还有许多理论和应用问题有待解决。

人机语音交互包括语音识别和语音合成两部分。前者是让计算机听懂人说话，涉及到模式识别方面的知识；后者是让计算机说话，这主要是由文本语音合

成系统（TTS）来完成。传统的 TTS 系统中合成语音都是单一话者的语音，这就使得合成语音显得单调，缺乏个性，要想得到多样的发音则必须建立多套语音数据库。声音转换技术则较容易实现多种音色的个性化发音，使传统的耗时庞大的语音数据库的采集得以简化为仅需采集一个说话人（源）的语音数据库，对于其他音色的声音，只需少量的训练语音，便可从源说话人的语音库通过声音转换技术获得，节约了大量工作量与存储空间，且使系统变得更加灵活。还有，未来的系统会在人们接收 E-mail 或手机短信息时自动将信件内容用模仿发信人的声音读出来。

1.2 源－目标说话人语音转换

1.2.1 语音转换简介

语音是人类用于信息交流最普遍、最实用的工具,也正是由于其具有的重要性,语音信号处理这门学科才能长时间且广泛的吸引着许多科研工作者对其进行不断的研究和深入探讨。在创新性和应用性方面,其涉及到了一系列的前沿科研课题,并且具有极大的商业价值。虽然语音转换这个话题已经提出有近 30 年的历史,但也就是在最近几年,才有了突破性的科研进展,但是也存在着许多待解决的问题,在语音信号处理这个大的研究领域中,语音转换技术仍处于相对基础的水平。

语音转换(VC, voice conversion)是指改变一个说话人(源说话人,source speaker)的语音个性特征,使之具有另外一个说话人(目标说话人,target speaker)的语音个性特征。在一句说话人的语音中，包含了两个最重要的信息点：一是语义，即这句语音要提供给我们的内容信息；二是源说话人的特征信息，即个性特征信息，表征了该说话人的身份特征。根据以上两个方面，语音转换主要的目的就是保持第一点语义信息不变，在此基础上,在一定可控的范围内改变第二点源说话人的个性化信息。使源说话人的语音经过语音转换技术的处理后，听起来像另外一个人，也就是目标人的说话声音。在这里，还需要提到另一个概念，语音变换(Voice Morphing)，语音变换和语音转换是两个非常相似的研究领域。语音变换不需要修改语音使其具有某个特定的说话人的个性特征，而是对语音信号本身的某个固定因子进行修改，如时长因子等。这项技术可以应用在放慢或是加快说话人的发音速率。其中，放慢发音速率可以使得模糊不清的语音也能让人听得懂；

加快发音速率则方便人们对语音的信息的检索和减少语音在存储空间上的占有量。如一些声音处理软件通过改变基音频率来实现男女、小孩声音的变换，又如一些情感类的广播节目，为防止透露听众的身份，将声音进行处理使声音变形。本文主要实现的是对语音转换系统的构建。

1.2.2 语音转换的意义

语音转换技术的迅速发展对我们的生活有极大的意义，也创造了很多方便，有着广泛的实用价值,具体有如下几个方面：

(1) 声音修复。一方面可以用于医学领域，对于声道受损发音不易听懂的病人，利用语音转换技术能对受损的声音进行增强和修复,加强语音的可懂度和说话人个人特征信息，保证交流的通常。另一方面，对于有研究价值或有实用性的受损音频，也可以尽可能的恢复其源发音人的发音特征信息。

(2) 多媒体娱乐。例如在配音领域,可以有效的减少电影配音工作者的工作量,避免了配音人对不同角色同一句话的反复录音，语音转换技术能很好的将配音员的一句话转换为不同目标说话人的语音。与语音识别技术相结合，还能为儿童玩具增加配音功能,可以在玩具中实现将大人的语音转换为儿童的语音，或者反过来将儿童的语音转换为大人的语音。另外，在网络游戏等方面可以将玩家的声音进行少量的录制并训练，使得游戏角色的发音能转换成该角色玩家自己的声音，增加了玩家身临其境的感觉，有实际的吸引力。

(3) 声音伪装。在不便透露说话人的身份时可以通过语音转换系统进行伪装,用伪装后的声音进行通信；相反的，在刑侦认证时亦可能对经过伪装的声音进行源说话人的声音特征信息恢复，为侦查提供了很好的依据。

(4) 通信领域。目前的语音通信系统，如果语音的编码率等于或低于 2.4bit/s。将会导致语音解码后将不带有源说话人的个性特征信息。这虽然对通信本身十分有利，但将会让使用双方感觉不便，所以我们设想可以在语音解码后增加一个语音转换的模块，还原源说话人的个性信息。这样既可以兼顾通信的便利，也平衡了使用者的舒适性。

总之，语音转换技术是对语音合成、语音识别等技术的延续和拓展。随着语音转换技术的不断成熟，以及人们对语音交流需求的增强，更加方便、更加实用的语音相关产品必会深入到寻常百姓的生活，深入到各行各业。其带来的经济、

社会效益将十分可观。

1.2.3 语音转换技术现状

语音转换技术源于语音识别与语音合成技术，在过去的二三十年间，语音转换技术才慢慢地得到研究工作者们的重视。总体来说，国外的研究起步早，成果多。早在 1970 年代初， Atal 等人就研究了使用 LPC 声码器改变声音特性的可行性。

1988 年，Abe 等人提出了一种基于矢量量化(VQ)的码本映射技术，并在此基础采用模糊(VQ)法提高了转换性能。1992 年，Valbret 等人使用 LMR(线性多变量回归)和 DFW (动态频率调整)的方法进行了说话人语音转换的研究。其中，LMR 的方法考虑到了人耳的听觉特性，可以在转换过程当中加感觉性系数，有效的提高了转换后的语音质量。1995 年，H.Kuwabara 引入模糊矢量量化的方法用于说话人语音转换，在一定程度上提高了语音转换的质量。同时，为了解决矢量量化的不连续性，Stylianou 引入 GMM (高斯混合模型)的算法，通过加权平均的方法有效的解决了不连续性，这也是目前比较成熟，应用相对广泛的方法。后来，许多研究者根据 GMM 模型的“过平滑”等缺点对 GMM 模型进行了进一步的改进。像在 2001 年，Toda 运用 GMM 和 DFW 加权的方法进行了说话人语音转换的研究，使得转换后的语音相比传统的 GMM 模型的方法有了进一步的提高。

国内对于说话人语音转换的研究虽然起步较晚，但成果也很丰富。如刘立等采用矢量量化(VQ)结合动态时间调整法(DTW)进行男女声转换；初敏等采用重采样声道相应特性和 TD-POSLA (时域基音同步叠加)法进行基音周期的变换来实现男女声的转换；王聪修对嗓音源的特性进行了研究，基于嗓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，以此来实现男女生之间的语音转换。微软亚洲研究院的 Yining Chen 等采用 GMM 和 MAP 自适应法来实现语音转换，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诊参数的概率分布上，较好的回避了“过平滑”的问题。Chung-HsienWu 等人提出了将 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因素的时长信息，并用 GAMMA 函数分布来描述状态持续时间变量。除了一些专门的科研机构外，清华大学，哈尔滨工业大学，南京邮电大学等也都

有相关的语音信号处理实验室对语音转换技术进行实验研究。

1.3 论文研究的主要内容和目标

传统的 TTS 系统中合成语音都是单一话者的语音，这就使得合成语音显得单调，缺乏个性，要想得到多样的发音则必须建立多套语音数据库。声音转换技术则较容易实现多种音色的个性化发音，使传统的耗时庞大的语音数据库的采集得以简化为仅需采集一个说话人（源）的语音数据库，对于其他音色的声音，只需少量的训练语音，便可从源说话人的语音库通过语音转换技术获得。目前国内外的 TTS 实现技术有很多，但考虑到其成熟性和可开发性，本文采用微软公司提出的 Microsoft Speech SDK（语音软件开发工具包）。它提供了关于语音处理的一套应用程序编程接口 SAPI。SAPI 提供了实现文字—语音转换（Text To Speech）和语音识别（Speech Recognition）程序的基本函数，大大简化了语音编程的难度和工作量。而使用其提供的接口只能调用 windows 系统语音，即只能用系统语音来朗读文字。因此若想实现个性化的 TTS，即用自己的声音或身边人的声音朗读文字，需要进入源－目标说话人语音转换技术。即将系统语音作为源语音，自己的声音作为目标语音，进行转换合成。

综上所述，可将个性化 TTS 系统的设计归纳为以下几个步骤：

（1）了解和掌握 Microsoft Speech SDK 结构和使用，利用其提供的 SAPI 接口和 COM 接口，在 VC++6.0 编程环境下实现文本和语音的转换。

（2）在 Matlab 上实现源—目标说话人语音转换算法。Matlab 是一种交互式的矢量语音系统，其基本数据单元是不需要指定维数的矩阵，这使得用 Matlab 可以解决许多科学与工程计算问题。并且 Matlab 系统自带很多处理语音信号的工具箱（Voice Box），为处理语音信号带来了极大的便利。查找相关文献比较诸多源—目标说话人语音转换算法的优劣，选择其一进行实现。将某一系统语音作为源语音，自己的声音则为目标语音。进行一系列的建模和训练，得出源语音和目标语音的转换函数，最后再合成出具有目标说话人特性的声音。

（3）实现以上两个步骤的整合，及 VC++和 Matlab 的混合编程（VC++调用 Matlab 引擎）。最后使用 MFC 编写一个带有界面的个性化 TTS 系统。其功能大致包括：文本朗读、朗读角色选择、试听、音量及语速调节、保存语音、不同角色语音转换等。

1.4 论文的组织结构

本文主要旨在开发一个基于 Microsoft Speech SDK 的个性化文本语音转换系统，主要实现 TTS 基本功能及语音转换算法两部分内容。其中考虑了一些语音情感因素，使朗读出的语音具有一定的情感色彩。并且设计一个说话人识别系统对语音转换的结果进行评价。本文的结构安排如下：

第一章分别介绍了 TTS 的实际应用背景及语音转换的实际意义和技术现状。同时指出了本课题所做的内容是一个个性化的 TTS 系统，并说明了传统 TTS 的不足及加入个性化实现的实际应用意义。

第二章介绍了 Microsoft Speech SDK 可应用于文本—语音转换和语音识别及它提供的一套语音应用程序接口。对于本文中主要使用的 TTS 的工作原理、基本框架和 COM 技术的主要接口类进行了详细的说明。

第三章介绍了语音信号处理、语音特征参数的提取、高斯混合模型及语音转换结果的评价标准，为下一章语音转换和语音合成做知识铺垫。此外，对于文中主要使用的客观评价方法中的说话人识别方法的核心思想进行了介绍。

第四章对语音转换算法和语音合成进行了详细的分步说明。介绍了基频目标模型的建模过程，利用 GMM 模型对基音频率和频谱两个语音重要的特征参数进行转换。对于目前使用较为广泛和相对成熟的 STRAIGHT 语音合成算法进行了简要的介绍。此外，简要介绍了不同语音情感特征及情感间转换的相关理论知识。

第五章以四个说话人（两男、两女）的语音样本为例，给出了应用说话人识别方法对语音转换效果进行评价的实验结果。并且使用 C++和 Matalb 混合编程将文本—语音转换和语音转换结合起来实现了一个系统工具，分别介绍了系统的运行环境、主要功能和结果的展示。

第六章总结全文，对整个已实现的系统进行分析讨论，并对未来的改进提出设想。

第 2 章 Microsoft Speech SDK 及 TTS 实现

2.1Microsoft Speech SDK 简述

SDK 是 Software Development Kit 的缩写，中文意思就是“软件开发工具包”，它有很多优点：易用性很强、多接口平台；能运用到很多开发程序中如 VB、VC 等，同时也提供了大量的语音库与主要的英文和中文搜索引擎。

Microsoft Speech SDK 提供了关于语音处理的一套应用程序编程接口 SAPI（Speech Application Programming Interface）。SAPI 提供了实现文字—语音转换（Text To Speech）和语音识别（Speech Recognition）程序的基本函数，大大简化了语音编程的难度，降低了语音编程的工作量。SAPI 包括以下组件对象（接口）：

（1）Voice Commands API。对应用程序进行控制，一般用于语音识别系统中。识别某个命令后，会调用相关接口使应用程序完成对应的功能。如果程序想实现语音控制，必须使用此组对象。

（2）Voice Dictation API。听写输入，即语音识别接口。

（3）Voice Text API。完成从文字到语音的转换，即语音合成。

（4）Voice Telephone API。语音识别和语音合成综合运用到电话系统之上，利用此接口可以建立一个电话应答系统，甚至可以通过电话控制计算机。

（5）Audio Objects API。封装了计算机发音系统。

SAPI 是架构在 COM 基础上的，微软还提供了 ActiveX 控件，所以不仅可用于一般的 windows 程序，还可以用于网页、VBA 甚至 EXCEL 的图表中。如果对 COM 感到陌生，还可以使用微软的 C++ WRAPPERS，它用 C++类封装了语音 SDK COM 对象。

2.1.1 Microsoft Speech SDK 结构

在应用 Microsoft Speech SDK 的软件系统中，可以分为三层结构：最底层是两个语音引擎，分为语音识别引擎和语音合成引擎；中间是 SAPI 的运行库，为上层提供了 API 的运行环境；最上层是各种需要语音服务的应用程序。应用程序通过 API 函数来使用 SAPI 运行库，而 SAPI 则通过 DDI(设备驱动接口)来使

用语音引擎。Microsoft Speech SDK5.1 的结构图如图 2.1 所示。

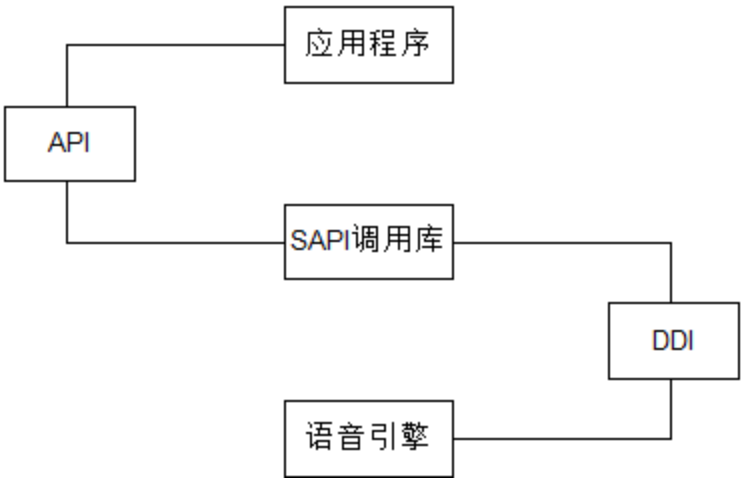

图 2.1　SDK 结构图

在微软提供的软件开发包 Microsoft Speech SDK 中，提供的 Speech API(SAPI)主要包含两大方面：

1．语音合成 API(Text To Speech)

2．语音识别(API for Speech Recognition)

其中 API For Text To Speech，就是微软 TTS 引擎的接口，通过它可以很容易地建立功能强大的文本语音程序，金山词霸的单词朗读功能就用到了这些 API。至于 API for Speech Recognition 就是与 TTS 相对应的语音识别，语音识别技术是一种令人振奋的技术，让用户能用声音给系统发出指令或输入信息，但由于目前语音识别技术准确度和识别速度不太理想，还未达到广泛应用的要求。为使 Speech API 具有更好的封装性和扩展性，同时方便各种不同用户的开发需求，SAPI 本身分为两个层次的结构，分别称为应用层接口和引擎层接口。应用层接口的对象为编程者提供较高级的接口形式，这个高级对象本身已经对一般语音应用提供了默认的设置，几乎不需要改变就能直接使用，使得编写程序更加方便。而对于语音质量有较高要求的应用，引擎层接口的对象则为编程者提供了较低级的控制方式，可以更灵活地控制语音引擎，但也使得程序编写变得较为复杂。在一般情况下，程序设计利用高级对象就足够了，但如果要更深层次地使用 Speech API，更精细地控制引擎的工作就必须使用低级对象。语音引擎通过 DDI 层和 SAPI 进行交互，应用程序通过 API 层和 SAPI 通信。通过使用这些 API，用户可以快速开发在语音识别或语音合成方面的应用程序。

在 Microsoft Speech SDK 中包含诸多接口供用户使用。其中应用层接口大致可以分为七类:语音接口类、事件接口类、语法编译接口类、词典接口类、资源接口类、语音识别接口类和文本语音转换接口类，但本文只需要文语转换类的接口。

2.1.2 Microsoft Speech SDK 的使用

要在应用程序中使用 Microsoft Speech SDK，应先进行 SDK 安装（windows7 系统已预安装了），其次还需要安装语音开发包。安装完微软的语音开发包后，系统会有几个不同的语音角色可供选择，如“Microsoft Mary”、“Microsoft Mike”，“Microsoft Sam”、“Microsoft Simplified Chinese”(英文与简体中文发音)、“Sample TTS Voice”，可以在控制面板的语音选项中看到。一些第三方企业也提供自己基于微软的语音库，安装注册后就能直接使用。

2.2Text To Speech 简介

“Text To Speech”即“文本到语音”的一种合成技术（简称 TTS)，它是同时运用语言学和心理学的杰作。在内置芯片的支持之下，通过神经网络的设计，把文字智能地转化为自然语音流；它是运用计算机技术对文本信息加以识别并转换为声音信息在计算机上通过声卡等多媒体设备将声音信息输出。也就是说把输入的文本能让计算机“读”出来。TTS 主要思想就是从有限的原始语音库中出发，合成具有无限词汇连续的语句。其工作流程如图 2.2 所示。

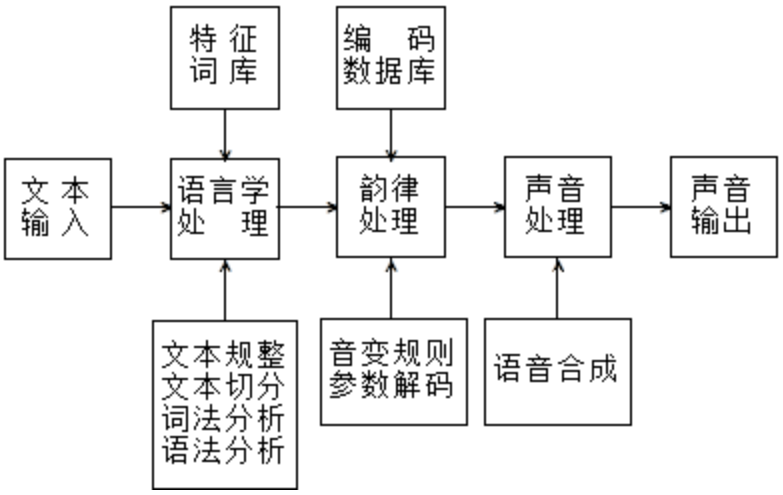

图 2.2 TTS 工作原理

TTS 技术对文本文件进行实时转换，转换时间之短可以用秒计算。在其特有智能语音控制器作用下，文本输出的语音音律流畅，使得听者在听取信息时感

觉自然，毫无机器语音输出的冷漠与生涩感。TTS 语音合成技术即将覆盖国标一、二级汉字，具有英文接口和自动识别中、英文及支持中英文混读等功能。

TTS 是语音合成应用的一种，它将储存于电脑中的文件，如帮助文件或者网页，转换成自然语音输出。TTS 可以帮助有视觉障碍的人阅读计算机上的信息，或者只是简单的用来增加文本文档的可读性。现在的 TTS 应用以包括语音驱动的插件形式以及声音敏感系统，TTS 经常与声音识别程序一起使用。

2.2.1 声学处理

系统产生的合成语音通过一个声学模块具体实现。建立声学的过程是：首先录制声音，然后提取出这些声音的声学参数，再整合成一个相对完整的语音库。对于语音合成过程来说，就是在发音过程中，先根据发音的需要从语音库中选择合适的声学参数，然后根据从韵律模型中得到的韵律参数，通过合成算法产生语音。通常我们称这种方法为参数合成方法。参数合成法相对来说难度有一点大，虽然经过运算能合成高清晰度的语音来，但近年来，波形拼接法(PSOLA)技术相对成熟一点，在合成过程中也容易形成清晰的语音。所以波形拼接合成语音的方法越来越被广泛应用。这种方法的主要核心思想是直接对存储于语音库的语音运用一种算法进行拼接，从而合成完整的语音。该系统的主要特点：首先在大量语音库中去选择最合适的语音单元或音素，其次是把选择出来的语音单元或音素进行参数拼接达到合成，关键技术就是在选音过程中往往采用很多复杂的技术来使合成的语音达到很高的音质。

2.2.2 基于语音库的 TTS 系统基本框架

在通常情况下，TTS 系统一般是基于大量的语音库进行的，它主要有文本分析、韵律的产生、单元的选择、波形合成法、语音库几部分组成的。基本示意图如图 2.3 所示：

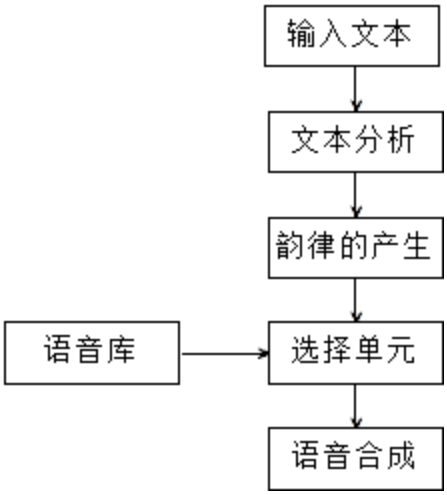

图 2.3　基于语音库的 TTS 系统框架

2.3COM 技术相关介绍

2.3.1COM 技术的主要优点

通过上面的内容，大概了解了 Microsoft Speech SDK，不难发现 Microsoft Speech SDK 是基于 COM 技术的。接下来，概括介绍一下 COM 技术的有关知识。所谓 COM（Component Object Model，即组件对象模型），是在程序设计中一种说明如何建立可动态互变组件的规范，提供了为保证能够互操作，客户和组件应遵循的标准。通过这种标准，程序员将可以在任意两个组件之间进行通信而不用考虑其所处的操作环境是否相同、使用的开发语言是否一致以及是否运行于同一台计算机，大大提高了程序开发效率。

COM 技术的主要优点如下：

（1）用户一般希望能够定制所用的应用程序，而组件技术从本质上讲就是可被定制的，用户可以用更能满足他们需要的某个组件来替换原来的组件。

（2）由于组件是相对应用程序独立的部件，可以在不同的程序中使用同一个组件而不会产生任何问题。软件的可重用性将大大的得到增强。

（3）随着网络带宽及其重要性的提高，分布式网络应用程序毫无疑问的成为软件市场上越来越重要的卖点。组件架构可以使得开发这类应用程序的过程得以简化。

（4）COM 是一种跨应用和语言共享二进制代码的方法。COM 通过定义二进制标准解决了这些问题，即 COM 明确指出二进制模块（DLLS 和 EXES）必须被编译成与指定的结构匹配。这个标准也明确规定了在内存中如何组织 COM

对象。COM 定义的二进制标准还必须独立于任何编程语言（如 C++中的命名修饰）。一旦满足了这些条件，就可以轻松地从任何编程语言中存取这些模块。由编译器负责所产生的二进制代码与标准兼容。这样使后来的人就能更容易地使用这些二进制代码。

2.3.2 COM 技术的主要接口类

SpVoice 类是支持文本语音合成(TTS)的核心类。通过 SpVoice 对象调用 TTS 引擎，从而实现朗读功能。SpVoice 类有以下主要属性：

Voice：表示发音类型，相当于进行朗读的人，包括 Microsoft Mary，Microsoft Mike，Microsoft Sam 和 Microsoft Simplified Chinese 四种。其中前三种只能读英文，最后一种可以读中文，也可以读英文，但对于英文单词只能将其包括的各个字母逐一朗读出来。

Rate：语音朗读速度。取值范围为-10 到+10。数值越大，速度越快。

Volume：音量。取值范围为 0 到 100。数值越大，音量越大。

SpVoice 有以下主要方法：

(1)Speak：完成将文本信息转换为语音并按照指定的参数进行朗读。该方法有 Text 和 Flags 两个参数，分别指定要朗读的文本和朗读方式（同步或异步等）。

(2)Pause：暂停使用该对象的所有朗读进程。该方法没有参数。

(3)Resume：恢复该对象所对应的被暂停的朗读进程。该方法没有参数。

2.4TTS 实现

（1）初始化语音接口

```
ISpVoice* pVoice;
::CoInitialize(NULL);
HRESULT hr = CoCreateInstance(CLSID_SpVoice, NULL, CLSCTX_ALL, IID_ISpVoice,(void **)&pVoice);
```

使用 pVoice 指针调用 SAPI 函数。

（2）获取/设置输出频率

SAPI 朗读文字的时候，可以采用多种频率方式输出声音，比如：8kHz 8Bit Mono、8kHz 8Bit Stereo、44kHz 16Bit Mono、44kHz 16Bit Stereo 等。在音调上

有所差别。具体可以参考 sapi.h。

（3）获取/设置播放所用语音。

引擎中所用的语音数据文件一般保存在 SpeechEngines 下的 spd 或者 vce 文件中。安装 SDK 后，在注册表中保存了可用的语音，比如英文的男/女，简体中文的男音等。

位置是：HKEY_LOCAL_MACHINE\Software\Microsoft\Speech\Voices\Tokens。如果安装在中文操作系统下，则缺省所用的朗读语音是简体中文。SAPI 的缺点是不能支持中英文混读，在朗读中文的时候，遇到英文，只能逐个字母读出。所以需要程序自己进行语音切换。

（4）开始/暂停/恢复/结束当前的朗读

要朗读的文字必须位于宽字符串中，假设位于 szWTextString 中，则：

开始朗读：

```
hr = m_cpVoice->Speak( szWTextString, SPF_ASYNC | SPF_IS_NOT_XML, 0 );
```

如果要解读一个 XML 文本，用：

```
hr = m_cpVoice->Speak( szWTextString, SPF_ASYNC | SPF_IS_XML, 0 );
```

暂停：m_cpVoice->Pause();

恢复：m_cpVoice->Resume();

结束：hr = m_cpVoice->Speak(NULL, SPF_PURGEBEFORESPEAK, 0);

（5）跳过部分朗读的文字

在朗读的过程中，可以跳过部分文字继续后面的朗读，代码如下：

```
ULONG ulGarbage = 0;
WCHAR szGarbage[] = L"Sentence";
hr = m_cpVoice->Skip( szGarbage, SkipNum, &ulGarbage );
```

SkipNum 是设置要跳过的句子数量，值可以是正/负。根据 sdk 的说明，目前 SAPI 仅仅支持 SENTENCE 这个类型。SAPI 是通过标点符号来区分句子的。

（6）播放 wav 文件

SAPI 可以播放已存在的 wav 文件，ISpStream 接口实现的。

（7）将朗读的结果保存成 wav 文件

再获取朗读者、声音、语速及输出频率等信息后保存为.wav 格式。

（8）设置朗读音量和速度

设置音量，范围是 0-100：m_cpVoice->SetVolume((USHORT)hpos);

设置速度，范围是-10-10：m_cpVoice->SetRate(hpos);

（9）设置 SAPI 通知消息。SAPI 在朗读的过程中，会给指定窗口发送消息，窗口收到消息后，可以主动获取 SAPI 的事件。根据事件的不同，用户可以得到当前 SAPI 的一些信息，比如正在朗读的单词的位置，当前的朗读口型值（用于显示动画口型，中文语音的情况下并不提供这个事件）等等。下图 2.4 为动画口型示例：

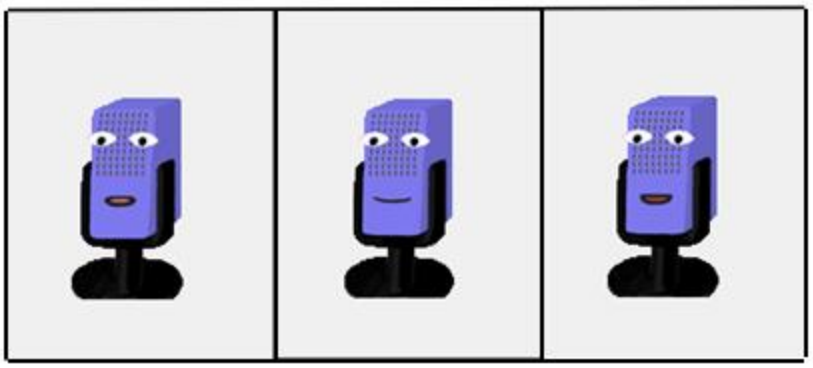

图 2.4　朗读动画口型示例

第 3 章 语音转换理论基础

本章中，我们从语音转换和语音信号处理的基础知识入手，介绍了语音信号的预处理、特征参数提取及 GMM 模型的理论知识，从而为下一章的语音转换算法奠定理论基础。

语音是人类社会交流信息最自然、最有效、最方便的工具，也是知识和信息的重要载体。作为人类的自然属性之一，由于人类个体声学器官的不同，说话人的声音特征也不相同。就像人的指纹或者 DNA 一样，人的声音也具有可识别的特征。

3.1 语音信号的基本特征

语音是说话人和听者之间互相传递的信息，是人类一种自然、方便、准确、高效的最重要的交流工具。人类的发音过程如下：首先，人产生出想要用语言表达的信息；然后把这些想要表达的信息变为由音素序列、韵律、响度、基音周期构成的语言编码；完成编码后，说话人会在合适的时候用脑神经发出信号命令声带振动，并改变声道的形状来发出编码中指定的声音序列。这样，说话人产生、发出语音信号，并把信号传递给听者时，就有了语音的感知过程。

声道可以视为一个谐振腔，其谐振频率称为共振峰频率，简称共振峰(Formant)。声门振动产生的脉冲称为声源，声源经过声道时，犹如通过一个具有某种谐振特性的腔体，被调制之后辐射出来。因此输出气流的频率特性既取决于声门脉冲串的特性，又取决于声道的特性。语音信号的形成和发音器官的运动密切相关，由于声道形状的改变和激励方式的变化相对于声源振动的速度要缓慢得多，故通常假设语音信号在 10-30ms 内认为短时平稳的。

语音的产生模型可简单分为激励模型、声道模型和辐射模型。如图 3.1 所示：

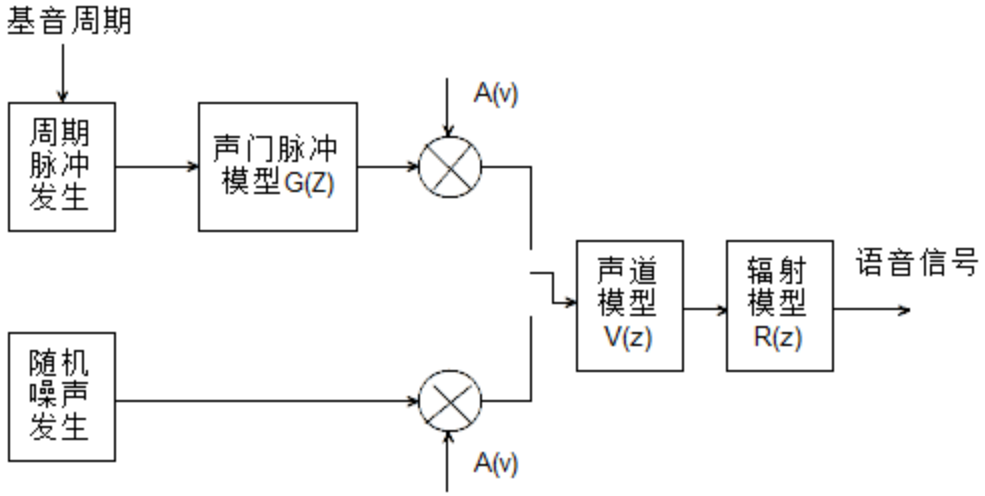

图 3.1　语音信号产生模型

激励发生器产生一串准周期脉冲序列（浊音）或者随机变化的噪声信号(清音)，激励模型参数的选取随着不同的语音信号输入而有所变化。

发浊音时，气流冲击绷紧的声带而产生振动，通过形状变化的声门时，形成准周期脉冲。脉冲随着声带的绷紧程度不同，振动的周期有所不同，即基音周期不一样。这样，在发浊音时，由声带的不断张开和关闭而产生了类似于三角脉冲的脉冲波。对浊音激励建模时，利用周期信号产生器发出的冲激序列来模拟激励信号，其周期等于激励信号的基音周期。然后将产生的冲激序列通过滤波器来模拟通过声门的实际信号。滤波器如公式(3.1)所示：

$$G(z) = \frac{1}{(1 - g_1 z^{-1})(1 - g_2 z^{-1})} \tag{3.1}$$

其中，g_1，g_2 都很接近 1。

对于清音，激励模型输出一个随机白噪音，实际上可以用均值为 0，方差为 1，并在时间或幅值上白色分布的序列来表示。

声道模型被定义为一个带有许多共振峰的滤波器，共振峰频率随着声道形状的变化而变化。通常用声道模型来体现声道的传输特性。声道模型一般采用如下的全极点函数$V(z)$：

$$V(z) = \frac{1}{\sum_{t=0}^{p} a_t z^{-t}} \tag{3.2}$$

其中$a_0 = 1$，a_t为实数，为声道模型参数。 p 为全极点滤波器的阶数。函数

的每一对极点对应一个共振峰。

语音信号的这种模型应该是“短时”的，因为语音信号是缓慢变化的，因此可以看成是短时平稳的，这就是语音信号短时分析的理论依据之一。这个模型对大多数语音来说都能很好的模拟，然而对理论要求有零点的鼻音和摩擦音等进行模拟时受到一些限制，对于鼻音和摩擦音的情况可以提高全极点模型的阶数，更好地逼近有零点的传递函数。

声道的终端是口和唇，辐射模型 $R(z)$ 与嘴型有关。研究表明，唇部的辐射效应在高频段较为明显，而在低频段影响较弱。因此，一般采用一个高通滤波器来表示辐射模型。其对应的表达式为：

$$R(z) = (1 - rz^{-1}), r \approx 1 \tag{3.3}$$

综上所述，一个语音信号完整的模型可以用三个子模型串联而成，其对应的传递函数为：

$$H(Z) = A(V)\,G(Z)\,V(Z)\,R(Z)$$
$$H(z) = AG(z)V(z)R(z) \tag{3.4}$$

3.2 语音信号的预处理

3.2.1 语音信号的采集与数字化

将原始语音信号变为数字信号，需要经过采样和量化两个步骤。采样频率是指每秒钟取得声音样本的次数，即采样率 f_s。采样率越高，声音的质量也就越好。将采样后的语音信号进行量化，即得到了数字语音信号。

录制一段语音信号，先将其用话筒转换为电信号，再用 A/D 转换器将其转换为数字信号，最后存入计算机中。设置录制语音的属性，获得 16KHz、16bit 的单声道音频格式以及标准 PCM 编码格式的 wav 文件，用于进行特征提取和模型训练。

3.2.2 语音信号的预加重和加窗

语音信号频谱的高频部分大约在 800Hz 以上按 6db/倍频程跌落，为了使信号频谱变得平坦，而便于进行频谱参数分析，需要提升语音的高频部分，即进行预加重处理。预加重一般用具有 6db/倍频的一阶数字滤波器来实现，如式（3.5）所示：

$$H(z) = 1 - \mu z^{-1} \tag{3.5}$$

式中，μ 值接近于 1。

若要恢复信号，即从做过预加重处理的信号频谱求实际频谱时，要进行去加重处理，即加上 6db/oct 的功率谱下降来还原原来的频谱特性。

语音信号是随时间而变化的，是一个非平稳态过程，但是在短时间范围内（一般在 10-30ms 内），其特性基本保持不变，即相对稳定，所以语音信号具有短时平稳性。任何语音信号的分析和处理必须建立在“短时”的基础上，即进行分帧处理。分帧是通过可移动的有限长度窗函数进行加权的方法实现的。

窗函数 $w(n)$ 乘以语音信号 $s(n)$，得到加窗信号 $s_w(n) = s(n) * w(n)$。分帧一般采用交叠分段的方法，这是为了使帧与帧之间平缓过度，从而保持其连续性。前一帧与后一帧的交叠部分称为帧移。帧移与帧长的比值一般取 $0-1/2$。

最常用的是汉明窗与矩形窗。其表达式如下（其中 N 为帧长）：

矩形窗：

$$w(n) = \begin{cases} 1, 0 \le n \le (\mathrm{N}-1) \\ 0, \mathrm{n} = \text{其他值} \end{cases} \tag{3.6}$$

汉明窗：

$$w(n) = \begin{cases} 0.54 - 0.46 * \cos(\dfrac{2\pi n}{N-1}), 0 \le n \le (N-1) \\ 0, n = \text{其他值} \end{cases} \tag{3.7}$$

3.2.3 语音信号的端点检测

对语音信号进行端点检测[1]的目的是准确地确定语音的起始点和终止点，区分语音信号和非语音信号。经过端点检测后，不仅能减少语音情感特征的采集量，节约处理时间，还能排除无声段或噪声段的干扰。

本文才用了短时帧能量和短时过零率相结合的双门限端点检测法来进行端点检测。在基于短时能量和短时过零率的双门限端点检测算法中，端点检测可以分为四个阶段：静音段、过渡段、语音段、结束段。通常设一个较高的门限 T_h 和一个比 T_H 稍低的门限 T_L，用以确定语音的起始点个结束点。在静音段，如果能

量或过零率超越了 T_L，就标记起始点。语音在起始点之后就进入了过渡段，然后将能量与过零率的数值进行比较，如果两个数值都在 T_L 以下，则认为当前处在静音状态，如果任何一个参数超过了 T_H，就可以确信进入了语音段。当前状态处于语音段时，如果短时能量和过零率的数值降到了 T_L 以下，而且时间长度小于最低时间门限，则认为是一段噪音，继续扫描后面的语音数据，否则标记好结束点[2]。

短时帧能量用每帧采样点值的加权平方和 E_n 来表示。即：

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)w(n-m)\right]^2 = \sum_{m=n-N+1}^{n} \left[x(m)w(n-m)\right]^2 \tag{3.8}$$

其中，$s(n)$ 为离散语音信号时间序列，$w(n)$ 为汉明窗函数，N 为窗长。语音信号的短时过零率 Z_n：

$$Z_n = \frac{1}{2} * \sum_{m=-\infty}^{\infty} \left| \operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)] \right| w(n-m) \tag{3.9}$$

其中，sgn 是语音信号 $x(n)$ 符号函数，$\operatorname{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ 0 & x(n) < 0 \end{cases}$。

图 3.2 为一段语音信号时域频谱、短时能量及过零率示例。 图 3.3 为端点检测后所得的有效语音。

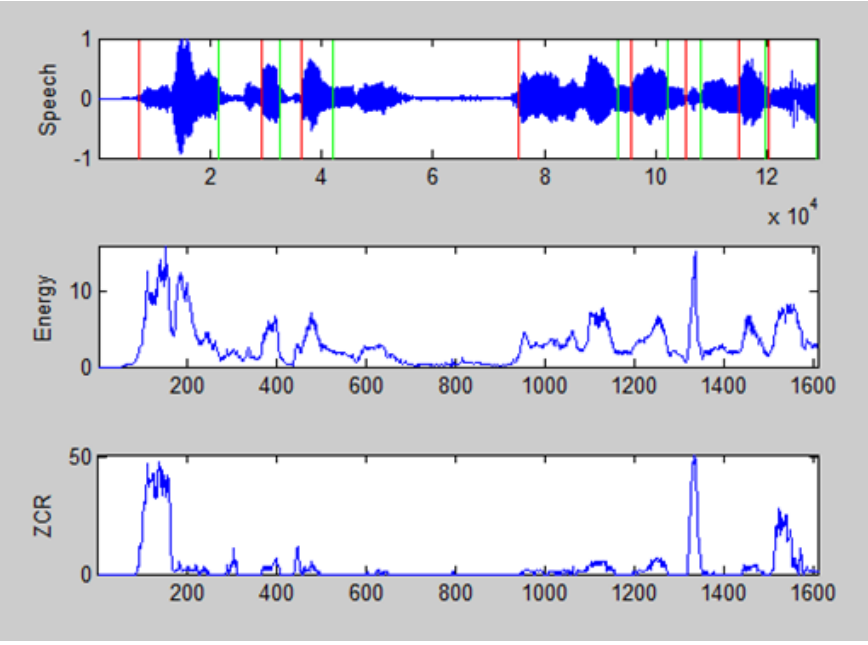

图 3.2 原始语音信号时域频谱、短时能量及过零率

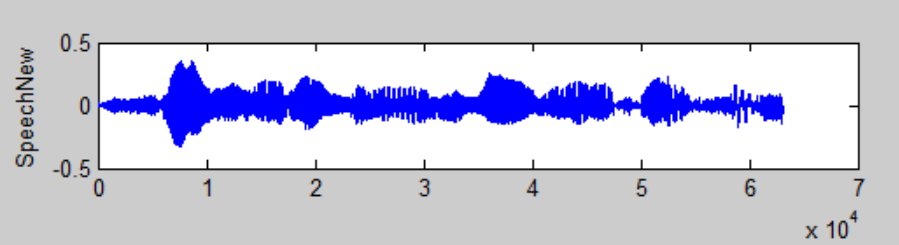

图 3.3 端点检测后的有效语音

3.3 特征参数提取

语音是具有声学特征的信号，声学特征是指音色、音高、音长和音强。语音的韵律特征是指音高、音强和音长方面所显示出来的抑扬顿挫。

3.3.1 基音频率检测

发浊音时，声带振动而产生准周期激励脉冲串，激励脉冲的振动周期就是基音周期。基音周期的倒数叫做基音频率，简称基频 F_0。采用短时分析的方法来估计基音周期的过程称为基音检测。基音频率的具体范围与个人的特征有关。一般来说，男性说话人的基音频率大致分布在 50-200Hz 范围内，而女性说话人和小孩的基音频率在 200-450Hz 之间。基音周期是语音信号处理中极为重要的参数之一，对于汉语来说更为重要。汉语是一种有调语言，基音的变化模式称为声调，不同的声调对应着不同的基频变化模式，声调携带着对辨别语义非常重要的信息。因此，正确有效的基频检测对于语音信号的研究具有非常重要的意义。

基音检测的方法大致可分为：(1)时频估计法；(2)变换域法，主要包括短时自相关法、平均幅度差法、倒谱法等。

权衡几种基频检测的方法，本文采取倒谱法。下面对倒谱法进行基音检测的原理介绍如下。

当信号序列为 $s(n)$，它的傅里叶变化为

$$X(w) = FT[x(n)] \tag{3.10}$$

则序列

$$\hat{x(n)} = FT^{-1}[\ln | X(w) |] \tag{3.11}$$

称 $\hat{x(n)}$ 为倒频谱，简称为倒谱，即 $s(n)$ 的倒谱序列 $\hat{x(n)}$ 是 $s(n)$ 幅值谱对数的傅里叶逆变换。其中 FT 和 FT^{-1} 分别表示傅里叶变换和傅里叶逆变换。$\hat{x(n)}$ 的量纲

是 Quefrency，又被称为倒频，它实际的单位还是时间单位。

语音 $s(n)$ 是由声门脉冲激励 $u(n)$ 经声道响应 $v(n)$ 滤波而得（在不考虑口唇辐射的条件下），即

$$x(n)=u(n)*v(n) \tag{3.12}$$

设这三个量的倒谱分别为 $\hat{x}(n)$、$\hat{u}(n)$ 及 $\hat{v}(n)$，则有

$$\hat{x}(n) = \hat{u}(n)+ \hat{v}(n) \tag{3.13}$$

可见，在倒频谱域中 $\hat{u}(n)$ 及 $\hat{v}(n)$ 是相对分离的，说明包含有基音信息的声脉冲倒谱可与声道响应倒谱分离，因此从倒频谱域分离 $\hat{u}(n)$ 后恢复出 $u(n)$，从中求出基音周期。

已知基音频率范围为 60-500Hz 之间，当采样频率为 f_s 时，在倒频率域上 60Hz 对应的基音周期（样点值）为 $P_{\max}=f_s/60$，而 500Hz 对应的基音周期（样点值）为 $P_{\min}=f_s/500$。所以在计算出倒谱后，就在倒频率为 $P_{\min}-P_{\max}$ 之间寻找倒谱函数的最大值，倒谱函数最大值对应的样点数就是该 i 帧语音信号的基音周期 $T_0(i)$（以样点为单位，如果要转成秒，则要乘 $\Delta t=1/f_s$），基音频率为 $F_0(i)=f_s/T_0(i)$。

录制一段语音，在 Matlab 上编写程序，得到下图为用倒谱法提取得到的基音周期图。

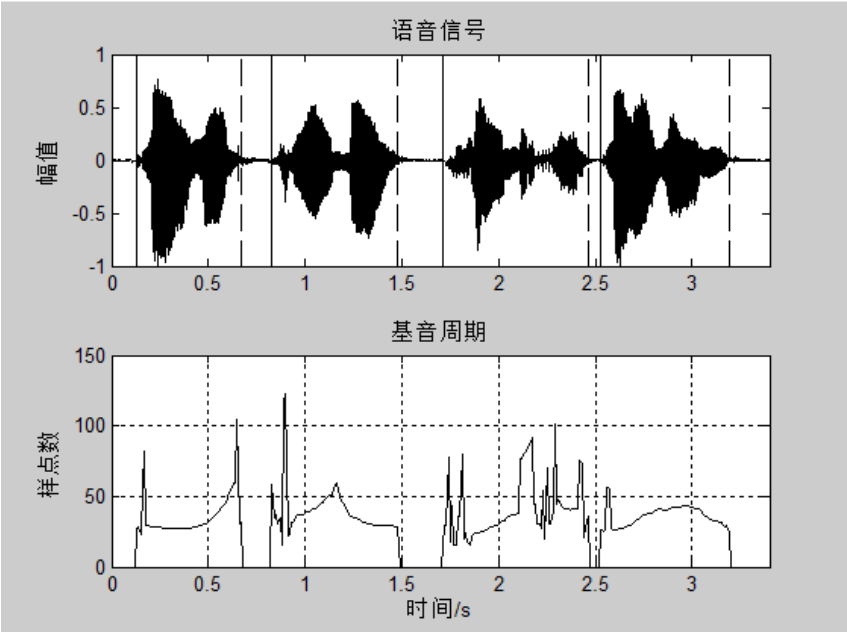

图 3.4　　倒谱法提取得到的基音周期图

从图 3.4 中可以看出，在基音周期轨迹中出现了偏离实际轨迹的数值，图中出现了突然跳变的偏离值，一般是实际值的 2 倍、3 倍或 $1/2$。这说明，在基音周期检测之后还需要进行后处理，以消除这些偏离值点。

3.3.2 基音检测后的平滑处理

基音检测算法很难做到处处准确可靠，基音频率值落在实际基音的倍频或分频所对应的频率等情况也时有发生。这种在求得的基音轨迹中有一个或几个基音频率偏离了正常轨迹（通常是偏离到实际值的 2 倍、3 倍或 $1/2$）的偏离点称为基音轨迹的“野点”。

为了去除这些野点，可以采用各种平滑算法，其中最常见的是中值滤波算法和线性平滑算法。中值滤波其基本原理是：设 $x(n)$ 为输入信号， $y(n)$ 为中值滤波器的输出。在 n_0 点的左右各取 L 个样点，连同 n_0 点一共取得 $(2L+1)$ 个样点，取这 $(2L+1)$ 个样值的中间值为平滑器的输出。即 n_0 处的输出值 $y(n_0)$ 就是将滑动窗的中心移到 n_0 处时窗内输入样点的中值。L 值一般取为 1 或 2，即中值平滑的“窗口”一般为 3 或 5 个样值，称为 3 点或 5 点中值平滑[3]。线性平滑是用滑动窗进行线性滤波处理，即

$$y(n) = \sum_{m=-L}^{L} x(n-m)w(m) \tag{3.14}$$

式中，$\{w(m)\}$，$m=-L,-L+1,...,0,1,2,...,L$，为$(2L+1)$点平滑窗，满足

$$\sum_{m=-L}^{L} w(m) = 1 \tag{3.15}$$

例如三点窗的权值可取值为$\{0.25,0.5,0.25\}$。线性平滑在纠正输入信号中不平滑处样点值的同时，也使附近各样点值得到了修改。本文采用线性平滑处理方法，得到了平滑处理后的基音周期和基音频率图如下图 3.5 所示。

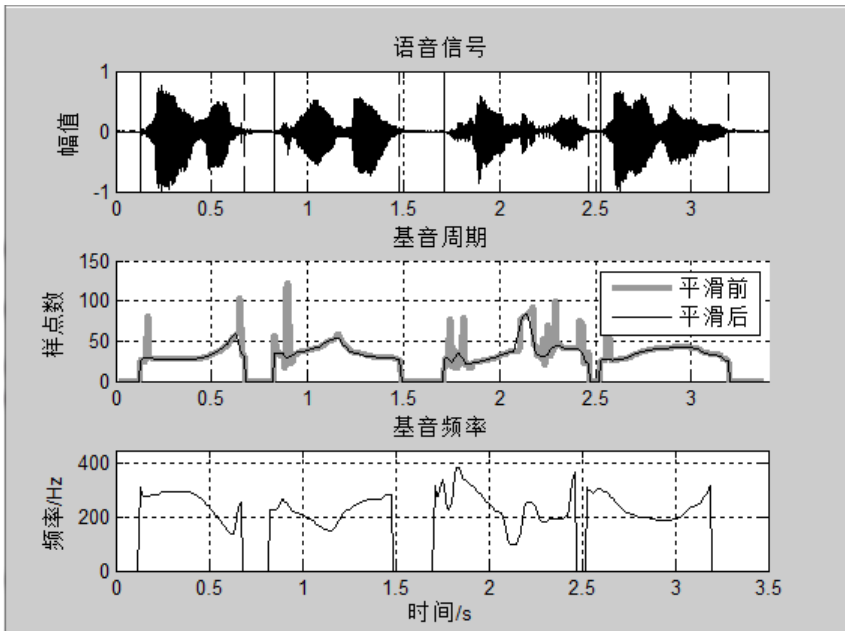

图 3.5　平滑处理后的语音信号、基音周期和基音频率图

3.3.3 频谱参数

频谱参数能够反映出一些重要的语音特征。实验表明，语音的感知过程与人类听觉系统具有频谱分析功能是紧密关联的。因此，对语音信号进行频谱分析是对语音信号进行分析和处理的重要方法。语音频谱是在频域中，语音信号的能量与频率的分布关系。从广义上讲，语音信号的频谱分析包括语音信号的频谱、LPC 谱、倒频谱、复倒谱分析等，常用的频谱分析方法有短时傅里叶变换法、线性预测分析和同态分析等。

1.短时谱

由于语音信号是非平稳的，但语音特性短时间内保持不变，因此可以对某一

帧语音进行傅里叶变换，即短时傅里叶变换。定义如下：

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwm} \tag{3.16}$$

其中，$w(n-m)$是窗函数序列。不同的窗函数，将得到不同的傅里叶变换结果。

2.倒谱

语音信号由激励信号与声道响应相卷积产生，要提取声道谱包络，需要解卷积去掉激励信息。倒谱分析（同态处理）是把激励源和声道的冲激响应分离的较好的解卷积方法。只需十几个倒谱系数就能相当好的描述语音信号的声道响应，因此在语音信号中占有重要地位。语音倒谱$c(n)$是语音信号$x(n)$的傅里叶变换的模的对数的逆傅里叶变换，由式(3.17)给出：

$$c(n) = DFT^{-1}\left\{\ln\left|DFT\left[x(n)\right]\right|\right\} \tag{3.17}$$

3.LPC 谱

线性预测分析（LPC）是 1947 年，维纳提出的一项技术。现在在语音信号处理中，LPC 主要应用在语音参数的分析和提取方面。线性预测分析（LPC）的基本思想是：由于语音信号取样值之间具有相关性，所以可以用过去若干个取样值的线性组合来预测或者逼近现在或将来的取样值。利用预测均方误差最小的准则，求得唯一的一组预测系数，使样值点与预测值之间的差异达到最小。线性预测主要有预测和建模两个作用。对一帧语音求出一组预测系数a_i后，这一组预测系数a_i就是语音的模型参数，LPC 谱由下式给出：

$$H(e^{jw}) = \frac{G}{1-\sum_{i=1}^{p} a_i e^{-jwi}} \tag{3.18}$$

由式可知，只要知道预测系数a_i就可得到 LPC 谱。

4.LPCC 谱

从语音的线性预测系数a_i可以直接求出其倒谱系数$\hat{h}(n)$：

$$\hat{h}(0) = 0 \tag{3.19}$$

$$\hat{h}(1) = -a_1 \tag{3.20}$$

$$\hat{h}(n) = -a_n - \sum_{k=1}^{n-1}(1-k/n)a_k\,\hat{h}(n-k) \quad 1<n\le p \tag{3.21}$$

$$\hat{h}(n) = -\sum_{k=1}^{p}(1 - k/n)a_k \hat{h}(n-k) \quad n > p \tag{3.22}$$

按式求得的$\hat{h}(n)$称之为 LPCC 系数，将其做傅里叶变换得到 LPCC 谱。

3.4GMM 模型基本知识

为了实现与文本无关的语音转换，首先要对源语音进行分析，提取源语音的特征参数，从而获取映射规则。然后再根据映射规则将源语音的特征参数转换为目标语音的特征参数，最后将其合成出转换语音。本研究主要提取源语音和目标语音的韵律特征，利用 *GMM* 模型[4,5]建立映射规则，最终实现不同说话人之间的转换。

高斯密度函数估计是一种参数化模型。有单高斯模型(Single Gaussian Model SGM)和高斯混合模型(Gaussian Mixture Model *GMM*)两类。类似于聚类，根据高斯概率密度函数(PDF)参数的不同，每一个高斯模型可以看作一种类别，输入一个样本 X，即可通过 PDF 计算其值，然后通过一个阈值来判断该样本是否属于高斯模型。很明显，SGM 适合于仅有两类别问题的划分，而 *GMM* 由于具有多个模型,划分更为精细，适用于多类别的划分，可以应用于复杂对象建模。

1.单高斯混合模型

多维高斯（正态）分布概率密度函数 PDF 定义如下：

$$N(x;\mu;\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \tag{3.23}$$

与一维高斯分布不同，其中 x 是维数为 d 的样本向量（列向量），μ 是模型期望，Σ 是模型方差。

2.高斯混合模型

高斯混合模型是单一高斯概率密度函数的延伸，由于 *GMM* 能够平滑地近似任意形状的密度分布，因此近年来常被用在语音、图像识别等方面，得到不错的效果。

设一组特征参数 $X=\{X_1,...,X_n\}$，参数个数为 n，假设每个点均由一个单高斯分布生成（参数 μ_i 和 Σ_i 未知），而这一组参数共由 M 个单高斯模型生成，具体哪个参数 X_i 属于哪个单高斯模型未知，且每个单高斯模型在混合模型中占的

比例α_i未知，将所有来自不同分布的数据点混在一起所构成的分布称为高斯混合分布。

从数学上讲，我们认为这些数据的概率分布密度函数可以通过加权函数表示：

$$p(x_i) = \sum_{j=1}^{M} \alpha_j N_j(x_i; \mu_j; \Sigma_j) \tag{3.24}$$

上式即称为*GMM*，$\sum_{j=1}^{M}\alpha_j = 1$，其中

$$N_j(x_i;\mu_j;\Sigma_j) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_j|}} \exp\left[-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)\right] \tag{3.25}$$

表示第 j 个 SGM 的 PDF。

令$\theta_j = (\alpha_j;\mu_j;\Sigma_j)$，*GMM* 共有 M 个 SGM 模型，现在，我们就需要通过样本集 X 来估计 *GMM* 的所有样本参数$\Theta = (\theta_1,\theta_2,...,\theta_M)^T$，样本 X 的概率公式为：

$$L(X \mid \Theta) = \log \prod_{i=1}^{N} \sum_{j=1}^{M} \alpha_j N_j(x_i; \mu_j; \Sigma_j) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} \alpha_j N_j(x_i; \mu_j; \Sigma_j) \tag{3.26}$$

采用 *EM* 估计 *GMM* 的参数：

通常采用 *EM* 算法（*Expectation Maximum*）对 *GMM* 参数进行估计。具体方法如下：

1) 初始值

运用 EM 算法训练 *GMM* 之前需要得到模型参数的初始化，而模型参数的初始值对系统的识别性能影响很大，因此如何确定模型参数的初始值非常关键。常用的初始化方法有：

随机选择法：均值的初始值是在语音数据中随机选取 M 个向量，初始化为单位矩阵。

平均分段法：将语音均分为 M 段，求各个分段的平均值，然后根据均值来求出方差，权重初始化为 1/ M 。

聚类选择法：聚类法用的最多的就是 LBG 以及 K-means 算法，这种方法与

高斯混合模型的组成原理相吻合，根据样本先验概率分布的理论，将特征矢量分为 M 个聚类，进行初始化。本文采用的是 K-means 算法。

K-means 算法是一种聚类算法，将簇中对象的均值作为簇的中心，可以是一个虚点，计算其他点与各个簇中心距离，归入距离最近的簇中。具体实现步骤如下：

随机选取 M 个初始聚类中心 $(K_1(1), K_2(1), ..., K_M(1))$；

将训练数据分配到这 M 个聚类中，对于每一个数据 x_i，根据最小距离准则，计算出距离最小的那个聚类，计算公式如下：

$$\left|x_i - k_m(j)\right| \le \left|x_i - k_n(j)\right| \quad \forall m \ne n 且 m,n = 1,2,...,M \tag{3.27}$$

从而将 x_i 归于第 m 类 S_m；

更新聚类中心：

$$k_m(j+1) = \frac{1}{N_m}\sum_{X_i \in k_m(j)} x_i \quad i = 1,2,...,M \tag{3.28}$$

其中 N_m 为聚类 S_m 中的样本个数。反复迭代更新直到聚类中心收敛。

根据聚类所得结果可得模型参数初始值：

$$p_m = \frac{N_m}{T} \tag{3.29}$$

$$\mu_m = \frac{1}{N_m}\sum_{x_i \in k_m} x_i \tag{3.30}$$

$$\Sigma_m = \frac{1}{N}\sum_{x_i \in k_m}(x_{mi} - \mu_{mi})^2, i = 0,1,\ldots,T-1 \tag{3.31}$$

图 3.6 给出了 K-means 算法的流程图，由于算法开始的时候用到的中心值为随机选取的，如果开始选取的值不理想，就可能造成聚类结果的局部最优，从而得不到想要的结果。但是由于 K-means 算法比较容易实现，在实际应用中依然比较广泛。

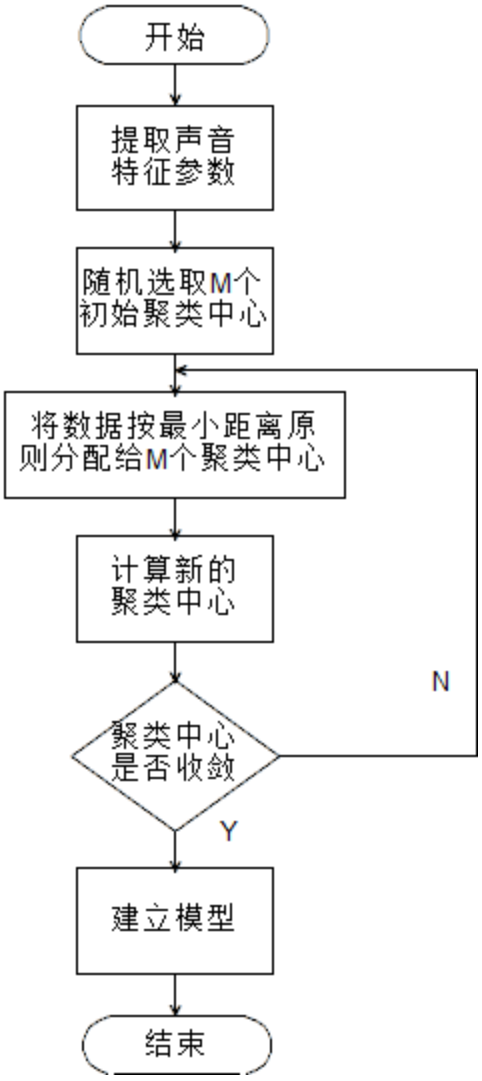

图 3.6　K-means 算法的流程图

2)　算法流程

(1)估计步骤(E-Step)：

令α_j的后验概率为：

$$\beta_j = E(\alpha_j \mid x_i;\Theta) = \frac{\alpha_j N_j(x_i;\Theta)}{\sum_{j=1}^{M}\alpha_j N_j(x_i;\Theta)} \qquad 1\le i\le n, 1\le j\le M \tag{3.32}$$

(2)最大化步骤(M-Step)：

更新权值：

$$\alpha_j = \sum_{i=1}^{N}\beta_{ij} \tag{3.33}$$

更新均值：

$$\mu_j = \frac{\sum_{i=1}^{n} \beta_{ij} x_i}{\sum_{i=1}^{n} \beta_{ij}} \tag{3.34}$$

更新方差矩阵：

$$\sum_j = \frac{\sum_{i=1}^{n} \beta_{ij} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} \beta_{ij}} \tag{3.35}$$

(3)收敛条件

不断地迭代 E 和 M 步骤，重复更新上面的三个值，直到参数的变化不显著，即$\left|\Theta - \Theta'\right| < \varepsilon$，$\Theta$为更新后的参数，通常$\varepsilon = 10^{-5}$[6]。

3.5 语音转换的评价标准

3.5.1 客观评价标准

一种客观评估方法是采用谱包络的误差（SD：Spectral Distortion）来评价系统性能。在语音转换的研究中，谱包络的误差主要采用两种客观方法来测量：

1.测量源说话人语音、转换后的语音和目标说话人语音之间的平均谱失真。例如，采用的是两个平均谱失真的比值，分别是转换后语音与目标说话人语音，源说话人语音与目标说话人语音之间的平均谱失真。

2.语音编码中最常用的信号噪声比（SNR）也被用于谱包络转换的客观评测中，SNR 的值越大，则转换的效果越好，SNR 公式如下：

$$SNR(s_1, s_2) = 10 \log_{10} \frac{\sum \left|FFT(s_1(n))\right|^2}{\sum \left(\left|FFT(s_2(n))\right| - \left|FFT(s_1(n))\right|\right)^2} \tag{3.36}$$

另一种方法是采用说话人识别系统，将转换后的语音输入说话人识别系统中，比较转换后的语音与源和目标说话人的相似度，若与目标说话人相似度比与源说话人的相似度大就表明转换成功。

在下一节中将介绍说话人识别系统的详细设计过程。

3.5.2 主观评价标准

ABX 测试方法是用来检测说话人识别度的一种方法。在这个测试中，测听

者测听语音 A，B 和 X，并判断在语音的个性特征方面语音 A 还是 B 更接近于 X。此处 X 是经语音转换后得到的语音，而 A 和 B 分别为源语音和目标语音。实验通过百分制对合成语音与目标语音的相似性进行打分。若 X 与 B 越相似，百分比越大。通过多次实验统计判断百分比来评价系统性能，百分比越大系统性能越好。

MOS 得分是测试者在听到合成的语音样本之后对声音进行打分，分值从 1 到 5，标准如下：优（5 分），良（4 分），中（3 分），差（2 分），劣（1 分）。最后全体测试者给出的平均分就是所测语音质量的 MOS 分。

本文对转换结果的评价标准主要以客观评价为主，一方面若采用主观评价需要有大量的测听者对声音进行打分，在这一点上无法得到满足。另一方面，主观评价存在着一定的主观因素，而客观评价将评价标准数值化，大小的比较一目了然，更具有说服力。

3.6 说话人识别

说话人识别和指纹识别、虹膜识别等一样，属于生物识别的一种，被认为是最自然的生物特征识别身份鉴定方式，因此又被称为“声纹”识别。声纹识别实际上是一个模式识别的过程，大致可以分为两个阶段，即训练阶段和识别阶段。训练阶段，通过采集用户的语音数据，经过相关的算法处理，为每一个用户建立一个特定的模型（本文采用 GMM 模型[7,8]），然后将这些用户的模型存放在声纹模型库中，这一过程可称为“注册”。识别阶段，根据待识别的用户的语音（测试语音），提取出特征参数，然后与模型库中的模型参量进行比较，按照一定的相似性准则或概率似然度进行比较，距离最小或是概率得分最高的为识别结果。

在本文中则使用说话人识别系统作为语音转换结果的评价标准，源语音和目标语音为模型库，转换后的语音作为测试语音，与二者进行匹配，若与目标语音的相似概率得分较高则表明转换成功。

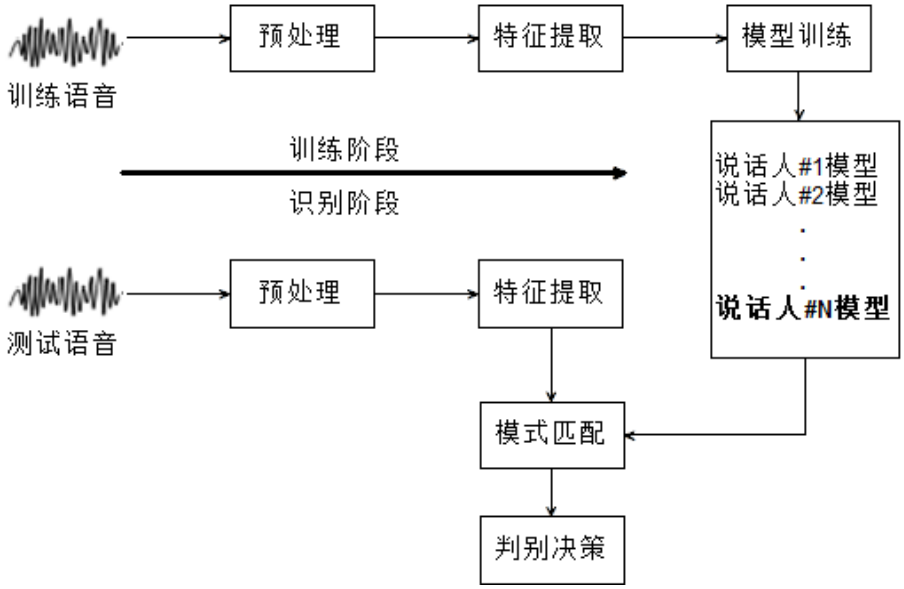

图 3.7 说话人识别系统结构图

如图 3.7 所示，说话人识别系统分为如下几个基本步骤：

(1)语音信号的预处理和特征提取，其中预处理包括预加重，端点检测等。提取得到的特征参数是能够有效表征说话人语音特征的；

(2)说话人模型的建立和模型参数的训练；

(3)测试语音与说话人模型的相似度计算；

(4)识别与判决策略，即根据匹配计算的结果，采用某种判决准则判定说话人到底是谁（说话人辨认）。

从总体上讲，说话人识别方法两个主要的研究重点是语音特征参数提取，声音模型的建立及识别决策策略。下面将分别介绍这两个方面。

- MFCC 特征参数提取

语音信号的特征矢量在频域上主要有线性预测倒谱系数（LPCC）和梅尔频率倒谱系数（MFCC）。LPCC 逼近人类发声机理但对于辅音的描述能力较差，抗噪声性能较差。MFCC 充分考虑人耳听觉结构和人类发声和接受声音的机理特性具有很好的鲁棒性。而且在没有任何假设前提条件，其具有较好的识别性能和抗噪能力。因此本文提取 MFCC 特征参数。MFCC 特征参数提取原理框图如图 3.8 所示。

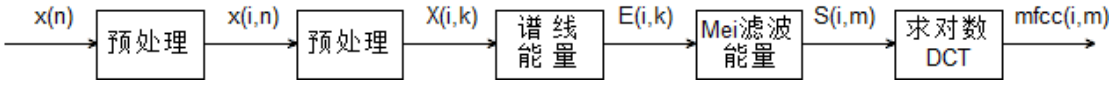

图 3.8 MFCC 特征参数提取原理框图

(1)预处理

预处理包括预加重、分帧、加窗函数。

预加重：声门脉冲的频率响应曲线接近于一个二阶低通滤波器，而口腔的辐射响应也接近于一个一阶高通滤波器。预加重的目的是为了补偿高频的损失，提升高频分量。

分帧处理：在 3.3.2 中指出，由于语音信号是一个准稳态的信号，把它分成较短的帧，在每帧中可将其看作稳态信号，可用处理稳态信号的方法来处理。同时，为了使一帧与另一帧之间的参数能较平稳地过渡，在相邻两帧之间相互有部分重叠。

加窗函数：加窗函数的目的是减少频谱中的泄露，将对每一帧语音乘以矩形窗或汉明窗。（详细说明见 3.3.2）

语音信号 $x(n)$ 经预处理后为 $x_i(m)$，其中下标 i 表示分帧后的第 i 帧。

(2)快速傅里叶变换

对每一帧信号进行 FFT 变换，从时域数据转变为频域数据：

$$X(i,k) = FFT\left[x_i(m)\right] \tag{3.37}$$

(3)计算谱线能量

对每一帧 FFT 后的数据计算谱线的能量：

$$E(i,k) = \left[X(i,k)\right]^2 \tag{3.38}$$

(4)计算通过 Mel 滤波器的能量

把求出的每帧谱线能量谱通过 Mel 滤波器，并计算在该 Mel 滤波器中的能量。在频域中相当于把每帧的能量谱 $E(i,k)$（i 表示第 i 帧，k 表示频域中的第 k 条谱线）与 Mel 滤波器的频域响应 $H_m(k)$ 相乘并相加：

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k)H_m(k), 0 \le m < M \tag{3.39}$$

(5)计算 DCT 倒谱

序列 $x(n)$ 的 FFT 倒谱 $\hat{x}(n)$ 为

$$\hat{x}(n) = FT^{-1}\left[\hat{X}(k)\right] \tag{3.40}$$

式中，$\hat{X}(k)=\ln\{FT[x(n)]\}=\ln\{X(k)\}$，$FT$ 和 FT^{-1} 表示傅里叶变换和傅里叶逆变换。序列 $x(n)$ 的 DCT 为

$$X(k)=\sqrt{\frac{2}{N}}\sum_{n=0}^{N-1}C(k)x(n)\cos\left[\frac{\pi(2n+1)k}{2N}\right],k=0,1,\ldots,N-1 \quad (3.41)$$

式中，参数 N 是序列 $x(n)$ 的长度；$C(k)$ 是正交因子，可表示为

$$C(k)=\begin{cases}\sqrt{2}/2, k=0\\ 1, k=1,2,\ldots,N-1\end{cases} \quad (3.42)$$

在式（3.40）中求取 FFT 的倒谱是把 $X(k)$ 取对数后计算 FFT 的逆变换。而这里求 DCT 的倒谱和求 FFT 的倒谱相类似，把 Mel 滤波器的能量取对数后计算 DCT：

$$mfcc(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log\left[S(i,m)\right]\cos\left(\frac{\pi n(2m-1)}{2M}\right) \quad (3.43)$$

式中，$S(i,m)$ 是由式（3.39）求出的 Mel 滤波器能量；m 是指第 m 个 Mel 滤波器（共有 M 个）；n 是 DCT 后的谱线。

这样就计算出了 MFCC 参数。

- GMM 声音模型及识别决策方法

本文采用的是 GMM 声音模型，高斯混合模型可以从概率统计的角度，对每一个声音类别进行表示，即每一种声音类别都被指定了一个不同的高斯概率密度函数。为了避免在模型训练中出现局部最优的情况，我们采用 EM 算法来自适应参数的优化与调整。GMM 和 EM 算法的相关知识已在 3.4 节介绍过。以下将着重介绍语音匹配判断说话者的过程。

说话人辨认是将测试语音与模型库中所有模型进行匹配计算，根据似然函数从模型集合中找出和测试语音最相似的模型人[9]。

假设说话人模型集合中有 N 个人，我们用高斯混合模型表示模型集合，对于一段测试语音 X，对比高斯混合模型集合中的所有模型，找到一个最大后验概率值得模型：

$$L=\arg\max_{1\le k\le N}P(\theta_k\mid X) \quad (3.44)$$

其中 L 表示识别出的说话人，根据最大后验概率与贝叶斯定理上式可以改写为：

$$L = \arg\max_{1\le k\le N} \frac{P(X \mid \theta_k)P(\theta_k)}{P(X)} = \arg\max_{1\le k\le N} \frac{P(X \mid \theta_k)P(\theta_k)}{\sum_{M=1}^{N} P(X \mid \theta_m)P(\theta_m)} \tag{3.45}$$

假设每个说话人模型出现的概率相同（都为 1/N）则上式可以简化近似为：

$$L = \arg\max_{1\le k\le N} P(X \mid \theta_k) \tag{3.46}$$

此时最后验概率就变为了最大似然估计，为了简化运算可以将上式两边进行取对数：

$$\log L = \arg\max_{1\le k\le N} \log(X \mid \theta_k) \tag{3.47}$$

第 4 章 语音转换和语音合成

4.1 基频目标模型

汉语是有调语言，既包含声调，也包含语调。声调是基于音节的固定的音高模式。语调则是基于语句的音高的变化轮廓。基频曲线轮廓的变化既代表了基于音节的声调变化，也代表了基于语句的语调变化。因此，对于语音转换，基频轮廓起着重要的作用。要想合成的目标语音有更好的可懂度和自然度，就需要建立更加合理的基频模型[10]。

4.1.1 传统 Pitch Target 模型

Pitch Target 模型是由许毅教授提出的基于汉语音节的基频模型。他认为，基频曲线轮廓并不只是底层功能的直接体现。事实上，基频曲线与底层功能之间有着很多的不同。对于汉语语音，对于语句的音节都有一个音高目标（target），但是音高目标并不是固定的，音高目标的不固定性会使得底层功能与基频曲线轮廓有着很大的不同。在实际发音过程中，基频曲线以渐近线的形式不断逼近音高目标，通常会在音节的前半部分出现较长的过渡段，这也造成了底层功能与基频曲线的不同。

人之所以能发出自然清晰的语言，一部分是因为基频曲线包含了超音段信息。由于基频曲线轮廓的变化体现了说话人的韵律特征变化趋势，因此它的斜率非常重要。汉语语言中有阴、阳、上、去四种声调及其各自的变化形式，许毅教授针对这四种声调及变化形式将 target 模型分为动态和静态两种；静态的 target 模型分别有高的和低的 target，高的 target 对应了声调中的阴平，低的 target 对应了声调中的上声。动态的 target 则是上升或下降的直线，上升的直线是指阳声的 target，下降的直线是指去声的 target。每个 target 都是在语音的音节基础上生成的。而求得的 Pitch Target 模型就是各个音节的基频曲线不断与之逼近的所有 target 的综合。基频渐近逼近 target 的过程可以用一个指数项来表示，即基频的实际基频曲线可表示为：

$$pitch(t) = C + \beta \exp(-\alpha t) \tag{4.1}$$

其中 C 是 target 的等式，一般取直线：$C = mt + b$。从公式中我们可以看出，

当 t 趋于正无穷时，等式的指数项部分趋于 0，基频值即为 target 值。这里，我们假设 t 到达音节的末尾时，指数项部分为 0。即假设基频曲线在音节的末尾和 target 重合[11]。

由于在该模型中，各个基频曲线的 target 只是简单地分成上升、下降和横直的直线，这样无法准确地体现原来的基频曲线，语音合成效果也受到影响。因此，本文在应用 Pitch Target 模型进行转换之后，在语音合成时，又对合成的语音进行了基音频率和频谱的修改，使其影响尽量减小[12]。

4.1.2 改进的 Pitch Target 模型

在实现的过程中，本文以音节为单位的 Pitch Target 模型进行了一些修正，使得 Pitch Target 模型更加的实用化，求取也更加方便。如图 4.1 所示，基频曲线是在前后音节的韵律曲线作用下不断向 target 逼近的过程。本节的主要任务就是对 Pitch Target 模型进行合理的调整和修正，以实现基频的 target 参数快速提取。

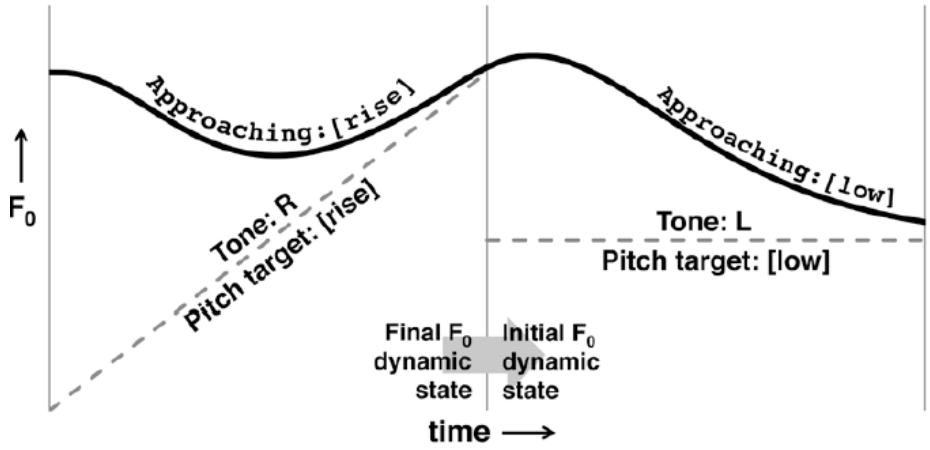

图 4.1　Pitch Target 模型

我们在前面的各个音调单音节基频检测试验的基础上，对 target 模型中的 target 进行重新设定，新的 target 的表达式如下：

$$y(t) = T(t) + \beta \exp(-\lambda t) \tag{4.2}$$

$$T(t) = at + b \tag{4.3}$$

其中，$0 \le t \le D, \lambda \ge 0$。

本文主要采用基于音节的 Pitch Target 模型和 *GMM* 进行语音转换，$[0, D]$为

每个音节的时间范围。其中，$T(t)$为我们所求的 Pitch Target，是基频曲线的近似值，一般为直线。$y(t)$代表了音节的实际基音轮廓。参数 a 和 b 分别是所求的 Pitch Target 直线的斜率和截距。系数β的值为$t=0$时 Pitch Target 直线与F_0表面轮廓的距离。参数λ是一个正数，代表了指数衰减率，代表了所求的 Pitch Target 直线逼近F_0表面轮廓的速率。λ值越大，代表了两者越接近[13]。由于人类一些发音器官生理上的极限，这些参数都受到了一定的限制，如最大间距范围和最大速度，音调的变化。可以应用非线性回归算法来估计这些参数。然而，如引文 13 所说，类似上述的模型并不一贯具有这种良好的估计性能，其中一个良好的解决方案是用一些值代替所谓的预期值参数。用(t_0, y_0)表示F_0轮廓的第一个点，将这个点插入等式来代替β，可以得到：

$$y(t) = (y_0 - b)\exp(-\lambda t) + at + b \tag{4.4}$$

其中，$t_0 = 0$。

(t_1, y_1)代表了一个指数分量衰减为 0 时的点，也是 Pitch Target 达到或者接近F_0轮廓的位置。请注意，虽然指数函数不能变为 0，但可以无限接近于 0。在这里，我们强制指数部分变为 0，是为了简化模型。因此，等式变为：

$$y_1 = at_1 + b \tag{4.5}$$

将参数a或b带入式中，得到：

$$y(t) = (y_0 - b)\exp(-\lambda t) + (\frac{y_1 - b}{t_1})t + b \tag{4.6}$$

或

$$y(t) = (y_0 - y_1 + at_1)\exp(-\lambda t) + at + y_1 - at_1 \tag{4.7}$$

用于估计的非线性回归方程是被广泛使用的 Levenberg-Marquardt 算法的一种实现。当非线性回归失败时，我们便使用了线性回归，其中β和λ参数被设置为 0。在实践中，前两个F_0的值的平均值用于估计(t_0, y_0)，因为第一点有可能是异常点。对于(t_1, y_1)这个中间点可以凭经验进行选择。这个中间点的选择是基于

假设基音目标和实际基音轮廓在音节的中间点重叠。因此，每一个音节关联一个基音目标。

4.2 基于基频目标模型的语音转换

本文设计的语音转换系统主要由 4 大模块组成：(1)语音信号预处理、划分标注和提取特征参数模块；(2)模型训练模块；(3)特征转换模块；(4)语音合成及后期处理模块。

其中预处理、划分标注和参数提取模块主要是对语音信号进行分帧、加窗、预加重，端点检测，并按音节来对语音段进行标注，记录每个音节的起始和终止时间点，以及分别提取源和目标语音的 Pitch Target 模型的特征参数。模型训练模块主要是对提取的 Pitch Target 特征参数进行 *GMM* 建模，并训练得到转换规则。在转换阶段，首先对源语音进行分析并提取特征参数，再根据在训练阶段得到的语音转换规则进行转换。语音合成及后处理模块主要是对转换后的特征参数进行语音重建，并对合成的语音进行平滑处理以及时长和能量修改，并使之尽可能接近所要转换成的目标说话人。

其结构如下图：

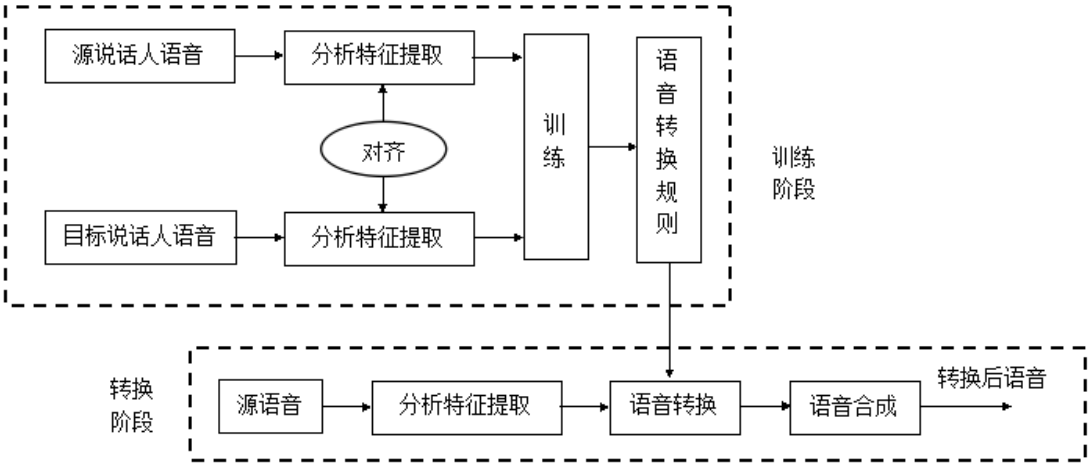

图 4.2　语音转换系统结构图

4.2.1 基音频率转换

基音频率转换之前，我们首先对语音进行分帧、端点检测等前端预处理，得出适合转换处理的语音帧，然后提取语音帧的特征参数。这里，我们首先将语句切分为音节，分别以音节为单位对音节进行分帧，提取音节的基音频率(F_0)轮廓，

并建立语音的基频参数化描述模型，本论文采用了 Pitch Target 模型。在该模型中，一个音节的基频曲线可以使用一组模型参数 (a,b,λ,β) 来表示。本文先将连续语句分解成离散的音节，将源说话人语音的音节韵律特征参数 (a,b,λ,β) 和目标说话人语音的音节韵律特征 (a,b,λ,β) 对齐，作为源参数和目标参数输入到 GMM 模型中进行训练。在语音转换过程中，通过将源说话人语音语句中各个音节的 Pitch Target 参数 (a,b,λ,β) 作为输入，而获得目标说话人语音的 Pitch Target 参数 (a,b,λ,β) 输出。

(1)音节划分和语音标注

由于 Pitch Target 模型是以音节为单位进行分析的，所以需要将一段连续语音按音节进行划分和标注。音节边界的标注主要为基于“双限门法”[14]的端点检测方法和基于 HMM 的强制对齐法[15]。“双限门法”算法简单，但是精度不高，无法检测相连的音节；强制对齐方法是一种可训练的方法，虽然能够较为精确地检测音节边界，但是需要标注好的语料进行训练。为简单起见，本文采用双限门法加手工修改的方法进行音节边界的标注。首先利用双限门法进行初步标注，然后借助 Praat 软件工具手工修改。通过 praat 显示的语音波形图和人工反复听辨，在保证两字之间达到最优发音的情况下，在边界处点击鼠标，以确保得到最佳的边界切分线并记录每个音节的起始和终止的时间点。Praat 语音学软件是一款跨平台的多功能语音学专业软件，主要用于对数字化的语音信号进行分析、标注、处理及合成等实验，同时可以生成各种语图和文字报表。图 4.3 为标注一段语音的示例。

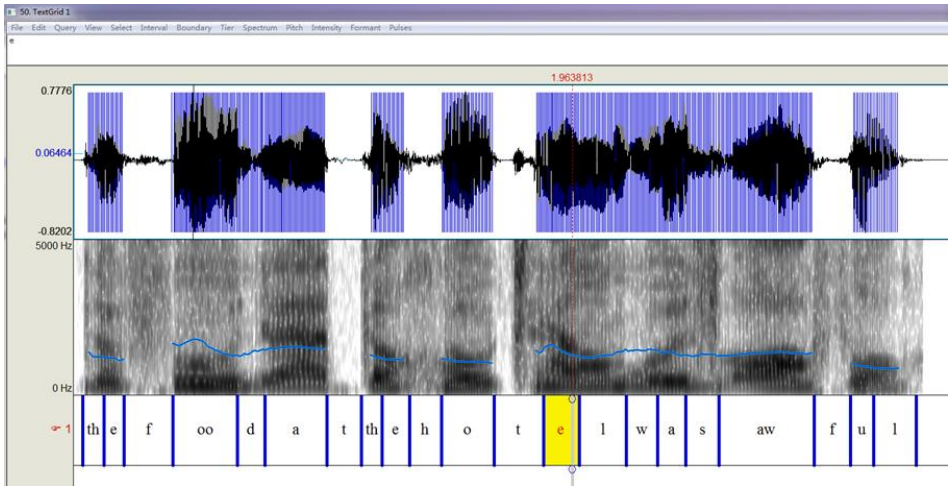

图 4.3 语音划分标注示例

(2)特征参数提取

语音转换之前首先要对语音进行分帧、端点检测等前端预处理，得出适合转换处理的语音帧。然后提取语音帧的特征参数，这里，我们分别提取源和目标说话人语句的基音频率(F_0)，并建立语音的基频参数化描述模型，本论文采用 Pitch Target 模型。在该模型中，一个音节的基频曲线可以使用一组模型参数 (a,b,λ,β) 来表示。图 4.4 为一段语音提取特征参数的示例，其中蓝色曲线为实际的基频曲线，红色曲线为使用 Pitch Target 模型拟合成的基频曲线（即 4.2 式），绿色直线为近似后的结果（即 4.3 式）。在公式 4.2 中，随着时间 t 的增加，公式中的指数项部分逐渐趋近于 0，与公式 4.3 趋于等价。即，对于划分后的每一段语音在其尾部处绿色直线和红色曲线是趋近于重合的。从图 4.4 的结果可看出确实如此，说明特征提取的结果比较理想。

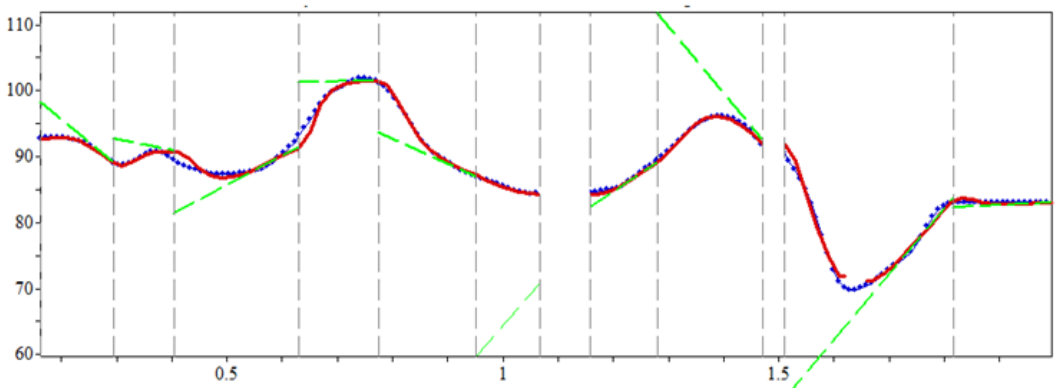

图 4.4 Pitch Target 模型的特征参数提取示例图

(3)基于 GMM 训练模型参数

本文先将连续语音语句分解成离散的音节，将源说话人语音的音节韵律特征参数 (a,b,λ,β) 和目标说话人语音的音节韵律特征 (a,b,λ,β) 对齐，作为源参数和目标参数输入到 GMM 模型中进行训练。经过训练，得到一个将源语音的特征参数映射成目标语音的特征参数的转换函数。源语音参数矢量 X 与目标语音参数矢量 Y 共同构成一个联合参数矢量 Z，$Z=\begin{bmatrix}X & Y\end{bmatrix}^T$，Z 的概率分布函数用 Q 个单高斯分布函数的加权和来表示。Z 的概率分布函数表示为：

$$P(x)=\sum_{q=1}^{Q}\alpha_q N(x;\mu_q;\sum_q) \tag{4.8}$$

$$\sum_{q}^{Q}\alpha_q=1,\alpha_q\geq 0 \tag{4.9}$$

式中，α_q 是加权系数，$N(x;\mu_q;\sum_q)$ 是 n 维的正态分布，即

$$p\left[{}^{X}\!/_{\mu_q},\sum_q\right]=\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}}\left|\sum_q\right|^{\frac{1}{2}}}}\exp\left[-\frac{1}{2}(x-\mu_q)^T\sum_q{}^{-1}(x-\mu_q)\right] \tag{4.10}$$

由贝叶斯准则可知，特征矢量属于第 f 类的概率为：

$$p({}^{C_q}\!/_{X})=\frac{\alpha_q N(x;\mu_q;\sum_q)}{\sum_{p=1}^{Q}\alpha_p N(x;\mu_q;\sum_q)} \tag{4.11}$$

运用 EM 算法，可以得到 M 组 GMM 模型参数 $(\alpha_q;\mu_q;\sum_q)$，其中 M 为 GMM 模型的高斯分量[16]。

(4)参数转换

在语音转换过程中，通过将源说话人语音语句中各个音节的 Pitch Target 参数 (a,b,λ,β) 作为输入，而获得目标说话人语音的参数输出。具体的转换流程如图 4.5 所示。

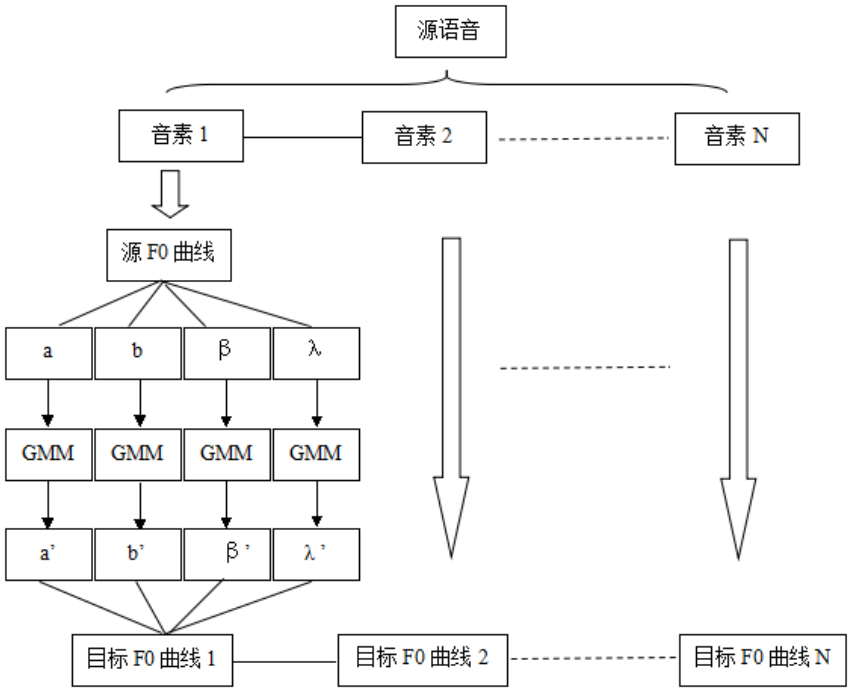

图 4.5　基于 GMM 的语音转换模型

基于 GMM 方法的转换函数为：

$$F(x)=\sum_{q=1}^{Q} p(C_q/x)\left[\mu_q^{\,y}+\sum_q^{YX}(\sum_q^{XX})^{-1}*(x-\mu_q^{\,x})\right] \tag{4.12}$$

式中，$\mu_q^{\,X}$，$\mu_q^{\,Y}$，$\sum_q^{YX}$，$\sum_q^{XX}$ 由均值矢量 μ_q 和协方差 $\sum_q$ 分解得到的：

$$\mu_q=\begin{bmatrix}\mu_q^{\,x}\\ \mu_q^{\,y}\end{bmatrix} \tag{4.13}$$

$$\sum_q=\begin{bmatrix}\sum_q^{XX} & \sum_q^{XY}\\ \sum_q^{YX} & \sum_q^{YY}\end{bmatrix} \tag{4.14}$$

利用式(4.12)得到的转换函数将源语音的特征矢量转换成目标语音的特征矢量，最后合成目标语音[17,18]。

4.2.2 频谱转换

(1)频谱参数训练

训练阶段我们使用了基于并行语料的源语音和目标语音的联合建模方法。并行语料是指源语音和目标语音提取参数进行训练的语料库是相同的。这样就能确保提取的参数经过动态时间规整 DTW 后，能够对齐并进行 GMM 联合训练。在训练阶段，首先对源语音和目标语音进行参数分析，并分别提取二者的参数

LPCC 倒谱系数，设系数阶数 P 为 24。将源语音和目标语音对应于相同语句内容的 LPCC 系数利用动态时间规整 DTW 对齐。对齐的目的是使得通过调整源和目标语音的参数矢量，使源语音和目标语音用同样长度的序列来对同一语料进行处理。对齐后的源和目标语音的每个语句的 LPCC 系数有了相同的帧数，并且每帧 LPCC 系数阶数为 24。

经过 DTW 对齐后源语音的 LPCC 参数序列为：

$$X_{N\times p} = \left[X_1; X_2; \ldots; X_N\right] \tag{4.15}$$

目标语音的 LPCC 参数序列为：

$$Y_{N\times p} = \left[Y_1; Y_2; \ldots; Y_N\right] \tag{4.16}$$

其中，N 为对齐后的基于同一语料的源语音和目标语音的 LPCC 参数序列的帧数，每一帧 LPCC 系数的阶数为 p。

建立联合矢量：

$$Z_{N\times 2p} = \left[F_1; F_2; \ldots; F_N\right] \tag{4.17}$$

其中每个联合矢量由源和目标 LPCC 参数矢量组成：

$$F_i = \left[x_1, x_2, \ldots, x_p, y_1, y_2, \ldots, y_p\right] \quad i = 1,2,...,N \tag{4.18}$$

通过这些联合矢量来训练 GMM 模型参数，即得到相应的频谱参数转换函数[19]。

训练过程如下：

将要训练的联合序列记作：$Z_{N\times 2p} = \left[X_{N\times p} \quad Y_{N\times p}\right]$。一个 M 阶的 GMM 模型的概率密度函数由 M 个单高斯概率密度函数加权得到，

$$p(w/\lambda) = \sum_{i=1}^{M} \alpha_i b_i(w) \tag{4.19}$$

其中，w 是要训练的 2p 维联合矢量。α_i 是混合权重，满足 $\sum_{i=1}^{M}\alpha_i = 1$。$b_i(w)$ 为正态分布概率密度函数，如下：

$$b_i(w) = \frac{1}{\sqrt{2\pi\left|\sum_i\right|}} \exp\left[-\frac{1}{2}(w-\mu_i)^T \sum_i^{-1}(w-\mu_i)\right], i = 1, 2, \ldots, M \tag{4.20}$$

其中 μ_i 为均值向量，$\sum_i$ 为协方差矩阵，完整的 GMM 模型用下式表示：

$$\lambda = (\alpha_i; \mu_i; \sum_i), i = 1, 2, \ldots, M \tag{4.21}$$

根据贝叶斯准则，特征矢量 x 属于第 i 类声学空间的概率为：

$$p(^{C_i}/_{X}) = \frac{\alpha_i N(x; \mu_i; \sum_i)}{\sum_{i=1}^{M} \alpha_i N(x; \mu_i; \sum_i)} \tag{4.22}$$

用 EM 算法对 M 组 GMM 模型参数进行估计后，即得到转换函数：

$$F(x) = \sum_{i=1}^{M} p(^{C_i}/_{X}) \left[\mu_i^{Y} + \sum_i^{YX} (\sum_i^{XX})^{-1} * (x - \mu_i^{X}) \right] \tag{4.23}$$

(2)频谱转换

在转换阶段，首先对源语音进行 LPC 分析并提取其 LPCC 参数，再根据在训练阶段得到的语音转换规则进行转换，得到转换的频谱特征，由这些转换的语音特征合成出最终的转换语音。

4.3 语音合成 STRAIGHT 模型

在 4.2 中的提到训练和转换阶段，我们使用 Pitch Taregt 模型训练特征参数，来获得 GMM 转换函数。并对 LPC 进行分析，利用 LPCC 倒谱系数对 GMM 联合建模，训练源和目标说话人的 GMM 模型参数，获得转换函数。最后利用 STRAIGHT 对韵律参数和频谱参数进行语音合成，得到新的具有目标说话人特性的语音。下面将对 STRAIGHT 算法进行详细的介绍。

STRAIGHT 算法是由同本和歌山大学的 Kawahara 教授提出的一种针对语音信号的分析合成算法，算法全称为 Speech Transformation and Represetation using Adaptive Interpolation of weighted spectrum（基于自适应加权谱内插的语音转换和重构）。STRAIGHT 算法强调的是完全去除激励对频谱的影响，最终将语音分解为相互独立的频谱参数和一系列脉冲激励的卷积。其优点有：通过对语音的短时谱自适应内插平滑方法可以提取精确的谱包络；利用提取的语音特征参数以及谱包络参数能够合成出高质量的语音信号；合成时可以对时长、基频以及谱参数进行高灵活的调整，同时调整后合成的语音音质不会有明显的下降。

STRAIGHT 分析合成算法主要分为下面三个部分：(1)去除周期性影响的谱估计；(2)精确的 F0 提取；(3)语音的重构。下面，将介绍 STRAIGHT 分析合成算法的三个核心问题：

1.去除周期性影响的谱估计

(1)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；

(2)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。

STRAIGHT 中采用了卷积二维三角窗的平滑方法。

$$S(w,t) = \sqrt{g^{-1}(\iint_D h_t(\lambda,\tau)g(\left|F(w-\lambda,t-\tau)\right|^2 d\lambda d\tau)} \tag{4.24}$$

$$h_t(\lambda,\tau) = \frac{1}{4}(1-\left|\frac{\lambda}{w_0(t)}\right|)(1-\left|\frac{\tau}{\tau_0(t)}\right|) \tag{4.25}$$

其中，$F(w,t)$表示计算得到的短时谱，$S(w,t)$表示平滑后得到的谱包络。函数$g()$定义平滑时保留谱参数的何种特性。例如$g(x)=x$保留的是信号的能量特性，而$g(x)=x^{1/3}$则保留的是信号的听感响度特性。

2.平滑可靠的基频轨迹的提取

STRAIGHT 中用小波分析的方法对语音信号的基频进行分析。首先寻找语音信号的基频成分，从中计算出瞬时频率，然后通过在频谱上进行谐波分析，并进行频率轴上的平滑，最终得到基频轨迹[20]。

3.语音重构

合成语音的参数包括基音频率，经过时间轴和频率轴平滑后的语音二维谱包络。在合成时使用基于基音同步叠加和最小相位冲激响应的方法，并且在合成的过程中可以对时长、基频和频谱参数进行修改和调整。使用公式如下：

$$y(t) = \sum_{t_1\in Q}\frac{1}{\sqrt{G(f_0(t_1))}}\upsilon(t-T(t_1)) \tag{4.26}$$

$$\upsilon = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}V(w,t_1)\varphi(w)e^{jwt}dw \tag{4.27}$$

$$T(t_1) = \sum_{t_1\in Q,k<1}\frac{1}{G(f_0(t_k))} \tag{4.28}$$

其中，式(4.26)反映的是一个基音同步叠加的过程，$y(t)$表示回复的语音信号，Q 表示用于合成的基音同步位置的集合，函数$G()$表示基频的调整，它可以

试任意形式的映射关系。式(4.27)反映的是每一帧对应的冲激响应的求取过程，$V(w,t_1)$表示最小相位冲激响应的傅里叶变换，$\phi(w)$为具有附加的控制相位的激励，用来改善听感。式(4.28)反映的是基音同步位置的确定过程。从先前分析得到的平滑谱$S(w,t)$可以计算得到$V(w,t_1)$，即可以将一般相位的谱转化为最小相位的谱，这里我们采用基于倒谱的变化方法，即

$$V(w,t) = \exp\left(\frac{1}{\sqrt{2\pi}}\int_0^{\infty} h_t(q)e^{jwq}dq\right) \tag{4.29}$$

$$h_t(q) = \begin{cases} 0, q < 0 \\ c_t(0), q = 0 \\ 2c_t(q), q > 0 \end{cases} \tag{4.30}$$

$$c_t(q) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-jwq}A(S(\mu(w), r(t)))dw \tag{4.31}$$

其中，q 表示倒频，$A()$，$\mu()$，$r()$分别表示对平滑谱$S(w,t)$在幅度、频率和时间轴上的调整[21,22]。

4.4 情感语音特征分析

4.4.1 语音的情感定义及分类

情感是人类的一项重要的本能，它在我们的生活、工作中有举足轻重的地位。由于环境和心理状态不同而引起的不同的情感状态，可以引起语音、表情以及行为上的不同表现。情感语音就是说话人在一定的情感状态下所产生的具有特定语义的发音。

要研究情感语音信号，首先需要针对某些特性和标准对情感做出一个合理的分类，然后针对不同类别的情感分别研究特征参数的性质。然而对于情感的分类，学术界并没有一个统一的认识，也没有一个定性定量的测量评价标准，所以，对于情感的具体分类是由研究的特定目的决定的。

在语音情感研究中常用的情感分类大多是如图 4.6 所示的八情感模型或者四情感模型，即喜、怒、惊、悲。八种情感共同分布在一个圆上。其中，圆心被称为自然原点，各情感分布在自然原点的周围，并且围绕自然原点排成了圆形，所以这种情感分类的方法叫“情感轮”。

图 4.6 情感轮

本研究所用到的情感主要是四情感模型中的喜、怒、悲，以及与它们作对比的平静情感。

4.4.2 情感语音特征参数分析

情感是人类的一种很重要的本能，是人们感知事物的必不可少的信息资源。语音信号中的情感，不仅与体现情感的韵律特征有很大的关系，也与说话人发音的音质，声道形状的改变以及肌肉张力等有关，这些特征的变化体现了各种情感的差异。因此，提取这些反映情感的特征并进行分析、研究，对于语音情感转换具有极其重要的意义[23]。

语音的韵律特征是语音的一个极其重要的信息。语音的韵律特征主要表现为音高、音长、音强等。反映情感的韵律特征参数主要表现在基频构造、时间构造、振幅构造、共振峰构造等方面[24,25]。

在时间构造上，同一语音在不同的情感状态下，语速会表现出一定的差异。在高兴、愤怒等激动的状态下，语速要比平静状态时高。而悲伤状态下，语速要比平静状态时低。在振幅构造上，同一语音在不同的情感下振幅能量差异也比较大。对于喜、怒等情感，语音的振幅能量往往比平静时要大，而悲伤情感的幅值则要比平静时低。在基频构造上，基音频率是情感语音的一个非常重要的参数。语音的基音曲线反映了语音的音高变化。基于音节的基频曲线的变化趋势与汉语的四种声调模式相对应，而基于语音的基频曲线与汉语的语调相对应。在共振峰构造上，由于人的发音不同情感状态下，声道的形状和肌肉的张力，音质也发生了变化。而共振峰频率与声道的形状和大小有关，每种声道的形状都有一套共振峰频率特征参数。因此，共振峰频率也是情感的特征参数[26]。通常在语音情感

转换时使用的主要特征参数包括以下内容（如表 4.1 所示）。

表 4.1　常用语音情感参数

特征参数	意义
Rate	语速，单位时间内音节通过的速率
Pitch Average	基音的均值
Pitch Range	基音的变化范围
Intensity	强度，语音信号的振幅方差
Pitch change	基音的平均变化率
FI Average	第一共振峰均值
FI Range	第一共振峰变化范围

在实际中，不同的情感对应着不同的语音声道特征和激励源的统计特征。通过研究，Murray 和 Arnott 总结了情感和语音参数的关系（如表 4.2 所示）。

表 4.2　情感和语音参数之间的关系

规律	愤怒	高兴	悲伤
语速	略快	快或慢	略慢
平均基音	非常高	很高	略低
基音范围	很宽	很宽	略窄
强度	高	高	低
声音质量	有呼吸声，胸腔声	有呼吸声，共鸣音调	有共鸣声
基音变化	重音处突变	光滑，向上弯曲	向下弯曲
清晰度	含糊	正常	含糊

浊音的声带振动频率称为基音频率，简称基音（或基频），是语音信号重要特征。在产生情绪时，由于生理方面的影响，基音也会产生相应的变化。因此，其也是表征情绪的重要特征。基音检测是一个比较复杂的问题，由于声带振动并不是完全周期性的，有些清浊音的过渡很难判定它应归属于周期或非周期性。

基频的生理特性决定了男性基频和女性基频的范围不同，男性的基频大约在 50—200Hz 之间，而女性基频范围在 200—450Hz 之间。基频随着语句的变化，

形成了语调。基频变化特性的作用随着语种的不同而有所差异，在普通话语句中，基频还有鉴别语义的作用。普通话是声调语言，语句声调的不同表达的意思也不同。同一语句，不同的声调表达了不同的情感。

第 5 章 实验验证与系统实现

本章中，建立一定数量的语音库，作为样本进行一些实验来验证语音转换算法结果的好坏，同时求取分析不同语音情感特征参数，总结概括不同语音情感的特点。另一方面将会介绍个性化的 TTS 系统工具的运行环境及最终实现。

5.1 语音转换

在该小节中对于语音转换后的结果进行评价，本文主要采用的是说话人识别的方法。说话人识别的核心思想是建立训练语音库，使用 GMM 为其建模训练，分别求得 GMM 的特征参数。当输入一段测试语音时，将其分别带入之前求得的特征参数中进行匹配，根据似然函数计算最大后验概率，后验概率值最大者则为对应的说话人。为了验证转换结果的好坏，我们以一位男性和一位女性为例进行测试，分别为其录制 20 段语音。以下将依次讨论男—男，女—女，男—女三种情况下转换结果的好坏。

表 6.1 男—男转换后验概率值

编号	源语音	目标语音
1	-6.5169e+04	-6.6048e+04
2	-7.8512e+04	-7.6880e+04
3	-5.7471e+04	-5.6912e+04
4	-5.6217e+04	-5.4794e+04
5	-1.0826e+05	-1.0962e+05
6	-5.1046e+04	-4.9878e+04
7	-3.7636e+04	-3.8546e+04
8	-4.9900e+04	-4.9878e+04
9	-5.1046e+04	-4.9153e+04
10	-6.1695e+04	-6.0057e+04
11	-5.3951e+04	-5.6217e+04
12	-8.1852e+04	-7.8429e+04
13	-7.0275e+04	-7.0833e+04
14	-5.1945e+04	-5.0472e+04

15	-4.1289e+04	-3.9948e+04
16	-4.7845e+04	-4.8082e+04
17	-3.9984e+04	-3.9124e+04
18	-5.3951e+04	-5.4146e+04
19	-4.8082e+04	-4.7845e+04
20	-5.0132e+04	-5.0029e+04

表 6.2 女—女后验概率值

编号	源语音	目标语音
1	-6.5948e+04	-6.3148e+04
2	-6.4412e+04	-6.2880e+04
3	-5.7174e+04	-5.7269e+04
4	-4.1762e+04	-4.3749e+04
5	-4.8499e+04	-4.8291e+04
6	-4.9878e+04	-5.1278e+04
7	-3.4663e+04	-3.2646e+04
8	-5.1046e+04	-4.9153e+04
9	-6.1695e+04	-6.0157e+04
10	-5.4951e+04	-5.6217e+04
11	-8.1525e+04	-7.8493e+04
12	-4.9931e+04	-4.9878e+04
13	-7.2075e+04	-7.0133e+04
14	-4.7485e+04	-4.8082e+04
15	-3.9784e+04	-3.9214e+04
16	-5.1945e+04	-5.0472e+04
17	-4.1928e+04	-3.9948e+04
18	-7.0133e+04	-6.8845e+04
19	-7.4231e+04	-7.4493e+04
20	-5.1294e+04	-5.0497e+04

表 6.3 男—女后验概率值

编号	源语音	目标语音
1	-8.1852e+04	-8.4243e+04
2	-7.2071e+04	-6.9554e+04
3	-5.5139e+04	-5.4882e+04
4	-6.1569e+04	-6.0228e+04
5	-2.1625e+04	-1.9138e+04
6	-3.0785e+04	-2.9650e+04
7	-3.3478e+04	-3.1844e+04
8	-1.6290e+04	-1.8505e+04
9	-1.4752e+04	-1.7478e+04
10	-3.9764e+04	-3.1265e+04
11	-2.6573e+04	-3.0238e+04
12	-4.4307e+04	-3.9807e+04
13	-5.6912e+04	-5.0946e+04
14	-4.1289e+04	-4.4461e+04
15	-5.1032e+04	-5.0209e+04
16	-3.4948e+04	-3.2166e+04
17	-2.7789e+04	-2.6212e+04
18	-5.2165e+04	-4.6155e+04
19	-4.2268e+04	-4.0364e+04
20	-1.8945e+04	-1.4566e+04

从以上三个表中的实验结果可以得出结论：男声—男声间语音转换成功率为65%，女生—女生间语音转换成功率为 70%，男声—女声语音转换成功率为 75%。从转换率来看语音转换的效果是比较理想的，在可接受范围内，而且可知男女间的转换比同性间（男—男，女—女）转换的结果更好一点。

5.2 情感特征

（1）基音频率的分析

为了分析不同情感的基音频率特征，需要建立一个情感语音数据库，本文采

用的是 CASIA 汉语情感语料库，对基频的动态范围及轨迹进行了分析。基频的动态范围包括基频的最大值，最小值，平均值及标准差。基频的生理特性决定了男性基频和女性基频的范围不同，因而建立男性和女性语音库，分别进行实验。我们选取中性、喜悦、悲伤和愤怒四种情感的语音对基音频率进行分析，每种情感选取 30 条语句，实验结果如图 6.1,6.2 所示。

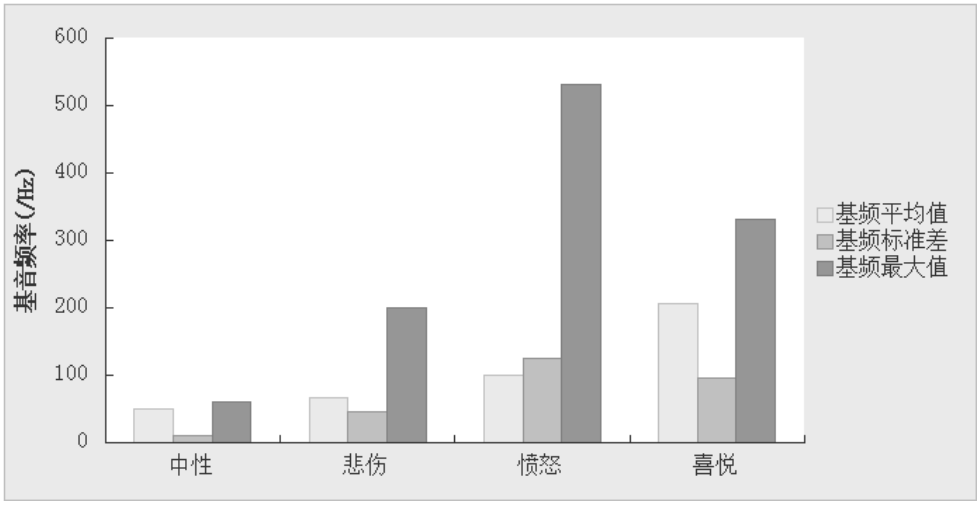

图 6.1 男声的不同情感基频对比图

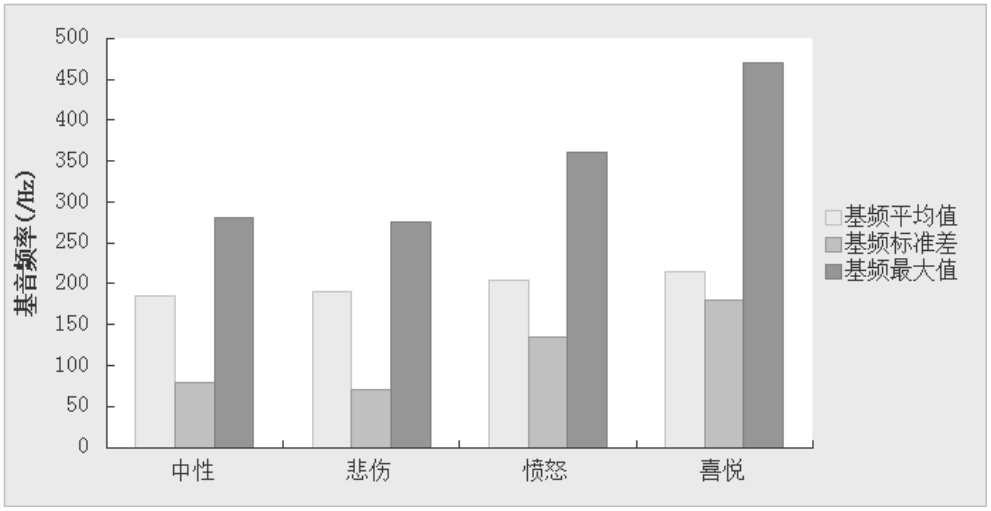

图 6.2 女生的不同情感基频对比图

由图 6.3 可知，本实验中，男声的基频平均值在 100Hz 左右，其中，中性和悲伤的语句基频平均值在 50Hz 左右，而愤怒的平均值在 100Hz 以下，喜悦的平均值在 200Hz 以下。由图 3.3 可知，女声的基频平均值在 160Hz 以上，其中，愤怒和喜悦的基频平均值在 200Hz 以上。

实验得出，所有语句的基频最小值为零，所以基频范围即为基频最大值。男

声的情感语句情感不同，基频变化不同。其中，喜悦与愤怒的基频范围跨度最大，而其基频平均值也最大。中性情感语句的基频范围和基频平均值都最小。悲伤的基频平均值较小，但基频范围较大。所有情感语句中，愤怒和喜悦的基频标准差比其他两类的值要大，中性的基频标准差最小。女声的基频变化类似，不同的是，悲伤的基频平均值和基频范围跨度与中性情感类似。

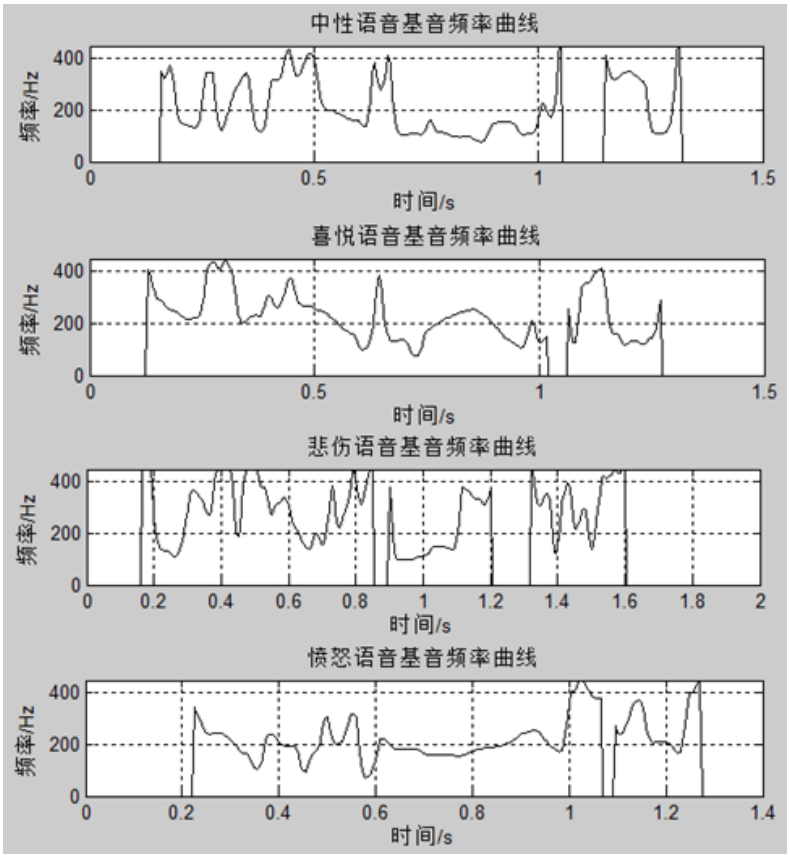

图 6.3　男声的不同情感的基频曲线

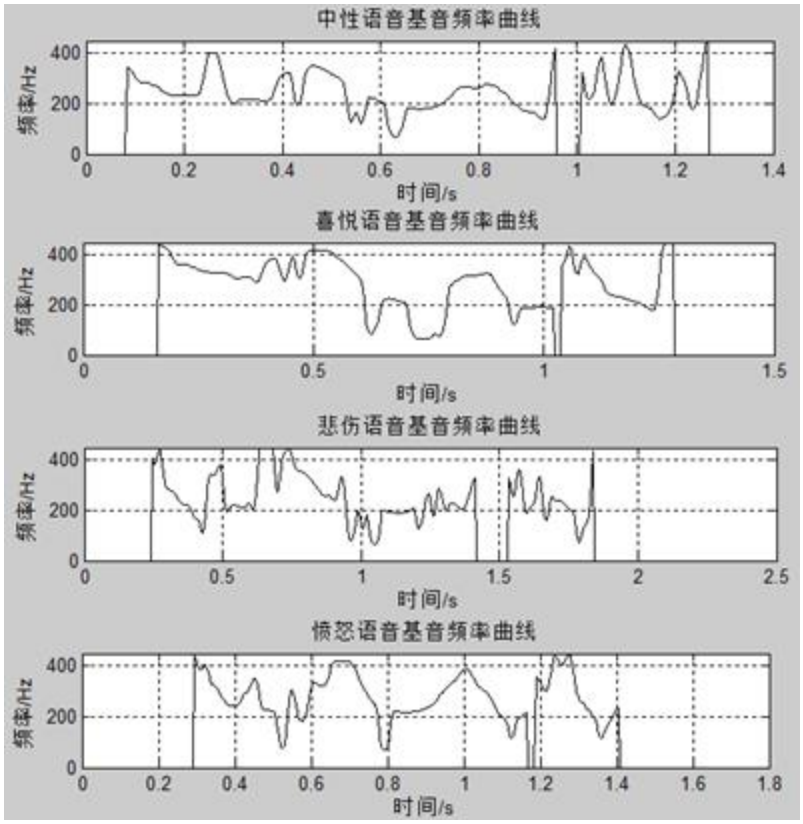

图 6.3　女声的不同情感的基频曲线

通过观察语音信号的基频轨迹曲线，发现悲伤语音与中性语音的基频曲线时长比较接近，而愤怒和喜悦语音的基频时长都比中性语音的要短。这说明语音的时长与基频时长有关。其中，喜悦语音的基频幅值比其他三种情感要高。

（2）情感语音时间长度的特征

情感语音持续时间长度的特征主要用来说明说话人语速的快慢。分析情感语音的时间构造主要分析不同情感语音下说话时间的差别。通过计算每一种情感下的同一语料从开始到结束的持续时间，我们可以对情感引起的时间变化进行研究。因为无声部分对情感也是有贡献的，所以计算的时间也包含了无声部分。

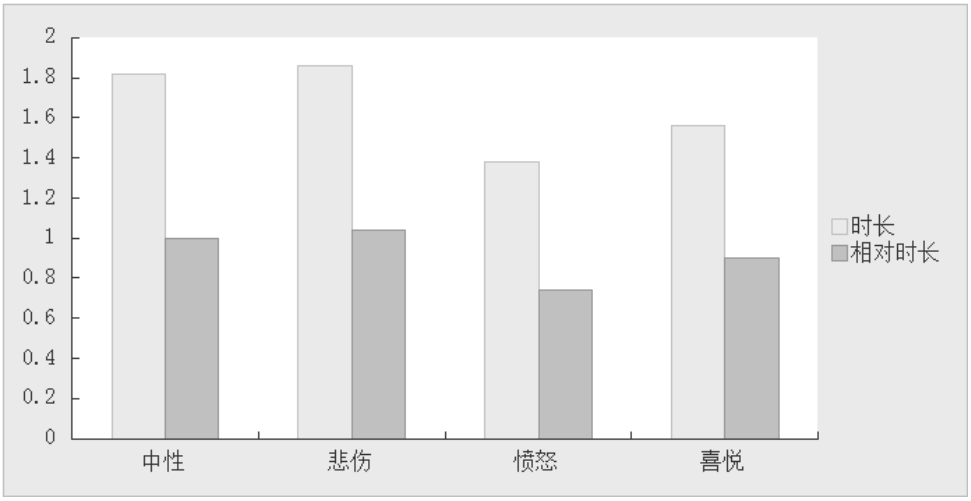

图 6.4 男声情感语句时间长度对比图

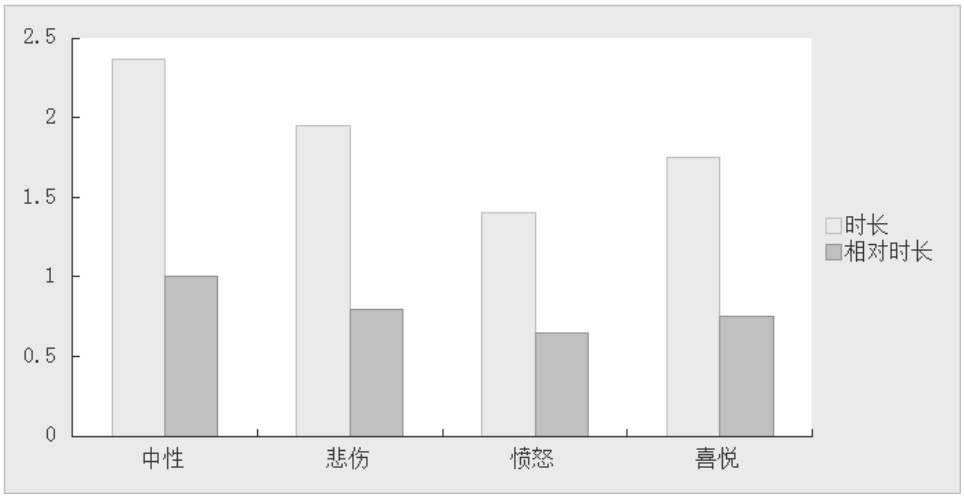

图 6.5　女声情感语句时间长度对比图

从语音库中的数据分析得出，不考虑文本的影响，男声情感语音时间长度排列为：愤怒<喜悦<中性<悲伤。女声情感语音时间长度排列为：愤怒<喜悦<悲伤<中性。所有情感语音中，悲伤和中性语音持续时间比较接近，喜悦语音的持续时间较短，愤怒语音最短。所以在本实验中，进行情感语音的转换时，适当的压缩处理后的愤怒语音的发音长度，而伸长悲伤的发音长度。使其与现实中的语音更为接近。实际语句中，悲伤地时间长度伸长了很多。通过观察得到，这些是由于和中性语句相比，情感语句中的一些音素被模糊的发音、拖长或者省略掉了的缘故。

通过以上对于不同情感基音频率和时长的分析，我们可以定性地归纳出中性、喜悦、悲伤和愤怒四种情感间的关系。通过 4.3 节中介绍的 STRAIGHT 语音合成算法，修正基音频率和时长这两个参数，可以使得原来中性的朗读声音具

有喜悦、悲伤和愤怒的情感色彩。

5.3 系统运行环境

在前面的章节中分别介绍了基于 Microsoft Speech SDK 开发的 TTS 系统与语音转换算法。两者分别在 VC++6.0 和 Matlab7 平台上编码实现，若要设计实现一个个性化 TTS 系统，需要将两者结合起来，使用 VC 和 Matlab 的混合编程 VC 调用 Matlab 的方法有很多种，在本文中出于使用和学习方便，选择 VC 调用 Matlab 引擎的方法。

5.4MFC 界面设计

本系统使用 MFC 进行界面上的设计，对 TTS 工具和语音转换算法进行整合设计完成一个完整的带有界面的个性化文本语音转换系统。下面介绍一些 MFC 常用的标准控件。

表 5.1　使用的 Windows 标准控件

控件	MFC 类	描述
按钮	CButton	用来产生某种行为的按钮，以及复选框、单选钮和组框
组合框	CComboBox	编辑框和列表框的组合
编辑框	CEdit	用于键入文本
图象列表	CImageList	一系列图象(典型情况下是一系列图标或位图)的集合。图象列表本身不是一种控件，它常常是和其它控件一起工作，为其它控件提供所用的图象列表
列表	CListCtrl	显示文本及其图标列表的窗口
列表框	CListBox	包括一系列字符串的列表
滑块	CSliderCtrl	包括一个有可选标记的滑块的窗口
静态文本	CStatic	常用于为其它控件提供标签

基于 MFC 的常用控件和 C++编程，实现了文本转换成语音以及不同说话人角色之间的语音转换。界面的设计如图 5.1 所示。

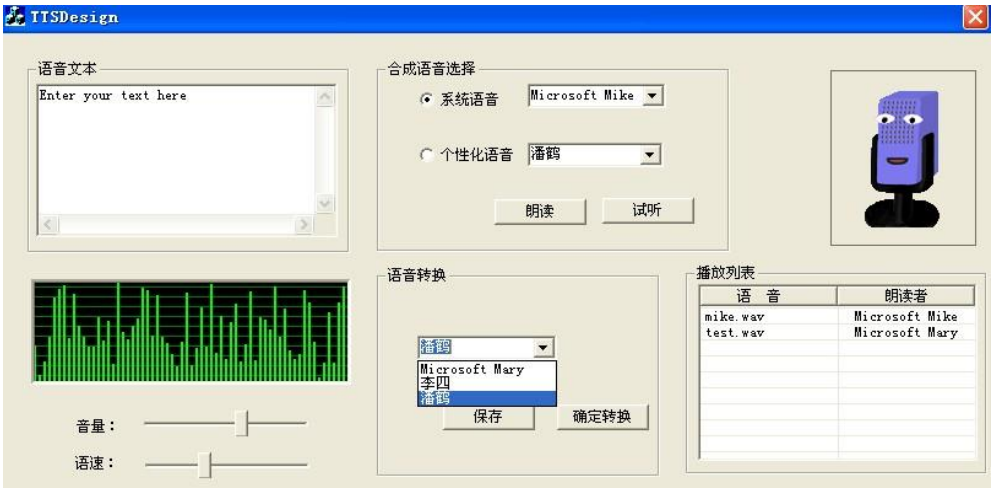

图 5.1　系统界面设计图

（1）在文本框中输入想要朗读的文字，选择语音朗读角色，分为系统语音和个性化语音，系统语音则依不同的操作系统版本而不同，个性化角色可根据自己的喜好进行选择，可以自己、身边同学老师或喜爱的明星的声音。

（2）试听功能则是使用户感受一下朗读者的声音，方便用户进行选择设置。

（3）点击朗读即可将文本框中的文字转换成语音，并有跳动的波形显示和不同角色对应头像的口型的张合变化。

（4）朗读的过程中可实时地调节音量和语速的大小。

（5）对于朗读后的声音可将其保存至播放列表中，记录其文件名和朗读者，双击可听取声音。

（6）语音间的转换，选择播放列表中的一条记录，可选择除该语音朗读者之外的朗读者进行转换，点击“确定转换”则可将相同的文本内容转换成具有另外一个朗读者声音特性的语音段。

（7）设置朗读语音具有不同的情感（高兴、悲伤、喜悦），默认为中性语音。

第 6 章 结论与展望

6.1 本文工作总结

当前的 TTS 仅能实现系统自带的朗读角色进行语音朗读且不具有感情色彩，在用户体验上稍稍欠缺了一些。本文加入了个性化的功能，完成了一个个性化文本语音转换系统的设计与实现。具体的工作包括：

（1）本文首先详尽地阐述了基于微软提供的 Speech SDK 语音工具包实现 TTS 的基本功能；然后，利用 Speech SDK 的 API 接口，将输入的文本内容准确朗读并且实现了中英文的混合朗读。

（2）为了实现个性化语音处理，本文提取了说话人的语音特征，建立了一个源说话人和目标说话人之间的语音转换模型。

（3）针对声音的情感因素进行了研究，提取分析不同情感语音的特征参数，获得了不同情感和中性语音间的转换关系，使朗读出来的语音具有一定的情感色彩。

最后，通过实际的语音角色转换实验，验证了算法的有效性和实用性。

6.2 进一步工作进展

本课题中设计的个性化文本语音转换工具目前仅处于初步的理论研究阶段，若使其很好地应用于日常生产生活中，在技术上仍存在一些限制因素。

（1）个性化的实现是是将系统角色的语音作为源语音，使用语音转换算法将其转换成自定义角色的语音。语音转换算法其中一个重要的步骤就是语音的音节划分，目前已有的方法为手工划分和半手工划分 HTK 工具，因此无法实现实时的给出转换的结果。对此，仍需要对音节划分算法进行研究，以实现完全的自动化划分。

（2）对于语音情感的研究仍处于定性的分析阶段，目前了解了不同情感在韵律特征和频谱特征上曲线分布的情况和特点及大致的转换关系。但仍无法给出不同情感语音和中性语音在具体参数上定量的转换关系。

（3）由于目前语音转换和语音合成技术上的局限，使得将系统语音进行转换后的清晰度并不是非常的理想，但仍在可接受范围内。

对于以上的设计和实现的不足之处，希望在以后更加系统的学习中将其逐一

解决。

参考文献

1. 刘羽. 语音端点检测及其在Matlab中的实现[J]. 计算机时代, 2005, (8):25-26.
2. 潘秀林.汉语普通话基频模式研究[D].南京:南京航空航天大学,2008:pp33.
3. 江太辉.一种改进的语音基频轮廓提取算法[J].五邑大学学报,2002,16(2):pp27-30.
4. 胡益平. 基于GMM的说话人识别技术研究与实现[D]. 厦门:厦门大学, 2007.
5. 霍春宝, 张彩娟, 赵红敏. 基于 GMM-UBM 的说话人确认系统的研究[J]. 辽宁工业大学学报(自然科学版), 2012, 32(2): 98-101.
6. G.McLachlan and T. Krishnan, “The EM Algorithm and Extensions” in Wiley Series in Probability and Statistics, New York: Wiley, 1997.
7. 朱建伟, 孙水发, 刘晓丽. 基于 MFCC 等组合特征的说话人识别模型[J]. 三峡大学学报(自然科学版), 2009, 31(6): 77-79.
8. 鲍福良, 方志刚, 徐洁. 基于MFCC和GMM的说话人确认研究[J]. 仪器仪表学报, 2008, 29(4): 73-76.
9. 吴晓娟, 韩先花, 聂开宝. 模糊 C-均值(FCM)聚类法与矢量量化法相结合用于说话人识别[J]. 电子与信息学报, 2002, 24(6): 845-848.
10. 陶建华,蔡莲红.基于音节韵律特征分类的汉语语音合成中韵律模型的研究[J].声学学报,2003, 28(5):396-402.
11. 陈高鹏,胡郁,王仁华.考虑语速和前后环境的基频 Target 模型及实现[D].安徽合肥,中国科学技术大学:191-194.
12. Y. Xu and Q. E. Wang, Pitch Targets and Their Realization: Evidence from Mandarin Chinese[J], Speech Commun. 2001.
13. Jianhua Tao, etc, Prosody Conversion from Neutral Speech to Emotional Speech[J], IEEE Transcatoin on Speech and Audio Processing, 2006, 14(4):1145-1154.
14. LAMEL L, LABINER L, ROSENBERG A, et al.An lmproved endpoint detector for isolated wordrecognition [J].IEEE ASSP Magazine, 1981,29: 777-785.
15. ACERO A, CRESPO C, de la TORRE C, et al.Robust HMM-based endpoint detector [C] //Proceedings of Eurospeech93, 1993: 1551-1554.

16. 单振宇,杨莹春.基于多项式拟合的中性-情感模型转换算法[J].计算机工程与应用,2008.44(21):206-208.
17. 赵义正.改进 GMM 谱包络转换性能的语音转换算法研究[J].科学技术与工程,2010, 10(17):4172-4174.
18. 康永国,双志伟.陶建华.张维,徐波.高斯混合模型和码本映射相结合的语音转换算法[A].第八届全国人机语音通讯学术会议(NCMMSC8):293-297.
19. 张炳,俞一彪.基于改进 GMM 和韵律联合短时谱的说话人转换[J].信号处理,2009, 25(4):548-552.
20. 张正军,杨卫英,陈赞.基于 STRA 丨 GHT 模型和人工神经网络的语音转换[J].语音技术,2010, 34(9):49-52.
21. 刘震,景新幸.汉语情感语晋合成的研冗[J]. Science&Technology Information,2008,9:78-79.
22. H.Kawahra and R. Akahane-Yamada,Perceptual Effects of Spectral Envelope and F0 Manipulations Using STRAIGHT Method[J], J. Acoust. Soc. Amer, 1998:1-10.
23. 谢波,陈岭,陈根才,陈纯.普通话语音情感识别的特征选择技术[J].浙江大学学报,2007,41(11):1816-1822.
24. 张立华,杨莹春.情感语音变化规律的特征分析[J].清华大学学报(自然科学版),2008,48(Sl):652-657.
25. 张石清,赵知劲,雷必成,杨广映.结合音质特征和韵律特征的语音情感识别[J].电路与系统学报,2009,14(4):120-123.
26. 蔡莲红,崔丹丹,蒋丹宁,杨鸿武.语音的情感信息分析与编辑[J].声学技术,2005,29(3): pp209-212.

致　谢

时光荏苒，转眼大学四年就这样匆匆过去。回想在大学求学的四年，心中充满无限感激和留恋之情。感谢母校为我们提供的良好学习环境，使我们能够在此专心学习，同时发展兴趣爱好、陶冶情操。

这篇论文的完成要感谢很多人，首先要感谢我的论文指导老师魏阳杰老师，魏老师在毕业设计中给予了我悉心的指导，尤其在我所不熟悉的信号处理部分，魏老师总是能一步步地指导我，给我方向，给我力量。并且帮助我确定了论文的选题，指引着我的研究方向。同时还要感谢关楠老师，关老师以其严谨的科研作风、多元的思维方式深深影响着我们的研究思维。同时，关老师以其幽默憨厚的形象关怀着我们每个学生，给予我们莫大的鼓励和支持，他坚持个人主观的能动性，他相信我们每个人能合理安排时间。其次我还要感谢实验室中的同学们，不论是在学习上还是生活上，他们总是无私地给予我帮助。大家一起讨论，营造了很好的学习氛围。

最后，我谨向所有关心和帮助过我的老师、同学、领导、同事和家人致以真诚的谢意。