目 录

1 绪论

语言是人类特有的功能，它是记载几千年人类文明史的根本手段，没有语言就没有今天的人类文明。声音是人类常用的工具，用语言进行信息相互之间的传递是我们人类最重要的基本功能之一。语言是我们人类进行思维和交流的形式，是从众多人的言语中概括总结出来的具有规律性的一种符号系统。而语音则是语言的声学表现形式，是声音和它所能表达的意思的一种结合，是相互传递信息的最重要手段，是人类最重要、最有效、最常用和最方便的交换信息的形式。语音中除了包括实际说话人发出的语音内容所表达的语言信息外，还包括说话者即发音者是谁和所带有的情感因素如喜怒哀乐等等各种信息。在我们人类今天已经构成了的通信系统中，语音通信方式（比如日常的电话通信、如今时兴的微信等）由于其非常方便和十分便捷的特点，早已经在现今最主要的信息传递途径中占据主导地位。语言和语音是人类文明的产物，是人类思维活动的一种表现及依托方式。人类的智力活动在一定程度上外在反映在个人语言和语音上，语言和语音包含有最丰富的信息量和智能的最高水平，因此，语言和语音与人类文明，与人类社会的进步有着密切的联系。

语音信号处理是采用数字信号处理技术来处理语音信号的一门新兴的学科，但它还是多学科的集成，是一门涉及领域非常广泛的交叉性学科。尽管在这一领域的研究人员之前从事的可能主要是信号与信息处理以及计算机应用等学科的研究，实际上，语音信号处理与其它的一些学科，像是语音学，语言学，声学，认知科学，生理学，还有心理学等学科都是紧密相连的。这诸多学科之间是一个相辅相成的关系，语音信号处理技术的发展需要依赖于这些学科的发展，而语音信号处理技术的进步同时也可以促进这些学科的进步。

1.1 说话人语音转换的定义

说话人语音转换就是使用语音信号处理技术对说话人语音信号进行处理，改变一个说话人(源说话人，source speaker)的语音个性特征，使其转换为具有另外一个说话人(目标说话人，target speaker)的语音个性特征，即 A 说话人的语音转换为像是 B 说话人在说话一样，具有 B 说话人的发音特征，但语音内容是没有变化的，仍是 A 说话人表达的语义信息。

说话人语音信号中包含了非常多的信息，除了其中非常重要的语义信息以外，还有能代表说话人身份信息的的个性特征、说话人的情感状态、说话人的说话态度以及说话人当时所处的场景等信息。说话人语音转换就是要使原有语音中的语义信息保持下来，

不发生改变，只是改变语音所具有的个性化的信息，使一个说话人的语音通过语音转换后听起来就像是另外一个说话人在说话一样。

1.2 研究语音转换的意义

科学领域中的研究与发展很多都是相辅相成，互相促进的。从理论的角度来看，语音转换就是一门涉及声学研究、信号处理以及模式识别等多个学科领域的典型性交叉学科。对语音转换技术进行研究时可以使用或学习各个领域的知识，开展调研；反过来，通过研究语音转换技术又可以促进这些科学领域的发展。另外，由于语音通信方式的重要性，对语音信号的研究已经发展到一定阶段，其中对语音转换的研究是当前对语音信号处理研究中继语音识别技术、说话人识别技术和语音合成技术之后又一新的研究方向。从实际应用角度来看，语音转换技术具备有广阔的应用前景。具体应用如下所示[1]：

1、在语音识别领域的应用：我们知道由于各方面因素的影响，每个人都有各自的发音特点，因此不同人纵使发同一个音其语音特征参数也不一样。这样，在语音识别领域，说话人个性特征参数是对语音识别的一个非常重要的研究依据。那语音转换同样也是对说话人个性特征参数的一个研究，因此可以为语音识别技术提供依据。另外在非特定人语音识别中，还可以通过语音转换实现说话人的归一化。

2、在 TTS 文语转换系统中的应用：众所周知现有的 TTS 系统由于现有语音合成方法及技术的局限性，使合成出来的语音缺失了其特有的个性化特征，以致听起来很是单调。但是，如果通过语音转换系统则可以根据需要选择某一个特定人，使 TTS 合成出来的语音通过语音转换系统转换后再进行语音合成，这样最终合成出来的语音就不再单调，而是具备了选择的特定人的说话语音特征。合成语音不再单调，可以根据实际需要满足各方面不同的需求。

3、在信号传输中的应用：由于语音信号的存储容量是非常大的，若在低码率的语音信号中传输，传输速度会非常慢。这时就可以利用到语音转换系统，在传输前，先提取只与说话内容相关的信号，在信道中则只传输这部分信号，在接收端再加入个性化特征，这样就可以既提高了传输速度也提高了传输有效性。

4、在医疗方面的应用：当说话人的发音声道受到损伤时，其发出来的语音的可懂度比较低，那么此时可以使用语音转换系统将受损的语音复原过来，使得语音的可懂度得以提高。

5、在刑侦方面的应用：当说话人需要被保护但又要传递信息时，可以使用语音转换系统将保密通信中说话人的个性化特征进行伪装，然后再进行通信。

6、在娱乐方面的应用：现有的电影、电视节目的配音都是让特定配音员根据画面所示实时进行配音。如果语音转换系统在其中得到使用，就可以将具有原演员个性化声

音特征的语音加入到语音库里面，当另外的配音员进行配音时通过语音转换系统进行实时转换，使其具有原来演员的个性特征。

语音转换还可以控制单一说话人的语音质量。因为人在长时间录音的情况下，很可能会产生疲劳以至于后来的录音质量有所下降，那么在这个时候就可以使用语音转换系统来纠正质量有所下降的语音。

1.3 语音转换的历史及研究现状

到今天为止，人们对语音转换技术的研究已经有四十几个年头了，特别是近二三十年，语音转换技术越来越引起研究人员广泛的注意。实际上，在更早以前，人们就在研究语音技术，只是将更多的注意力放在语音识别和语音合成以及语音编码等语音技术的研究上，所以，可以说语音转换技术是语音技术中的一个新的研究方向，语音识别技术和语音合成技术等是语音转换技术的起源。从国内外对语音转换技术的研究来看，国外比较早就在研究这门新兴的学科，因此，研究得比较深入，自然也就取得了比较大的研究成果，而我们国内对语音转换技术的研究则相对国外来说起步比较晚，但我们国内的研究技术发展比较快，经过这么些年的研究也取得了不错的研究成果。现今，频谱特征参数和基音周期的转换是语音转换技术研究人员的主要关注点。

最早对说话人语音转换技术进行研究是在二十世纪八十年代末期，Abe[2]由于受到说话人自适应技术的启发，提出矢量量化的频谱包络语音转换方法，但转换效果并非理想的，主要是因为矢量量化方法是让语音转换发生在每一个特征子空间，这样就忽视了各特征子空间之间的联系，使得特征空间不连续，引起语音转换效果不佳的结果。九十年代初期，Vallbret[3]提出基于线性多变量回归（LMR）和动态频率调整（DFW）的语音转换方法，还采用基音同步叠加法(PSOLA)针对激励信号来调整其韵律特征；LMR 转换方法是在一个独立的特征子空间进行语音转换，这样容易丢失与其他特征子空间的有关信息，同样会因为特征空间不连续，造成语音转换效果不佳；DFW 方法分为线性频率调整和非线性频率调整，由于线性频率调整的丢失信息和补零现象，目前比较少用到线性频率调整，可以使用分段-非线性的函数进行非线性频率调整。同时，Narendranath[4]提出基于神经网络的语音转换方法，主要是对语音的共振峰特性实现了转换，Baukoin 还采用了 BP 神经网络的实验。九十年代中期，Kuwabara[5]提出模糊矢量量化的语音转换方法，语音转换效果在一定程度上得到提升。九十年代末期，Stylianou[6,7]提出高斯混合模型（GMM）的语音转换算法，其加权求平均的方法解决了特征空间不连续的问题，语音转换技术向前迈进了一大步；但这种方法也有其局限性，即引起了语音过平滑现象。二十一世纪初，Toda[8,9]针对过平滑问题，采用 Kain[10,11]联合特征矢量的直接估计方法，提出基于 DFW 的 GMM 语音转换方法，由于通过了动态频率调整，语音转换质量得到很大改

善。Yelui[12]在 2003 年提出创新点，即因为认识到人耳是非线性感知频谱的，在 Kain 联合特征矢量的方法上通过增加加权感知距离测度也使转换后重建的语音效果得到改善。

以上主要是基于频谱的特征转换，还有对基因周期转换的研究，主要是利用参差信号进行语音转换。K.S.Rao[13]就提出有关强激励脉冲的残差信号韵律转换算法；R.Muralishankar[14]提出了离散余弦变换（DCT）在残差信号中的语音转换；K.S.Lee[15]提出采用快速傅里叶变换（FFT）和快速傅里叶逆变换（IFFT）的基于残差信号基因周期转换；Stylianou[6,7]提出 GMM 模型的同时还提出谐波加噪声模型（HNM）来建模叠加形成新的基因周期的方法。另外，在九十年代末期，Kawahara[5,16,17]还针对语音参数的修改和恢复提出自适应加权谱内插（STRAIGHT）语音分析合成系统。

这些是国外取得的研究成果及现状，国内对语音转换的研究同样有一定的佳绩。初敏[18]等人提出了针对男女声语音转换的时域基因同步叠加(TD-PSOLA)方法；双志伟[19]提出基于汉语音素的码本映射算法；陈一宁[20]也针对过平滑问题，进行了概率分布的转移，从而提出了基于 GMM 和 MAP 的语音转换方法；左国玉[21]提出了采用线谱对（LSP）特征参数和遗传算法的径向基函数（RBF）网络的语音转换方法，转换效果得到很大的改善，系统稳定性也得到很大提高。Chung-Hsien Wu[22]针对语音韵律特性，采用隐马尔科夫模型（HMM）进行语音转换，其中，语音中因素时长用 HMM 的状态持续时间表征，还将 HMM 的状态持续时间变量用 Gamma 函数的分布来描述，在语音情感信息上得到比较好的控制和转换。

到现今，语音转换技术在各个方面的研究都取得了比较好的进展，但技术有无限的发展空间，对语音转换的研究有技术成熟的方面，也有些方面的技术尚不成熟，比如说语音实时转换的实现技术等，仍需我们不断改进技术，挖掘创新，以期进行实际开发。

1.4 论文组织结构

本论文主要是对说话人语音转换技术进行研究，全文共分为五章，具体的章节内容安排如下：

第一章：绪论

本章主要论述说话人语音转换的定义及研究意义，并介绍语音转换的历史和国内外研究现状。

第二章：语音转换基础及原理

.本章主要阐述与语音转换有关的基础知识和原理。首先介绍语音产生的机理，并根据语音产生机理对语音信号从激励模型、声道模型和辐射模型进行数学建模。然后对语音转换原理进行阐述，最后介绍主观和客观两方面的语音转换效果评估方法。

第三章：语音信号分析与特征参数提取

本章首先介绍了对语音信号进行预处理的方法，然后从短时能量及短时平均幅度、短时过零率、短时自相关函数和短时平均幅度差函数几个特征参数论述了对说话人语音信号的时频分析，并对语音信号进行了线性预测分析，得到 LPC 的推演参数 LSP，知道其与共振峰有紧密联系，最后本章还对基因周期和共振峰这二个重要特征参数的提取方法进行了重点论述。

第四章：语音转换算法

本章先是简单介绍了语音转换原理、语音信号特征参数训练前的动态时间规整处理方法和 STRAIGHT 语音分析合成模型。然后重点论述基于 ANN 中采用 SC 和 PSO 算法的 RBF 网络和基于改进的后知概率 GMM 模型这两种语音转换算法，并进行了实验比对。

第五章：总结与展望

本章主要是对全文的一个工作总结，并对今后语音转换技术的一个展望。

1.5 本章小结

本章先是对语音信号处理作了引出，然后对本文研究的重点语音转换技术的定义进行了阐述，再重点阐述了语音转换的意义，分析了语音转换的研究历史以及国内外研究现状，最后再简单介绍了本文全部章节内容安排。

2 语音转换基础及原理

2.1 语音产生机理

声音是一种振动频率在 20～2000Hz 之间的能被人耳听到的波。大自然中包含有各种不同的声音，如风声、雨声、雷声、机械声、不同乐器声等等。而人说话的语音是各种不同声音中的其中一种，它是从人的发声器官发出的，是具有一定的规律性语法和语义的声音[23]。语音的震动频率最高可以达到 15000Hz 左右。

人类生成语音过程的第一个阶段是决定想要给对方传递什么内容，然后将内容转换成语言的形式。选择能够表达其内容的适当语句，将其按既定的语法规则排列，便能构成语言的形式。由大脑对发声器官发出运动指令，发声器官各种肌肉运动，振动空气而形成语音波。

又人类的语音是由人体发声器官在大脑控制下的生理运动产生的。人的发音器官由三部分组成：①肺和气管产生气源；②喉和声带组成声门；③咽腔、口腔和鼻腔组成声道。如图 2.1 所示，这些器官共同构成一条形状复杂的管道。空气由肺部排入喉部，经过声带进入声道，最后由嘴辐射出声波，这就形成了语音[23-25]。语音由声带振动或不经声带振动来产生，其中由声带振动产生的音统称为浊音，而不由声带振动产生的音统称为清音。在发音器官中，肺和气管是整个系统的能源，喉是主要的声音生成机构，而声道则对生成的声音进行调制。

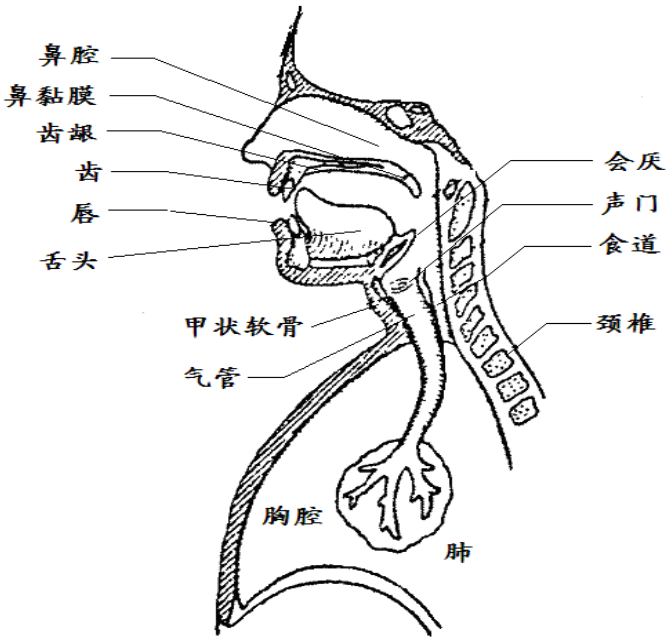


图 2.1　人的发声器官剖面图

当声音产生后，便沿着声道进行传播。声道可以看成一根具有非均匀截面的声管，在发声时起着共鸣器的作用。声音进入声道后，犹如进入一个具有某种谐振特性的腔体，声音的频谱特性必然就会受到声道共振特性的影响。声道内具有一组谐振点，称为共振峰频率或共振峰，声道的频谱特性便主要是反映出每个峰的共振峰位置和这些个共振峰的频带宽度[24]。共振峰所在的位置及其频带宽度主要取决于声道的形状和大小，因此，不同的语音和不同的共振峰参数相对应。

2.2 语音信号生成的数学模型

建立了语音信号的数学模型才能够用计算机来定量地对语音信号进行模拟和处理。从人体发声器官的发声机理这方面来看，声道情况会因为发出声音的性质不同而有所不同。另外，由于声道和声门的相互耦合，还形成了语音信号的非线性特性。由此可知，语音信号实际上是一个非平稳的随机过程，具有随着时间而发生改变的特性，所以数学模型中的信号参数应该也是随着时间而改变的[23]。但语音信号的这一特性是非常缓慢的，因此可以将语音信号划分成一些连续的短时段进行处理，在这些短时段内语音信号特性可以看作是固定不变的，是不会随着时间而发生改变的平稳随机过程。从而，可以将短时间段内的语音信号采用线性时不变模型来表示。

通过对人体发声器官进行剖析和对语音信号产生机理进行分析，可以知道首先是由肺部和气管里的气流激励声道，然后从嘴唇或鼻孔，或者从嘴唇和鼻孔同时辐射出来而形成语音声波。我们将声道入口声门以下的部分，称为“声门子系统”，主要功能是用来产生激励振动，因此是“激励系统”；而声门到嘴唇或是鼻孔的呼气通道是声道，称之为“声道系统”；最后语音从嘴唇或是鼻孔辐射出去，所以嘴唇或是鼻孔之外就称之为“辐射系统”[23,26]。

激励系统、声道系统和辐射系统各自对应着气流冲击声带产生振动形成激励效应，声道中各器官对语音的调音作用，嘴唇和鼻孔辐射语音的效应，因此，可以对这三个系统分别进行建模，成为激励模型、声道模型以及辐射模型。这样，就可以将激励模型、声道模型和辐射模型这三个子模型级联起来表示成一个完整的语音信号数学模型。

2.2.1 激励模型

浊音是由声带的不断开启和关闭产生的脉冲波，仪器测试其类似于斜三角脉冲波，也就是这时的激励波可以看作是具有周期性的斜三角脉冲波。

单个三角脉冲波可以用数学表达式（2-1）表示成：

$$g(n)=\begin{cases}\frac{1}{2}\left[1-\cos\frac{n\pi}{N_1}\right] & 0\le n\le N_1\\ \cos\frac{n-N_1}{2N_2}\pi & N_1\le n\le N_1+N_2\\ 0 & \text{其他}\end{cases} \tag{2.1}$$

式中，N1 为斜三角波的上升部分的时间，N2 为其下降部分的时间，观察图 2.2 中单个斜三角波的频谱$G\left(e^{jw}\right)$，可以发现，它表现出一个低通滤波器的特性。其 z 变换的全极模型表示形式如下：

$$G(z)=\frac{1}{\left(1-e^{-cT}\cdot z^{-1}\right)^2} \tag{2.2}$$

其中，c 是一个常数，并且 $T=N_1+N_2$。显然上式表示的斜三角波可以描述为一个二级点模型，所以，可以认为单个斜三角波模型被加权单位脉冲序列激励产生的结果就是得到斜三角波串。

这个单位脉冲串和幅值因子的 z 变换形式可以表示成如下所示：

$$E(z)=\frac{A_v}{1-z^{-1}} \tag{2.3}$$

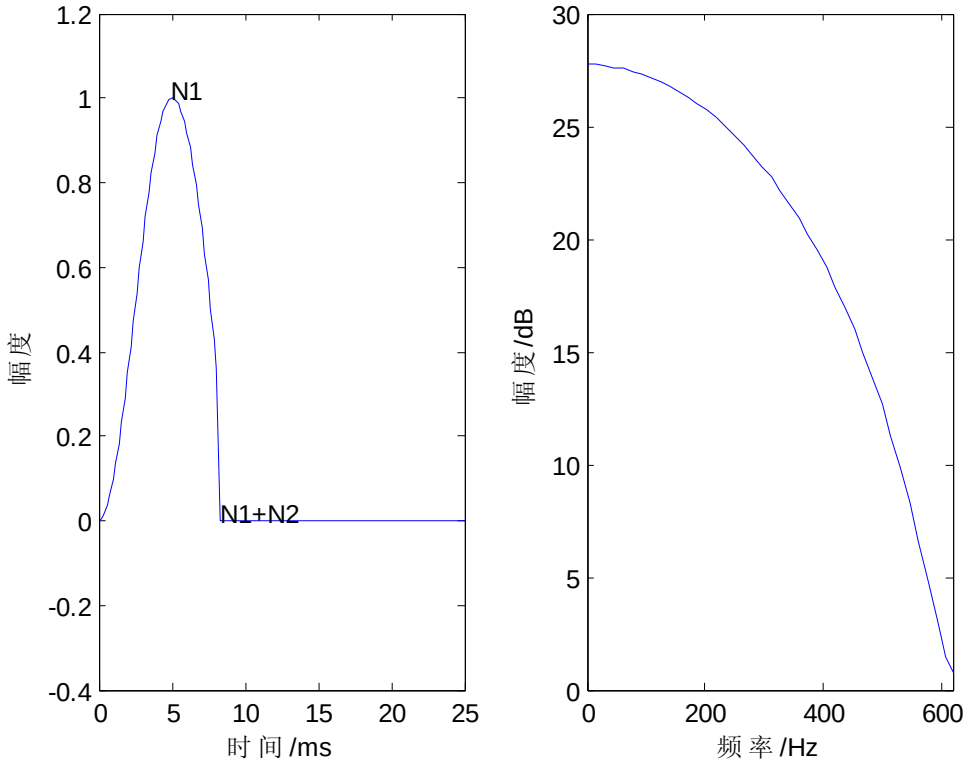


图 2.2 单个的斜三角波图和其频谱图

所以完整的激励模型表示如下：

$$U(z)=G(z)\cdot E(z)=\frac{A_v}{1-z^{-1}}\cdot\frac{1}{\left(1-e^{-cT}\cdot z^{-1}\right)^2} \tag{2.4}$$

在发清音的场合，声道被阻碍形成湍流，所以可以模拟成随机白噪声。

2.2.2 声道模型

当声波通过声道时，受到声腔共振的影响，在声波的其中某些频率处会产生谐振现象。谐振现象在信号频谱图上的表现就是其谱线包络在谐振频率处产生峰值，这种峰值一般就被称作为共振峰[26,27]。如图 2.3 所示为一段语音信号的频谱图，具有明显的峰起，即为共振峰，一般元音可以有 3～5 个共振峰。我们将从这个角度描述出的声道模型称为共振峰模型。由于人耳听觉的柯蒂氏器官的纤毛细胞的位置是按着频率感受去排列的，所以用共振峰的方法来表示这种声道模型是行之有效的，因而经常被拿来使用。

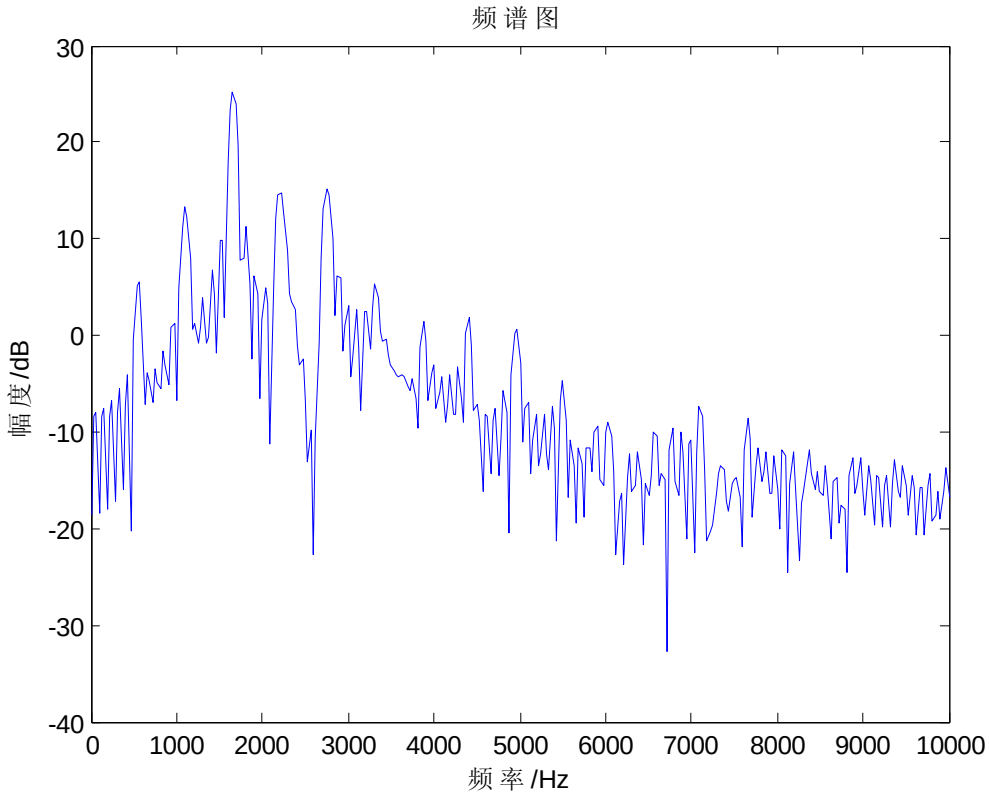


图 2.3　语音信号频谱图

实践表明，一个元音用前 3 个共振峰进行表示就足够了，而辅音或鼻音，因为比较复杂，对其表示可能要用到至少 5 个共振峰才行。一个二阶谐振器的传输函数可表示成：

$$V_i(z)=\frac{A_i}{1-B_iz^{-1}-C_iz^{-2}} \tag{2.5}$$

多个V_i叠加就可形成声道的共振峰模型，即声道模型可以表示成：

$$V(z)=\sum_{i=1}^{M}V_i(z)=V_i(z)=\sum_{i=1}^{M}\frac{A_i}{1-B_iz^{-1}-C_iz^{-2}}=\frac{\sum_{r=0}^{R}b_iz^{-r}}{1-\sum_{k=1}^{N}a_kz^{-k}} \tag{2.6}$$

通常 N>R，还有分子和分母无公共因子，及分母无重根。可见，声道模型的传递函数是一个零极点模型。

另外，语音信号还可以用语谱图来直观的表示信号随时间变化的频谱特性。时间量用横轴表示，纵轴则表示语音信号的频率，语音信号的能量用图像的黑白度来表示，这样就构成了语谱图。如图 2.4 为“She had your dark suit in greasy wash water all year.”的语谱图。黑带部分表示声道的谐振频率，条纹图形表示浊音部分，这是因为此时的时域波形有周期性，在浊音的时间间隔内图形显得很紧密。

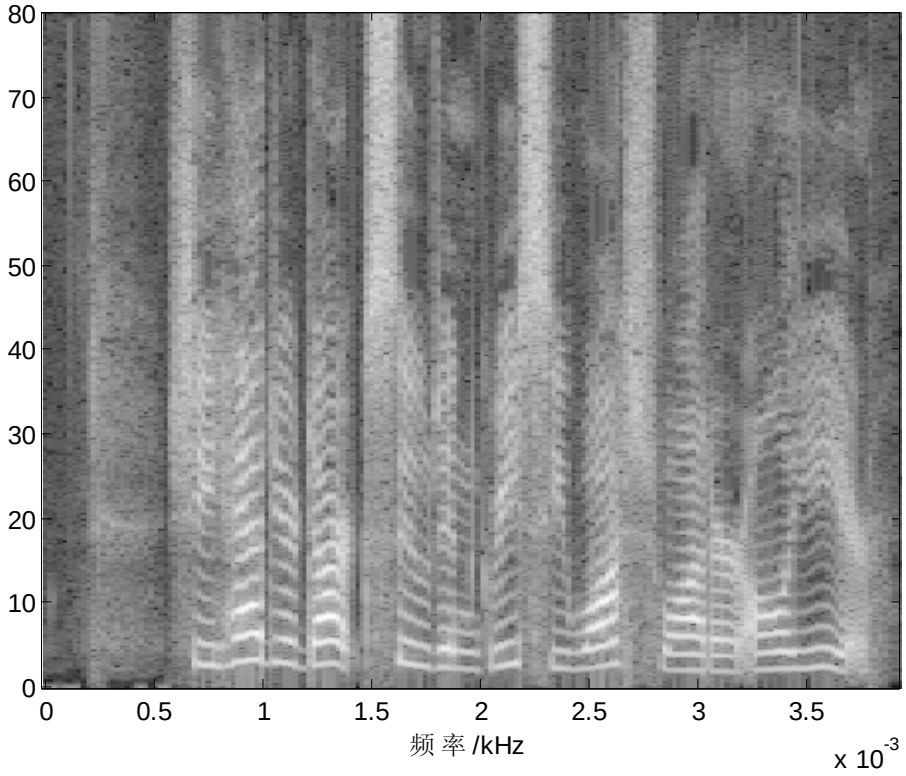


图 2.4 “She had your dark suit in greasy wash water all year.”的语谱图

2.2.3 辐射模型

声道的终端是口和唇，速度波会从声道模型输出，而语音信号则为声压波，辐射阻抗就是速度波和声压波的倒比，口唇的辐射效应就可以用辐射阻抗来表征。如果认为口唇张开的面积非常小，头部的表面积远远大于口唇张开的面积，则可推测出下面的辐射阻抗的公式[23]：

$$z(\Omega)=\frac{j\Omega LR}{R+j\Omega L} \tag{2.7}$$

其中，$R=\frac{128}{9\pi^2}$，$L=\frac{8a}{3\pi c}$，这里a表示成口唇开口半径，c作为声波传播速度。

由于辐射阻抗实部和因辐射而产生的能量损耗成正比例关系，并且研究表明，口唇端的辐射效应在高频段影响较为明显，而在低频段影响则较小。因此，辐射模型可以用高通滤波器来表示成：

$$R(z)=1-rz^{-1} \tag{2.8}$$

其中，$r\approx1$。

2.2.4 数字模型

前面分别讨论得到语音信号激励模型$U(z)$，声道模型$V(z)$和辐射模型$R(z)$，并且知道其级联组合形式为零极点模型。因此，语音信号产生的完整模型可以用 3 个子模型级联而成，如图 2.5 所示即为语音信号的数学模型表示。这样语音信号数学模型的传递函数$H(z)$可以用下列式子表示为：

$$H(z)=A_V U(z)V(z)R(z) \tag{2.9}$$

其中，A 是加权系数，$Av+An=1$。

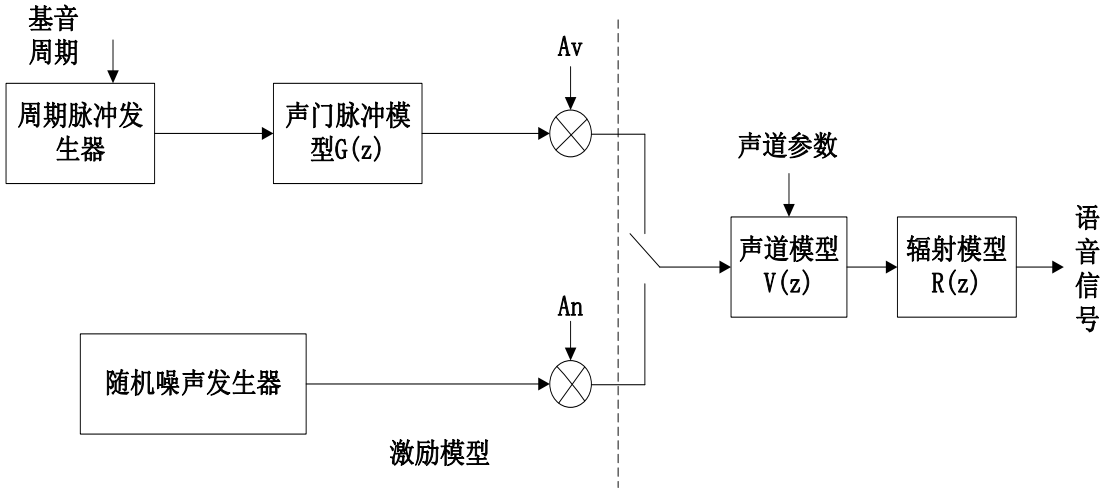


图 2.5 语音信号生成的数学模型

2.3 语音转换的系统框架

说话人语音信号中含有诸多不同的信息，如说话的内容、说话人的个性化特征、以及说话人所处的说话环境等等。其中，说话人的个性化特征是指与说话人自身身份相关的声音方面的特征，而与具体的说话内容和说话人所处的说话环境没有关系。前面我们讲到说话人语音转换的目的就是要保持说话人原有的语义信息不变，而改变说话人语音

中所具有的个性化的信息，使其听起来像是另一个人在说话。

我们要达到这样一个语音转换效果，首先要提取能表征说话人各方面特点的声学特征参数。然后，对声学特征参数进行转换，再用转换后的声学特征参数合成出新的，接近于目标说话人的语音。为了能很好地完成这样一个语音转换，一般我们将这样一个转换过程分成两部分，训练和转换两个部分，如图 2.6 所示语音转换系统的转换原理。

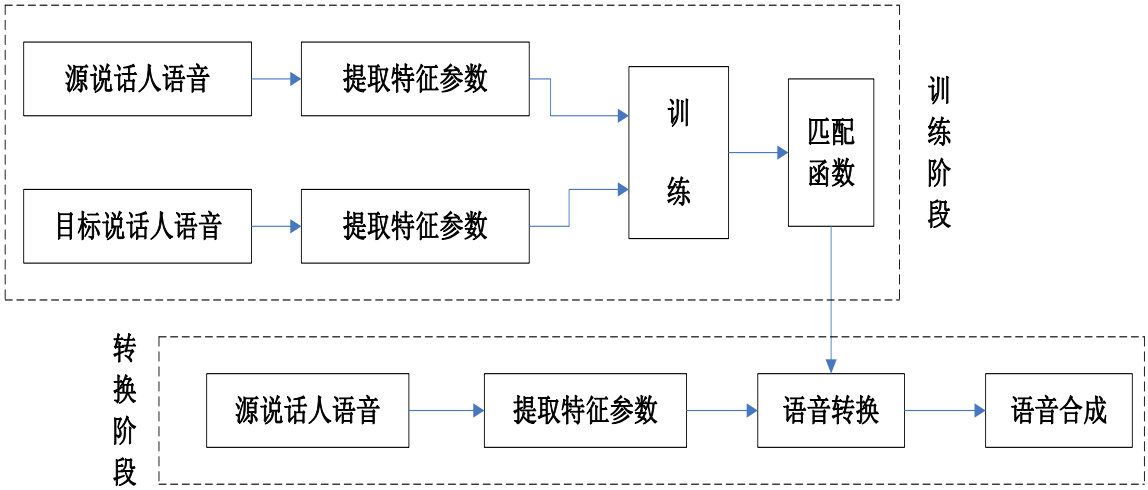


图 2.6 语音转换系统框架

其中训练阶段的主要任务就是要找出源说话人与目标说话人语音的个性化特征参数，并找出两者特征参数序列之间的对应关系。其过程为先是分析源说话人的训练语音和目标说话人的训练语音，然后对分析后的语音进行特征参数的提取。最后对每个特征参数序列分别进行训练，得到转换规则，即建立源说话人和目标说话人之间的匹配函数。

在转换阶段，首先对源说话人语音进行分析并提取特征参数，然后使用训练阶段得到的匹配函数，对源说话人的个性化特征参数进行转换。最后再使用语音合成方法对转换后的特征参数进行语音合成，从而使得合成后的语音具有目标说话人的个性化特征。

由语音转换的系统框架图中可以知道，语音转换的关键技术主要表现在以下几个方面：

①选取说话人语音特征参数。无论在训练阶段还是在转换阶段都必须对语音特征参数进行提取，并且在以后的转换和合成中都要用到这些语音特征参数，所以，提取说话人语音特征参数这一点需引起足够的重视。若是说话人的个性化特征不能由选取的说话人语音特征参数很好的反映出来，就会出现转换后合成语音跟目标语音出现大的偏差的问题，因此必须选取影响说话人音色的主要特征参数来进行转换。

②训练出匹配函数。据上所知，之所以需要利用训练阶段对特征参数进行训练，其目的就是为了找出源说话人个性特征参数与目标说话人个性特征参数之间的转换规则，建立两者之间的匹配函数，这就是语音转换的核心所在。其实，两者之间的转换规则实际上就是它们的个性化特征参数集之间的一种映射关系，其中，源语音特征参数集是原

像，目标语音参数集就是像。通过建立不同模型对参数集进行训练，寻找到最优的映射函数来确定原像和像之间的对应关系，也即两者的匹配函数。

③合成语音。因为转换的是代表语音个性化的特征参数，所以需要对转换后的特征参数进行语音合成，使合成后的语音具有目标说话人的个性特征，这就是语音合成。合成后的语音不仅要求不单调，具有目标说话人的个性化特征，还要有较好的语音质量，即要求语音相对清晰自然。这就对语音合成模型要求比较高，需要合成模型尽可能的精确。现今的语音合成技术用得比较多的是 PSOLA 算法，相较于 LPC 和共振峰合成器，其语音合成效果更好。另外 STRAIGHT 语音分析合成模型也受到越来越多的关注。

2.4 说话人语音转换效果评价方法

对于评价说话人语音转换的效果也是整个语音转换系统的其中一部分，评价语音转换的效果可以帮助不断的改进构建的语音转换系统。语音中是包含很多不同因素的，经过转换后合成出的语音的效果自然就可以有很多的评判因素，针对不同的评判因素就设定出了不同的评价标准。只有根据不同的评价标准改进语音转换系统各方面性能进行，才能从整体上提高语音转换的质量。语音转换质量的评价方法主要分为主观评价方法和客观评价方法[21]。

2.4.1 主观评价

语音是说话人发出的声音，也是说给人听，让人辨识的。所以纵使是经过转换的语音也终将是为了服务于人的，那人对转换后语音的主观评价自然是一个非常重要的标准。主观评价语音转换后的效果用得比较普遍的主要有以下三种方法：

1、ABX 测试

在主观评价方法中，最常用的一种主观测试手段就是 ABX 测试方法。ABX 测试方法是针对语音本身的，在 ABX 测试中，A 代表源说话人的语音，B 代表与之相对应的目标说话人的语音，X 表示通过语音转换系统转换后得到的语音。

ABX 测试方法通过众多测听人员主观听觉判断转换后的语音在个性特征方面是更接近于源说话人的语音还是更接近于目标说话人的语音，最后对测听人员的主观判断结果进行统计，以计算转换后语音更接近于目标说话人的概率，从而判断语音转换系统的效果。尽管可能会有所有的测听人员都选择转换后的语音更接近于目标说话人的语音，但是转换语音也不能被认为就是目标说话人自己说出来的，只能说更像目标说话人发出的声音，主观测试出语音转换效果是非常不错的。

2、倾向性测试

我们知道语音转换的算法不是单一的，自对语音的转换研究提出以来研究人

员提出过许多不同的转换算法。倾向性测试就是针对语音的不同转换算法的，主要用来评价两种不同语音转换算法孰优孰劣的一种主观评价方法。倾向性测试过程类似于 ABX 测试，其中 A 表示采用其中一种转换算法的获得的转换语音，B 表示采用另一种转换算法获得的转换语音，对比 AB，看哪种转换语音有更高的转换质量，X 表示获得较好转换语音所采用的那种转换算法，是 AB 中的其中一种。测试时，要求测试人员评价由哪种语音转换算法转换出来的语音效果更好，更加接近于目标语音的个性特征。由此可见，倾向性测试是一种横向的语音转换效果比较评估方法。

3、MOS 平均意见得分方法

MOS 方法针对因素其实类似于 ABX 测试方法，只是多了语音质量这一项。因为转换语音最终会被应用到一些实际的场合，转换语音的质量必须从实际使用方面进行评估。传统的 MOS 方法主要是针对语音质量，衡量语音的可懂度、清晰度、自然度等，而采用 MOS 方法对转换语音效果进行评价，不仅需要衡量转换后的语音质量，还需要衡量个性化特征这一指标，该指标用来评价转换语音是否更接近于目标说话人的语音，根据以上指标，MOS 方法将转换语音分为“很差”、“差”、“一般”、“好”、“很好”五等，5 分代表转换效果“很好”，表示转换后的语音十分逼近目标语音，且转换语音听觉质量也好，1 分则代表转换效果“很差”，表示虽然经过转换，但转换后的语音依然十分接近于源语音，和目标语音相差较大，转换语音听觉质量很差。

2.4.2 客观评价方法

研究语音信号的频谱包络这些客观存在的语音特征参数，可以根据其在语音转换前后的变化差异进行客观的评价，还能给出具体的数字。现今，用得比较多的语音转换客观评价方法主要有以下三种：

1、频谱失真程度[29]

前面讲过对语音进行转换时，选择语音频谱包络参数作为转换的重要特征参数。实际上频谱包络参数作为语音信号的重要特征参数，不单只是表征语义信息，也可以表征出说话人个性化特征信息。不同人说同样的语义内容由于其个性化特征不同，其语音信号频谱也会有所不同，因此可以用频谱参数的差异衡量语音个性特征的差异程度。具体来讲，可以分为绝对距离测试方法和相对距离测度方法。

绝对距离测试方法测试的是转换后的语音和目标语音的平均谱失真。可以表示为：

$$D=\frac{1}{L}\sum_{l=1}^{L}d(x_l,b_l) \tag{2.10}$$

其中，x_l 表示的是经转换后得到的语音频谱参数，b_l 表示的则是目标语音频谱参数，d 表示某种谱失真测度，L 为语音帧的数目。

相对距离测度方法测试的是转换后的语音和目标语音的频谱距离与转换后的语音和源语音的频谱距离的比值。可以表示为：

$$D=\frac{\sum_{l=1}^{L}d(x_l,b_l)}{\sum_{l=1}^{L}d(x_l,a_l)}\times 100\% \tag{2.11}$$

其中，a_l、b_l、x_l 分别表示源说话人语音频谱参数、目标语音频谱参数和转换后语音频谱参数。

D 比值越小，说明转换系统性能越好，也就是转换后的语音更加接近于目标语音。D 值的大小与频谱参数的时间规整能力有很大的关系[30]，因此 D 值通常情况下比较大，实际上 50％左右就可以获得可接受的转换语音质量。目前技术上用得最多的就是相对距离测度方法。

2、信噪比

在对语音编码的性能评估中会用到信噪比这一检测因素，同样，在语音转换的性能评估中也可以借用信噪比这一检测因素。信噪比 $SNR(S_1,S_2)$ 具体表示如下：

$$SNR(S_1,S_2)=10\lg\frac{\sum\left|FFT(S_1(n))\right|^2}{\sum\left(\left|FFT(S_2(n))\right|-\left|FFT(S_1(n))\right|\right)^2} \tag{2.12}$$

其中，S_1 表示转换语音矢量，S_2 表示实际目标语音矢量，信噪比值越大，说明转换效果越好。

3、说话人辨识

说话人辨识也被用来进行转换语音效果的评估[31]，其主要思路是：进行语音识别，将源说话人语音和目标说话人语音都作为识别的目标语音，而将经语音转换系统转换后合成出来的语音作为需要进行语音识别的对象，测试转换合成出来的语音更倾向于哪一个目标语音。

测试模型的数学表达式为：

$$S=\arg\max_{q\in(S\ T)}\sum_{n=1}^{N}\log p(x_n\mid\lambda_q) \tag{2.13}$$

其中，λ_S 和 λ_T 分别代表源说话人和目标说话人的语音识别模型。计算出转换语音分别基于源说话人语音识别和目标说话人语音识别的最大似然和，比较这两种识别的和值，哪一种语音识别模型的和值大，则代表转换后的语音属于该说话人。通常对于说话人决策的置信度测量可以采用基于目标说话人和源说话人的对数似然比，表达公式如下：

$$\theta_{ST} = \log \frac{\sum_{n=1}^{N} p(X_n / \lambda_T)}{\sum_{n=1}^{N} p(X_n / \lambda_S)} \tag{2.14}$$

其中，θ_{ST} 似然比值越大，则表示转换算法的性能就较好。

上面从主观和客观方面分别介绍了几种语音转换系统的评估方法，当系统评估应用于实际中时，为了能给予语音转换系统好的改进建议，需要将各种主观和客观的语音评价方法充分利用起来，从各个方面准确评价语音转换系统的性能。由于人类听觉系统的特殊复杂性，往往客观的频谱差别不一定表明存在主观感觉在性能上的差异，这表明客观语音转换的质量和主观感官评定之间的联系薄弱。若是完全依靠主观评价方法对语音转换进行评估需要花费较大的人力物力，因此，为了对语音转换系统进行合理评估可以采用主观和客观的评价方法相结合使用。

2.5 本章小结

本章简单介绍了语音信号的产生机理，并对语音信号进行数学建模，然后阐述了语音信号的转换原理，还介绍了对语音转换效果的主观和客观评估方法。

3 语音信号分析及特征参数提取

3.1 语音信号预处理

初采录的语音信号或是语音库中的语音信号都是模拟信号，在对语音信号进行数字处理之前，先要将模拟语音信号采样离散化，以避免信号的频域混叠失真。然后再进行数字处理。

3.1.1 预加重处理

当语音信号的平均功率谱受到声门激励和口鼻辐射的影响时，高频端大约在 800Hz 以上会按 6dB/倍频程跌落，因此语音信号需要在预处理中进行预加重处理，以加重语音信号的高频部分，从而使语音信号的频谱变得平坦，增加语音的高频分辨率[26]。预加重可以采用一阶 FIR 高通数字滤波器来实现，数字表示为：

$$H(z)=1-uz^{-1} \tag{3.1}$$

其中，u 为预加重系数，其值一般近似于 1。

将预加重系数 u 设为 0.98 时，高通滤波器的特性表示如图 3.1。

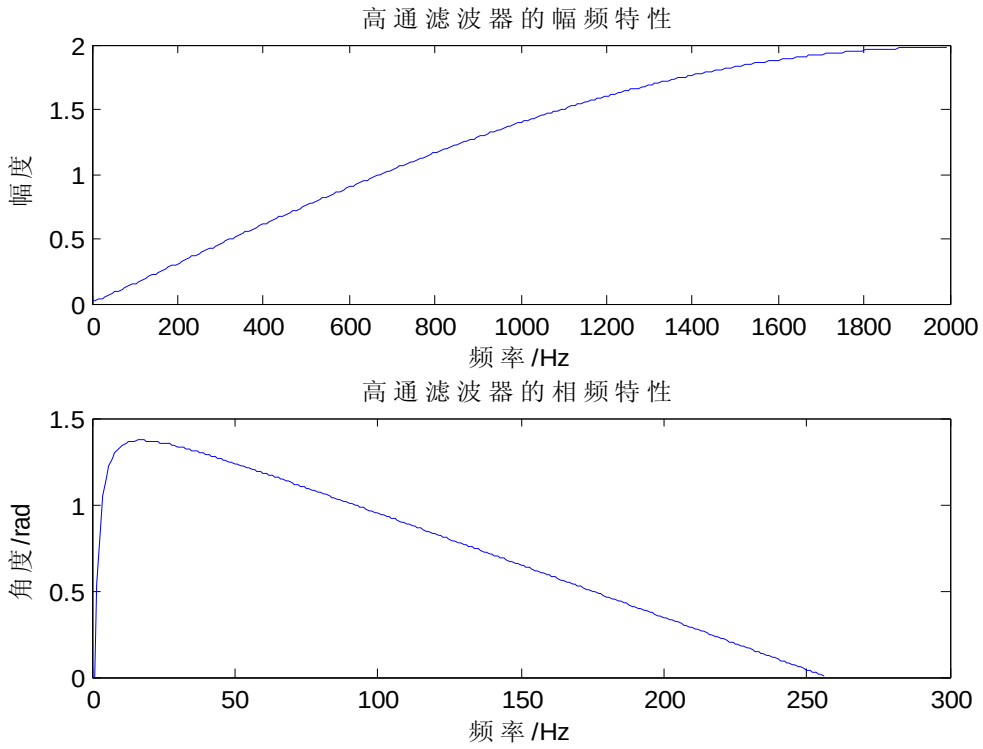


图 3.1 高通滤波器的幅频和相频特性

对一段语音信号进行上述高通滤波器预加重后，信号的频谱在高频部分的幅度得到了提升。图 3.2 为语音信号预加重前后的信号波形及频谱图。

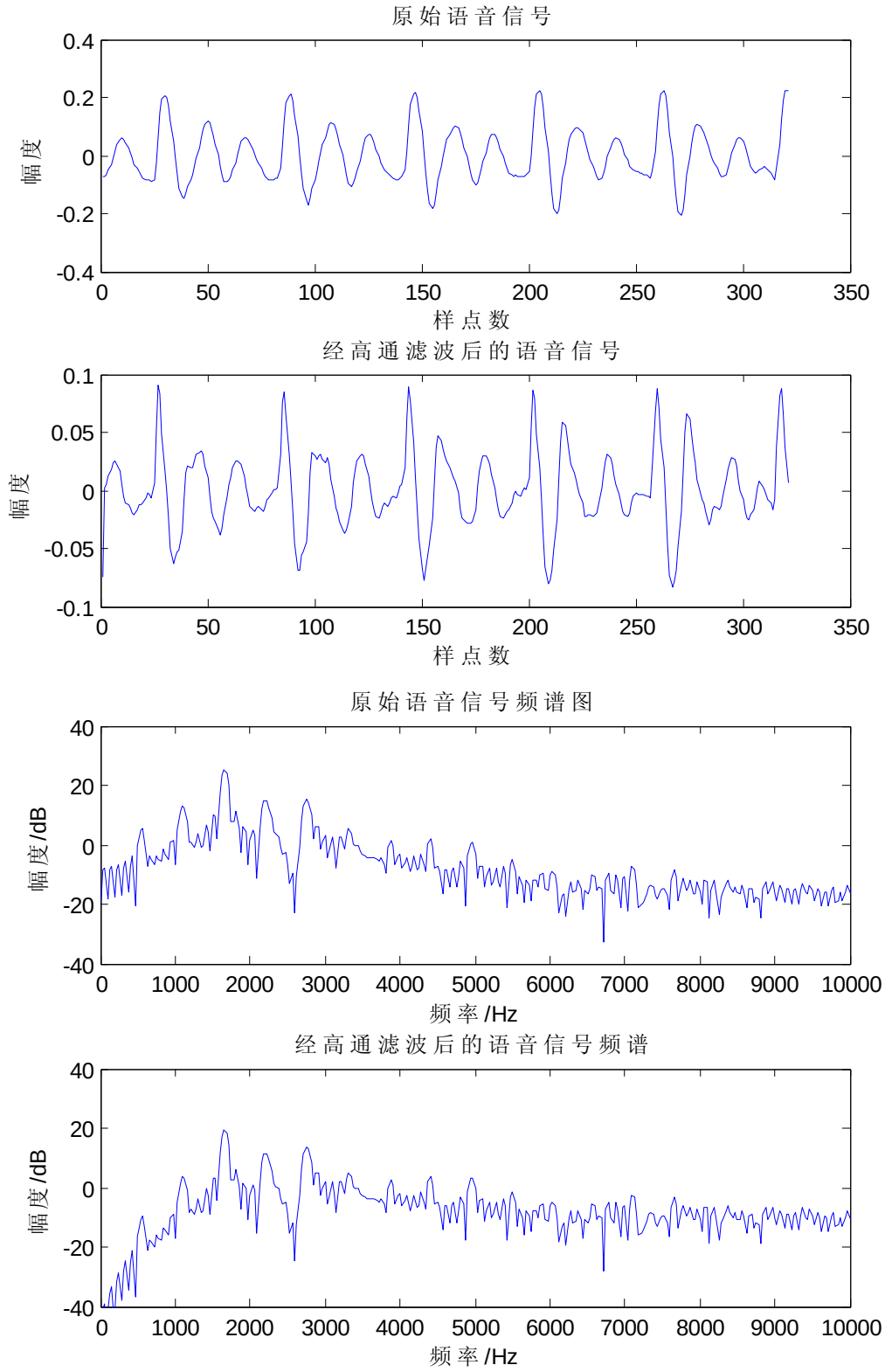


图 3.2 语音信号预加重前后的信号波形及频谱图

3.1.2 分帧处理

前面我们提过语音信号是一种随时间而变化的信号，主要分为两大类——浊音和清音。

浊音的基音周期、清浊音信号幅度和声道参数等都随时间而缓慢变化。并且实际的语音信号是很长的，我们不能也不必对非常长的数据进行一次性处理。明智的解决办法就是每次取一段数据，进行分析，然后再取下一段数据，再进行分析。

而语音信号又具有短时平稳性，即由于发音器官的惯性运动，可以认为在小段时间内（一般为 10～30ms）语音信号是近似不变的。因此，取段也就是我们所说的分帧，可以把语音信号划分为一些短时间段即分析帧来进行处理，其中每一段信号就是一帧，每一段的长度就是帧长。分帧虽然可以采用连续分段的方法，但为了能保持每一帧的连续性，即使帧与帧之间平滑过渡，可以选择交叠分段的方法。交叠分段就是对语音信号分帧时，每相邻的帧之间都有一部分信号是重叠的，前后两帧信号的交叠部分就称为帧移[23]。一般，帧移长度取帧长的 0～1/2 长。至于分帧的方法，则是采用后面我们将提到的加窗的方法来实现。

3.1.3 加窗处理

分帧可以通过对可移动的有限长度窗口进行加权的方法来实现。窗每次移动的距离如果恰好与窗的宽度相等，相当于各帧语音信号是相互衔接的，而如果窗的移动距离比窗宽要小，那么相邻窗之间将会有一部分是重叠的。

在语音数字信号处理中常用的窗函数有直角窗也称矩形窗和 Hamming 窗两种，窗序列沿着语音样点值序列逐帧从左向右移动，N 为窗长度。

矩形窗属于时间变量的零次幂窗。在信号处理中对矩形窗的使用是最为频繁的，习惯上不加窗即是使信号通过了矩形窗处理。矩形窗的优点是其主瓣相对集中，其缺点则是具有较高的旁瓣，还伴有负旁瓣。由于频谱能量的泄漏与窗函数两侧的旁瓣有关，因此，矩形窗的高旁瓣和负旁瓣会导致高频干扰和频谱泄漏，甚至可能出现负谱现象。矩形窗函数表达式为：

$$w(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & \text{其他}\end{cases} \tag{3.2}$$

Hamming 窗又称余弦窗，其频谱可以认为是 3 个矩形时间窗的频谱合成，或者说是 3 个 sinc(t)型函数之和。Hamming 窗主瓣相较于矩形窗有所加宽并降低，旁瓣幅度显著减小，在减小频谱泄漏方面，Hamming 窗比矩形窗更适合。但 Hamming 窗主瓣加宽，相当于分析带宽加宽，频率分辨力下降。其窗函数表达式为：

$$w(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)], & 0\le n\le N-1\\ 0, & \text{其他}\end{cases} \tag{3.3}$$

另外还有其他形式的窗函数，如 Hanning 窗，Blackman 窗等，均是中心对称的。

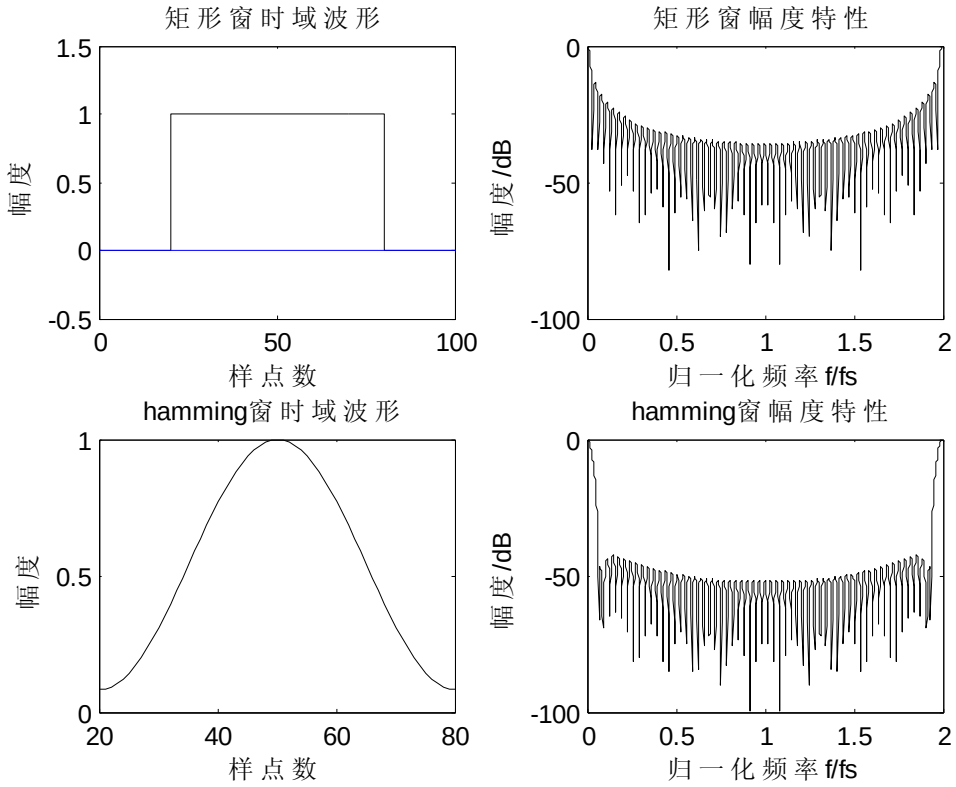


图 3.3 矩形窗和 hamming 窗的时域波形及幅频特性

不同的窗函数对信号频谱的影响是不一样的，这主要是因为不同的窗函数，产生泄漏的大小不一样，频率分辨能力也不一样。图 3.3 为矩形窗和 hamming 窗的时域波形及幅频特性，对比可以看出，矩形窗的主瓣宽度小于 hamming 窗，具有较高的频率分辨率，但是矩形窗的旁瓣峰值较大，因此其频谱泄露比较严重。相比较，虽然 hamming 窗的主瓣宽度较宽，约大于矩形窗的一倍，但是它的旁瓣衰减较大，具有更平滑的低通特性，能够在较高的程度上反映短时信号的频率特性。

经过上述预处理后，语音信号就被窗函数分割成很多帧的短时信号，并且每一帧短时信号都可以被视作为一个平稳随机信号。如图 3.4 所示为语音信号分帧加 hamming 后信号波形。针对分帧后的语音信号就可以对其数字信号处理以进行特征参数的提取。进行处理时，先是按帧取出数据区中的数据，一帧处理完后再取出下一帧的数据进行处理，最后得到按时间排列的特征参数序列，这个序列是由每一帧的语音特征参数共同组成的。

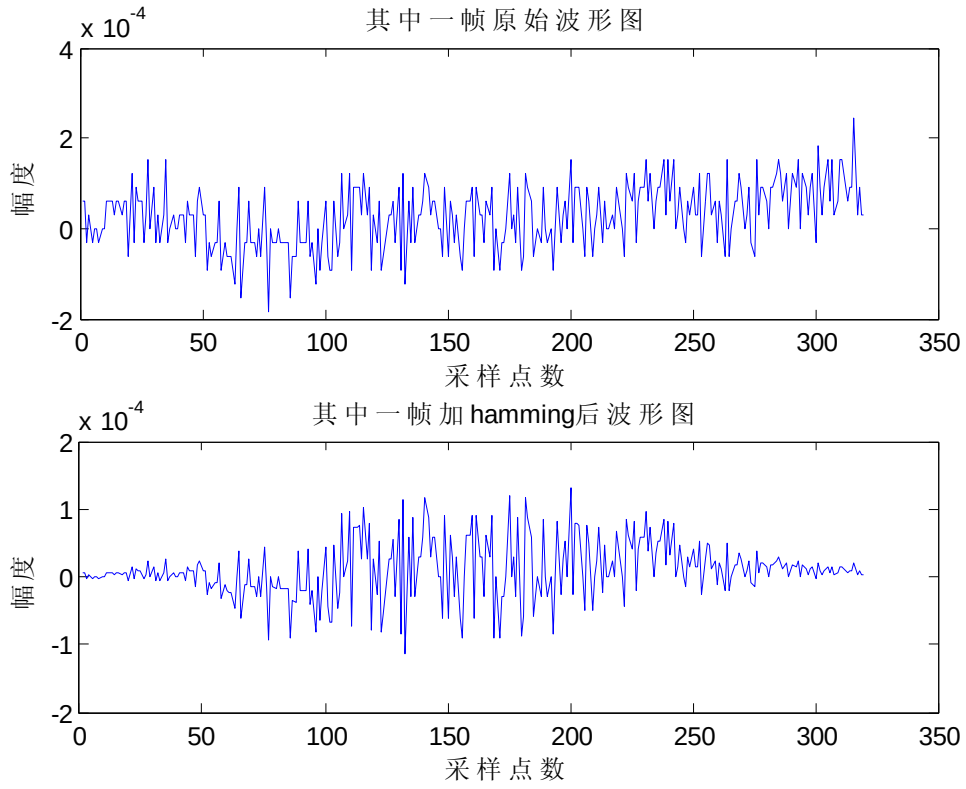


图 3.4 语音信号加 hamming 前后信号波形

3.2 语音信号时域分析

语音信号特性是随着时间变化而改变的，对语音信号分析最直接方法就是以时间为自变量进行分析，称为语音信号时域分析。语音信号时域特征参数有短时能量及短时平均幅度、短时过零率、短时相关函数和短时平均幅度差函数等。

3.2.1 短时能量及短时平均幅度

因为语音信号的能量随时间变化比较明显，一般清音信号和浊音信号之间的能量差异特别突出，像是一段信号中浊音部分的能量就比清音部分的能量大得多。因此，可以通过对短时能量的分析描述语音的这种特征变化。

定义 n 时刻某语音信号的短时平均能量[32] En 为：

$$E\text{n}=\sum_{m=-\infty}^{+\infty}[x(m)w(n-m)]^2=\sum_{m=n-(N-1)}^{n}[x(m)w(n-m)]^2 \qquad (3.4)$$

窗函数的选择直接影响着短时能量的计算，因此窗函数的选择显得尤为重要，若是窗长过长，会使计算出的短时能量变化很小，没法将语音信号的特性表征出来，若是窗长过短，又会保留下语音信号的振幅瞬间变化细节，无法反映出振幅包络的变化。

由于短时能量对信号的高低电平非常敏感，在进行信号样值计算时容易溢出，因此在度量语音信号幅度时可以采用另一个参数，短时平均幅度函数[32]来表示：

$$\mathrm{Mn}=\sum_{m=-\infty}^{+\infty}|x(m)|w(n-m)=\sum_{m=n-(N-1)}^{n}|x(n)|w(n-m) \qquad (3.5)$$

通过表达式可以发现，短时平均幅度相较于短时能量计算得到简化。

图 3.5 是对语音信号进行短时能量和短时平均幅度的处理后得到的。

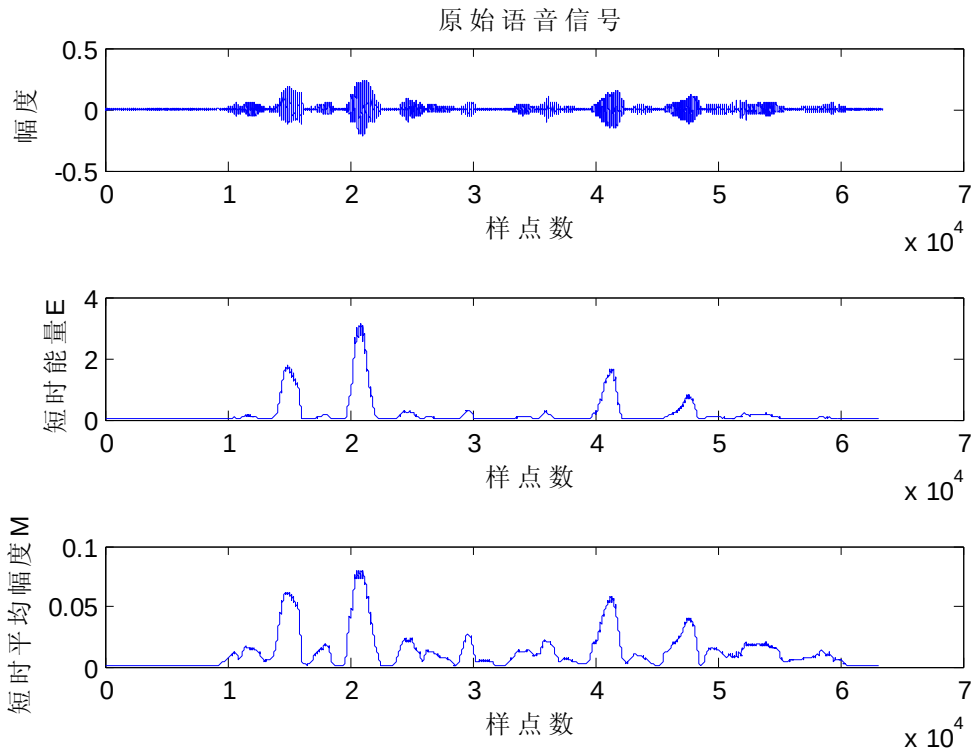


图 3.5 语音信号短时能量与短时平均幅度

前面讲过语音中的清音和浊音之间的能量差别相当突出，由图 3.5 中原语音信号波形和其短时能量波形及短时平均幅度波形的对比中，我们也可以发现，浊音的短时能量和短时平均幅度明显高于清音。所以可以选取短时能量和短时平均幅度函数区分清浊音，另外还可以用来区分声母与韵母，无声与有声，连字分界等。

3.2.2 短时过零率分析

每一帧语音中信号通过零电平的次数称为短时平均过零率。因而对于连续语音信号，语音信号的时域波形通过横轴意味着过零，而对于离散语音信号，则若是相邻的采样值的代数符号不同则称为过零。因此，单位时间内信号过零的次数称为过零率。一段长时间内的过零率就称为平均过零率了。

对语音信号短时过零率[27]的定义为：

$$Zn=\frac{1}{2}\sum_{m=0}^{N-1}\left|\mathrm{sgn}[x_n(m)]-\mathrm{sgn}[x_n(m-1)]\right| \tag{3.6}$$

其中，sgn[ ]为符号函数，即

$$\mathrm{sgn}[x(n)]=\begin{cases}1, & x(n)\geq 0\\ -1, & x(n)<0\end{cases} \tag{3.7}$$

过零率在某种程度上来说是可以反映语音信号的频率信息的。对于浊音来说，尽管声道有多个共振峰，但因为声门会引起谱的高频跌落，使其语音能量集中在 3kHz 以下；而对清音来说，语音能量集中在高频上。高频意味着语音信号具有较高的平均过零率，低频则相反。因此，可以认为，相较于清音信号，浊音信号具有较低的平均过零率，而清音信号则具有较高的平均过零率。图 3.6 是对一段加入 15dB 噪声后的语音信号画出的短时过零率变化曲线，从图中可以看出清音与浊音的短时过零率区别还是比较明显的。

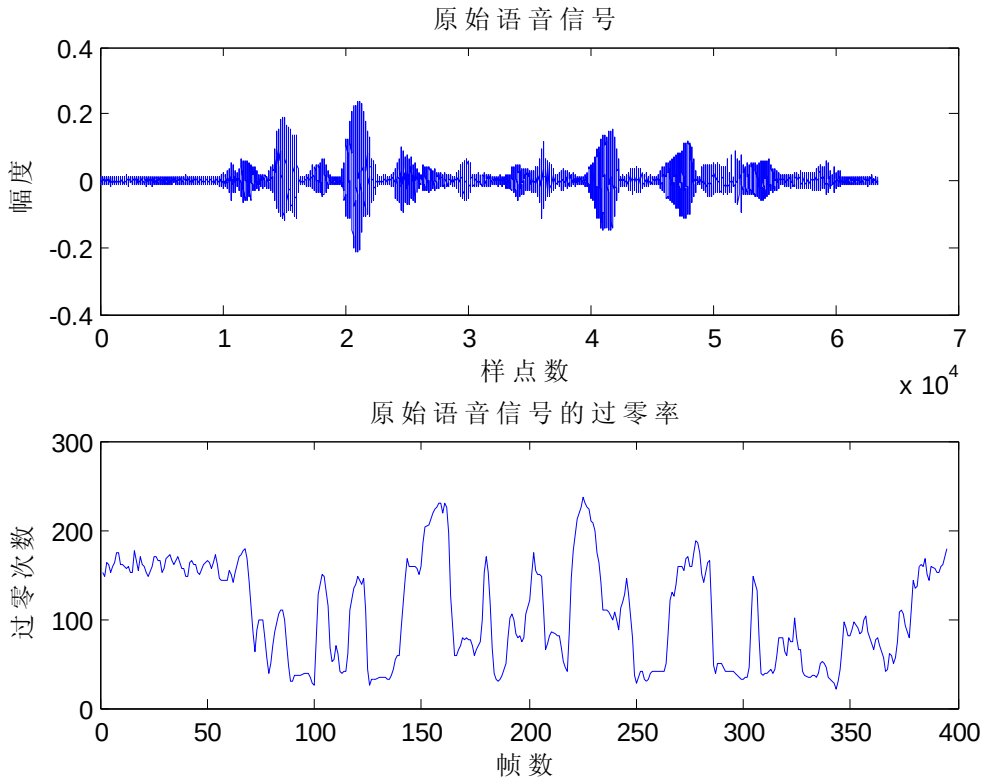


图 3.6 语音信号短时过零率

3.2.3 短时自相关分析

语音信号的短时相关分析分为自相关和互相关，一般互相关是研究两个信号之间的相关性，而自相关则是研究一个信号自身时间波形的相关性。由于不同语音的发音机理不同，因此不同语音的时间波形的差异性较大，像是浊音信号的时间波形具有一定的周

期性，波形之间的差异性自然就比较小，相似性则比较好；而清音信号的时间波形是随机白噪声的特性，没有什么规律，则波形之间的差异性自然就比较大，相关性也就差。因此对于不同语音信号的相似特性可以用短时自相关函数来进行表示。对其定义为：

$$R_n(k)=\sum_{m=0}^{N-1-k} x_n(m)x_n(m+k) \quad 0\le k\le K \tag{3.8}$$

其中，K 是最大的延迟点数。

由此可以知道，短时自相关函数可以很好的反映出浊音信号的周期性，但不同窗函数对又会影响自相关函数的结果。因此，不同的窗长会直接影响浊音信号的短时自相关性。由于语音信号随时间变化这个特性，需要窗长 N 尽可能的小些，才不至于影响信号的短时性；而为了明显反映出语音信号变化的周期性，又必须使窗长 N 足够宽，能够匹配预期的基音周期。这两者显然是相互矛盾的，要同时满足显然是不可能，这就需要能够让窗长自适应基音周期的变化，但计算起来会增加复杂度。为了解决这个问题就有了修正的短时自相关函数来代替：

$$\hat{R}_n(k)=\sum_{m=0}^{N-1} x(n+m)x(n+m+k) \quad 0\le k\le K \tag{3.9}$$

其实，若严格来说，$\hat{R}_n(k)$ 并不是真正的自相关函数，只是具有互相关函数的性质，但是 $\hat{R}_n(0)$ 与最接近的第一个最大值点仍然代表了基音周期的位置，可以用来估算浊音信号的基音周期。

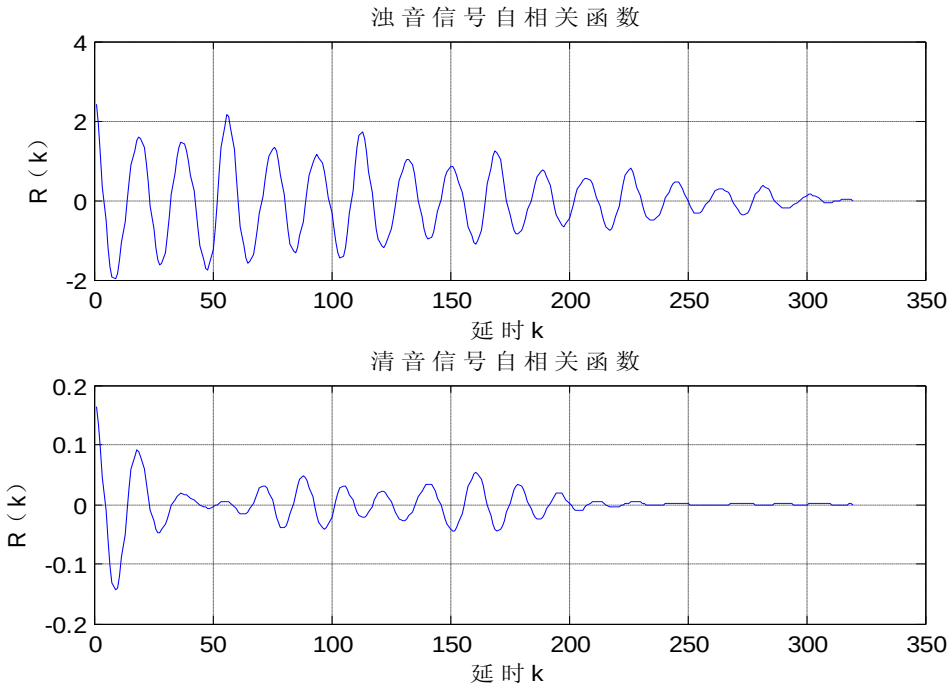


图 3.7 浊音和清音信号的短时自相关函数

如图 3.7，为一帧浊音信号和一帧清音信号的短时自相关函数波形。可以观察发现，

浊音信号的短时自相关函数波形有明显的峰值突出，周期性也非常明显，而清音信号的自相关函数没有周期性，也不具有明显突出的峰起，其性质类似于白噪声。

3.2.4 短时平均幅度差函数

短时平均幅度差函数是与自相关函数所起作用类似的参量，但是可以避免自相关函数中的乘法，从而减少运算时间。短时平均幅度差函数主要是利用差值，因为如果说语音信号是完全的周期信号，则距离为周期的整数倍数上的样点的幅值是相等的，其差值为零。实际的语音信号虽然不是完全的周期信号，差值不为零，但其值也很小，因此，可以将短时平均幅度差函数定义为：

$$F_n(k)=\sum_{m=0}^{N-1-k}\left|x_n(m)-x_n(m+k)\right| \tag{3.10}$$

如图 3.8，即为一帧浊音信号和一帧清音信号的短时平均幅度差函数波形。

短时平均幅度差函数也可用于基音周期的检测，并且因为没有了需要长时间处理的乘法，相较于短时自相关方法计算起来更为简便。

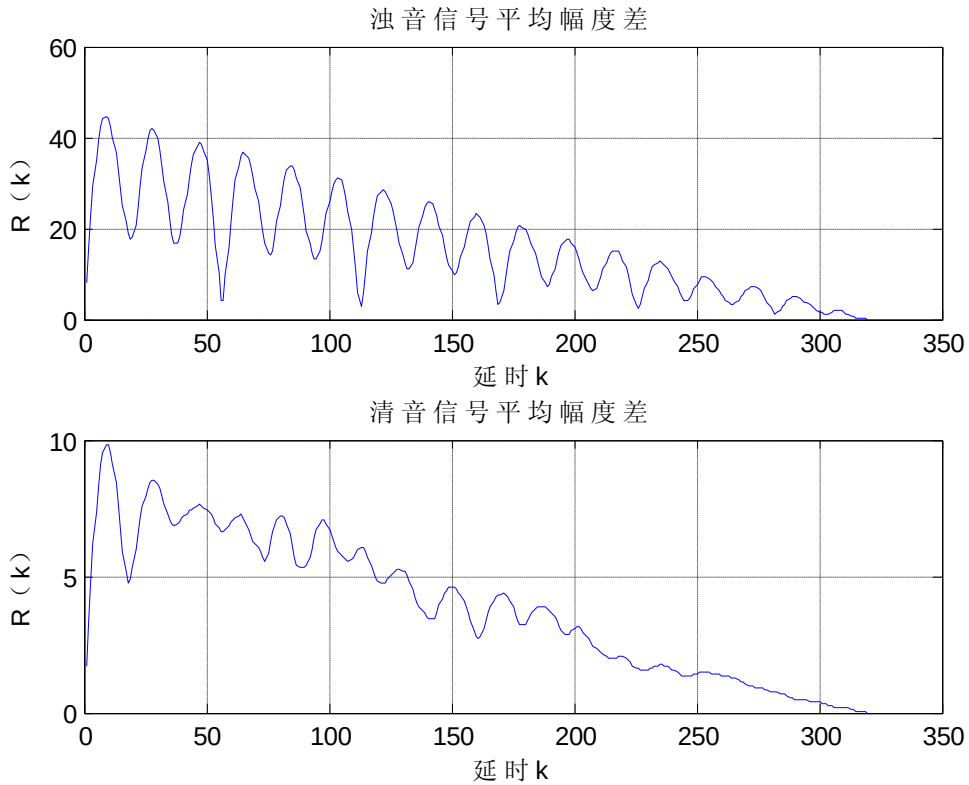


图 3.8 浊音和清音信号的短时平均幅度差函数

3.3 语音信号线性预测分析

线性预测分析（LPC）是最有效的语音分析技术之一，在语音编码、语音合成、语

音识别和说话人识别、语音转换等语音处理领域中得到了广泛的应用。语音线性预测的基本思想是：一个语音信号的抽样值可以用过去若干个取样值的线性组合来逼近。通过使实际语音抽样值与线性预测抽样值的均方误差达到最小，可以确定唯一的一组线性预测系数。

采用线性预测分析不仅能够得到语音信号的预测波形，而且能够提供一个非常好的声道模型。如果将语音模型看作激励源通过一个线性时不变系统产生的输出，那么可以利用 LPC 分析对声道参数进行估值，以少量低信息率的时变参数精确地描述语音波形及其频谱的性质。此外，LPC 分析还能够对共振峰、功率谱等语音参数进行精确估计，LPC 分析得到的参数可以作为语音转换的重要参数之一。

3.3.1 LPC 分析基本原理

LPC 分析为线性时不变因果稳定系统$V(z)$建立一个全极点模型，并利用均方误差准则，对已知的语音信号$S(n)$进行模型参数估计。

如果利用 p 个取样值来进行预测，则称为 p 阶线性预测。假设用过去 p 个取样值$\{s(n-1),s(n-2),\cdots s(n-p)\}$的加权之和来预测信号当前取样值$S(n)$，则预测信号$\hat{S}(n)$为：

$$\hat{S}(n)=\sum_{k=1}^{p}a_k(n-k) \tag{3.11}$$

其中加权系数用a_k表示，称为预测系数，则预测系数误差为：

$$e(n)=s(n)-\hat{S}(n)=s(n)-\sum_{k=1}^{p}a_k(n-k) \tag{3.12}$$

要使预测最佳，则要使短时平均预测误差最小有：

$$\varepsilon=E\left[e^2(n)\right]=\min \tag{3.13}$$

$$\frac{\partial\left[e^2(n)\right]}{\partial a_k}=0,\ 1\le k\le p \tag{3.14}$$

令

$$\phi(i,k)=E\left[s(n-i),S(n-k)\right] \tag{3.15}$$

最小的ε可表示成：

$$\varepsilon_{\min}=\phi(0,0)-\sum_{k=1}^{p}a_k\phi(0,k) \tag{3.16}$$

显然，误差越接近于零，线性预测的准确度在均方误差最小的意义上为最佳，由此可以计算出预测系数。

通过 LPC 分析，由若干帧语音可以得到若干组 LPC 参数，每组参数形成一个描绘该帧语音特征的矢量，即 LPC 特征矢量。由 LPC 特征矢量可以进一步得到很多种派生特征矢量，例如线性预测倒谱系数、线谱对特征、部分相关系数、对数面积比等等。不同的特征矢量具有不同的特点，它们在语音编码和识别领域有着不同的应用价值。

使用一段采样频率为11000Hz的语音进行测试、运行，取其中第30帧进行观察，线性预测阶数为12，看到图3.9所示的原始语音帧的波形，预测语音帧波形和它们之间预测误差的波形。

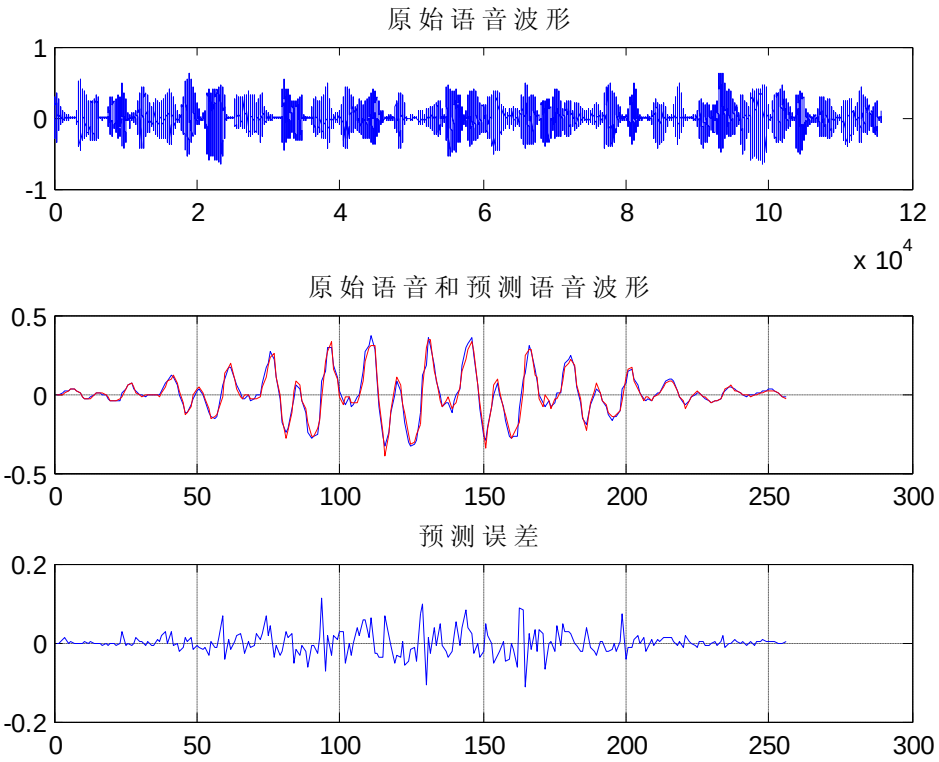


图 3.9 原始语音、预测语音及预测误差波形

从图3.10我们可以看出，原始语音帧和预测语音帧的短时谱和LPC谱的波形。

3.11是原始语音和预测误差的倒谱波形，我们可以从中计算出原始语音的基音周期。从图中看出两峰值之间的间隔为40点左右，因此语音信号的基音周期为40/11000=3.6ms，频率为278Hz左右。

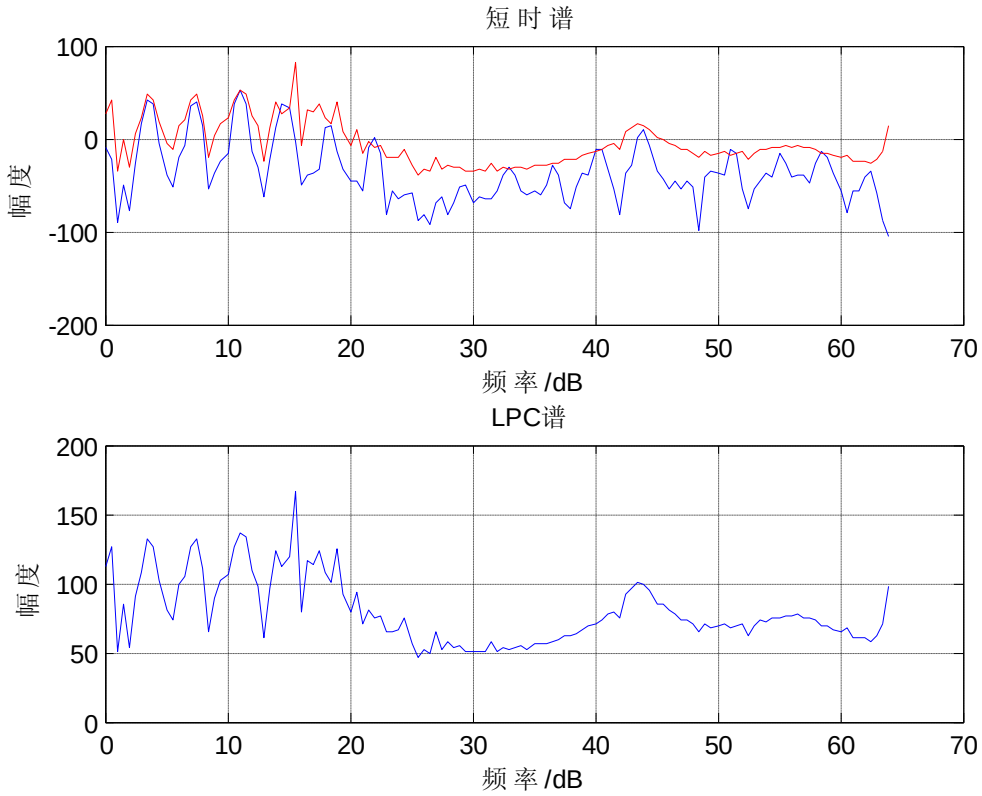


图 3.10 原始语音和预测语音的短时谱和 LPC 谱

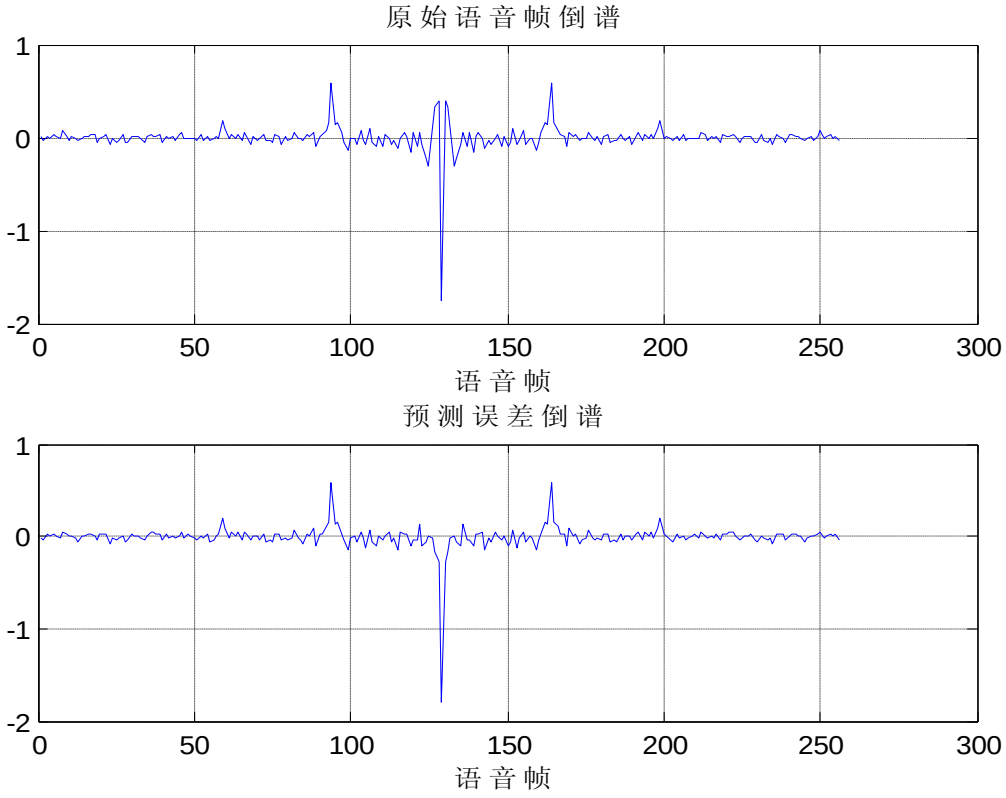


图 3.11 原始语音和预测误差的倒谱

3.3.2 LSP 参数

线谱对（LSP）参数[39]是语音信号频谱包络中的其中一个特征参数，和频谱包络的共振峰有非常紧密的联系。是线性预测系数的一种推演参数，具有非常好的量化特性和插值特性，因而在语音转换的研究中获得了非常广泛的应用[40]。

对LSP参数分析中，仍然采用全极点模型为基础，设p阶线性预测误差滤波器传递函数为A(z)， $A(z)=A^{(p)}(z)$，由线性预测推倒可得[23]：

$$A^{(p)}(z)=A^{(p-1)}(z)-k_p z^{-1}A^{(p-1)}z^{-1} \tag{3.17}$$

令 $P(z)=A(z)-z^{-(p+1)}A(z^{-1})$，$Q(z)=A(z)-z^{-(p+1)}A(z^{-1})$

则 $$A(z)=\frac{1}{2}[P(z)+Q(z) \tag{3.18}$$

$A(z)$ 和合成滤波器 $H(z)$ 之间满足 $H(z)*A(z)=1$。也就是说当 $A(z)$ 的零点在 z 平面单位圆内时，$P(z)$ 和 $Q(z)$ 的零点都在单位圆上，并且其零点沿着单位圆随着 ω 的增加而交替出现。

设 $P(z)$ 和 $Q(z)$ 的第 i 个零点分别为 ω_i 和 θ_i，ω_i 和 θ_i 的排列关系是：

$$0<\omega 1<\theta 1<\cdots<\omega_{p/2}<\theta_{p/2}<\pi$$

ω_i 和 θ_i 就是和LSP系数对应的线谱频率LSF，分别为P(z)和Q(z)的第i个根，$\cos\omega_i$、$\cos\theta_i$ 就是LSP系数在余弦域的表示。

$P(z)$ 和 $Q(z)$ 可进行因式分解成：

$$\begin{aligned}P(z)&=(1+z^{-1})\prod_{i=1}^{p/2}(1-2\cos\omega_i z^{-1}+z^{-2})\\ Q(z)&=(1-z^{-1})\prod_{i=1}^{p/2}(1-2\cos\theta_i z^{-1}+z^{-2})\end{aligned} \tag{3.19}$$

因为因式分解中的系数 ω_i、θ_i 成对出现，反映出了谱的特性，所以称之为“线谱对”，就是线谱对进行分析要求的系数。线谱对参数能够很好地反映声道幅度谱的特点，在语音信号幅度谱较大的地方LSP分布较密，反之较疏，这就相当于反映出幅度谱的共振峰特性，并且每对零点 (ω_i,θ_i) 对应于一个共振峰。

求解线谱对参数就是求解 $P(z)$ 和 $Q(z)$ 关于 z 的根。当线性预测系数求出后，就可以通过代数方程式求根法和DFT法两种方法求解LSP参数。

1、代数方程式求根

因为

$$\prod_{j=1}^{m}(1-2z^{-1}\cos\omega_j+z^{-2})=(2z^{-1})^m\prod_{i=1}^{m}\left(\frac{z+z^{-1}}{2}-\cos\omega_j\right) \tag{3.20}$$

令 $\left(z+z^{-1}\right)/2\big|_{z=e^{j\omega}}=\cos\omega=y$，则可以通过变换，使 $P(z)/(1+z^{-1})=0$ 和 $Q(z)/(1-z^{-1})=0$ 表示成关于 y 的一对 $p/2$ 次代数方程组。这对代数方程可以用牛顿迭代法求解得到方程的根，进一步求出 ω_i 和 θ_i。

2、DFT法

对 $P(z)$ 和 $Q(z)$ 的系数求DFT，得到 $z_k=e^{-j\frac{k\pi}{N}},k=0,1,\cdots,N$ 各点的值，搜索极小值点的位置，即可能的零点位置。用这种方法可以直接得到LSF参数，而其码长取决于N的取值。

3.4 语音信号特征参数

由诸多不同因素导致每个说话人都有各自特有的说话方式，即每个人的语音特征是不尽相同的。人们可以从说话人发出语音的音色、音高等信息中来感受不同说话人的个性化特征，一般我们在考虑选择用来表征语音特征的参数时会从不同特征的形成因素来考虑：

① 说话人先天性的特征：每个人的发声器官是先天就形成了的，由于不同人生理构造的不同，同一语音经不同人发出也会不同。主要反映在语音的频谱结构上，其中主要包含频谱包络和频谱细节构造都会有所不同，因为频谱包络是声道共振和声道反共振特性的反映，频谱细节构造是声带振动等一些音源特性的反映；

② 说话人后天性说话习惯形成的特征：由于说话人生活在不同环境，有不同的生活方式，这样就形成了不同的说话习惯，自然发音习惯也会有所不同。主要表现在语音的频谱结构的时间变化上，具有一种动态特性；

③ 在语音的识别、转换、合成等阶段由于采用不同模型，合理选取特征参数。

总的来说，选取用于语音转换的特征参数可以从短时声学特征、声学特征的时变和语言学特征这三个方面选取：

① 短时声学特征：短时声学特征也即音段特征信息，在短时间段内不会产生改变。其与各人的发声器官有关，像是LSP参数、共振峰参数、倒谱系数、基音周期等都是短时声学特征参数。

② 时变声学特征：声学特征的时变也即超音段信息，具有时变性，不稳定。与说话人一贯的风格和说话人的韵律特征有关，像是基音频率、因素时长以及基音轨迹等都是时变声学特征参数。

③ 语言学特征：语言学特征与个人的生活环境、文化水平等因素有关。像是说话时的选词造句、方言和口音等都属于语言学特征，现今的说话人语音转换技术还没有达到对这个程度的转换。

说话人的声学特征参数在很大程度上决定了说话人的个性化特征。对说话人发出的语音信号进行分析，并找出表征其本质特征的参数加以利用才能进行高效的语音通信。语音信号的特性是随着时间而不断变化的，是一个非平稳的随机过程。换一个角度看，语音信号尽管随着时间而不断改变，但是，语音信号具有在一个短时间段内时保持不变的基本特性。这是由于人体肌肉的惯性运动所致，不可能瞬时完成一个状态的转变，需要一定的时间过程。所以可以假设在没有完成状态转变时，可以近似的认为它是不变的，只要时间够短，这个设想就是成立的。在一个较短的时间内语音信号的特征基本保持不变，即语音的“短时平稳性”。因此，贯穿于整个语音分析过程采用的是“短时分析技术”。

短时方法是用平稳信号的处理方法处理非平稳信号的关键。虽然短时处理方法是语音处理的基本，但对于一些要求较高的研究领域或应用方面，应该要考虑语音信号的是时变或非平稳的，此时则可以采用HMM进行分析处理。

那语音信号最重要的感知特性在功率谱中比较突出，而相位信号只是起着非常小的作用，所以相较于时域特征参数，频域特征参数则显得更为重要。

对于历年研究人员对语音信号特征参数的分析，可以看出各种特征参数对语音信号的个性化特征都有不同的反应，但总结各项研究结果可以发现决定说话人语音个性特征参数的最主要因素中一定少不了基音周期和共振峰参数，另外，频谱包络中的还有一个重要的特征参数就是线对参数LSP，LSP可以很好地反映幅度谱中的共振峰的特性，所以，LSP也是研究语音转换的一个重要参数。

3.4.1 基音周期估计

在语音产生的数学模型中基音是激励源的一个重要参数，是语音信号最重要的参数之一。基音是一种周期性的声带振动，声带振动频率的倒数就是语音信号的基音周期。对于基音周期的提取和估计是语音信号处理研究中非常需要引起重视的问题，尤其是研究汉语语言时。众所周知汉语是一种有调语言，其声调就反映了基音的变化，携带有极其重要的辨义信息，有区别意义的功能。准确地检测语音信号的基音周期对于高质量的语音分析合成、语音识别、语音转换和语音编码等具有重要的意义。例如在低速率语音编码中，准确地基音检测是非常关键的，它直接影响到整个系统的性能。基于种种影响，语音信号的基音周期是反映说话人个性特征的重要参数，对于说话人语音转换选择基音周期作为转换参数之一是极其必要的，基音周期的提取与估计则显得尤为重要。

因为不同人的声道特征是不同的，就算是同一个人其声道随时间变化也会发生改变，而基因周期的范围比较宽，且同一个人在不同情况下发出语音的基音周期也是有所不同的，另外基音周期还会受到发音声调的影响，所以要精确地对基因周期进行检测实

际上是一件比较困难的事。对其提取与估计的困难主要表现在一下几个方面：

（1）基音周期的变化范围很大[35]，男声女声，成人小孩的基音周期都有很大的区别，低音男声80Hz到高音女声的500Hz，这个范围是非常广的，难以进行基音检测。

（2）声道共振峰有时会严重影响到语音中激励信号的谐波结构，若是想要只是提取与声道有关的声源信息而去掉声道共振峰的影响，这一点是非常不容易的。

（3）语音信号本身是准周期的，信号波形的峰值会受到共振峰结构及随机噪声信号的影响，这样就难以精确地确定出浊音信号每个基音周期开始和结束的位置。

（4）另外，我们知道语音信号的变化是非常繁杂的，声门的激励波形并不就是一个完全的周期序列。特别是语音信号的头和尾部，并不具有声带振动那样的周期性。因地，对于语音信号的过渡部分是很难判断到底是周期还是非周期性的，无法判断其周期性，自然也就无法估量出其基音周期，

我们知道能够准确地检测基音周期是非常重要的，尽管对于基音周期的估计有如此多的困难，但由于其重要性，历来对于提取基音周期的研究一直在进行中，为此提出了各种不同的基因检测算法。

3.4.1.1 基于短时自相关法的基音周期估值

由前面对于短时自相关的分析[25]知，清音信号的时间波形没有什么规律性，自然就没有峰值，自相关性比较差，而浊音信号的时间波形具有一定的周期性，在基音周期的整数倍上会出现峰值，相似性是比较好的。根据这一个特点，我们可以采用短时自相关法来对基音周期进行检测。当语音信号有明显的峰值时，可以判断是浊音信号，而没有峰值时，则可以判断是清音信号。另外根据浊音信号时间波形上相邻两个峰值的位置就可以通过计算估计出其基音周期值。

在进行浊音信号的自相关处理中，由于受到声道共振峰特性的干扰，基音的周期性和共振峰的周期性可能会出现混叠，从而会加大基音周期检测的难度，纵使检测出来也会有一定的误差。这就需要在估算基音周期前去掉声道共振峰对基音周期的干扰。因此为了突出基音的周期性，减少无关信息的干扰，我们可以对语音信号先进行预处理再进行自相关计算。

我们知道信号的共振峰信息出现在语音信号的低幅度部分，而基音信息则出现在语音信号的高幅度部分，因此可以对语音信号采用非线性变换，其中一种有效地变换方式就是“中心削波”[26]，用来消除语音信号的低幅度部分。

中心削波函数表示如下：

$$f(x)=\begin{cases} x-x_L & x>x_L \\ 0 & -x_L\le x\le x_L \\ x+x_L & x<-x_L \end{cases} \tag{2.21}$$

削波电平 x_L 由语音信号的峰值幅度来确定，作为削波的一个门限值，它等于语音段最大幅度 A_{max} 的一个固定百分数，一般取信号最大幅度的60%～70%。为了达到较好的削波效果，对这个门限值得设定是非常重要的，一般是最大可能地设定高一点，当然，需要满足不损失基音信息这个前提。

语音信号经中心削波的结果是削去了很多与声道有关的波动响应，将超过削波电平的信号保留下来。再对其求自相关函数，这样，语音信号的基音周期所在位置会出现大并且陡的峰值，在其它位置的次要峰值，其幅度却小很多，最终结果是有效减少了倍频和半频错误。图3.12和图3.13分别是语音信号在进行削波前后的对比图和削波前后的修正自相关对比图。

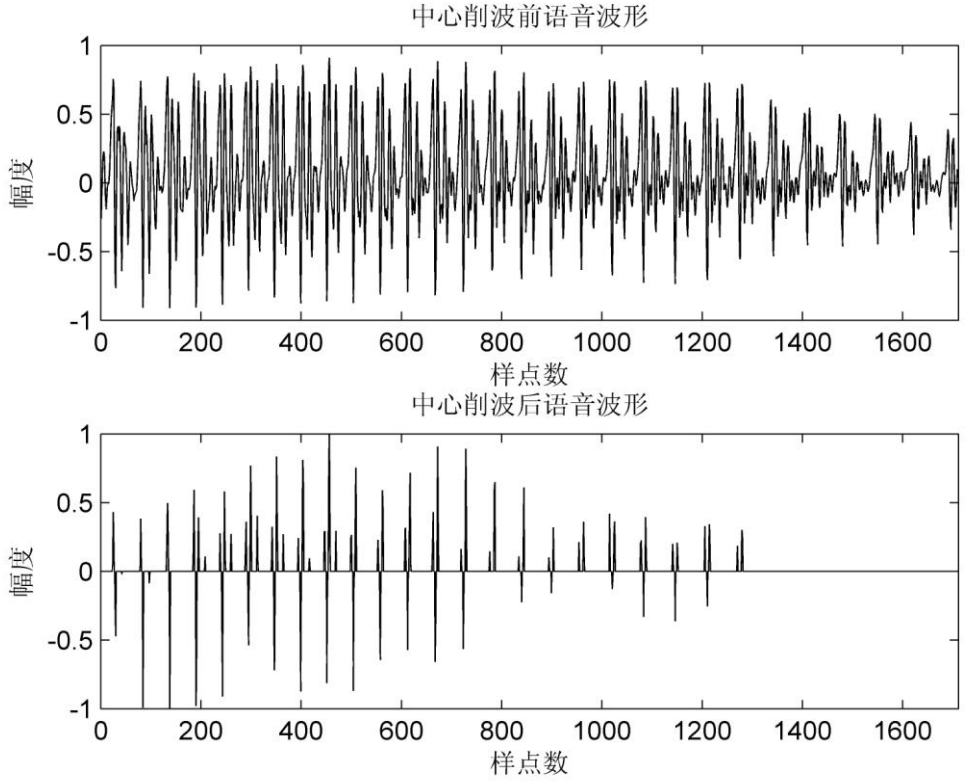


图 3.12 中心削波前后语音信号对比图

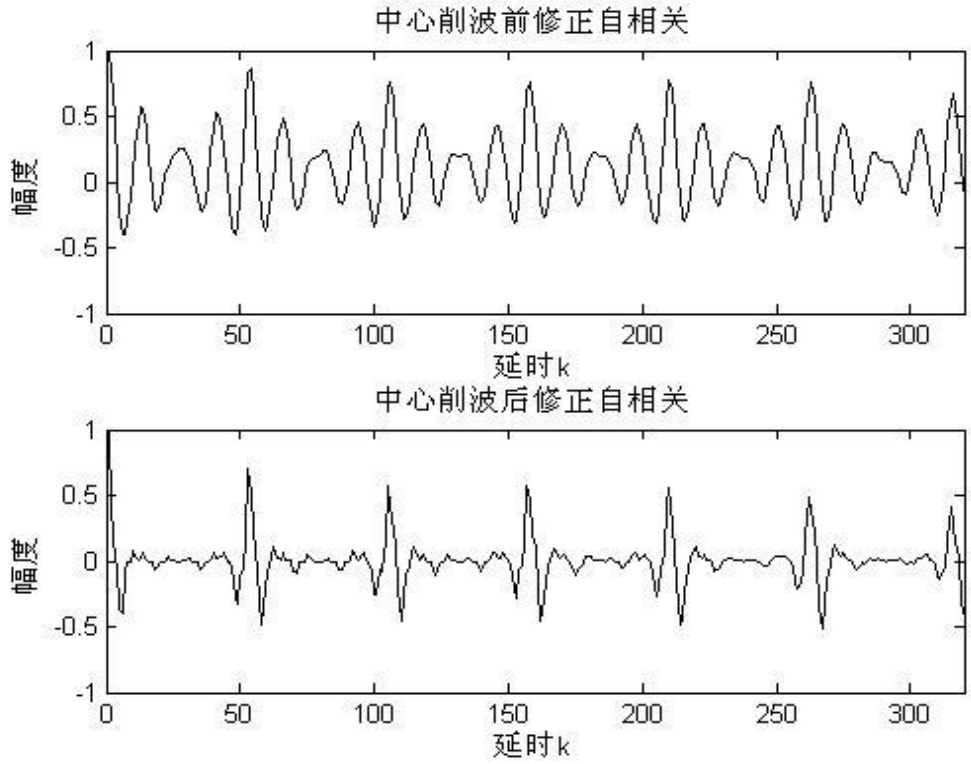


图 3.13 中心削波前后修正自相关对比图

3.4.1.2 基于 AMDF 法的基音周期估值

对于短时平均幅度差函数AMDF我们知道，如果说语音信号是完全的周期信号，则距离为周期的整数倍数上的样点的幅值是相等的，其差值为零。对于实际的浊音语音信号虽然不是完全的周期信号，在基音周期的整数倍上，这个差值不为零，但其值也很小。因而，可以通过计算短时平均幅度差函数中相邻谷值间的距离来对基音周期进行估值。需要注意的是，基于自相关法进行基音周期估值时采用的是两相邻峰值点间的距离，而基于AMDF法[24]进行估值时采用的是两相邻谷值点间的距离。

估值步骤如下：

1、先将语音信号经过低通滤波器处理；

2、计算滤波后的语音信号的平均幅度差函数，并求出最小值的下标，作为语音信号基音周期的初步值；

3、搜索平均幅度差函数的若干局部极小值点作为基音周期的备选点；

4、在某个极小值点左右各几个点范围内对其AMDF值取平均，若该极小值点与此平均值得差距大于阈值则称为清晰谷点，否则为不清晰谷点，舍掉不清晰点。

5、在所有清晰谷点中选择最左边的那个店作为语音信号的基音周期。

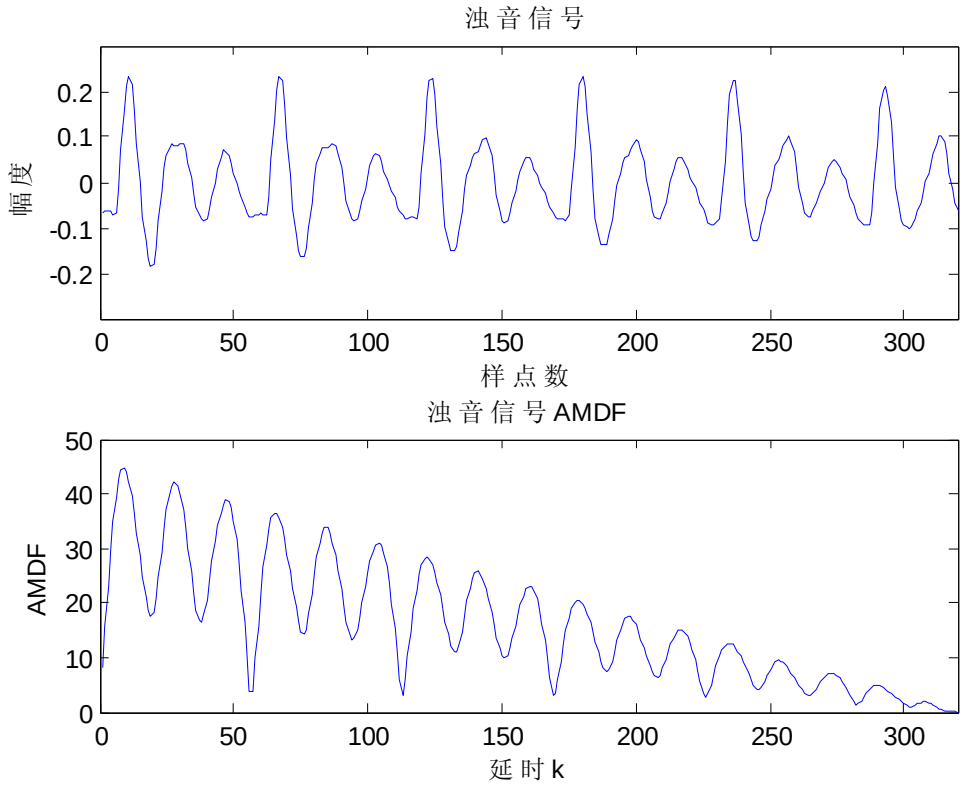


图 3.14 语音信号的短时平均幅度差函数

短时平均幅度差函数中是没有乘法运算的，因此其算法更为简便。并且在基音周期的估值点处，短时平均幅度差函数的谷值点比自相关函数的峰值点更为尖锐，对基音周期的估值也就更为精确。图3.14 所示为语音信号的短时平均幅度差函数。

3.4.1.3 基于小波变换法的基音周期估值

一个信号的小波变换[36]具有这样一个性质：信号小波变换模的极大值点与信号的突变点相对应，至于信号规则变换或是缓慢变化部分会使得小波变换模取得较小值。

由前面对人体发声器官进行剖析和对语音产生机理的分析，可以知道语音是由气流经过声门再激励声道，最后从嘴唇或鼻孔，或同时从嘴唇和鼻孔辐射出来而形成声音。对于浊音语音，语音由声带振动或不经声带振动来产生，其中由声带振动产生的音统称为浊音，并且气流冲击声门时，声门会周期性的开启或是关闭。声门的这种周期性的开启关闭会在语音信号中引起锐变。对语音信号作小波变换，则其极值点与声门的开启闭合相对应，相邻极值点之间的距离就对应着基音周期。这就是采用小波变换法对语音信号进行基音周期检测的原理。

在基音周期的检测估值中应用到的小波变换一般都是采用二进小波变换。如图3.15 所示为语音信号的多级小波分解。

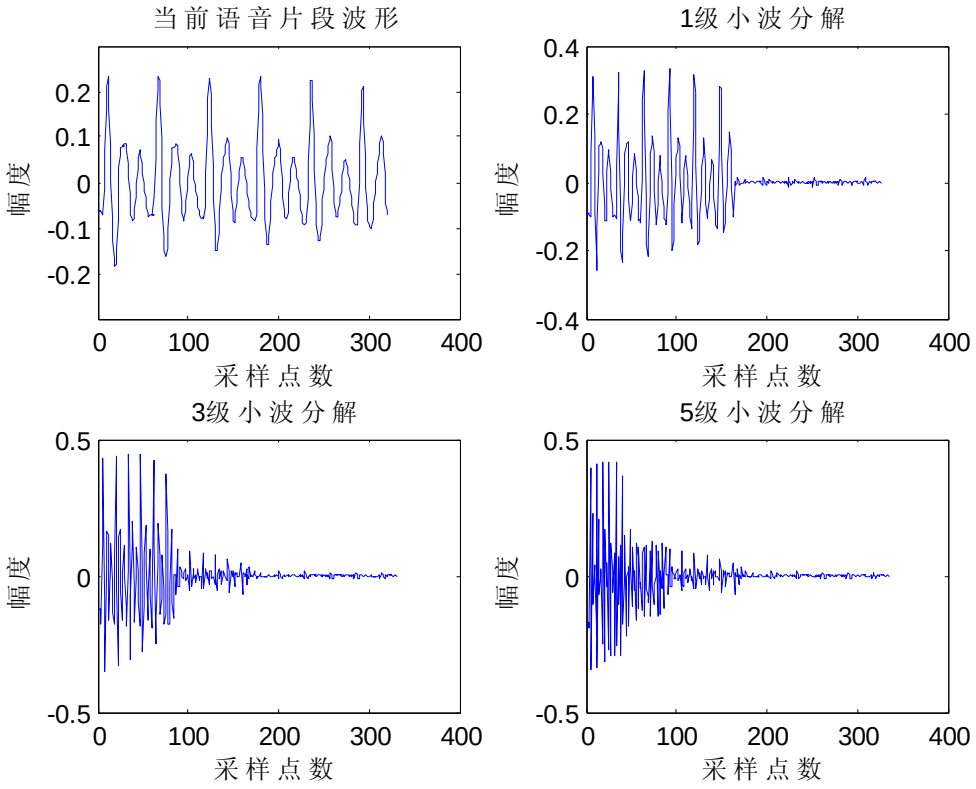


图 3.15 语音信号的多级小波分解

3.4.2 基音周期估值后处理

观察浊音语音信号的波形，可以发现其周期性还是比较明显的，但是形状比较复杂，这使得基因检测算法很难做到处处准确可靠。不管采用哪种方法进行提取，提取出的基音频率轨迹与实际的基音频率轨迹都会有偏差，不可能完全一致。不过大部分还是相吻合的，只是在一些局部区域中会有一个或是几个基音频率估计值产生偏差，甚至远离正常的轨迹，这种偏离出去的点就称为基音轨迹的“野点”[37]。那这些“野点”既然有比较大的偏差就需要将其去除。因此，为了去除这些“野点”，对求得的基音轨迹进行平滑处理就显得非常必要了。常用的平滑处理方法有中值平滑处理、线性平滑处理。

1、中值平滑处理

中值平滑的主要思想是：假设x(n)是输入信号，y(n)则是经中值滤波器滤波后的输出，利用一个滑动窗，则n0处的输出值y(n0)就是将滑动窗的中心移到n0处时，窗内输入样点在这时的中值。也就是在中点n0的左右分别取L个样点，这样连同被平滑点一起构成一组具有（2L+1）个采样值的信号，并且将这些采样值按照大小顺序一字排开，平滑器最后的输出就是排列序列的中间值[37]。其中，一般取滑动窗的窗长为三到五个样值点，这样就可以说明一般L的值为一或者是二。中值平滑处理具有既可以将少量野点去除掉，又不至于损害基音周期轨迹中相邻平滑段跃变型变化的优点。

2、线性平滑处理

线性平滑处理实际上就是一种滤波处理，只是这种滤波处理是线性的，并且是利用一个滑动窗来线性滤波的，其表达式为：

$$y(n)=\sum_{m=-L}^{L} x(n-m)w(m) \tag{3.22}$$

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

$$\sum_{m=-L}^{L} w(m)=1 \tag{3.23}$$

线性平滑处理的一个特点是既将输入信号中的非平滑样点值纠正了过来，又修改了在非平滑样点值附近的样点值。研究发现，加大滑动窗口的长度大尽管可以使信号更为平滑，但是，也有可能面临相邻平滑段之间跃变模糊被加重了的问题。

另外，为了达到平滑的最优效果，可以将中值平滑和线性平滑这两种平滑方法进行组合平滑处理，为了使平滑的基音轨迹更为贴近，还可以采用二次平滑的算法。下图3.16即为经过各种平滑处理后的基音周期轨迹，图中分别采用了五点中值平滑、线性平滑、五点中值平滑和三点中值平滑的组合以及五点中值平滑和五点线性平滑的组合。可以看出，线性平滑的效果相对中值平滑效果要差些，两次中值平滑的组合平滑效果优于一次中值平滑效果，也比中值平滑和线性平滑的组合平滑效果要好。

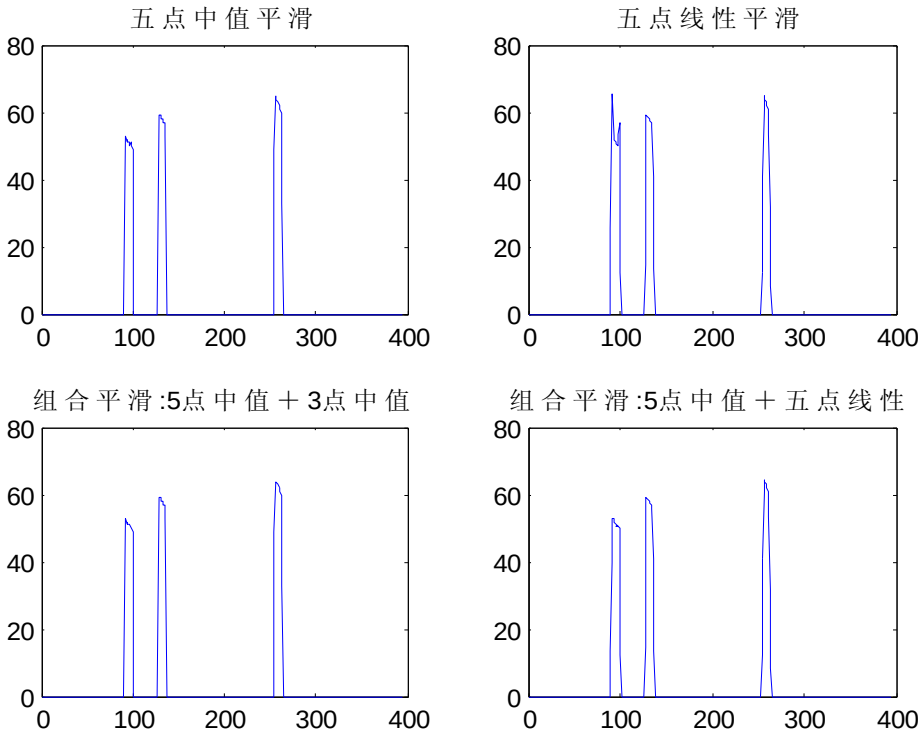


图 3.16 各种平滑算法平滑效果对比图

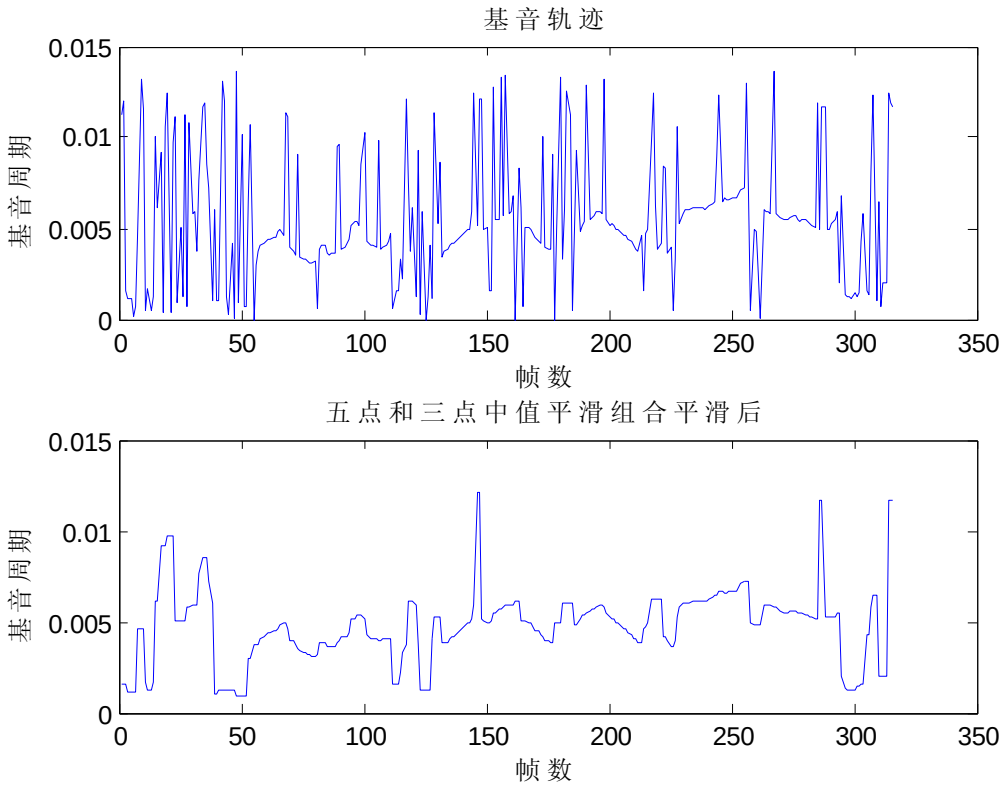


图 3.17 五点中值平滑处理前后基音轨迹对比图

如图 3.17 为一段语音进行五点中值平滑处理前后的基音轨迹对比图，对语音通过短时自相关法进行基音周期估值后得到基音轨迹图，然后再采用五点中值平滑处理。

3.4.3 共振峰估计

在前面研究声道模型时，一般是将声道看成是一段非均匀截面的声管，在发音时声道主要是起到共振器的作用。当声波通过声道时，受到声腔共振的影响，在某些频率附近形成谐振。反映在信号频谱图上，在谐振频率处其谱线包络产生峰值，一般称其为共振峰。共振峰频率及共振峰频带宽度都属于共振峰参数，而共振峰信息则又包含于语音频谱包络中，是频谱包络中的其中一个参数。并且研究认为，共振峰参数是语音频谱包络中最为突出的一个特征参数，所以对共振峰参数的估计关键就是对语音频谱包络的估计。

与对基音周期的估计类似，要精确地对共振峰进行估值实际上也是一件比较困难的事。对其提取与估计的困难主要表现在共振峰产生叠加，出现虚假峰值和高音调语音样点差几个方面[23-27]：

1. 共振峰产生叠加。在语音信号的频谱中可以看到，有些相邻共振峰可能靠得比较近，这样就难以对其进行分辨。而在实际中，要能够寻找一种有效的算法对叠加的共振

峰进行分辨是是有一定困难的。

2. 出现虚假峰值。一般正常情况下，认为语音频谱包络中的最大值就是共振峰。但是，因为一些因素的影响，在语音频谱包络中有时会出现虚假峰值。一般，对语音频谱包络的预测分析方法不同，出现虚假峰值的情况也会有所不同。像是利用非线性处理时，出现虚假情况就比较多，而在采用线性预测方法时，出现虚假峰值情况较少。

3. 高音调语音样点差。因生理个体差异，女声和童声的音调是比较高的，但高音调语音的谐波间隔比较宽，提供给频谱包络的样点比较少。而传统的频谱包络估计方法都是利用谐波峰值提供样点，因为高音调语音无法提供较多的样点就无法进行共振峰的选取。高音调语音中的线性预测包络峰值会有离开其本身实际的位置的倾向性，并且会向与之最为相近的谐波峰值方向移动，所以就算是采用线性预测的方法对频谱包络进行估值也还是存在这个问题。

我们知道能够准确地检测基音周期是非常重要的，对共振峰的提取同样也是非常重要的，纵使有这么些困难，但基于其重要性，对其研究仍然不断，下面是几种不同的算法。

3.4.3.1 基于带通滤波器组的共振峰估值

带通滤波器组法类似于语谱仪，但由于使用了计算机，使滤波器特性的选取更具灵活性，实现框图如图 1 所示。带通滤波器组是最早提取共振峰参数使用的方法，相较于线性预测方法稍微差一些， 但从匹配度来这一方面来看，滤波器组法又优于线性预测方法，这是因为人耳有个灵敏度，使用滤波器组法估计出的共振峰频率与这个灵敏度更为匹配。

带通滤波器由于其中心频率的分布不同具有两种不同的设计。其中一种是将带通滤波器组设计成同样地带宽，来确保滤波器各通道具有一样的群延时，这种设计是在其中心频率在分析频段上的分布是等间距的前提下设计的。另外一种是将带通滤波器组设计成不同的带宽，随高低频变化，高频端的带宽就设计得比低频端的宽，还要使滤波器阶数也和带宽是正比例关系，这样的设计也是为了确保一样的群延时，避免失真波形的产生。

在实际研究中经常使用带通滤波器组分析信号中的共振峰走向，这是因为带通滤波器组具有分辨较高频率的优点。事实上优良的截止特性才能提高频率的分辨率，这就需要带通滤波器取得足够大的阶数，但这又导致每一个滤波器的冲激响应都比较长。而较长冲激响应又会模糊语音的时变特性，所以时间分辨率和频率分辨率是一对矛盾体。尽管如此，还是依然需要使用带通滤波器组的高频率分辨率来对共振峰进行估值。

图 3.18 是利用带通滤波器组的高频率分辨率来对共振峰进行估值的结构示意图。滤波器的中心频率从 150～7000Hz，分析带宽从 100～1000Hz，频率是递增的，并且遵循

对数规律。频谱包络是由滤波器输出并经过全波整流后得到的，峰值检测部分是用来选择在一定频率范围里的峰值,逻辑部分则辨识检测出的峰值并确定出前面3个共振峰值。这样顺序指定出频谱峰值，并且每一个峰值都高于之前的共振峰频率，还被自身已知的频率范围所约束。

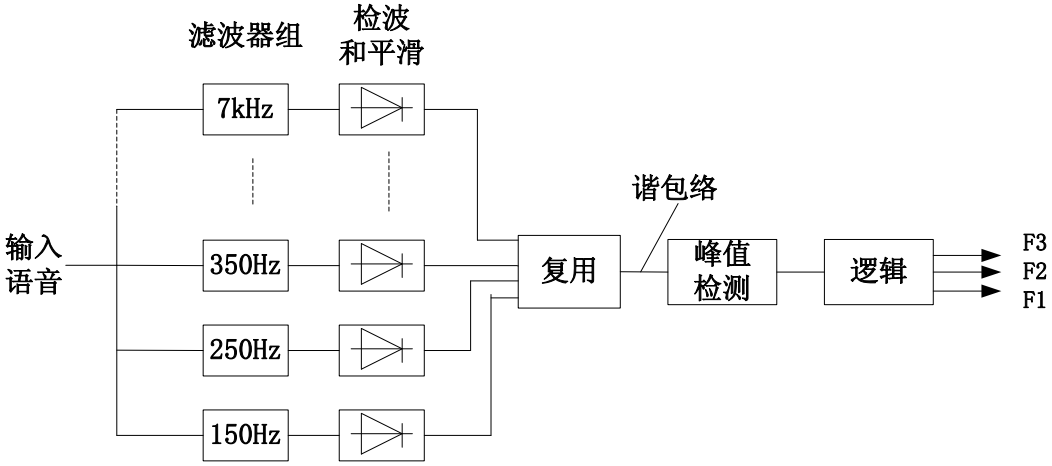


图 3.18 带通滤波器组法提取共振峰

3.4.3.2 基于倒谱法的共振峰估值

虽然可以直接对语音信号求离散傅里叶变换 DFT，然后用 DFT 谱来提取语音信号的共振峰参数，但是这种方法测得的共振峰会有比较大的误差。这是因为基频谐波会影响到由 DFT 得到的谱，共振峰的最大值只会在谐波频率上出现，这就需要找到一种方法将基频谐波给消除掉，倒谱法就可以达到这样一种目的。

倒谱分析也即同态解卷分析，可以分离出基因谐波以及声道频谱包络，主要是利用二次时域和频域之间的变换及对数运算方法进行分离。这样在同态滤波后就可以去除基频谐波的影响，得到较平滑的谱包络，然后就可以较精确地提取出语音信号的共振峰参数。倒谱低时部分用来进行声道、声门以及辐射信息的分析，而高频部分则用来进行激励源信息的分析。首先对倒谱进行低时窗选，然后进入语音倒谱分析系统最后一级进行离散傅里叶变换，最后的输出就是平滑后的对数模函数。一定语音段输入的谐振就可以从这个平滑后的对数谱中反映出来，定位其中的峰值点就可以估计出语音信号的共振峰，这是因为共振峰的频率和对数谱的峰值点基本上是相互对应的。

图 3.19 所示为采用倒谱法进行语音频谱包络求取的原理图。实验表明，倒谱法因为其频谱曲线的波动比较小，所以估计共振峰参数的效果是较好的，但其运算量太大。采

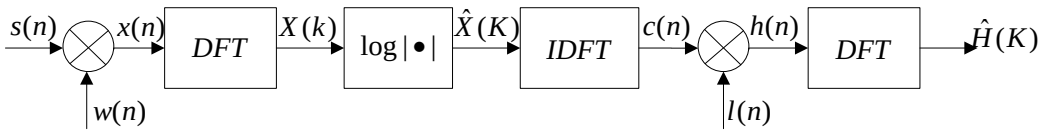


图 3.19 倒谱法语音频谱包络求取原理图

用 MATLAB 经过倒谱法进行共振峰检测后运行出的图 3.20 所示。

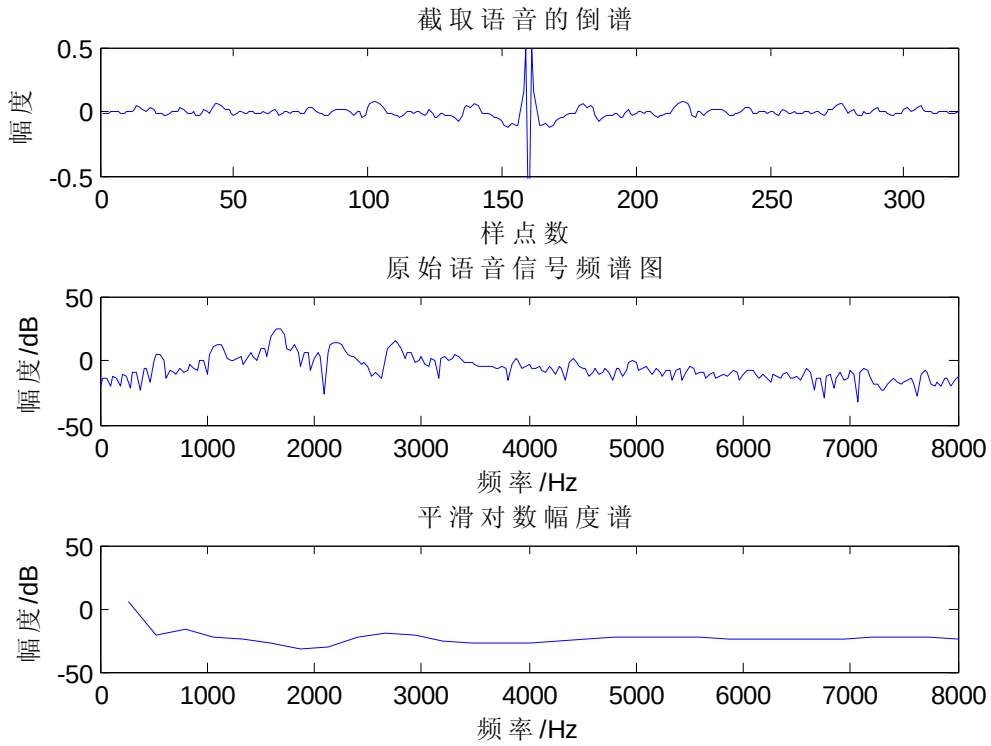


图 3.20　共振峰检测运行结果

3.4.3.3 基于 LPC 法的共振峰估值

从线性预测导出的声道滤波器是频谱包络估计器的最新形式，根据这个声道滤波器可以找出共振峰线性预测提供了一个优良的声道模型(条件是语音不含噪声)。尽管线性预测法的频率灵敏度和人耳不相匹配，但它仍然是最廉价、最优良的行之有效的方法。当然 LPC 法也有其缺点，用一个全极点模型逼近语音谱，对于含有零点的某些音来说全极模型分母多项式的根反映了零极点的复合效应，无法区分这些根是相应于零点还是极点，或完全与声道的谐振极点有关。但如上所述，LPC 法仍行之有效。

用线性预测可对语音信号进行解卷：即把激励分量归入预测残差中，得到声道响应的全极模型 H(z)的分量，从而得到这个分量的 a_i 参数。解卷有一定的缺陷，那就是由于逼近误差的存在而使信号声道响应分量精度稍微降低了，但其非常重要的作用是可以消除激励分量对信号的影响。这时，对声道响应分量进行谱峰求解，就可以得到语音信号的共振峰值。对谱峰的求解可以采用求根法和离散傅立叶变换（DFT）这两种方法。求根法主要是需要计算出全级模型分量中分母多项式的根植；而 DFT 法则是定位分母多项式中离散频率响应谷点值以确定共振峰的位置，DFT 法的运算量相对来说要小一点。

下面详细介绍求根法和 DFT 法。

1、DFT 法

因为 $A(z)=1-\sum_{i=1}^{p}a_i z^{-1}$，所以若求此多项式系数序列$(1,a_1,a_2,\cdots,a_p)$的 DFT，就可得到 $A(k)$。但是，预测阶数 p 一般不大，这就会影响到对离散频率响应谷点值的求解的精度，也就是共振峰频率值精度会受到影响。当频率分辨率具有较高的 DFT 时，对序列时间长度的增加上可以使用后边补零的方法，即用$(1,a_1,a_2,\cdots,a_p,0,0,\cdots,0)$进行 DFT，为了也能使用快速傅立叶变换 FFT，将长度尽量取为 64 点、128 点、256 点和 512 点等 2n 点上。另外，当频率分辨率比较低时，还可以利用抛物线内插技术，来对共振峰频率值进行求取。

设抛物线函数为：$y(\lambda)=a\lambda^2+b\lambda+c$，另 $A(k)$ 的某个谷值为 $y(0)$，而其相邻值分别为 $y(-1)$ 和 $y(1)$，则可求得系数如下：

$$\begin{cases} c=y(0), \\ b=[y(1)-y(-1)]/2, \\ a=[y(1)+y(-1)]/2-y(0) \end{cases} \tag{3.24}$$

接着，求 $dy(\lambda)/d\lambda=0$，就可以得到抛物线极小点位置。

$$\lambda_p=-b/2a \tag{3.25}$$

由此得共振峰频率：

$$F=\frac{f_s}{N}(k_p+\lambda_p) \tag{3.26}$$

式中，f_s 为 $A(k)$ 的取样率；N 为其长度的点数；$\frac{f_s}{N}$ 是频率分辨率，即幅频特性上的一个点所相当的频率数；k_p 为 $y(0)$ 处得离散频率，即在幅频特性上的点数，它相当于频率为 $\frac{fs}{N}k_p$；而 $\lambda_p=-\frac{b}{2a}$ 是相对于图 3 上的点数，也即是到 k_p 的距离。共振峰带宽可先由条件 $y(\lambda_W)/y(\lambda_p)=0.5$ 得出 λ

$$\lambda_W=\frac{-b+\sqrt{b^2-4a(c-0.5y(\lambda_p))}}{2a} \tag{3.27}$$

再由此得

$$B_F=\frac{2(\lambda_W-\lambda_p)f_s}{N}=\frac{f_s\sqrt{b^2-4a[c-0.5y(\lambda_p)]}}{aN} \tag{3.28}$$

2、求根法

求根法是找出多项式的复根，根据找出的复根来确定共振峰的方法。通过对预测多项式系数的分解可以精确地确定出共振峰是求根法的优越之处。找出多项式复根的过程

比较常用的是采用牛顿——拉夫逊{ Newton-Raphson)算法。这个算法的过程是首先猜测出一个根值，并且事先设定一个阈值，然后就这个猜测值进行多项式计算以及求导，最后利用结果找出一个改进了的猜测值，不断重复这个过程，直到前后两个猜测值的差值小于事先设定的这个阈值时，求根过程结束。

假如每一帧的原始猜测值和前一帧根的位置相重合，那么一般来说根的帧帧之间的移动是足够小的，这样经过比较少的重复运算之后，可以使新的根值聚在一起。并且当初始求根时，第一帧最初猜测值在单位圆上可以等间隔放置。

具体的过程是：设某个点 z_i 为一个根，其共轭值也会是一个根。则与 i 相应的共振峰频率 F_i 和 3dB 带宽 B_i 分别为：

$$F_i = \frac{\theta_i}{2\pi T_s} \tag{3.29}$$

$$B_i = \frac{\ln|z_i|}{\pi T_s} \tag{3.30}$$

其中，T_s 是取样周期。

因为 p 是预先选定的预测器阶数，所以复共轭对的数量最多是 $\frac{p}{2}$。这样就很容易排除掉额外极点，因为这些额外极点并不属于共振峰，根据是额外极点的带宽比共振峰的带宽大很多。所以，要解决某一个极点的归属问题就没那么复杂了。

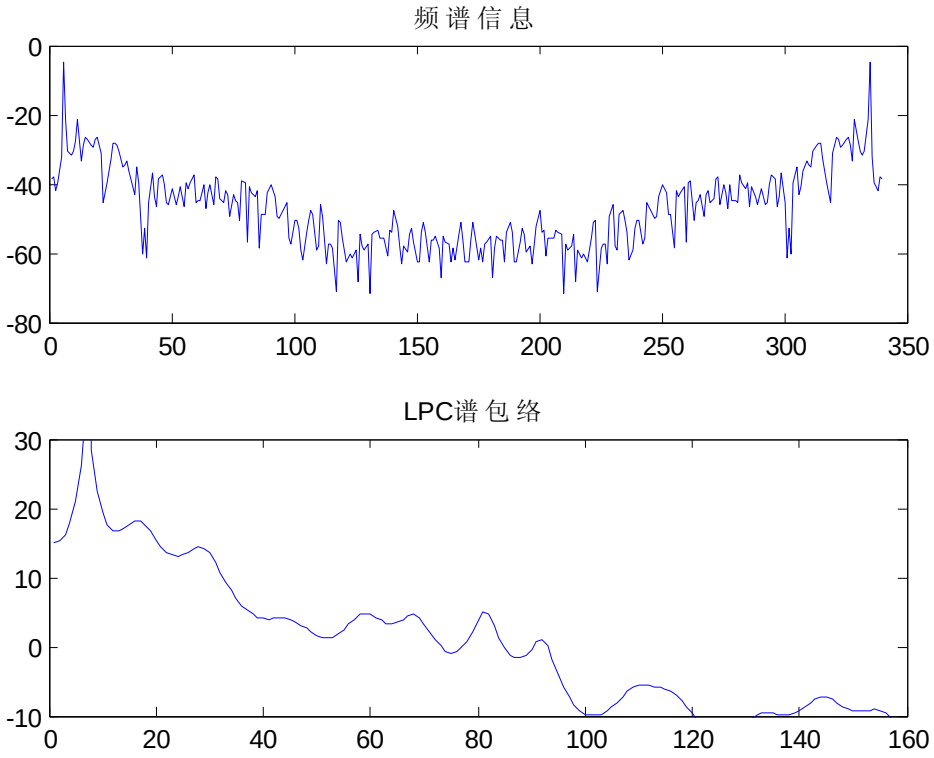


图 3.21 语音信号的 LPC 谱包络信息

图 3.21 为由 LPC 参数表征的语音信号共振峰谱包络信息。

3.5 本章小结

本章对语音信号预处理方法进行了介绍，分析了语音信号的时频特征，还对语音信号的线性预测方法进行了分析，并详细论述了语音信号主要特征参数的提取方法。

4 语音转换算法

在第二章中讨论到语音转换的过程包含训练和转换两部分的内容。在训练阶段，需要将源说话人和目标说话人语音的个性化特征放在一起进行训练，找出两者之间参数的对应规则，在转换阶段，则是将源说话人的个性化特征通过训练阶段得到的对应规则进行转化，然后再合成就得到具有目标说话人个性化特征的语音。

由此语音转换的过程我可以知道语音转换中的重点内容是找到一种最优方法对语音的特征参数进行训练，以得到最优的匹配规则，另外就是找到合适的方法对语音进行特征提取和对转换后的语音进行合成。对语音的转换一般是选取韵律信息和频谱包络进行转换，本章就基于 ANN 的语音转换算法和基于 GMM 的语音转换算法进行研究。

4.1 转换过程

在语音转换系统中，最常用的特征参数就是频谱包络，而频谱包络中的其中一个特征参数就是线谱对参数 LSP，LSP 可以很好地反映共振峰的特性。因此选择 LSP 作为语音频谱包络的转换参数。

将语音转换分为训练和转换两个阶段进行。在训练阶段，将源说话人和目标说话人的语音信号进行预处理，采用 STRAIGHT 语音分析——合成模型，提取源——目标说话人的频谱参数，并将提取出的频谱参数转换为比较好计算的 LSP 参数。接着使用动态时间规整(Dynamic Time Warping，DTW)算法对 LSP 参数进行对齐，然后采用 ANN 算法或是 GMM 算法进行训练映射，建立频谱转换规则。在转换阶段，仍然采用 STRAIGHT 语音分析——合成模型，提取源说话人的频谱参数，并将提取出的频谱参数转换为 LSP 参数，利用前面已经训练好得到的转换规则转换源说话人的 LSP 参数。最后，把转换后的 LSP 参数进行逆向滤波得到频谱包络，并通过 STRAIGHT 模型合成转换后的语音。

4.2 动态时间规整

在进行语音信号特征参数转换时，需要将源说话人和目标说话人的特征矢量时间序列进行训练得到匹配函数。但是，语音信号具有非常大的随机性，即便是同一个人在不同时刻所讲的同一句话、发的同一个音，也不可能具有完全相同的时间长度，在进行特征参数训练匹配时，这些时间长度的变化必然会影响到函数匹配，因此，进行动态时间规整（DTW）就显得尤为必要。

DTW[41-43]是把时间规整和距离测度计算结合起来的一种非线性规整技术。语音转换中使用 DTW 就是将源说话人和目标说话人的特征矢量序列进行时间规整，使得规整后

的每一帧参数描述的都是同一个音节。

设源说话人和目标说话人语音特征矢量序列分别为 $X=\{x_1,x_2,\cdots,x_M\}$ 和 $Y=\{y_1,y_2,\cdots,y_N\}$，$M\neq N$。DTW 算法就是要寻找到一个最佳的时间规整函数，使得源说话人语音的时间轴非线性地映射到目标说话人语音时间轴上，使得总计累计失真量最小。即通过局部优化的方法实现加权距离总和是最小的：

$$D=\min_{f}\sum_{i=1}^{M}[d(x(i),y(f)] \tag{4.1}$$

其中，$f=w(i)$ 为时间规整函数，$d[x(i),y(f)]$ 为第 i 帧源矢量和目标矢量之间的距离测度。现在所要做的就是寻找到这样一个时间规整函数 f 使得 D 值最小，即处于最优时间规整情况下的两矢量之间的匹配路径。这样就可以通过时间规整函数 f 计算出与目标矢量相对应的源矢量参数序列，继而用于语音转换的特征参数训练。

DWT 算法步骤如下：

1、初始化。规整函数起始点为 $(1,1)$，任一点为 (i,j)，终止点为 (I,J)。

2、目标函数。定义 $D(i,j)$ 是 $x(1,i)$ 和 $y(1,j)$ 之间的 DWT 距离，对应的最佳路径由 $(1,1)$ 到 (i,j)。

3、递推关系。$D(i,j)=D(1,1)+\min\{D(i-1,j),D(i-1,j-1)D(i,j-1)$ （4.2）

4、得到最佳路径。$D(m,n)$

图 4.1 是源语音帧数为 350，目标语音帧数为 430 的时间规整函数走势图。

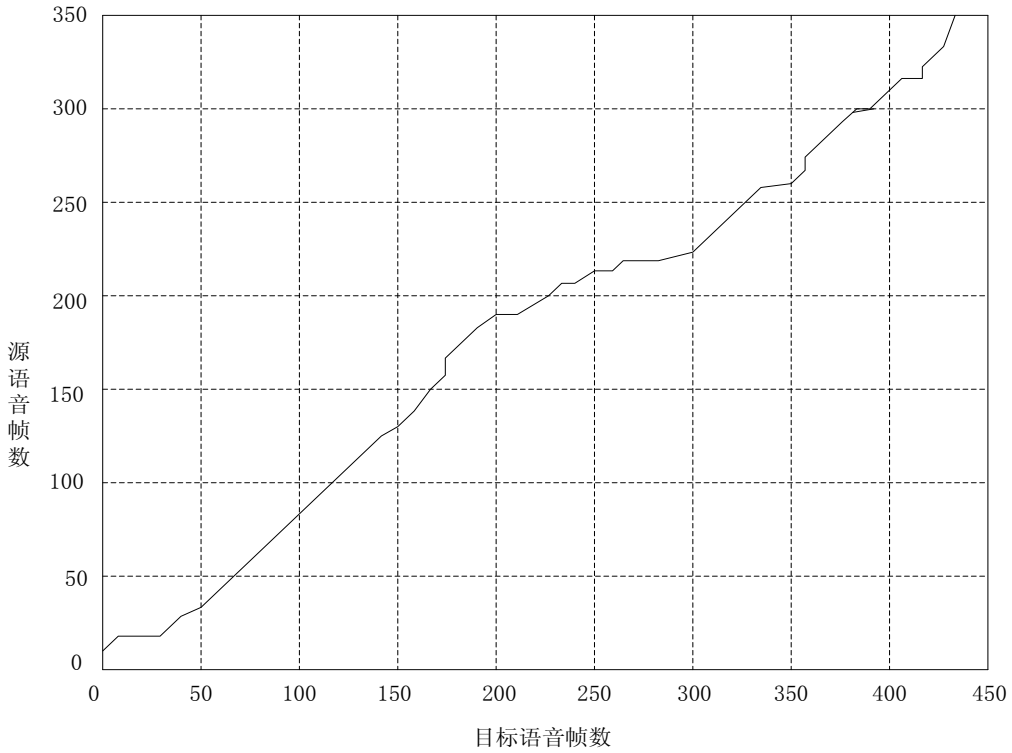


图 4.1 DWT 时间搜索路径

4.3 STRAIGHT 语音分析——合成模型

STRAIGHT（Speech Transformation and Representation using Adaptive Interpolation of weighted spectrum)是一种语音分析合成模型，可以对语音的转换以及重构进行实现，其实现的原理是通过自适应加权谱内插（STRAIGHT）。STRAIGHT 模型是在九十年代末期由 Kawahara 教授提出的语音分析合成系统，主要针对语音参数的修改和恢复。STRAIGHT 对频谱包络进行精确地提取时，需要自适应内插平滑语音信号的短时谱，并且可以将语音信号的频谱及基频参数分解成独立的两部分。STRAIGHT 还能够对语音信号基频参数、时长以及语速参数进行灵活调整，语音参数被调整后依然能够合成出质量效果优良的语音。

4.3.1 STRAIGHT 提取谱包络

STRAIGHT 模型在分析语音信号频谱时先对语音信号进行短时傅里叶变换，然后提取语音信号频谱包络。设源语音信号为 $x(t)$， f_g 为语音由模拟量转换为数字量时的抽样频率。

语音信号短时谱为：

$$F(w,t) = FFT[x(t)w(t)] = X(w)*W(w) \qquad (4.3)$$

令 $h_t(\lambda,\tau)=\frac{1}{4}(1-\left|\frac{\lambda}{\omega_0(t)}\right|)(1-\left|\frac{\tau}{\tau_0(t)}\right|)$

平滑后谱包络为：

$$s(w,t)=[g^{-1}(\int h_t(\lambda,\tau)g(|F(w-\lambda,t-\tau)|)^2 d\lambda d\tau]^{1/2} \quad (4.4)$$

其中，函数 $g(\cdot)$ 定义了插值操作时要保留的何种特性。

4.3.2 STRAIGHT 提取基频轨迹

使用小波变换分析基频，计算出瞬时基频，进行谐波分析。平滑频率轴，得到最终的基频。

将 $x(t)$ 分解为一系列经过滤波处理后的复合信号 $D(t,\tau_0)$，并采用复合 Gabor 滤波器得到的 $g_{NG}(t)$ 作为分析小波，η 为 Gabor 滤波器参数，代表滤波器频率分辨率。由其中关系可推导出：

$$g(t)=e^{-\pi(\frac{t}{\eta}-j\pi t}\quad \eta> \quad (4.5)$$

$$g_{NG}(t)=g\ t(-\ \ 0.2-5g)t+\ (\quad (4.6)$$

$$D(t,\tau_0)=|\tau_0|^{-\frac{1}{2}}\int_{-\infty}^{+\infty}x(t)\overline{g_{NG}(\frac{t-u}{\tau_0})}du \quad (4.7)$$

在实际语音中应用中，“基本性”指数 M 为：

$$M=\lg[\int_{\Omega}(\frac{d|D|}{du}-\frac{1}{\Omega}\int_{\Omega}\frac{d|D|}{du})^2 du]-\lg[\int_{\Omega}(\frac{d\arg D}{du}-\frac{1}{\Omega}\int_{\Omega}\frac{d^2\arg D}{du^2})^2 du]+\lg(\int_{\Omega}|D|^2 du)+\lg\Omega(\tau_0)+2\lg\tau_0 \quad (4.8)$$

M 的最大值对应语音信号的基频成分，此时 τ_0 的取值为计算瞬时频率所用到的值。因此，瞬时频率为：

$$f_0=w_0(t)/2\pi \quad (4.9)$$

其中，$w_0(t)=2f_g\arcsin\frac{|y_d(t)|}{2}$，而 $y_d(t)=\frac{D(t+\Delta t/2,\tau_0)}{|D(t+\Delta t/2,\tau_0)|}-\frac{D(t-\Delta t/2,\tau_0)}{|D(t-\Delta t/2,\tau_0)|}$。

4.3.3 STRAIGHT 合成器实现

用 STRAIGHT 模型进行语音的合成时是将语音信号的基频曲线数值和二维谱包络作为合成系统的输入数据，其中的二维谱包络是由时间轴以及频率轴都平滑后得到的。然后采用 PSOLA 技术与最小相位冲激响应的方法进行合成。另外，在语音合成时还需要对时长、基频和频谱特征参数进行调整。

设 $y(t)$ 为合成后的语音信号，语音合成过程如下：

①设 $A()$、$u(\omega)$、$r(t)$ 分别表示为对平滑后频谱包络 $s(w,t)$ 在幅度、频率和时间轴上的调整。

$$c_t(q)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-j\omega q}\ \ln A\ S(u\ (\omega\ (r\ t),u(\omega))r,t\ (d\omega) \tag{4.10}$$

②采用倒谱转换方法，将一般相位谱转化为最小相位。

$$h_t(q)=\begin{cases}0 & q<0\\ c_t(0) & q=0\\ c_t(q) & q>0\end{cases} \tag{4.11}$$

③对最小相位冲击响应进行傅里叶变换。

$$V(\omega t)=\exp\left(\frac{1}{\sqrt{2\pi}}\int_0^{\infty}h_t\ q\ e^{j\omega q}dq\right) \tag{4.12}$$

④求取每一帧语音应的冲激响应。

$$u_{ti}(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}V(\omega,t)\phi(\omega)e^{j\omega(t)}d\omega \tag{4.13}$$

其中，$\phi(\omega)$ 表示具有附加的控制相位的激励，主要目的是用来改善听觉。

⑤确定基音同步位置。

$$T(t_i)=\sum_{t_k\in Q,k<i}\frac{1}{G(f_0(t_k))} \tag{4.14}$$

其中，Q 表示用于语音合成的基音同步位置的集合函数，$G(\cdot)$ 表示基频的调整函数，可以是任意形式的映射关系。

⑥基音同步叠加。

$$y(t)=\sum_{t_i}\frac{1}{\sqrt{G(f_0(t_i))}}u_{ti}(t-T(t_i)) \tag{4.15}$$

这整个过程就是语音合成步骤，$y(t)$ 即为合成后的语音信号。

STRAIGHT 模型提取的频谱包络并不能直接用来进行参数的建模，需要对频谱包络进行特征参数化，然后再对参数进行建模。选择一种合适的频谱包络参数化方法是非常重要的，因为它将直接影响到转换语音的合成的效果。

4.4 基于 ANN 的语音转换算法

人工神经网络（Artificial Neural Networks，简记作 ANN）[45]，是对人类大脑系统的一阶特性的一种描述，由大量简单的处理单元即人工神经元广泛地相互连接而组成的一个并行处理网络系统。尽管神经元的结构和功能是非常简单的，但由其构成的网络系统对知识的存储方式是分布式的，使得神经网络具有很强的自组织和自学习能力以及很高

的容错力和顽健性。神经元、训练及学习算法和网络的连接方式这三个基本要素构成了神经网络[23]。运用 ANN 进行语音转换主要就是利用 ANN 的自学习能力对源说话人和目标说话人的语音个性特征参数进行训练，以期得到它们之间的匹配函数。对于利用 ANN 进行语音的转换，前人也做过不少的研究。

学者 NarendraIlath[4]提出了使用神经网络对说话人语音进行转换的算法。他将神经网络分为 4 层结构，其中 3 个输入单元、2 个隐含层以及 3 个输出单元。输入单元的输入为源说话人语音特征参数中的前 3 个共振峰，输出单元的输出则为目标说话人语音特征参数的前 2 个共振峰，而中间的隐含层使用 8 个神经元，利用误差反向传播（BP）算法对其训练。利用训练得到的转换函数进行语音转换，然后合成通过转换得到的共振峰频率与平均基音频率，最终得到转换语音。

学者 Baukoin 也是利用 BP 神经网络对语音信号进行转换。在他的实验中，神经网络的隐含层选择的有所不同，选择了分别含有 2 个隐含层和 3 个隐含层的二种类型神经网络。在含 2 个隐含层的网络中，每个隐含层包含 15 个神经元，而在含 3 个隐含层的网络中，每个隐含层包含 12 个神经元。他选择倒谱参数作为训练的特征参数。他将实验分为训练和转换两个阶段，其步骤为：

训练阶段：先用均值与协方差归一化处理源说话人语音信号和目标说话人语音信号的频谱参数，然后对其分类，再将进行过动态时间调整的源说话人特征参数和目标说话人特征参数分别作为训练网络的输入输出序列。在这一个阶段，找到使输入输出序列具有最小平均距离的路径，这个路径就是输入和输出之间的一种最优联系，也即源说话人特征参数和目标说话人特征参数的匹配函数。

转换阶段：先归一化处理源特征矢量，然后进行归类，再用神经网络转换归类后的源特征矢量，再解归一化处理转换出的特征矢量。

国内，左国玉[21,22]提出了采用线谱对（LSP）特征参数为输入和利用遗传算法进行训练的径向基函数（RBF）网络的语音转换方法，这种算法使转换效果得到很大的改善，系统稳定性也得到很大提高。

4.4.1 RBF 网络结构

径向基函数网络（Radial Basis Function Network，简称 RBF 网络）是一种有关函数映射的人工神经网络，其函数映射的执行采用的是局部接受域。RBF 网络模型[45]具有两种模型：正规化网络和广义网络，本文在进行语音转换的实现时选用的是广义网络。广义径向基函数神经网络是一种具有单隐层的 3 层前馈网络，其 3 层分别为输入层、隐含层和输出层。信号源的节点为输入层的输入，隐含层中有多个隐单元，其变换函数是非线性函数，这个非线性函数的中心点是径向对称的，并且会产生衰减，输出层自然就是

对输入做出响应。

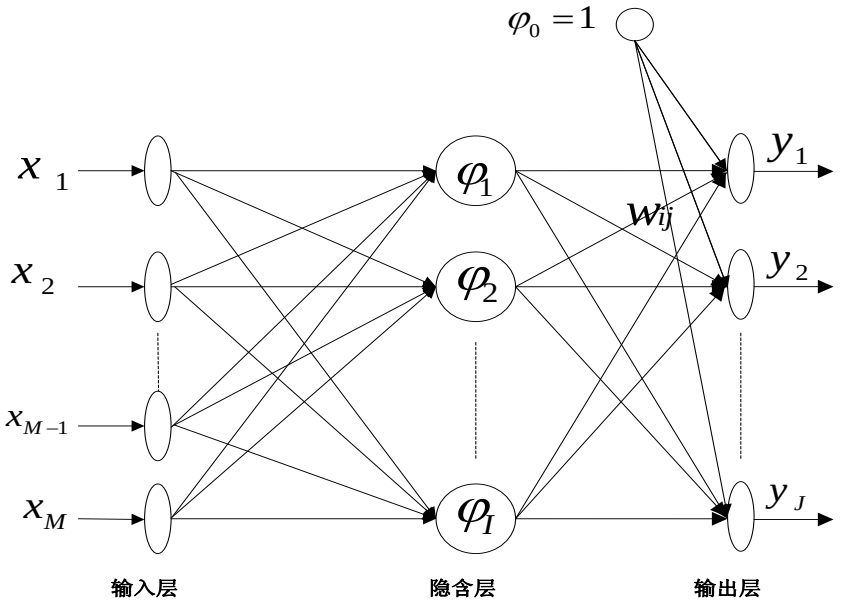


图 4.2　RBF 网络的拓扑结构

RBF 网络的工作原理[46]是：隐单元具有径向基函数，作为隐含层的空间，输入层输入信号，并传递给隐含层，对输入矢量作非线性变换，使低维输入变换到高维空间，这样就可以将非线性问题线性化 ，也就是隐含层和输出层之间存在一个线性关系。图 4.2 为 RBF 网络拓扑结构。

从 RBF 网络的拓扑结构中可以分析出：设定的输入层的神经元有 M 个，隐含层的神经元有 I 个，输出层的神经元有 J 个。隐含层和输出层之间用 $w_{ij}(i=1,2,\cdots,I;\ j=1,2,\cdots,J)$ 作为其连接权值，另还有一个阈值 φ_0 设置在输出层。因为隐含层的基函数要求选择其中心点是径向对称的，而高斯函数是中心对称函数，所以通常选择高斯函数作为 RBF 网络隐含层的基函数[46]。

假设有样本 X 作为 RBF 网络的训练样本集，$X=[X_1,X_2,\cdots,X_L]^T$，$x_k=[x_{k1},x_{k2},\cdots,x_{kM}]$，$k=1,2,\cdots L$ 为训练样本集中的任一样本。

隐含层的高斯函数可以表示如下：

$$\varphi(X_k,c_i)=\exp\left(\frac{1}{2\sigma_i^2}\left\|X_k-c_i\right\|^2\right) \tag{4.16}$$

其中，$\varphi(X_k,c_i)$ 表示第 i 个隐含层的输出，X_k 为第 k 个输入样本，c_i 为第 i 个隐含层的高斯函数的中心，c_i 为第 i 个隐含层变量，也称标准化常数或者是基宽度。

在 RBF 网络[46]中，隐含层的每一个神经元节点都有一个径向基函数，而每一个径向基函数都有一个中心向量 c_i，并且这个中心向量的维数和输入样本的维数保持一致，$c_i=[c_{i1},c_{i2},\cdots,c_{im}]^T, m=1,2,\cdots M$ 整个 RBF 网络有 M 个这种中心向量。因为隐含层神经节

点的净输入为每一训练样本与该节点的基函数中心向量 c_i 之间的欧几里得范数 $\|X-c_i\|_2$，也就是：

$$\delta i=\|X-c_i\|_2=\sqrt{\sum_{m=1}^{M}(x_m-c_{mi})^2} \tag{4.17}$$

隐含层神经元节点的输出为隐含层径向基函数对隐含层输入的非线性转换。所以有如下隐含层输出公式：

$$\varphi(X_k,c_i)=G(\|X_k-c_i\|)=\exp(-\frac{1}{2\sigma_i^2}\|X_k-c_i\|^2)=\exp(-\frac{1}{2\sigma_i^2}\sum_{m=1}^{M}(x_{km}-c_{im})^2) \tag{4.18}$$

$$\varphi(X_k,c_i)=\exp(\frac{1}{2\sigma_i^2}\|X_k-c_i\|^2 \tag{4.19}$$

其中，σ_i 为径向基函数的方差。

在 RBF 网络中，每一个隐含层的输出实际上表达的是输入样本 X 离开这个神经元节点的中心的程度。并且 RBF 网络中并没有一个隐含全矩阵来连接输入与输出，所以，RBF 网络隐含层训练学习的是基函数的中心向量。

RBF 网络的输出层为一组线性组合器，所以隐含层和输出层之间是线性的关系，不像输入层和隐含层之间是非线性的。也就是说输入层实现的是 $X_k\to\varphi(x,c_i)$ 的非线性映射，而输出层实现的是 $\varphi(x,c_i)\to y_k$ 的线性映射，因此可知 RBF 网络的输出有如下公式：

$$y_k(X_k)=w_{j0}+\sum_{i=1}^{I}w_{ji}\varphi_i(X_k,c_i),j=1,2, \tag{4.20}$$

由以上推导可知，RBF 网络对输入样本和输出之间可以建立如上公式 4 的关系式，我们将 RBF 用于语音转换时，这个关系式就可以认为是训练得到的匹配函数。分析此式子，可以知道有两类参数是待定的，一类为基函数中心 c_i 与宽度 σ_i，另一类是隐含层和输出层这两者之间的的联结权值 w_{ij}。一旦将确定出基函数中心 c_i 与宽度 σ_i，那就可以用公式 4 知道 RBF 的输出了，也就可以求解出来了，这样又联结权值 w_{ij} 是线性的，所以可以通过 PSO 算法求解出来。

4.4.2 RBF 网络隐含层学习算法——SC 算法

RBF 网络隐含层的训练是为了得到隐含层径向基函数的中心和方差。对于基函数的中心，我们采用 SC 算法去求；在确定基函数中心后，就可以确定出基函数的方差值。

SC（减法聚类）算法[49,50]是一种相对有效的确定基函数中心个数的聚类算法。考虑把数据归一化放到一个单位超立体中的 n 维空间的 p 个数据点 $(X_1,X_2,\cdots,Xp)$，首先由公式(4.21) 列出数据点 Xi 所在处的密度指标：

$$D_i = \sum_{j=1}^{p} \exp(-\frac{\|x_i - x_j\|^2}{(\gamma_a / 2)^2}) \tag{4.21}$$

其中 γ_a 是一个正数，假如一个数据点附近有多个相邻的数据点，则此数据点会有较高的密度值。半径 $D_{\max} < 0.15 D_{c1}$ 定义了这个点的一个邻域，半径之外的数据点对这个点的密度指标影响相对来说小一点。我们可以明显发现，假如一个数据点附近有多个相邻的相关数据点，那么此点会具有相对较高的密度值。通过计算出全部数据点的密度指标之后，选取具有最高密度指标的数据点作为第一个聚类的中心，令 x_{ci} 为该点， D_{c1} 为该点密度指标，则每一个数据点 x_i 的密度指标可以用公式（4.22）进行修正。

$$D_i = D_i - D_{c1} \exp(-\frac{\|x_i - x_{c1}\|^2}{(\gamma_b / 2)^2}) \tag{4.22}$$

从(4.22)式可以看出常数 γ_b 定义了一个领域，该领域的密度指标呈显著减少的趋势，一般选 $\gamma_b = 1.5\gamma_a$ ，这是为了防止出现聚类中心相距过近的情况。换句话说，降低了它四周训练数据被选作为隐含层节点中心的概率。由于距离第一个聚类中心 x_{ci} 较近的数据点的密度指标显著减小，当 $D_i \le 0$ 时，此点 x_i 将不可能变为聚类中心点。通过公式(4.22)修订全部数据点的密度指标后，得出下一个聚类中心，多次重复这个过程，利用顺序的削去已经存在的密度函数值来选取新的聚类中心，当新聚类中心相应的密度指标 $D_{\max}$ 与初始最高密度值 D_{c1} 相比满足公式(4.23)时则聚类过程结束。

$$D_{\max}/D_c \prec \lambda \tag{4.23}$$

通常设定 $D_{\max} < 0.15 D_{c1}$ 作为循环终止的条件。

1、确定基函数中心 x_{ci} [45]

把 SC 算法引到基函数中心的学习后，其学习算法为：令 $\{x_1, x_2, \cdots, x_n\} \in R^M$ 是一个训练样本集，而 M 是其输入样本空间的维数，本文中的样本空间 LSF 参数是 16 维的；n 是样本的个数。

第一步：计算参数 γ_a ， γ_b 。

$$\gamma_a = \frac{1}{2} \min_k (\max_i \|x_i - x_k\|) \tag{4.24}$$

$$\gamma_b = 1.25\gamma_a \tag{4.25}$$

第二步：对每个输入训练样本 x_i ，计算每个样本的初始密度指标：

$$D_i = \sum_{j=1}^{n} \exp(-\frac{\|x_i - x_j\|^2}{(\gamma_a / 2)^2}) \tag{4.26}$$

第三步：选择密度指标最大的一个样本点，并把该点作为第一个聚类中心 x_{c1} 。

第四步：假定 x_{ck} 为第 k 次选取出来的聚类中心，该中心的密度指标为 D_{ck} ，利用公式（4.27）对每一个数据点的密度指标进行修正，选取密度指标最高的样本点 x_{ck+1} 作为

一个新的聚类中心。

$$D_i = D_i - D_{c1}\exp(-\frac{\|x_i - x_{c1}\|^2}{(\gamma_b/2)^2}) \tag{4.27}$$

第五步：判断 $D_{\max}/D_{c1} < \lambda$ 是不是成立的。如果不等式成立就要退出来结束掉，如果不成立就要转到第三步；当然其中 λ 是事先已经给定好的参数，而最后生成的聚类中心数目便由该参数决定，λ 的大小与聚类数目成反比。这里取 $\lambda = 0.15$。

2、确定基函数方差 σ

神经网络隐含层中基函数的方差能够在求解出各个基函数的中心后通过下面的公式得到：

$$\sigma_1 = \sigma_2 = \cdots = \sigma_I = \frac{d_{\max}}{\sqrt{2I}} \tag{4.28}$$

其中 I 为聚类中心点的个数，是使用 SC 算法训练得到，$d_{\max}$ 则为每个聚类中心相互间距离的最大值。

4.4.3 RBF 网络输出层学习算法——PSO 算法

PSO（粒子群优化)算法是一种基于迭代的优化算法，通过粒子追随问题空间中最优粒子来进行问题空间搜索。在粒子群算法中，所有的粒子都有一个被优化的函数决定的适值，粒子在搜索空间中的位置是每个优化问题的解，同时，粒子的速度决定了粒子飞翔的距离和方向。种群规模也即粒子个数，若粒子群体规模为 m,假设在一个 d 维（由优化问题所决定）的空间中搜索，则第 i（i=1,2,…,m）个粒子在搜索空间中的位置可以表示为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{id})$，第 i 粒子的速度可以表示为 $V_i = (v_{i1}, v_{i2}, \cdots, v_{id})$。PSO 初始化为一群随机粒子，然后粒子根据两个极值来动态调整自己的位置和飞行速度，第一个极值是粒子本身找到的最优值即个体极值 p_b；第二个极值是整个种群目前找到的最优解即整体极值 g_b。找到两个最优值，粒子对其速度和位置的更新可以通过下面式子来计算：

$$v_i^{k+1} = \omega v_i^k + c_1 r_1 (P_b - x_i^k) + c_2 r_2 (g_b - x_i^k) \tag{4.29}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{4.30}$$

式中 v_i^{k+1} 表示第 i 个粒子的第 $k+1$ 次飞行速度；v_i^k 表示第 i 个粒子的第 k 次飞行速度；x_i^{k+1} 则表示第 i 个粒子在第 k+1 次飞行中的位置，x_i^k 是第 i 个粒子在第 k 次飞行中的位置；c1、c2 为学习因子，通常令 c1=c2，并且范围在 0 和 4 之间；r1 和 r2 是介于 0 到 1 之间的随机数；还有就是 ω 是惯性权重系数，粒子的当前速度受历史速度影响程度就由 ω 控制着，会影响算法的收敛性，通常设置在[0.4,1.2]之间，ω 的迭代过程表示为：

$$\omega = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{k_{\max}} \cdot k \tag{4.31}$$

其中，k 代表当前迭代次数，$k_{\max}$代表最大迭代次数。ω较大表示全局收敛能力越强，局部收敛能力弱；ω较小，则局部收敛能力强，全局收敛能力弱。

粒子在更新自己位置时，需要受到最大速度$v_{\max}$和最小速度$v_{\min}$的限制。当$v_i^{k+1} > v_{\max}$时，则设定为$v_i^{k+1} = v_{\max}$；当$v_i^{k+1} < v_{\min}$时，则设定为$v_i^{k+1} = v_{\min}$。$v_{\max}$的选择受到粒子宽度范围的限制，$v_{\max}$设置比较大时，才能够保证粒子种群具有全局搜索能力，而$v_{\max}$设置比较小时，就加强了种群在局部的搜索能力。

迭代过程重复进行，其终止条件有两种，一是达到了设置的最大迭代次数，二是粒子在解空间中相对静止。

公式(4.29)得出粒子 i 新的速度主要通过三部分来计算：粒子 i 前一时刻的速度、粒子 i 当前位置与群体最佳位置之间的间距、粒子 i 当前位置与本身最佳位置之间的间距。粒子 i 新位置的坐标可以通过公式(4.30)计算得来。粒子下一步运动的位置通过公式(4.29)、(4.30)来决定。下图 4.3 以两维空间为范例，并根据公式(4.29)、(4.30)，描述了粒子从位置移动的原理。

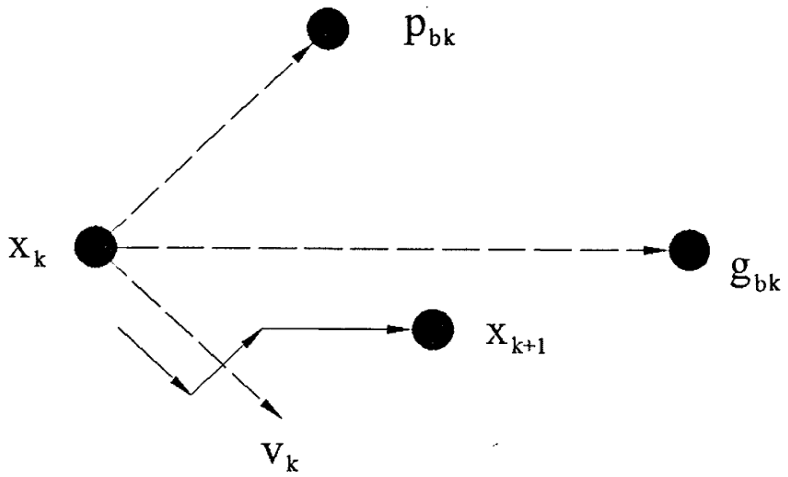


图 4.3 粒子移动示意图

在前面，对 RBF 网络选择是考虑用 SC 算法来进行的，也就是通过对训练样本采用 SC 算法聚类，得到聚类中心个数。若获得 I 个聚类中心个数，此时就可以选择 16-I-16 的 RBF 网络对其进行频谱包络转换。同时将各个聚类中心当作为 RBF 网络中隐含层的基函数中心，通过公式（4.32）求出各径向基函数的方差，

$$\sigma_1 = \sigma_2 = \cdots = \sigma_I = \frac{d_{\max}}{\sqrt{2I}} \tag{4.32}$$

然后就能够通过公式（4.33）将各隐含层节点的输出求解出来，

$$\varphi(X_k, c_i) = G(\|X_k - c_i\|) = \exp(-\frac{1}{2\sigma_i^2}\|X_k - c_i\|) = \exp(-\frac{1}{2\sigma_i^2}\sum_{m=1}^{M}(x_{km} - c_{im})^2) \tag{4.33}$$

再把隐含层节点的输出当作输出层的输入，也就是 PSO 算法采用的是隐含层节点的

输出作为其输入，同时将目标说话人的 LSP 参数作为目标输出。通过把 PSO 算法放到输出层进行训练，可以用如下步骤表示出 RBF 网络输出层的训练算法：

第一步：将期望响应设置为：

$d=[d_1,d_2,\cdots,d_J]^T$，本文中选择 J=16。

第二步：假定已知通过 SC 算法后得到的聚类中心为 I，再通过下面公式（4.34）计算得出每个隐含层节点的输出响应：

$$\varphi_i(x)=\exp(\frac{-\|x-x_{ci}\|^2}{2\delta_i^2},(i=1,2,I) \tag{4.34}$$

第三步：将每一个粒子的速度、位置和粒子个数 m 进行初始化。$t_{\max}$ 为其最大循环的迭代次数。

第四步：再将个体极值 p_{bf} 和全局极值 g_{bf} 以及位置的全局极值 g_b 和个体极值 p_b 进行初始化。

第五步：通过公式（4.35）计算得出输出层各神经元响应：

$$y_{kj}(X_k)=w_{0j}+\sum_{i=1}^{I}w_{ij}\varphi(X_k,c_i),(j=1,2,\cdots,16),(k=1,2,\cdots,K) \tag{4.35}$$

第六步：计算出误差向量：

$$e_k=[e_1,e_2,\cdots,e_J]=[(y_1-d_1),(y_2-d_2),\cdots,(y_J-d_J)] \tag{4.36}$$

第七步：计算得出粒子适应度：

$$MSE=\frac{1}{K}\sum_{k=1}^{K}e_k^2=\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{16}(y_{ki}-d_{kj})^2 \tag{4.37}$$

第八步：输入其它粒子，并带回到第五步进行迭代，直到计算出每一个粒子的适应度。

第九步：通过比较得到 p_b、p_{bf}、g_b、g_{bf}。

第十步：更新所有粒子的速度以及位置。

根据公式(4.29)、公式(4.30)更新粒子的位置以及速度，同时考虑更新后的位置和速度是否在设置的范围内。

考虑速度：若 $v_{ij}(t+1)>v_{\max}$，则 $v_{ij}(t+1)=v_{\max}$；

若 $v_{ij}(t+1)<-v_{\max}$，则 $v_{ij}(t+1)=-v_{\max}$；

否则 $v_{ij}(t+1)$ 不变。

考虑位置：若 $x_{ij}(t+1)>x_{\max}$，则 $x_{ij}(t+1)=x_{\max}$；

若 $x_{ij}(t+1)<x_{\min}$，则 $x_{ij}(t+1)=x_{\min}$；

否则 $x_{ij}(t+1)$ 不变。

其中，$x_{\min}$、$x_{\max}$、$v_{\max}$ 分别为最小位置、最大位置和最大速度。

第十一步：比较出迭代次数是不是满足最大迭代次数或者是预设的精度，若满足，则此时迭代结束。网络的最佳权值便是迭代得到的全局最优解值 g_b。

4.4.4 基于改进的 RBF 网络谱包络转换

根据本章第一节语音转换流程进行转换，在训练阶段，先对选取的语音特征参数 LSP 参数进行 DTW 对齐后，然后将源说话人的参数序列作为神经网络的输入，目标说话人的参数序列作为神经网络的输出，进行 RBF 神经网络训练，其中对网络隐含层采用 SC 算法训练，对其输出层用 PSO 算法进行训练。整个训练结束得到输入与输出之间的匹配函数。在转换阶段，将源说话人的 LSP 参数作为 RBF 网络的输入，通过匹配函数做映射，最后的得到转换后的谱包络。图 4.4 为 RBF 网络频谱包络转换。

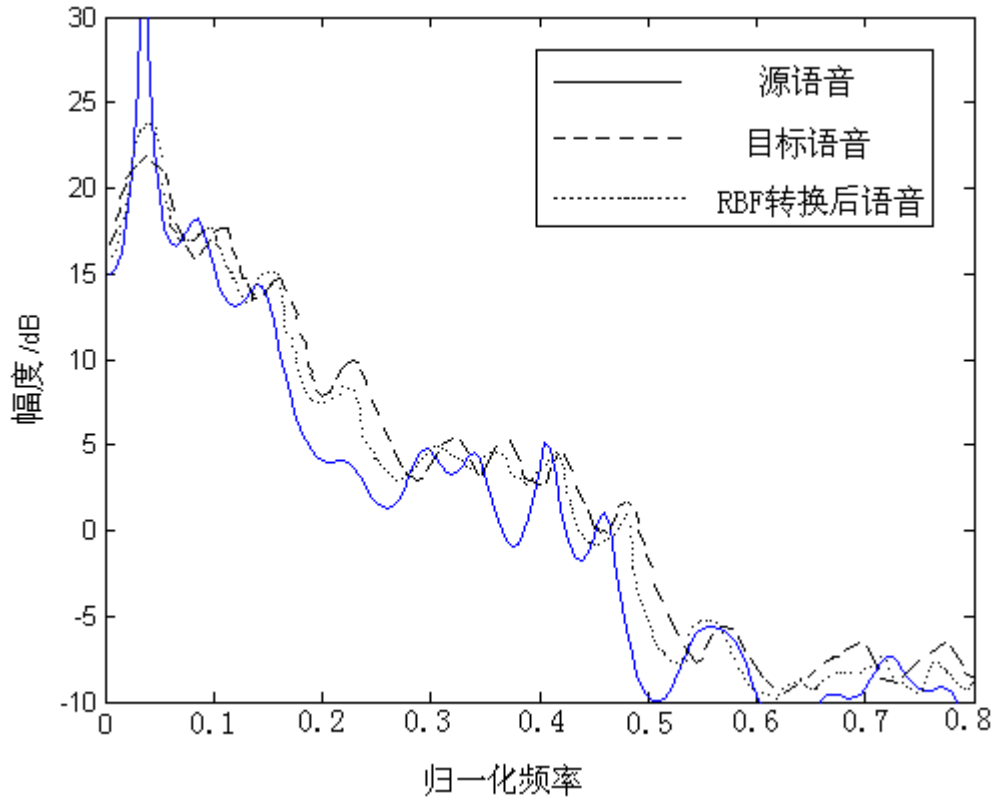


图 4.4 RBF 网络频谱包络转换

4.5 基于 GMM 的语音转换算法

现在常用到的语音转换算法之一还有高斯混合模型（GMM）算法。最早引入 GMM 算法的是 Stylianou、Kain 等，通过加权求取平均的方法解决了矢量算法中参数离散化的特点。Stylianou 对源特征参数和目标特征参数分别进行建模，采用最小二乘法估计未知矢量；Kain 将源特征参数和目标特征参数进行联合概率密度建模，并采用联合概率估计求解未知矢量。本文通过对基本 GMM 算法的改进，解决由于对特征矢量加权求平均引

起的参数过平滑问题。

4.5.1 GMM 建模

前面讲到对语音信号的转换分为训练和转换两个部分，在进行特征参数的转换之前需要将源说话人和目标说话人的特征矢量时间序列进行动态时间规整，使得规整后的每一帧参数描述的都是同一个音节。在对特征参数进行 DTW 动态时间规整后，下一步就是对规整后的特征参数进行 GMM 建模训练和转换。

GMM 模型是具有多个高斯混合分布函数的状态模型，将进行 DTW 规整后的语音特征参数矢量集的概率分布进行拟合。设源语音参数矢量表示为 X，目标语音参数矢量为 Y，构成一个联合矢量 Z，$Z=[XY]'$，由 Z 建立高斯模型。一个 M 阶的高斯混合模型的概率密度函数表示如下式：

$$P(z_n)=\sum_{m=1}^{M}\alpha_m N(z_n,\mu_m,\sum_m) \quad \sum_m^M(\alpha_m=\alpha_m\pm1, \tag{4.38}$$

其中，z_n 是任一样本，m 为单个高斯函数的个数，也称为混合度。α_m 是混合权重，即第 m 个高斯分布在整体高斯混合分布中权重值。μ_m 和 $\sum_m$ 为第 m 个高斯函数的均值和协方差。$N(z_n;\mu_m,\sum_m)$ 代表 d 维的正态分布，其数学表达式为：

$$N(z_n;\mu_m,\sum_m)=\frac{1}{(2\pi)^{d/2}\left|\sum_m\right|^{1/2}}\exp\left[-\frac{1}{2}(z_n-\mu_m)'\sum_m^{-1}(z_n-\mu_m)\right] \tag{4.39}$$

简单来说，混合高斯模型可用参数 ρ 简化表示为：

$$\rho=(m,\alpha,\mu\sum \tag{4.40}$$

4.5.2 GMM 模型训练

GMM 模型训练的结果就是要求出模型参数 ρ，这里采用最大似然估计中的期望最大（EM）迭代算法[54]。

上述建出的 GMM 模型最终的似然函数可以表示为 N 个 $P(z_n)$ 相乘：

$$P(Z/\rho)=\prod_{n=1}^{N}\sum_{m=1}^{M}\alpha_m p(z\ \mu_m\sum_m \tag{4.41}$$

EM 算法就是通过特征参数矢量的增加迭代使这个似然函数达到最大，进而求得最大时所对应的模型参数 ρ。采用 EM 算法对 GMM 模型的参数估计是从参数 ρ 的初始值开始，运用 EM 算法得到一个新的模型参数 $\hat{\rho}$，使得在新的模型参数下的似然度 $P(Z/\hat{\rho})>P(Z/\rho)$。新的模型参数 $\hat{\rho}$ 再作为新的初始值反复迭代，直到模型收敛。

混合权重的重估迭代公式如下：

$$\alpha_m = \frac{1}{N}\sum_{n=1}^{N} \log P(m \mid z_n, \rho) \tag{4.42}$$

混合均值的重估迭代公式如下：

$$\mu_m = \frac{\sum_{n=1}^{N} \log P(M \mid z_n, \rho, z_n)}{\sum_{n=1}^{N} \log P(M \mid z_n, \rho,} \tag{4.43}$$

混合协方差的重估迭代公式如下：

$$\sum_m = \frac{\sum_{n=1}^{N} \log P(M \mid z_n, \rho)(z_n - \mu_m)^2}{\log P(M \mid z_n, \rho)} \tag{4.44}$$

$P(m / z_{n,\rho})$ 表示第 n 帧特征矢量属于第 m 类的后置概率，可由贝叶斯准则计算得到：

$$P(m / z_{n,\rho}) = \frac{\alpha_m N(z_n; \mu_m, \sum_m)}{\sum_{m=1}^{M} \alpha_m N(z_n; \mu_m, \sum_m)} \tag{4.45}$$

4.5.3 GMM 模型的转换

把 GMM 看作一个分类器，对特征参数矢量进行分类，将每个高斯成分看作为一类。并没有把每一个特征矢量唯一的分给其中某一个高斯成分，而是可以将每一个特征矢量分给若干个不同的高斯成分，只是分给每一个高斯成分的概率不一样。在每一单个高斯分量里对源特征矢量和目标特征矢量之间建立起线性关系，即为转换函数，采用使时间对应的源和目标特征参数矢量间距离最小的方法估计转换函数。可以得到转换函数为：

$$\hat{y} = F(z_n) = \sum_{m=1}^{M} P(m / z_{n,\rho})[\mu_m^Y + z_n \sum_m^{YX} \sum_m^{XX-1} - \mu_m^X \sum_m^{YX} \sum_m^{XX-1}] \tag{4.46}$$

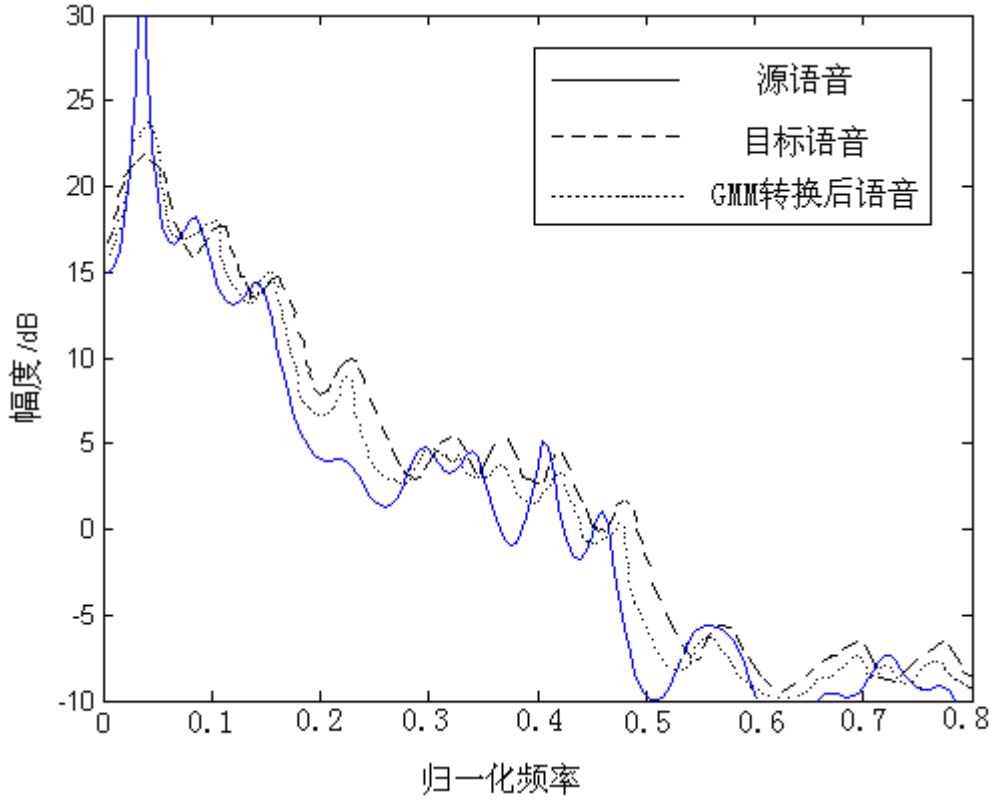
对上式分析可以发现，转换函数主要由均值项 $\mu_m^Y \sum_{m=1}^{M} P(m / z_{n,\rho})$ 和相关项 $\sum_{m=1}^{M} P(m / z_{n,\rho})[z_n \sum_m^{YX} \sum_m^{XX-1} - \mu_m^X \sum_m^{YX} \sum_m^{XX-1}]$ 两部分组成。因为 μ_m^Y 代表每一单个高斯分布的混合均值，当均值和相关项标准差同时增加各自相同比例时，均值项所造成的整个转换特征标准差的增加量比相关项所造成的整个转换特征标准差的增加量要大些，而且随着变化量的增大，差距逐渐增大。所以转换后语音信号频谱离散性会变差，转换特征比较集中，即出现过平滑现象，因此 μ_m^Y 是造成转换后特征参数过平滑现象的主要原因。为了防止过平滑现象的发生，可以考虑转换语音的协方差。我们在特征参数的转换中可

以保持足够大的协方差以制约混合均值的影响。但有一个现实问题是，源说话人语音特征参数的协方差很难得和目标说话人语音特征参数的协方差进行匹配，考虑到这一点，我们可以假设转换语音和目标语音有着相同的协方差，因此我们可以对原来的转换函数加以改进，得到新的转换函数为：

$$\hat{y} = F(z_n) = z_n + \sum_{m=1}^{M} P(m / z_{n,\rho})(\mu_m^Y - \mu_m^X) \tag{4.47}$$

对训练阶段中的混合均值矢量 μ_m 和混合协方差 $\sum_m$ 进行因式分解，可以得到上式转换函数中的 μ_m^X， μ_m^Y， $\sum_m^{XX}$ 和 $\sum_m^{YX}$ ：

$$\mu_m = \begin{bmatrix} \mu_m^X \\ \mu_m^Y \end{bmatrix}, \quad \sum_m = \begin{bmatrix} \sum_m^{XX} & \sum_m^{xx} \\ \sum_m^{YX} & \sum_m^{YY} \end{bmatrix} \tag{4.48}$$



4.5 GMM 频谱转换

这样得到转换函数后，就可以利用转换函数将源说话人语音的特征参数矢量进行转换，然后再利用我们前面介绍过的 STRAIGHT 模型合成算法，将转换所得特征参数合成目标语音，从而最终完成语音的转换。图 4.5 为 GMM 频谱转换后的语音波形，图中的转换波形相较于源语音波形更接近于目标语音波形，说明转换效果良好。

4.6 语音合成

由于清音帧属于高频信号，而人耳对低频信号比较敏感，清音帧对语音的贡献比较小，所以在语音合成时，对清音帧采用直接复制的方式，只对浊音帧进行转换合成，如果转换后相邻语音帧的 LSP 系数差距变大，可以需要对 LSP 进行平滑处理，所用的平滑方法与基音提取的平滑处理方法相同。用 STRAIGHT 模型对语音进行合成，实验表明这种方法是有效的，图 4.6 表示源语音的波形图，图 4.7 表示目标语音的波形，图 4.8 表示使用 RBF 网络转换方法得到的转换语音波形，图 4.9 表示使用 GMM 模型转换方法得到的转换语音的波形。从图 4.8 和图 4.9 中可以看出两种转换方法得到的转换语音从波形上都比较接近于目标语音，各有优劣，还需进行改进，以期得到更好的转换效果。

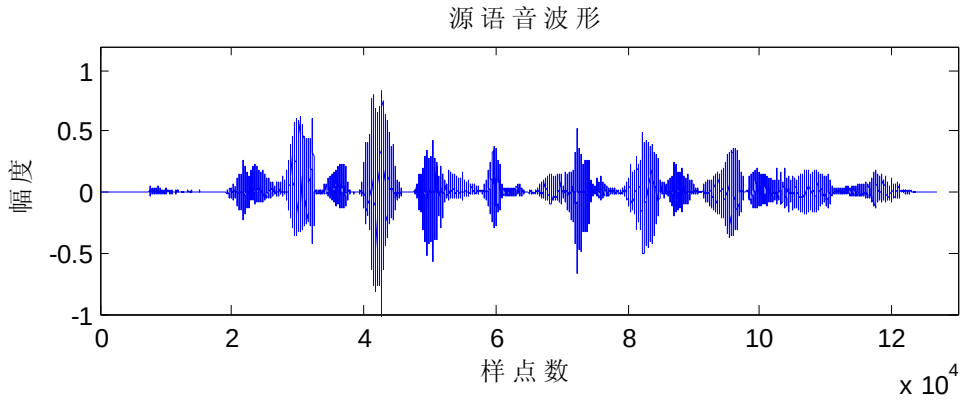


图 4.6 源语音时域波形

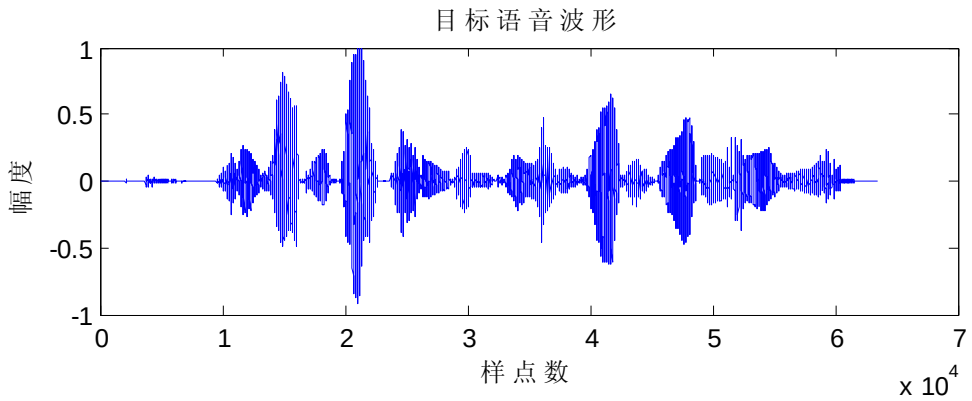


图 4.7 目标语音时域波形

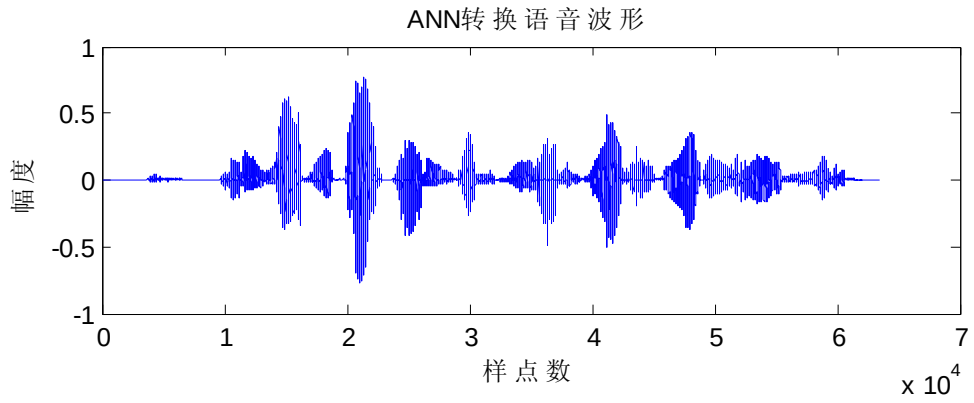


图 4.8 使用 RBF 转换方法得到的转换语音波形

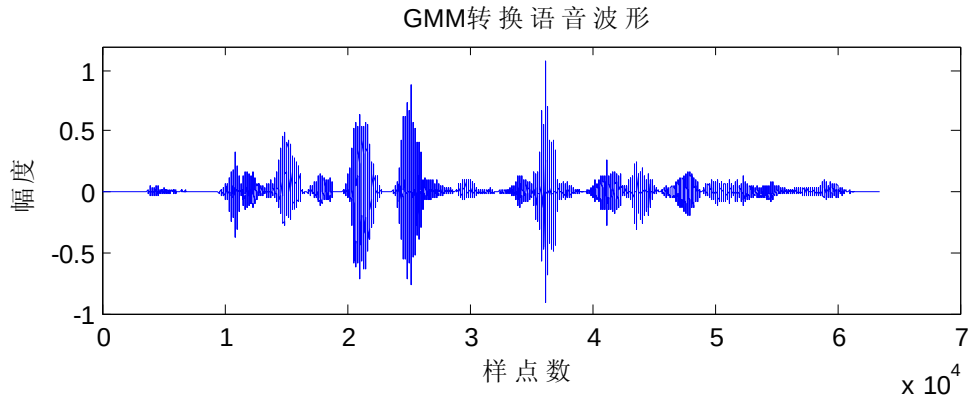


图 4.9 使用 ANN 转换方法得到的转换语音波形

4.7 本章小结

本章实现了基于改进的 ANN 模型和改进的 GMM 模型的语音转换，详细介绍了其实现过程，并对两种算法的转换效果作了比对。

5 总结与展望

5.1 总结

本文研究的主要对象是说话人语音转换技术。因为语音转换的重要意义，不断有人投入到语音转换技术的研究中，到现今，已经有不少的语音转换算法。本文研究的目的就是在现今算法的基础上，再加以研究与优化，意图找到一种更加有效的转换算法，从而达到转换的目的，使转换后的语音特征听起来与目标说话人的语音特征更为接近，并且使转换后的语音质量更为理想。

论文针对语音转换的两个重要阶段——训练和转换阶段进行了研究，分析这两个阶段，将重点放在语音信号的分析与合成、特征参数和匹配函数这三个点上。对于语音信号的分析及信号的合成，采用 STRAIGHT 语音分析合成模型。STRAIGHT 语音分析合成模型相较于其他的语音合成模型最大的优势是能够将语音信号分解为相互独立的频谱参数和基频参数，并能灵活调整语音信号的基频、时长、语速等参数，调整语音参数后仍能合成出高质量的语音。在特征参数的选择上，根据语音信号的产生机理，对语音信号进行了不同的分析，并介绍了不同特征参数的提取方法，选择了频谱包络和基因周期作为转换的重要特征参数。由于 LSP 参数非常好的量化特性和插值特性，作为频谱包络的重要特征进入到语音转换的训练和转换中。对于源语音信号和目标语音信号的匹配函数，本文采用了改进的 ANN 模型算法和 GMM 模型算法进行训练与转换。选择 ANN 中的 RBF 网络进行训练，并对 RBF 网络隐含层的学习采用 SC 算法代替原来的 K-均值聚类算法，对输出层的学习采用 PSO 算法代替原来的 LMS 算法，最终得到匹配函数，即转换规则。在 GMM 模型算法中主要是针对转换语音的过平滑现象进行协方差改进得到新的转换函数。最后将转换出的特征参数进行合成得到转换语音。另外，本文的转换系统均是在 MATLAB 平台上实现的。

由于本人主观方面专业知识有限，及一些客观原因，本文的系统实现还有存在着一些问题，致使转换效果不佳。虽然在一定程度上使得转换后的语音特征更为接近目标说话人的语音特征，但与其还是有一定的距离，有待进一步的学习与研究。

5.2 语音转换研究方向与展望

想要实现一个优良的说话人语音转换系统，找到有效的语音转换算法，达到理想的语音转换目的，目前的工作还远远不够，需要对现今相关语音转换算法加强研究并优化。目前的研究不足主要表现在以下几方面:

由于语音的复杂性，构建精确的语音发音模型，选取更多、更准确的语音特征参数作为语音转换的基本参数是非常必要的。只有更多的特征参数才能表整出更加清晰自然的语音。

2、现今的语音转换算法主要是对表征语音的特征参数进行线性转换，但语音发声受各方面因素影响，语音信号并不是线性平稳信号，仅仅采用线性转换，并不能达到期望的转换效果。如果能够研究出一种算法实现语音的非线性转换，语音的转换效果将会有较大的提高。

3、现今的说话人语音转换算法仍然停留在理论研究阶段，离实际开发应用还有一定的距离，需要加强相关算法的研究，以实现语音转换技术的实用性，根据需要做到语音的实时转换。

总之，说话人语音转换技术还有待进一步的研究与探索，相信随着科学技术的发展，说话人语音转换技术会有进一步的突破，会在社会生活的各个领域得到越来越广泛的应用。

5.3 本章小结

本章先是对全文以及研究的语音转换算法作了一个总结，最后对语音转换今后的研究方向作了展望。

致 谢

值此论文完稿之际，借此机会对论文完成过程中给予我指导和帮助的老师和同学们致以我最真诚的谢意!

首先，衷心感谢我的导师程建政教授。程老师渊博的知识、严谨的治学态度以及一丝不苟的敬业精神让我受益匪浅。本论文的研究自始至终得到了程老师的精心指导，程老师不但从学术上指导我完成课题调研、设计、调试等各个阶段的任务，而且帮助我树立科学严谨的研究态度。在程老师的帮助下，我得以顺利完成硕士论文、建立了良好的科研作风并提高了论文写作的水平。另外，程老师以其平易近人的作风从思想上、生活中都给予了无微不至的关怀。从程老师身上，我不仅学到了丰富的知识，更是学到了做人做事该有的态度。在此，谨向程老师致以我最崇高的敬意和最诚挚的谢意！

同时还要感谢武汉纺织大学电子与电气工程学院的老师们在我学习期间所给予的无私帮助与关心，感谢各位老师！

此外，还要感谢研究生期间的同学以及从本科阶段一起走到现在的同学们，我们在一起度过了三年甚至七年的美丽岁月。感谢各位师兄师姐以及我可爱的师弟师妹们，感谢你们对我的支持与帮助，感谢你们陪我度过了一段年轻、快乐而充实的美好时光，你们学习与生活的态度让我收获颇丰。

感谢室友蔡迪、易娟、张琪等，感谢远在他方的挚友们，和你们一起走过的日子很是美丽！

向参加论文审阅、答辩的专家和老师表示诚挚的谢意!

最后，还要感谢我的父母和家人，感谢他们多年来对我的培养和支持，是他们的爱和心血使我不断进步。

谨以此文献给所有支持、关心、帮助和教诲过我的人们。

参考文献

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

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

[3] H.Valbret, .Mulines and J.Tubach. Voice transfomation using PSOLA techniques, Speech Communication [J]. V01.11.No.2-3, 1992:75—187

[4] Narendranath M., rthyH.A. , endran S., gnanarayana B.. Transformation of formants for voice conversion using artificial neural networks，Speech Communication[J]. 1995, (2):207-216

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

[6] Y.Stylianou, O.Cappe and E.Moulines. Statisticl methods for voice quality transformation[J]. Proc.EUROSPEECH, l995:447-450

[7] Y.Stylianou, O.Cappe and E.Moulines. Continuous Probabilistic Transform For Voice Conversion[J], IEEE transactions of speech and audio processing, 1998, Vol.6.

[8] T.Toda, H.Saruwatari and K.Shikano. Voice conversion algorithm based on Gaussian mixture model with Dynamic frequency warping of STRAIGHT spectrum[J]. Proc.ICASS, 2001:84l-844

[9] T.Toda. STRAIGHT-based Voice Conversion Algorithm Based On Gaussian Mixture Model, Proc ICSLP, Beijng China ESCA,2000:279—282

[10] A.Kain and M.Macon. Spectral Voice Conversion For Text-To-Speech Synthesis, Proceedings Of ICASSP, 1998:282-288

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

[12] Ye Hui and Young Steve. High Quality Voice Morphing, Proceedings of ICASSP, Vol.1, 2004:9-12

[l3] K.S.Rao, B.Yegnanarayana. Deternlination of instants of significant excitation in speech using group deJay functjon[J]. ICASSP, Vol.3, No.5, l995:325-333

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

[15] K.S.Lee. Statical approach for Voice personality transformation[J]. ICASSP, Vol.15, No.2, 2007:654-65l

[16] Seneff.S.System to independently modify excitation and/or spectrum of speech wave

form without explicit pitch extraction. IEEE Transactions on Acoustics, Speech, and Signal Processing,1982, 30(4):566-578

[17] KuwabaraH.. A pitch-sychronous analysis/synthesis system to independently modify formant fiequeneies and bandwiths for voiced speech, Speech Communication, 1984: 211-220

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

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

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

[21] 左国玉，刘文举，阮晓钢．基于遗传径向基神经网络的声音转换[J]，中文信息学报，2004

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

[23] 胡航．语音信号处理[M]．哈尔滨:哈尔滨工业大学出版社．2004

[24] 韩纪庆，张磊，郑铁然．语音信号处理[M]．北京：清华大学出版社，2004

[25] 赵力．语音信号处理[M]．北京:机械工业出版社．2005

[26] 张雪英.数字语音处理及 MATLAB 仿真[M].北京：电子工业出版社，2010

[27] 张雄伟.陈亮，杨吉斌．现代语音处理技术及应用[M]．北京：机械工业出版社，2003

[28] Y.Stylianou, O.cuppe and E.Moulines. StaStical methods for voice quality transformation[J]. In Proceedings 0f Eurospeech, Madrid, Spain, Sep. 1995:447-450

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

[30] Hui Ye, S.Young,Quality-enhanced voice morphing using maximum likelihood transformations, IEEE Transactions on Audio, Speech and Language Processing, Vol.14, No.4, Jul.2006:1301-1312

[31] L.M.ArSlan. Speaker transfornlation algorithm using segmental codebooks[J]. Speech Communication, Vol.28.1999:2ll-226

[32] 张雄伟，陈亮，杨吉斌．现代语音处理技术及应用[M]．北京：机械工业出版社，2003

[33] KAIN A., High Resolution Voice Transformation[D]. OGI School of Science and Engineering at Oregon Health and Science University, 2001.

[34] 鲍长春．数字语音编码原理[M] ．西安：西安电子科技大学出版社，2007

[35] 朱廷韵，高文．基于数据挖掘的普通话韵律规则学习．计算机学报，2000，23(11)：1179 一 1183

[36] 王大凯，彭进业.小波分析及其在信号处理中的应用[M] ．北京：电子工业出版社，2006

[37] 张雄伟，陈亮，杨吉斌．现代语音处理技术及应用[M]．北京：机械工业出版社，2003

[38] 宋建华．基于 Matlab 的一种基音周期检测算法[J] ．信息技术．2009

[39] F.Itakura. Linear spectral representation of linear predictive coefficients[J]. Journal of Acoustic Society of Americ,87(4), 1990:1738-l990

[40] D.A.Reynolds and R.C.Rose, Robust text-independent speaker identification using Gaussian Mixture speaker models, IEEE Trans. Speech Audio Process., Vol.3, No.1, Jan.1995:72-83.

[41] S.Haltsonen. Improved dynamic Time warping methods for discrete utterance recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.33, No.2, 1985:449-450

[42] F.Itakura. Linear spectral representation of linear predictive coefficients[J]. Journal of Acoustic Society of Americ, 87(4),1990:1738-l990

[43] L.Rabiner, C.Schmidt, Application of dynamic time warping to connect digit recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.28, No.4, 1980:377—388

[44] 张正军等．基于 STRAIGHT 模型和人工神经网络的语音转换[J].电声技术.2010

[45] 高隽．人工神经网络原理及仿真实例．北京：机械工业出版社，2003

[46] 田景文，高美娟．人工神经网络算法研究及应用[M] ．北京：北京理工大学出版社，2006

[47] 阎平凡，张长水．人工神经网络与模拟进化计算[M] ．北京：清华大学出版社.2000

[48] 周开利，康耀红．神经网络及其 MATLAB 仿真程序设计[M] ．北京：清华大学出版社.2005

[49] S.Chiu. Fuzzy Model Identifieation Based on Cluster Estimation. Journal of Intelligent And Fuzzy Systems, 1994

[50] R.R.Paivaand A.Dourad. Intepretability and Learning in Neuor-Fuzzy Systems. Fuzzy Sets And Systems, 2004, 147:17—38

[51] 刘希玉，刘弘．人工神经网络与微粒群优化[M] ．北京：北京邮电大学出版社，2008

[52] 黄席樾，向长城，殷礼胜．现代智能算法理论及应用[M] ．北京：科学出版社，2009

[53] R.Eberhart, J.Kennedy.A New Optimizer Using Particle Swarm Theory[A]. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. Nagoya Japan, 4-6Oct, 1995:9-43

[54] C.K.Chow. Anoptimum character recognition system using decision functions[J]. IRE Trans.,1957:247—254