摘要

为实现对大型在线音乐数据进行自动分析、组织和检索，乐音分离在近几年越来越受到人们的关注。非立体声分离是希望从单一声道的多音音乐中恢复出每一个声源/乐器声线，这是一个非常具有挑战性的问题。而本文要研究的是歌声分离，也就是从单一声道的多音音乐中将歌唱声分离出来，并且能达到较好的分离效果。概括地说，现有的非立体声乐音分离系统都是基于传统的信号处理技术（主要是正弦模型），统计技术（如稀疏编码和非负矩阵分解），或者基于心理声学的研究（计算听觉场景分析，CASA）。

音乐信号是一种典型的非平稳信号，因此对于分析音乐信号，时-频（T-F）分析方法是有效的。T-F 分析方法是非平稳信号处理的一个重要分支，它是利用时间和频率的联合函数来表示非平稳信号，并对其进行分析和处理。T-F 分析方法按照 T-F 联合函数的不同可以分为线性和非线性两种。常见的线性 T-F 表示主要有短时傅立叶变换（STFT）、Gabor 展开以及小波变换等。非线性 T-F 方法主要有 Wigner-Ville 分布（WVD）和 Cohen 类。此外，听觉滤波器也逐渐成为对信号进行 T-F 分析的重要方式。

CASA 研究的主要目标是分析一个听觉场景，并识别场景中的各种声音。我们也是根据 CASA 的思想建立了基于 T-F 分析的歌声分离系统。该分离系统由四个模块组成：T-F 分解、主音高检测、歌声 T-F 信息的提取和歌声的合成。在 T-F 分解阶段我们采用了 STFT 和 Gammatone 滤波器两种 T-F 分析技术，由此设计出两种分离方法。基于 STFT 的歌声分离系统，在 T-F 分解阶段是利用 STFT 将输入的时域信号变换到 T-F 域。经过这样的处理，信号的频谱具有随时间变化的特点。基于 Gammatone 滤波器的歌声分离系统，是利用一个 Gammatone 滤波器组将输入信号分解成多个频带的时域信号，然后将每个滤波器的输出划分为重叠的帧。两种方法的主音高的检测模块是一致的。虽然基音提取的方法有很多，但是大多数乐音信号是多音信号，所以想

要从多音信号中提取出歌声的音高相当困难。而我们利用乐音的谐波特性进行音高的提取。第三阶段是进行歌声T-F信息的提取。基于STFT的方法，是根据检测到的主音高，将每一帧信号的频谱中的各次谐波结构提取出来。而在基于Gammatone滤波器的方法中，除了要用到主音高，还需要计算相关图谱和交叉通道互相关，以及端点检测。最后一步，STFT方法中是对提取的歌声的短时谱进行反变换。而在第二种分离方法中，对各通道进行叠加求和就可以得到分离的歌声。

关键词：时-频分析；歌声分离；主音高检测；听觉滤波器；计算听觉场景分析

ABSTRACT

As the demand for automatic analyzing, organizing, and retrieving a vast amount of online music data explodes, musical sound separation has attracted significant attention in recent years. Monaural separation that attempts to recover each source/instrument line from single-channel polyphonic music is a particularly challenging problem. We will separate the vocal from single-channel polyphonic music, and obtain good separated result. Broadly speaking, existed monaural musical sound separation systems are either based on traditional signal processing techniques (mainly sinusoidal modeling), statistical techniques (such as sparse coding and nonnegative matrix factorization), or psychoacoustic studies (computational auditory scene analysis, CASA).

Time-Frequency (T-F) analysis is very effective to research musical signal which is a typical non-stationary signal. T-F analysis method is an important ramification of non-stationary signal processing. It employs the joint function of time and frequency to represent, analyze and process the non-stationary signal. We can classify the T-F analysis methods to linear and nonlinear representations according to the joint function. The linear analysis includes short-time Fourier transform (STFT), Gabor transform and wavelet transform. The nonlinear analysis method contains Wigner-Ville distribution and Cohen's class. Furthermore, auditory filter has become an important T-F analysis technique.

Analyzing an auditory scene and identifying the various sounds present in it has been the primary focus of the research called CASA. We design the vocal separation system based on T-F analysis drawing inspiration from the CASA. The system consists of T-F decomposition, predominant pitch detection, extraction of vocal T-F information and synthesis of vocal. Because STFT and Gammatone filter are used to

decompose signal in T-F decomposition stage, we design two different separation methods. In the vocal separation method based on STFT, the time domain signal is transformed into time-frequency domains using STFT. So the spectrum varies with time after processing. But the system based on Gammatone filter uses a Gammatone filterbank to decompose the original signal into many time domain signals with different frequency bands, then each filtered output are divided into overlapping frames. The predominant pitch detection stage is consistent in the tow different separation method. Although some methods are used to detect pitch, it is very difficult to detect the pitch of vocal where the musical signals are polyphonic. We extract pitch of vocal employing the harmonic characteristic of music. In the third stage, T-F information of vocal is extracted. The STFT system extracts the harmonics in spectrum of each frame according to detected predominant pitch. In the second separation method, the correlogram, cross-channel correlation and onset detection features are computed besides predominant pitch. In the last stage, the vocal is synthesized. Inverse transform of extracted STFT of vocal is computed in the STFT method. Vocal is synthesized by adding all channels in the Gammatone filter method.

Key words: Time-Frequency analysis; vocal separation; predominant pitch detection; auditory filter; computational auditory scene analysis

原 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

论文作者签名：谢秀芳　　日 期：2009年5月20日

关于学位论文使用授权的声明

本人完全了解山东大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权山东大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

(保密论文在解密后应遵守此规定)

论文作者签名：谢秀芳　导师签名：刘光临　日　期：2009年5月20日

第一章 绪论

1.1. 课题研究的背景和意义

在从不同声源分离声音这一问题上，人类的听觉系统有着非凡的能力。这种能力的一个重要方面就是，人类可以听出由乐器伴奏的歌声。这样的任务对人来说是很容易的，但对于机器而言却很困难。虽然语音分离已经得到了广泛地研究，不过从多音音乐中分离歌声的研究却很少。因为歌声是由发音器官产生，所以对于歌声分离而言，探究语音分离技术是很重要的。从声音分离的观点看，歌声和语音之间最重要的不同是：与这两种声音同时存在的其它干扰声音在本质有所不同。在一个实际的声学环境中，语音通常被各种干扰污染，可以是谐波或非谐波，窄带或宽带。在语音和干扰的频谱不相关的意义上讲，多数情况下的干扰和语音是独立的。然而，对于唱片中的歌声几乎总是有乐器伴奏，而这些乐器在多数情况下是谐波的、宽带的，并且由于音乐信号是由乐器声音与歌声一起组成的一个统一整体，因此乐器声音与歌声还是相关的[1]。这样的差别使歌声从多音音乐中分离出来变得更加困难。所以，我们需要结合更多的时 T-F 分析技术，来研究更好的声音分离技术。

歌声分离系统在某些领域有广泛的应用，如自动歌词识别和校正。自动歌词识别通常要求输入系统的是单独的歌声[2]，这对于几乎所有的歌曲常常是不实际的，因为歌声是由乐器伴奏。不过，如果将成功的歌声分离用于预处理，这样的要求可以得到满足。给歌声校正歌词对于卡拉 OK 这样的应用是一个关键步骤[3]，而且通常这还是一项复杂的工作。因此，自动操作该过程将带来相当大的帮助。一个精确的歌词校正系统将允许听众容易地跟随歌声。可是，当伴奏存在时使歌声校正歌词的任务变得很困难，不过一个分离系统可以用于解决这样的问题。歌声分离系统还可以用于对歌手的辨认。几项研究已经解决了在真实唱片中歌手辨认的问题，但是到目前为止，所做的任何努力都不能分离一个歌手的声音。通过歌声分离，期望提高歌手辨认的精确

性。歌声分离的另一个应用领域是音乐信息检索[4]。歌声携带很多有用的信息，如旋律，歌声分离就可以用于旋律的提取。基于歌声携带主要旋律的假设，我们可以利用单频记谱技术从分离的歌声轨迹中提取旋律。因为单音记谱比多音记谱更简单，将歌声分离作为预处理就可以绕开直接利用多音记谱技术提取旋律的难题。

1.2. T-F 分析的发展

音乐信号的时域分析和频域分析是音乐分析的两种重要方法。但是这两种方法均有局限性：时域分析对音乐信号的频率特性没有直观的了解；而频域特性中又没有音乐信号随时间的变换关系[5]。分析与处理平稳信号最常用、最主要的方法是傅立叶变换，它建立了信号从时域到频域的变换桥梁，而傅立叶反变换则建立了信号从频域到时域的变换桥梁，它们之间是一一对应的映射关。因此在传统的信号分析预处理中，时域和频域构成了表征信号的两种主要方式。但是傅立叶变换的不足在于它是在整体上将信号分解为不同的频率分量，从而缺乏局域性信息，即对信号的表征要么完全在时域，要么完全在频域，它不能揭示某种频率分量出现在什么时候以及随时间的变化情况[6]。为了克服传统傅立叶变换的这种全局性变换的局限性，必须使用局部变换的方法，用时间和频率的联合函数来表示信号，这就是 T-F 分析思想的来源。T-F 分析方法按照 T-F 联合函数的不同可以分为线性和非线性两种[7]。此外，就是近年来常用的听觉滤波器。

1.2.1. 线性 T-F 表示

线性 T-F 表示是由傅立叶变换演化而来的，满足线性叠加性。假设 $x(t)=ax_1(t)+bx_2(t)$，记 $x(t)$，$x_1(t)$，$x_2(t)$的线性 T-F 表示分别为 $P(t,f)$，$P_1(t,f)$，$P_2(t,f)$，则有

$$P(t,f)=aP_1(t,f)+bP_2(t,f) \tag{1-1}$$

常见的线性 T-F 表示主要有 STFT、Gabor 展开以及小波变换等。STFT 的基本思想是用一个时间宽度足够窄的固定的窗函数乘时间信号，使取出的信号可以被看成平稳的，然后对取出的这一段信号进行傅立叶

变换，便可以反映出该时间宽度中的频谱变化规律。如果让这个固定的窗函数沿着时间轴移动，那就可以得到信号频谱随时间变化的规律。由于其算法简单，所以在很长一段时间里成为非平稳信号分析标准和有力的工具。Gabor 提出了一种同时使用频率和时间来表示一个时间函数的思想和方法，这种方法便是后来的 Gabor 展开。Gabor 展开的思想在很大程度上开创了 T-F 分析的先河。现在 Gabor 展开已经在暂态信号检测，时变滤波，图像信号处理等领域取得了成功的应用。在 STFT 和 Gabor 展开中都使用了固定的时间窗函数，这就引出了时间分辨率和频率分辨率的矛盾。小波变换是一种在时间-尺度平面内，利用多分辨率分析思想分析非平稳信号的方法。由于其本身分辨力的优良性能，一经提出便很快成了非平稳信号分析和处理的一大热点[8]。

1.2.2. 非线性 T-F 表示

非线性 T-F 表示又叫做二次型 T-F 表示，它反映的是信号能量的 T-F 分布，不满足叠加性。假设 $x(t)=ax_1(t)+bx_2(t)$，记 $x(t)$，$x_1(t)$，$x_2(t)$ 的线性时频表示分别为 $P(t,f)$，$P_1(t,f)$，$P_2(t,f)$，则有

$$P(t,f)=|a|^2 P_1(t,f)+|b|^2 P_2(t,f)+2R[abP_{12}(t,f)] \qquad (1-2)$$

式中最后一项称之为干扰项，这是非线性 T-F 表示固有的一个属性。非线性 T-F 表示主要有 Cohen 类和放射类（Affine），其中最著名的是 WVD。WVD 和 Cohen 类是采用对信号的双线性乘积进行核函数加权平均的方法来实现的非线性 T-F 表示，它们表示的是信号的能量密度分布。WVD，由于其本身满足的大部分所期望的数学性质，如实值性，对称性，边缘积分特性，能量守恒，T-F 移位等特性，所以它确实反映了非平稳信号的时变频谱特性，加之能作相关化解释，从而成为非平稳信号分析处理的一个有力的工具。但是由于其对多分量信号产生的难以抑制的“交叉项干扰”，从而限制了它的发展。后来，L. Cohen 将各种变形的 WVD 统一为双线性 T-F 分布理论，给出了一个统一的数学公式，通过选取不同的核函数，可以得到不同的 T-F 分布，其中 WVD 是最简单的形式。人们把 Cohen 公式所表示的 T-F 分布统称为 Cohen 类 T-F 分布。Cohen 类 T-F 表示的一个最大特点是时移不变与频移不变

特性自动满足。由于只是各种变形 WVD 的统一形式，Cohen 类仍避免不了交叉项干扰这个缺点[8]。

1.2.3. 听觉滤波器

耳蜗常被认为是由一组带通滤波器组成的，因为它具备频率选择功能，可将不同频率映射到基底膜上的不同位置。这种频率选择特性表现为：基底膜上的每一个位置都对应着一个特征频率，将这个特征频率下的纯音信号作为输人刺激耳蜗的时候，基底膜的这个位置会发生最大幅度的听觉神经冲动。作为一个滤波器组，耳蜗基底膜具有以下特征[9]：

（1）变化的滤波器带宽。低频处的频率分辨率较高，高频处的频率分辨率相对较低，这表明耳蜗基底膜的不同位置对应的滤波器的带宽是不一样的。

（2）单个滤波器的频率响应非对称。基底膜的每个位置都对应一个特征频率，该位置对此频率的信号响应幅度最大。特征频率的低频范围斜率比较平缓，特征频率的高频范围斜率较陡。

（3）单个滤波器的频率响应同刺激声的强度相关。基底膜的振动特性是呈压缩非线性的，这意味着双倍强度的声音刺激并不会引起双倍幅度的基底膜振动。从滤波器角度来看，这表明了单个听觉滤波器在特征频率处的响应幅度会随着刺激声强度的增加而增加，但是增加的速度会越来越慢。

目前主要有以下几种常见的听觉滤波器。Lyon等的共振滤波器比较容易实现，计算复杂度低，在语音识别领域有一定的应用，但是效果并不理想。Roex函数最早用于拟合人耳对噪声背景中识别出特定信号的频率阈值的掩蔽实验数据。Roex函数滤波器在听觉掩蔽实验中得到了较多的应用。但是Roex函数没有简单形式的冲激响应函数，这是它的一个最大缺点。因为Gammatone函数最早在拟合各种听神经的生理学冲激响应数据中取得了十分满意的效果，因此是一种应用很广的听觉滤波器。Gammachirp滤波器也是由冲激响应函数定义的。Gammachirp滤波器克服了Gammatone函数不能模拟基底膜滤波器非对

称、强度依赖特性的缺点。

1.3. 声音分离技术

一般地说，现有的非立体声乐音分离系统都是基于传统的信号处理技术（主要是正弦模型），统计技术（如稀疏编码和非负矩阵分解），或者基于心理声学的研究（计算听觉场景分析）[10]。

正弦模型的基本思想是，将一个声音模拟成具有时变频率、幅度和相位的正弦曲线的一个线性组合。要实现乐音分离，就需要为音乐中的每个声源估计所需的参数。Virtanen 利用最小二乘估计来计算这些参数，并利用谱包络进行建模以解决谐波重叠的问题[11]。Every 和 Szymanski 利用一种叫做谱滤波器的技术提取了正弦信号[12]。当能够准确估计每个声源的音高或知道先验知识的情况下，正弦信号建模对于有较少声音的多音音乐通常能得出很好的结果。

用于乐音分离的统计方法，一般是对声源的某一统计特性进行假设。稀疏编码方法就是假设一个声源就是来自一个过完备集的基的加权和。如果某一个基的概率高，权值就假设为 0，也就是说大部分的基在多数时间里是不起作用的（无效的）[13]。虽然非负矩阵分解也试图找到一个具有非负元素的混合矩阵和声源矩阵，这样使得重构的误差最小，但是一般要求混合加权和声源是稀疏的[14]。最近提出的几个系统[15,11]已经证明了统计方法的适用性。但是若将此类方法用于更多的问题上，还需要进一步研究。

最早提出的与声音分离有关的概念是听觉场景分析（ASA）。ASA 源自 Cherry 在 1953 年的发现，即人类听觉系统能够从复杂的混合声音中有效地选择并跟踪某一说话人的声音。Cherry 把这一现象称之为“鸡尾酒效应”。自此，声源分离一直是一个重要的研究课题。ASA 这一概念，首先是由著名心理听觉学家 Albert Bregman 提出的。Bregman 认为，听觉系统利用声音的各种特性（时域、频域、空间位置等），通过自下而上（分解）和自上而下（学习）的双向信息交流，对现实世界的混合声音进行分解，使各分量归属于各自的物理声源[16]。ASA 已经激发研究者为声音分离建立了 CASA。与其它声音分离方法相比，

CASA 做了最小的关于同时存在的声音的假设，并利用声音的固有特征来替代，因此展示了在非立体声的歌声分离中更大的潜力。CASA 的目标是为一般的声音分离建立计算系统。目前已经提出了几个 CASA 系统用于乐音分离。Mellinger 所作的工作描述了第一次将 CASA 用于乐音分离的尝试[17]。他的系统提取了端点信息和普通的频率变化，并利用它们组成了来自相同乐器的频率泛音。不过，这两个特征似乎不能充分地分离不同的声音。Godsmark 和 Brown 开发了一个 CASA 系统[18]，为了对各种声源进行分组，该系统在一个黑板体系结构中采用了调和性和其它原则。Goto 开发了一个音乐现场描述系统[19]，该系统为旋律检测采用了调和性原则。另一个由 Meron 和 Hirose 提出的系统是为了从钢琴伴奏中分离歌声[20]。为了使系统工作，需要有大量的先验知识，如预混合歌声和钢琴的泛音轨迹或钢琴声音的乐谱。这种先验知识在多数情况下是不可知的，因此该系统不能用于大多数真实的唱片。最近，由 Hu 和 Wang 开发的一个声音分离系统成功地从基于音高跟踪和调幅的声学干扰中分离了浊音语音[21]。Hu-Wang 系统首先使用了听觉滤波器作为 T-F 分解的工具，并为确定的和未确定的谐波使用了不同的分离方法。

1.4. 本文的研究内容及章节安排

本文主要研究基于T-F分析的歌声分离技术。根据对几种T-F分析方法的比较，最终确定两种分离方法：基于STFT的歌声分离和基于Gammatone滤波器的歌声分离。歌声分离过程分为四步：T-F分解、主音高检测、歌声T-F信息的提取以及歌声的合成。我们将会在下面的几个章节详细介绍这四个模块。论文的内容安排如下：

第一章绪论，介绍了论文的研究背景和意义，T-F分析的发展，以及目前提出的几种声音分离技术。

第二章介绍用于音乐信号处理的几种T-F分析方法，包括STFT、小波变换、WVD和Gammatone滤波器。

第三章介绍主音高检测。首先介绍了基音检测的几种常见方法，并介绍在此基础上发展的多音音高检测系统。最后详细分析了本文所

用到的方法。

第四章介绍两种分离方法。基于 STFT 进行歌声分离的方法，具体过程包括如何获得声音的 STFT，提取歌声的短时谱，歌声的合成。基于 Gammatone 滤波器进行歌声分离的方法，具体过程包括基于 Gammatone 滤波器的 T-F 分解，特征提取，提取歌声的 T-F 信息，以及歌声的合成。

第五章对本文所做的工作进行了总结。

第二章 音乐信号的时-频分析

T-F分析实际上是将一维时间信号映射到T-F(有的是时间-尺度)二维，很好地表示出信号的频率成分随时间的变化规律。在音乐声学中，傅立叶变换只能充分地表示我们能听见的单频信号的音高，不能简洁地表示音乐信号的感知现象，从而促进了T-F表示理论在音乐信号分析中的发展。T-F分析处理能直观的感觉音乐信号的在时间和频率上的变化，还在信号分离中起了非常重要的作用。不同的T-F方法在分离中的效果是不一样的。我们从两个角度衡量一种T-F分析方法是否适用于乐音信号的分离。一是某种T-F方法将时间信号映射到T-F域以后，还能否转换到时域。另一个就是，某种T-F方法能否很清楚的根据某一个或多个特征，提取要分离的歌声的T-F信息。根据这两个标准，我们分别对STFT、小波变换、WVD以及Gammatone滤波器进行分析和比较，选出适用于歌声分离的T-F分析方法。

2.1. 短时傅里叶变换

标准傅立叶变换只在频域里有局部分析的能力，而在时域里不具备这种能力。因此，为了研究信号在局部时间范围内的瞬时频率特性，1946年，D. Gabor引入了STFT或窗口傅里叶变换的概念。对于分析音乐信号，STFT的方法是有效的解决途径。由于音乐信号的特性是随时间缓慢变化的，因而可以假设它在一小段时间内保持不变。那么，将STFT用于分析乐音信号，就是认为音乐信号是局部平稳的，可以对某一帧音乐信号进行傅里叶变换，即STFT，其定义为

$$STFT_x(t,\omega)=\int_{-\infty}^{\infty}x(\tau)g^*(\tau-t)e^{-j\omega\tau}d\tau \qquad (2-1)$$

可见，STFT是窗选音乐信号的标准傅里叶变换。式（2-1）中 $g(\tau)$是窗函数，并且应取对称函数。选择不同的窗函数，将得到不同的STFT结果。由（2-1）式可以明显的看出STFT是用一个时间宽度足够窄的固定的窗函数乘时间信号，使取出的信号可以被看成平稳的，然后对取出

的这一段信号进行傅立叶变换，便可以反映出该时间宽度中的频谱变化规律。如果让这个固定的窗函数沿着时间轴移动，那就可以得到信号频谱随时间变化的规律。

如傅里叶变换一样，我们总是希望能由变换域重建出原信号，对STFT也是如此。STFT反映信号 $x(t)$ 在 $t=\tau$ 附近的频谱特征，即反映出一个信号在任意局部范围的频谱特征，其反变换定义为

$$x(\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}STFT_x(t,\omega)g(\tau-t)e^{j\omega\tau}dtd\omega \tag{2-2}$$

定义式（2-1）表明，STFT是一种线性T-F表示，它具有以下几个性质。

性质1：叠加性

$$\begin{aligned}&x_1(t)\rightarrow STFT_{x_1}(a,b),\quad x_2(t)\rightarrow STFT_{x_2}(a,b)\\&\lambda_1x_1(t)+\lambda_2x_2(t)\rightarrow\lambda_1STFT_{x_1}(a,b)+\lambda_2STFT_{x_2}(a,b)\end{aligned} \tag{2-3}$$

性质2：频移不变性

$$x(t)\rightarrow x(t)e^{j\omega_0t}\Rightarrow STFT_x(t,\omega)\rightarrow STFT_x(t,\omega-\omega_0) \tag{2-4}$$

性质3：不具有时移不变性

$$x(t)\rightarrow x(t-t_0)\Rightarrow STFT_x(t,\omega)\rightarrow STFT_x(t-t_0,\omega)e^{-j\omega t_0} \tag{2-5}$$

性质4：STFT的低通实现

$$STFT_x(t,\omega)=\int_{-\infty}^{\infty}X(\omega'+\omega)G^*(\omega)e^{j\omega't}d\omega' \tag{2-6}$$

式中 $X(\omega)$ 和 $G(\omega)$ 分别是函数 $x(t)$ 和窗函数 $g(t)$ 的傅里叶变换。

性质5：STFT的带通实现

$$STFT_x(t,\omega)=e^{-j\omega t}\int_{-\infty}^{\infty}X(\omega')G^*(\omega'-\omega)e^{j\omega't}d\omega' \tag{2-7}$$

2.2. 小波变换

小波变换是20世纪80年代后期发展起来的一门新兴的应用数学分支，近年来有学者将小波变换应用到工程振动信号分析等领域中。在理论上，构成小波变换比较系统框架的主要是法国数学家Y. Meyer、地质物理学家A. Grossman的贡献。而把这一理论引用到工程应用，特

别是信号处理领域，法国学者I. Daubechies和S. Mallat发挥了极为重要的作用。在工程应用领域，特别在信号处理、图像处理、语音分析、模式识别和量子物理等领域，小波变换被认为是信号分析工具和方法上的重大突破。

给定一个基本函数$\psi(t)$，令

$$\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{2-8}$$

式中a，b均为常数，且$a>0$。显然，$\psi_{a,b}(t)$是基本函数$\psi(t)$先作移位再作伸缩以后得到的。若a，b不断地变化，我们可得到一族函数$\psi_{a,b}(t)$。信号$x(t)$的小波变换（Wavelet Transform，WT）定义为

$$WT_x(a,b)=\frac{1}{\sqrt{a}}\int x(t)\psi^*(\frac{t-b}{a})dt = \int x(t)\psi^*_{a,b}(t)dt \tag{2-9}$$

信号$x(t)$的小波变换$WT_x(a,b)$是a和b的函数，b是时移，a是尺度因子。$\psi(t)$又称为基本小波，或母小波。$\psi_{a,b}(t)$是母小波经移位和伸缩所产生的一族函数，我们称之为小波基函数，或简称小波基[22]。

记$\Psi(\omega)$为$\psi(t)$的傅里叶变换，若

$$c_\psi \overset{\Delta}{=} \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega} < \infty \tag{2-10}$$

则$x(t)$可由其小波变换$WT_x(a,b)$来恢复，即

$$x(t)=\frac{1}{c_\psi}\int_0^\infty a^{-2}\int_{-\infty}^{\infty} WT_x(a,b)\psi_{a,b}(t)dadb \tag{2-11}$$

连续小波变换也是一种线性T-F表示，它具有以下几个基本性质。

性质1：叠加性

$$\begin{aligned} &x_1(t)\rightarrow WT_{x_1}(a,b), \quad x_2(t)\rightarrow WT_{x_2}(a,b) \\ &\lambda_1 x_1(t)+\lambda_2 x_2(t)\rightarrow \lambda_1 WT_{x_1}(a,b)+\lambda_2 WT_{x_2}(a,b) \end{aligned} \tag{2-12}$$

性质2：时移不变性

$$x(t)\rightarrow x(t-t_0)\Rightarrow WT_x(a,b)\rightarrow WT_x(a,b-t_0) \tag{2-13}$$

性质3：尺度转换

$$x(t)\rightarrow x(\frac{t}{\rho})\Rightarrow WT_x(a,b)\rightarrow \sqrt{\rho}\,WT_x(\frac{a}{\rho},\frac{b}{\rho}) \tag{2-14}$$

性质4：内积定理

$$< WT_{x_1}(a,b),\ WT_{x_2}(a,b) > = C_\psi < x_1(t),\ x_2(t) > \qquad (2-15)$$

2.3. Wigner-Ville 分布

1932年，Wigner提出了Wigner分布，最初应用于量子力学的研究。1948年，Ville将其引入信号分析领域。WVD是分析非平稳时变信号的重要工具。信号 $x(t)$ 的WVD定义为

$$W_x(t,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})e^{-j\omega\tau}d\tau \qquad (2-16)$$

WVD是对 $x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})$ 的傅里叶变换，其反变换表示为

$$x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})=\frac{1}{2\pi}\int_{-\infty}^{\infty}W_x(t,\omega)e^{j\omega\tau}d\omega \qquad (2-17)$$

首先取 $t=\frac{\tau}{2}$，然后进行变量替换，令 $\tau=t$，得

$$x(t)=\frac{1}{2\pi\, x^*(0)}\int_{-\infty}^{\infty}W_x(\frac{t}{2},\omega)e^{j\omega t}d\omega \qquad (2-18)$$

熟悉WVD的性质对于全面了解该T-F分布是十分必要的。下面介绍WVD的基本性质。

性质1：实性

无论信号是实的还是复的，其分布对所有 t 和 ω 值总是实的。

性质2：对称性

$$W_x(t,\omega)=W_x(t,-\omega),\quad W_x(t,\omega)=W_x(-t,\omega) \qquad (2-19)$$

性质3：边缘特性

时间边缘特性

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}W_x(t,\omega)d\omega=|x(t)|^2 \qquad (|x(t)|^2\text{ 为瞬时功率}) \qquad (2-20)$$

频率边缘特性

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}W_x(t,\omega)dt=|X(\omega)|^2 \qquad (|X(\omega|^2\text{ 为能量密度}) \qquad (2-21)$$

性质4：时移和频移不变性

$$x(t) \to x(t-t_0)e^{j\omega_0 t} \Rightarrow W_x(t,\omega) \to W_x(t-t_0,\omega-\omega_0) \qquad (2-22)$$

2.4. Gammatone 滤波器

我们在绪论中已经详细介绍了4种听觉滤波器，通过比较各自的特点，最后选择Gammatone滤波器作为分析的对象。Gammatone滤波器最早由Johannesma于1972年提出，用以描述猫的听神经的生理学冲激响应数据特性。Gammatone滤波器用一个因果的冲激响应来描述滤波器特性，其时域表达式为：

$$g(t) = b^n t^{(n-1)} e^{-2\pi bt} \cos(2\pi f_0 t + \phi) u(t) \qquad (2-23)$$

式中，$t<0$时$u(t)=0$；n为滤波器的阶数；参数b为 Gammatone 滤波器的等价矩形带宽（简称 ERB），它同滤波器的中心频率f_0的关系是

$$b = ERB(f_0) = 24.7 + 0.108 f_0 \qquad (2-24)$$

由公式（2-23）可以总结出 Gammatone 滤波器的三个优点：①需要的参数很少，实践表明 Gammatone 滤波器只需要 3 个主要参数就能够很好地模拟听觉实验中的生理数据；②需要的阶次较低，相关文献研究表明，阶数为 4 的 Gammatone 滤波器就能很好地模拟耳蜗的听觉滤波特性；③具有简单的冲激响应函数，能够由此推导出 Gammatone 函数的传递函数，进行各种滤波器性能分析，同时冲激响应函数有利于听觉系统模拟电路的实现。

经过 Gammatone 滤波的信号得到的是多个不同频带的时域波形。要想恢复出原始信号，只需将各个通道的时域波形叠加。

2.5. T-F 分析方法的比较

STFT的优点在于其物理意义明确，对于许多实际的信号，给出了与我们的直观感知相符的T-F构造。而且它不同于WVD，不会出现交叉项，由此成为历史上应用最多的一种T-F分析。但是由于测不准原理对窗函数T-F分辨能力的制约，在应用当中，必须对时窗与频窗宽度做出折中，而这种折中取决于窗函数和信号的T-F特性。当被分析信号是缓变和瞬变共存的信号类型时，任何折中都将变得没有意义。无论采用

任何宽度的时窗，要么照顾到缓变信号成分的要求而满足不了瞬变信号成分的需要，要么反之，或者是两种成分的分析结果都不能接受。

STFT虽然有着分辨率不高等明显缺陷，但由于其算法简单，实现容易，所以在很长一段时间里成为分析非平稳信号的有效方式。图2-1是一段音乐信号的STFT。首先我们知道乐音信号是一种典型的时变、非平稳信号，用STFT进行分析具有一定的局限性。但同时乐音信号又具有自身的特点。在时域，乐音信号具有明显的时间连续性，一个音符的持续时间比较长。在频域，一个音符的音高持续时间和时域一样，各谐波结构表现的很明显，而且不会经常出现频率突变的现象。因此，利用STFT对乐音信号进行分析能够充分地反映信号T-F特性。如果以音高作为分离的特征，那么由STFT得到的T-F图可以找到歌声的那一部分T-F信息。同时，我们还可以用STFT的反变换恢复出原始信号的时域波形。从这两点看，利用STFT进行歌声的分离是可行的。

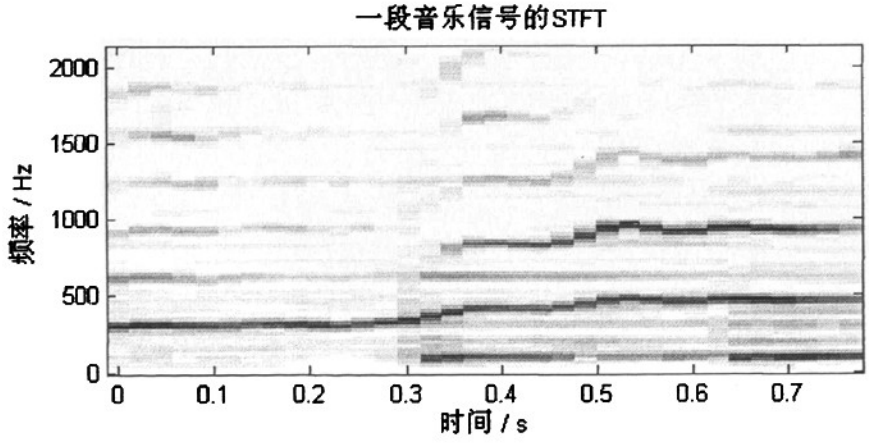

图2-1 一段音乐信号的STFT

与STFT不同，小波变换能较好地解决时间和频率分辨力的矛盾。小波变换的窗是可调T-F窗，在高频时使用短窗，在低频时则用宽窗，即以不同的尺度观察信号，以不同的分辨力分析信号，充分体现了多分辨率分析的思想，与时变、非平稳信号的特性一致。但是小波变换对T-F平面也是一种机械式的划分，在实际中选择能反映信号特征的小波不易，而且一旦选定小波就必须用同一个小波分析下去，因此并不具备自适应的特点。另外，小波变换引入的是尺度因子 a，由于尺度因子与频率之间没有直接的联系，而且频率在小波变换中没有明显地表

现出来，因此小波变换的结果不是一种真正的T-F谱。

图2-2是一段音乐信号的小波变换。与2-1图相比，小波变换在不同频率段T-F信息表现的不明显，有较大的重叠。我们是根据CASA的思想建立分离系统，以某一个或几个特征作为分离线索，但是利用小波变换不能有效地提取歌声的信息。所以，虽然小波分析有很多优点，但是在我们的系统中不采用该T-F分析方法。

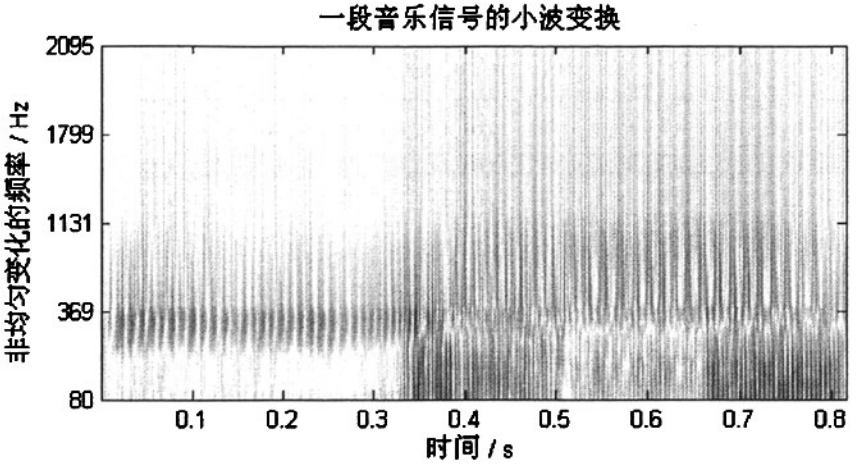

图2-2 一段音乐信号的小波变换

理论和实践均已表明，WV分布是表征线性调频信号的最好T-F分布，具有很好的T-F聚集性。而且WVD具有明显的物理意义，它可被看作信号能量在时域和频域中的分布。但由于它是一种双线性变换，交叉项是双线性T-F分布的固有结果，它来自于多分量信号中不同信号分量之间的交叉作用。交叉项通常是振荡的，而且幅度可以达到自主项的两倍，造成信号的T-F特征模糊不清[23]。交叉项的出现使得T-F分布图的可读性变差，特别当需要分析的信号为多分量信号时。如何有效抑制交叉项，对T-F分析非常重要。事实上，交叉项与T-F分布的有限支撑特性密切相关，而交叉项的抑制又主要通过核函数的设计来实现。常用的加核函数后的WVD有伪Wigner-Ville分布（PWVD）和平滑伪Wigner-Ville分布（SPWD）。虽然通过T-F平滑的方法抑制了部分交叉项，但这是以牺牲整个T-F分布的T-F分辨率为代价的。

图2-3（a）是音乐信号的WVD。从图中可以清楚地看到各个泛音之间存在交叉项，同时不同声源之间也存在交叉项。因此，WVD不能

直接用于信号分离。图2-4（b）为音乐信号的PWVD。PWVD可以抑制交叉项中的时域交叉项,但是无法改变频域中的交叉项。图2-4（c）为音乐信号的SPWD。分析乐音信号时，SPWD可以有效的抑制由于泛音以及多个声音引起的交叉项。但是，这又涉及到如何有由SPWD恢复出原信号的问题。若要由SPWD恢复出原信号很困难，这里面涉及到两次修正的问题。到目前为止，还没有提出一种明确的方法可以由SPWD进行信号的恢复。因此，即使WVD分布有很好的T-F特性，但是目前还不适用于信号的分离。

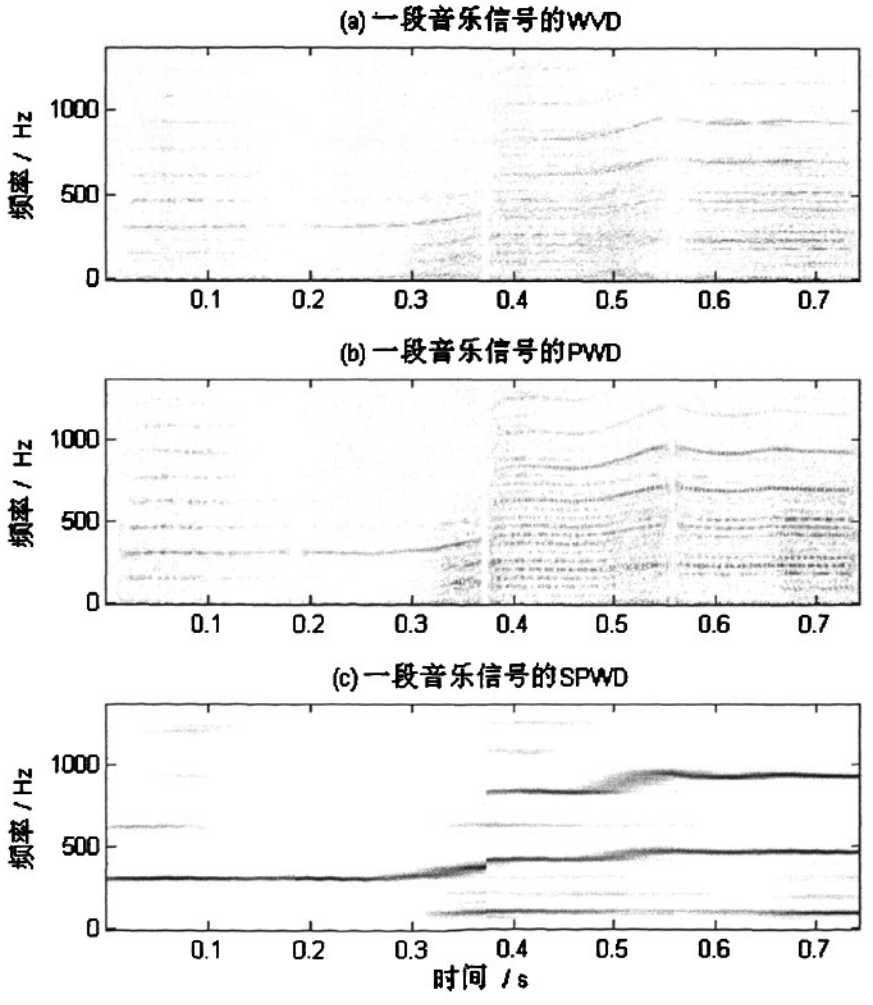

图2-3 一段音乐信号的WVD、PWV和SPWD

经Gammatone滤波器组滤波后将得到多个通道的时域波形。Gammatone滤波器组的中心频率按照ERB的关系呈准对数分布。基底膜

的每个位置都对应一个特征频率，该位置对此频率的信号响应幅度最大。特征频率的低频范围斜率比较平缓，特征频率的高频范围斜率较陡。低频处的频率分辨率较高，高频处的频率分辨率相对较低。图2-4展示了Gammatone滤波后的听觉T-F图。该听觉T-F分析与STFT一样，都能清楚的表现信号的T-F结构，这是小波变换所不具有的。两者的不同之处是它不受窗选择的限制。听觉T-F分析与WVD的相同之处在于都清楚地表现瞬时频率，不同之处是它不存在交叉项。实际上，听觉T-F表示集合了线性和双线性T-F表示的各自优点，它在T-F分析领域越来越受到关注。

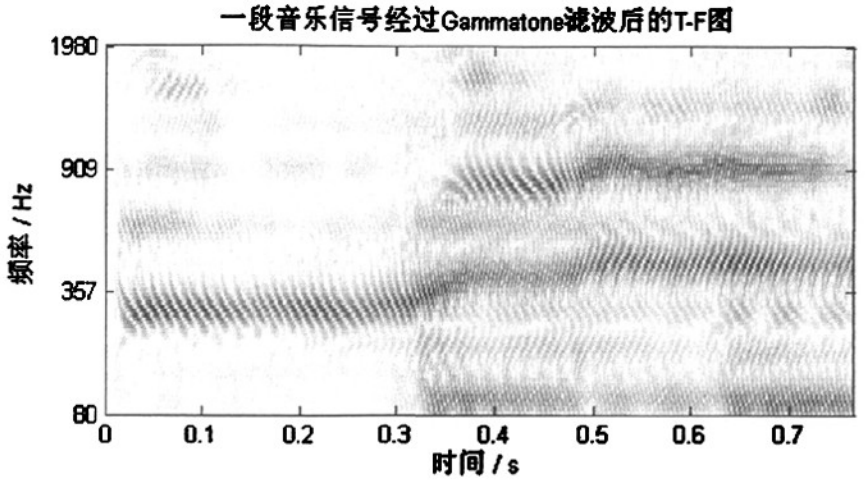

图 2-4 经过 Gammatone 滤波后的听觉 T-F 图

我们分析了 STFT、小波变换、WVD 和听觉 T-F 分析这四种 T-F 分析方法，并比较了各自的性能。最终选择 STFT 和听觉 T-F 分析作为建立歌声分离系统的 T-F 方法。

第三章 主音高检测

声带每开启或闭合一次的时间就是基音周期，它的倒数称为基音频率。基音频率取决于声带的大小、厚薄、松紧程度以及声门上下之间的气压差的效应等。一般基音频率越高，声带被拉得越长、越紧、越薄，声门的形状也变得越细长，而且这时声带在闭合时也未必是完全闭合。基音频率最低可达到 80Hz 左右，最高可到达 500Hz 左右。其范围随发音人的性别、年龄及具体情况而定[8]。音高检测其实质是音乐信号的基音检测。所谓音高，也称音调，它是人耳区分声音振动频率高低的一个度量。音高主要取决于声音频率的高低。与语音信号相比，音乐信号的周期性相对很稳定，每个音的音高是近似固定不变的，其频谱近似为离散谱[24]，可以检测到音乐信号比较精确的音高。在音乐中除了有歌声之外，通常还有多个乐器同时演奏。所以，我们称检测音乐中歌声音高的方式叫做主音高检测。

3.1. 基音检测方法介绍

单语音信号基音的提取经过几十年的研究，已经有了许多成熟的方法。基音检测的方法大致可分为波形估计法、相关处理法和变换法[9]。波形估计法是直接由语音波形来估计基音周期，分析出波形上的周期峰值。相关处理法在语音信号处理中广泛使用，这是因为相关处理法抗波形的相位失真能力强，另外它在硬件处理上结构简单。变换法是将语音信号变换到频域或倒谱域来估计基音周期，利用同态分析方法将声道的影响消除，得到属于激励部分的信息，进一步求取基音周期。

除了单语音信号外，现在还针对多音音乐提出了不少多音音高的检测。多音音高的检测是在单音音高的检测方法上发展起来的，不过更复杂，需要处理的问题也更多。这是因为，在音乐中通常多个乐器同时演奏。音乐的多音特性在检测音高中引起了谐波重叠问题，也就是一个音符的一个谐波所对应的频率与另一个同时演奏的音符的一个

谐波的频率相同或相近[21]。通常在没有具体乐器先验知识的情况下，很难区分哪一个谐波属于歌声。为了克服这些困难，许多研究人员通过不断的努力，提出了许多比较有效的方法。Goto 提出的方法[25]是首先对频谱进行滤波得到基频，然后利用期望-极大（EM）算法估计每个基频的概率，将概率最大的频率作为歌声的音高。由 Klapuri 提出的多音高检测算法[26]，利用调和性原则和谱光滑性在稳定乐音的混合声中检测每个乐器的音高。在该音高假设的得分是最高的情况下，第一个检测到的音高就是主音高。Li 和 Wang 提出了一种检测多音音频中歌声音高的方法[27]。他们的方法对 Wu 等人所提出的含噪语音的多音高检测算法[28]进行了扩展。首先用一个听觉外围对输入信号进行滤波，并计算一个相关图来提取每个通道的周期信息。然后利用通道和峰值选择获得有用的周期信息，这些信息是通过一个统计模型进行整合的。最后，利用一个 HMM 模拟音高产生过程，将最概率最大的音高轨迹看成歌声的音高包络。赵军和潘永湘通过分析基频平均谐波能量值的情况，提出一种基于谐波能量的混叠语音信号基音提取算法[29]。首先计算语音信号基频平均谐波能量曲线，然后寻找曲线中的所有极值点，最后经过分频及倍频过滤提取混叠语音所包含的基音频率。

根据文献[29]中介绍的算法，我们提出了自己的基于谐波能量比的多音音乐的主音高检测。我们的主音高检测算法是基于人声和乐器声音的谐波结构有明显差别，同时利用变换域的方法，将时间序列变换到频域，在频谱中寻找谐波结构。这一过程最经典的是离散傅里叶变换（DFT）。歌声有一个更宽的音高范围。正常语音的音高范围在 80 和 500Hz 之间，而歌声的音高范围在 80 和 1000Hz 之间。检测歌声的音高需要更好的频率分辨率。所以我们采用傅里叶变换改进的方法-非均匀离散傅里叶变换（NDFT）。

3.2. DFT 与 NDFT 的比较

3.2.1. 离散傅里叶变换

对于给定长度为 N 的序列 $x(n)$，其离散傅立叶变换 $X(k)$ 就定义为

Z平面单位圆上均匀分布的N个点处Z变换的值。利用快速傅立叶变换（FFT）算法，可以很方便地计算DFT，因而DFT得到了非常广泛的使用。

$$X(k)=\sum_{n=0}^{N-1}x(n)z^{-kn} \tag{3-1}$$

其中$z=e^{j2\pi/N}$，$k=0,1,\cdots,N-1$。

然而，因为DFT只能给出均匀分布点处Z变换的值，也就意味着它只能提供固定频谱精度[30]，而这只与采样点N的大小有关。要想得到更高的频谱精度必须增加采样点数目N，这就导致计算量迅速增加。为了克服上述DFT算法只能靠增加采样点的数目来提高频谱精度的缺陷，人们期望获得更为一般的采样点处的Z变换值。Rabiner等在1969年提出了一种线性调频变换算法(Chirp Transform Algorithm, CTA)[31]，该算法可以计算Z平面中从任意点开始的M个采样点处的Z变换值，但这M个点之间的距离(角度)还是均匀相等的。Oppenheim和Johnson在1971年提出了使用FFT来计算单位圆上非均匀分布采样点处Z变换的算法，从而可以用不均匀的频谱精度来进行频谱的计算分析[32]。这种方法叫做非均匀离散傅里叶变换（NDFT）。

3.2.2. 非均匀离散傅里叶变换

有限长度(N)序列$x(n)$的NDFT定义为：

$$X(z_k)=\sum_{n=0}^{N-1}x(n)z_k^{-n},\quad k=0,1,\cdots,N-1 \tag{3-2}$$

其中z_0，z_1，…，z_{N-1}是Z平面上任意分布的N个不同点，式（3-2）可以写成如下矩阵形式：

$$X=Dx \tag{3-3}$$

其中

$$X=\begin{bmatrix}X(z_0)\\X(z_1)\\\vdots\\X(z_{N-1})\end{bmatrix},\qquad x=\begin{bmatrix}x(0)\\x(1)\\\vdots\\x(N-1)\end{bmatrix} \tag{3-4}$$

$$D=\begin{bmatrix} 1 & z_0^{-1} & z_0^{-2} & \cdots & z_0^{-(N-1)} \\ 1 & z_1^{-1} & z_1^{-2} & \cdots & z_1^{-(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_{N-1}^{-1} & z_{N-1}^{-2} & \cdots & z_{N-1}^{-(N-1)} \end{bmatrix} \tag{3-5}$$

注意，NDFT矩阵 D 是范德蒙特矩阵，完全由 N 个点 z_k 来决定。

通常(均匀)的DFT就是NDFT的特例，这时 N 个采样点位于单位圆上，且是均匀分布的。NDFT所具有的在 Z 平面上任意选择频谱采样点的灵活性，可以用到语音信号处理等很多实际应用之中。非均匀频谱采样思想已经在频谱分析和滤波器设计中得到了极为广泛的应用。

3.3. 基于 NDFT 的主音高检测

一般而言，歌声的谐波结构和乐器的谐波结构有明显的区别。歌声的谐波能量变化缓慢，经常存在前几个泛音比基频处能量高的情况[33]。而乐器的第一个谐波的能量要比基频处的能量低。因此我们就根据泛音和基频能量的比值确定歌声的音高。我们的算法是基于经典的频域基音检测方法，分为三步。第一步是利用 NDFT 将时域信号变换到频域。第二步计算功率谱的谱包络。第三步找出频域中最突出的几个能量点，根据这几个突出的能量，寻找它们的倍频和二分频。然后计算二分频与该频率值或者该频率值与倍频处的能量值之比。最后根据能量的比值确定歌声的音高。

3.3.1. 设计 NDFT

我们在选择NDFT时，只需要改变 z_k，使其适用于乐音音高的变化特性。比如说，我们希望在低频处有较高的频率分辨率，而在高频处频率分辨率可以相对较低。实际歌声的音高在80Hz~1000Hz，那么我们希望这个范围内的频率分辨率越高越好，而对更高的频率范围，可以降低其分辨率。因此，我们就采用NDFT作为主音高检测的变换域方法，找出一个适合的频率变化曲线，从而得到我们期望的频率分辨率。我们选取的是指数形式的频率变化方式，图3-1（a）展示了NDFT频率采样点在单位圆上的分布情况，采样点是不均匀分布的，开始的几个

点相对密集，点与点之间的距离逐渐增大。图3-1（b）是根据 z_k 的取值而得到的NDFT的频率变化曲线。该曲线的频率范围是从80Hz到2000Hz，之所以选择这个频率范围，是因为歌声的基频大概在80Hz-1000Hz，至少可以包含第一个泛音。从图中可以清楚的看到，低频处频率变化缓慢，分辨率高，随着频率的增加变化加快，分辨率降低。

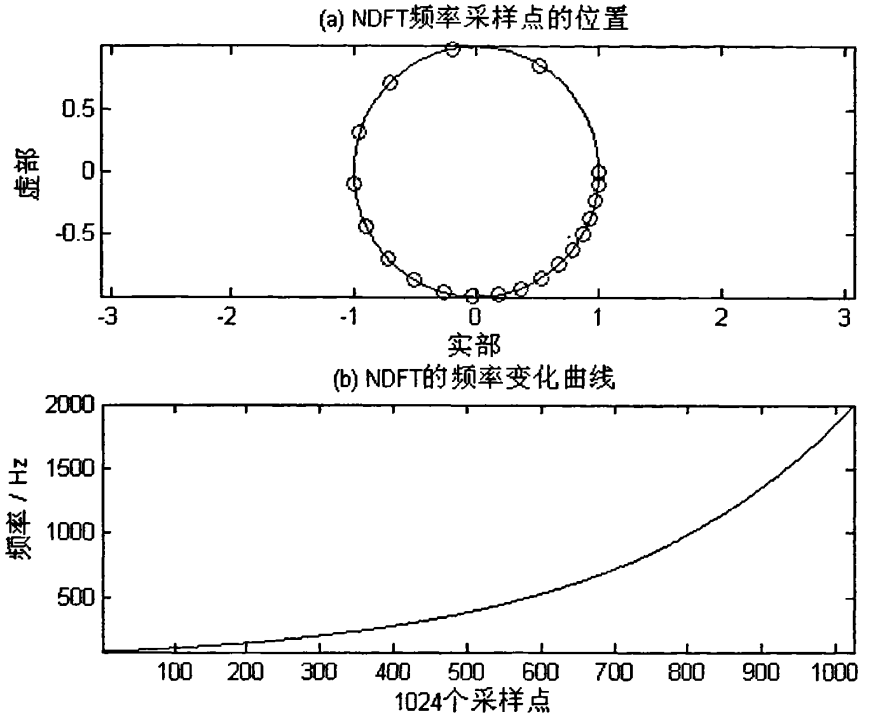

图3-1 （a）NDFT频率采样点在单位圆上的分布 （b）NDFT的频率变化曲线

我们已经确定了 z_k 的取值，就相当于确定了矩阵 $\boldsymbol{D}$。根据公式（3-3）就可以得到一段时间序列的 NDFT。图 3-2（a）是一段时间序列的 DFT 频谱图，图 3-2（b）是该序列的 NDFT 频谱图。对两个图比较可知，两者都是用 1024 个点进行频域变换，频率范围都为80Hz-2000Hz，但是它们的频谱分布情况却截然不同。图 3-2（a）的频谱是均匀变化的，而图 3-2（b）则是非均匀变化。

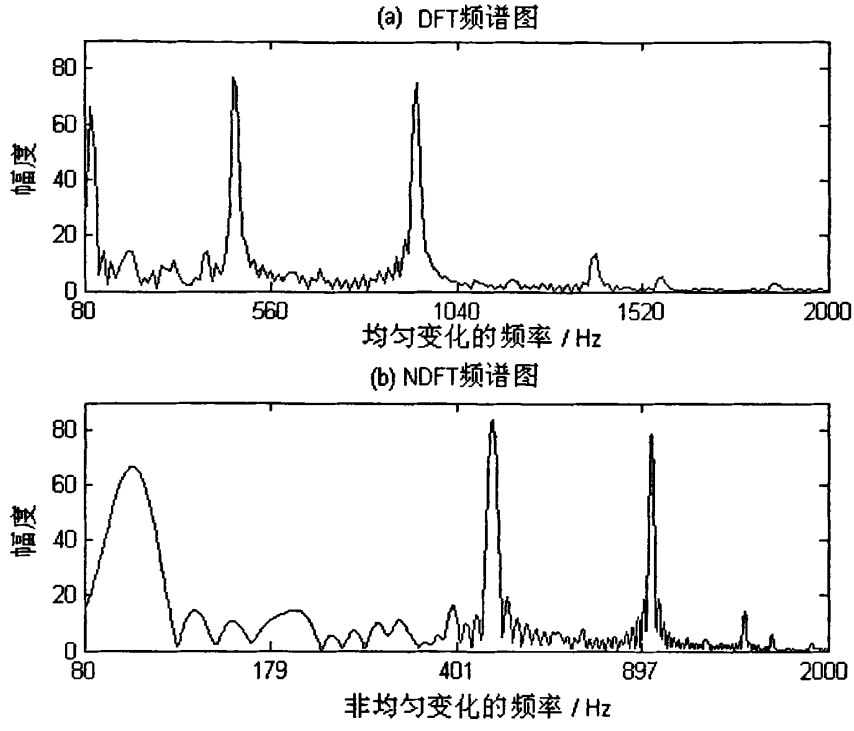

图 3-2 DFT 频谱图与 NDFT 频谱图的比较

3.3.2. 计算谱包络

为了能准确找到歌声的音高，在这一步要计算 NDFT 功率谱的谱包络。这样做是为了去除频谱图中不必要的极大值点——峰值点，保证下一步搜寻二分频和倍频的精确度。找出功率谱中所有的峰值，并对所有峰值点进行插值运算，重新得到 1024 个点。然后将所有的点连起来，就得到了一个初步的包络曲线，如图 3-3（a）所示。需要注意的问题是，若相邻两个峰值之间不存在第三个峰值，在能量都比较大的情况下，这样的处理会丢失某些重要的峰值，如图 3-3（b）所示。为了避免这种情况，我们在原始频谱中挑选出能量最大的 5 个峰值点，5 个峰值应该包含我们所需要的信息。然后在得到的初步包络曲线中查找这五个峰值是否存在，如果不存在，则在包络中恢复丢失的峰值。具体方法为，如果某个峰值丢失，那说明相邻处有一个能量较大的峰值，且两者之间不存在极大值，而是存在一个极小值。那么我们找到丢失峰值的位置后，去掉丢失峰值与极小值之间的包络值，用拟合的

方式代替。同理，对极小值与相邻峰值之间也进行相同的处理。这样就可以得到比较理想的包络曲线，如图 3-3（c）所示。

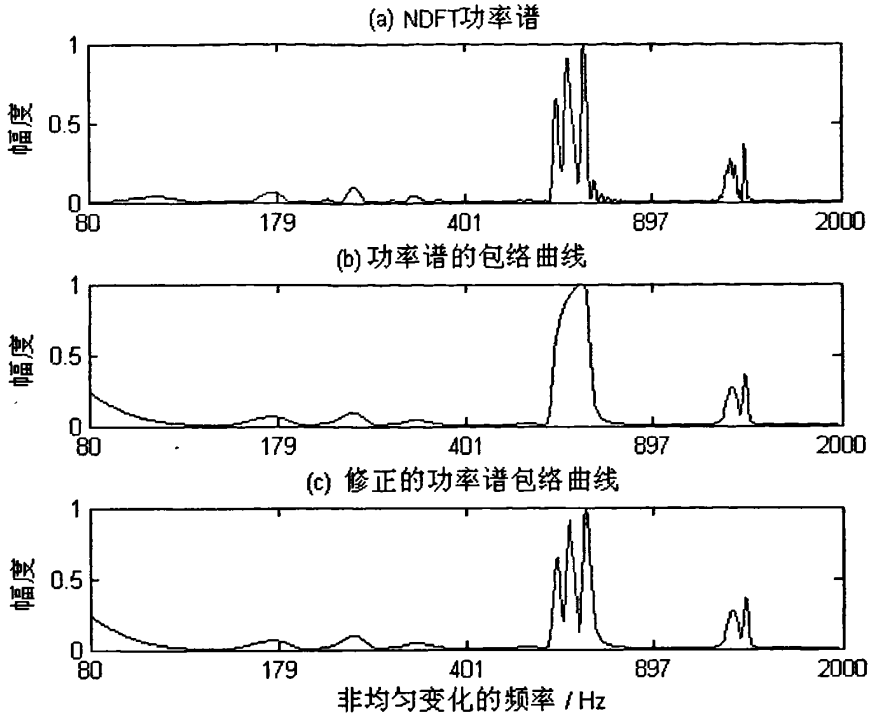

图 3-3 计算谱包络的过程

3.3.3. 主音高的搜索

在语音中，浊音信号可能包含三四十次谐波分量，而基波分量往往不是最强的分量。因为语音的第一共振峰通常在 300~1000Hz 范围内，这就是说，2~8 次谐波成分往往比基波分量还强。丰富的谐波成分使语音信号的波形变得复杂，给基音检测带来困难，经常发生基频估计结果为实际基频的二三次倍频或二次分频的情况。这种情况在歌声中也同样存在，但是我们可以将其作为歌声音高检测的一个优势。在音乐中，人声和乐器声音的谐波结构有所不同。尤其是从第 2 到第 5 个谐波，人声的谐波功率比乐器要高。从图 3-4 中，我们可以明显地看到。在图 3-4（a）歌声的谐波结构中，一次谐波能量要比基波能量高。在图 3-4（b）歌声的谐波结构中，一次谐波能量比基波能量低。在图 3-4（c）钢琴的谐波结构中，一次谐波能量要比基波能量小很多。

同样，在图 3-4（d）长笛的谐波结构中，一次谐波能量也比基波能量低。因此，我们就根据谐波和基波能量的比值来确定歌声的音高。

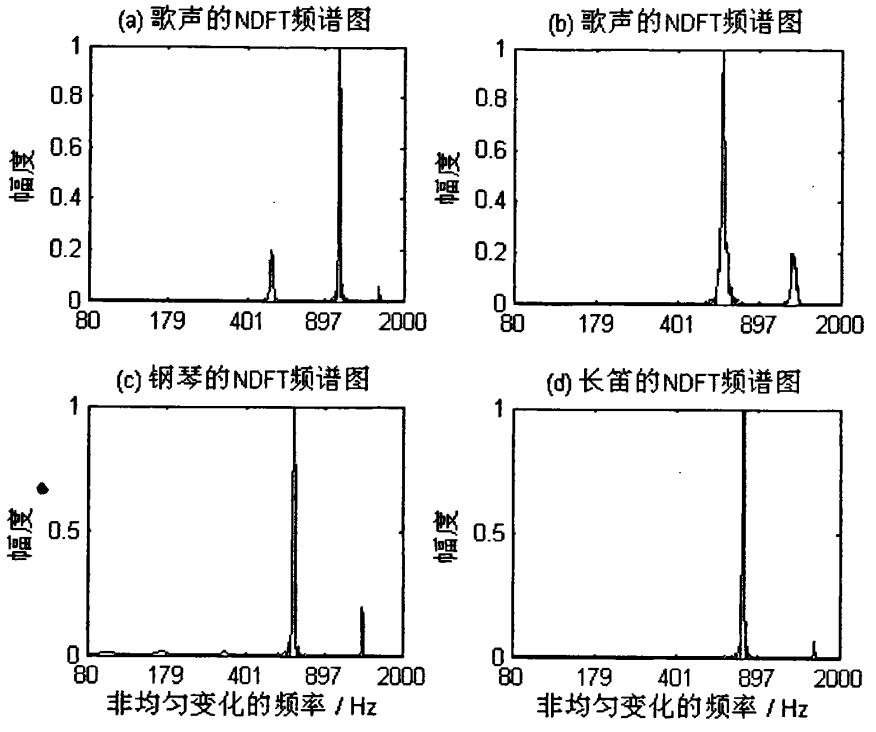

图 3-4 歌声、钢琴和长笛的 NDFT 频谱图比较

首先，我们从上一节得到的谱包络上选取能量较大的峰值点所对应的频率作为计算歌声音高的待选频率。能量门限设定为 E_{th}=0.2，对于大于该能量门限的峰值，通常已经包含了歌声的基波以及几次谐波。那么，就可以保证我们在进行频率搜索时准确地找到歌声的音高。对能量大于 E_{th} 的 n 个峰值按从大到小排列，表示为 P_i（i=1,2,…,n），各个峰值所对应的频率表示为 F_i（i=1,2,…,n）。这样做是将能量最大的峰值所对应的频率作为主音高的首选频率。接着搜索每个频率的二倍频和二次分频所对应的峰值。我们将以图 3-5 中的谱包络为例，详细介绍对二倍频和二次分频进行搜索的情况。

在谱包络中，大于能量门限 E_{th} 的峰值有三个，分别为 P_1=1，P_2=0.87 和 P_3=0.4，所对应的频率分别是 F_1=1146Hz，F_{21}=573.5Hz，F_{31}=660.7Hz。分别对这三个频率进行搜索，具体步骤如下：

（1）最大峰值 P_1=1，F_1=1146Hz。假设其二倍频为 2292Hz，二次分频为 573Hz。在图 3-5 中所有的峰值点中，没有与 2292Hz 接近的峰

值。因为我们的 NDFT 最大的频率为 2000Hz，这样就可以减少计算量，提高计算效率。因此，不需要对其二倍频进行搜索，直接将 P_{12} 和 F_{12} 的值置为 0。与 573Hz 最接近的是 573.5Hz 所对应的峰值，两者的差值为 0.5，小于 10Hz。然后我们寻找与 F_1 的二分之三倍频 1719Hz 最接近的频率对应的峰值，结果为 1719.6Hz，两者的差值为 0.6，也小于 10Hz。因此，我们认为 573.5Hz 是 F_1 的二分频，F_{10}=573.5Hz，P_{10}=0.87。反之，不予考虑。这样我们就得到了两个向量，频率向量 $[F_{10},F_1,F_{12}]=[573.5,1146,0]$ 和能量向量 $[P_{10},P_1,P_{12}]=[0.87,1,0]$。

（2）第二个峰值 P_2=0.87，F_2=573.5Hz。假设其二倍频为 1147Hz，二次分频为 286.8Hz。重复执行步骤（1）。在搜索二倍频时，找到了一个与 1147Hz 接近的峰值，其频率为 1146Hz，两者的差值小于 10，因此我们认为二倍频存在，F_{22}=1146Hz，P_{22}=1。而在搜索二次分频时，与 286.8 最近的峰值所对应的频率为 246，两者差距大于 10，因此我们认为其二次分频不存在。最后得到的频率向量 $[F_{20},F_2,F_{22}]=[0,573.5,1146]$，能量向量 $[P_{20},P_2,P_{22}]=[0,0.87,1]$。

（3）最后一个峰值 P_3=0.4，F_3=660.7Hz。重复执行步骤（1）。最后得到的频率向量 $[F_{30},F_3,F_{32}]=[660.7,1328.6,0]$，能量向量 $[P_{30},P_3,P_{32}]=[0.4,0.05,0]$。

我们得到了三个频率向量和能量向量。下面就是如何从这三个向量中选出我们需要的歌声音高。首先在三个频率向量中挑选至少包含两个频率值不为 0 的向量。然后根据所选频率向量的第一个值，检查在频率值之间是否存在成整数倍的情况。若存在这种情况，则只保留最小的频率。经过这样的挑选，还剩下两个向量 $P=[P_{10},P_1;\ P_{30},P_3]$，$F=[F_{10},F_1;\ F_{30},F_3]$。接下来是计算一次谐波与基波能量的比值。

$$R1=\frac{P_1}{P_{10}},\qquad R2=\frac{P_3}{P_{30}} \tag{3-6}$$

根据歌声与乐器谐波结构的差别，通过比较 $R1$ 和 $R2$，确定歌声的音高。

（1）$R1>1$，$R2>1$。在这种情况下，说明两个频率的一次谐波能量都要比基波能量大。通过上面分析可知，对歌声而言这属于正常情况，但是对乐器来说这是不正常的。造成这种情况的原因在于，在谱包络中选取的能量较大的频率分量时，可能包含了乐器的基频，但是

在搜索阶段又得到了它的二分频。这样的话，得到的二分频实际上是错误结果，可能是谱包络中二分频处正好存在的一个峰值，那么一般情况下该峰值的能量就比较低。而歌声也存在基波比一次谐波能量低的情况，但是相差是有限的。所以，如果 $R1$ 和 $R2$ 都大于 1，我们将较小的那个值确定为歌声的音高。

（2）$R1>1>R2$ 或 $R1<1<R2$。当 $R1$ 和 $R2$ 中只有一个大于 1 时，我们既要比较 $R1$ 和 $R2$ 的大小，还要给大于 1 的比值加一个限制条件。就像情况（1）所说的，若大于 1 的比值过大，可能是由于错误搜索到的二分频引起的。因此，我们为这种情况加一个门限值 R_{th}。经过实验结果的比较，我们最后选择 $R_{th}=10$。若大于 1 的比值小于 R_{th}，表示其代表歌声，否则选择小于 1 的比值。

（3）$R1<1$，$R2<1$。当 $R1$ 和 $R2$ 都小于 1 时，需要依据歌声的谐波能量变化比乐器的缓慢来判断哪一个属于歌声。也就是说，我们选择较大的那个比值为歌声所对应的谐波。

通过以上的三步，我们就可以确定歌声的音高。针对图 3-5 进行分析，得到 $R1=1.156$，$R2=0.134$。这属于第二种情况，因此我们最终确定 $F_{10}=573.5$Hz 是歌声的音高。对要分析的一段音乐信号所有帧进行以上的处理，就可以最终得到整段音乐中歌声的音高。图 3-6 表示的是我们对一段音乐进行主音高检测的结果，‘x’线对应的是歌声的实际音高，点线对应的是检测的音高。在误差允许范围内，也就是我们假设检测到的音高与实际音高的差值小于 5Hz，我们的基于 NDFT 的主音高方法正确率可达到 84.8%。这样有效的音高检测结果，为下一步要做的分离工作提供了保障。

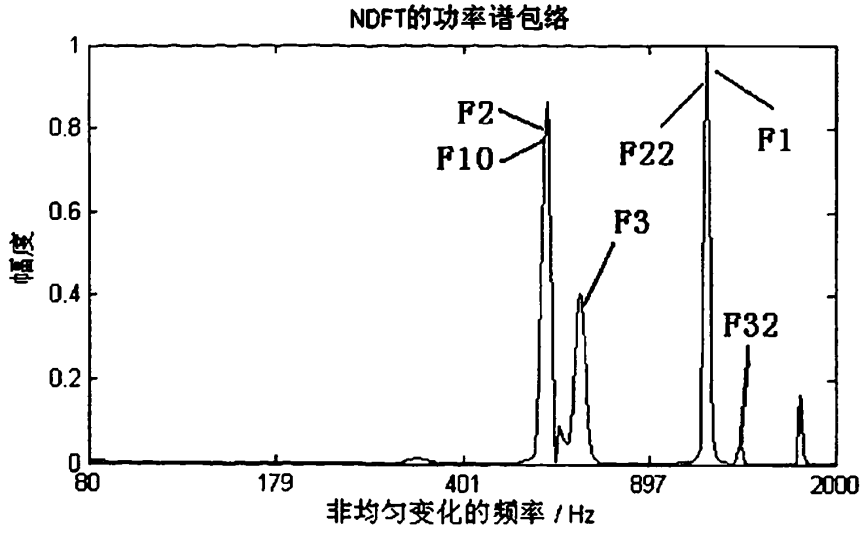

图 3-5 主音高的确定过程

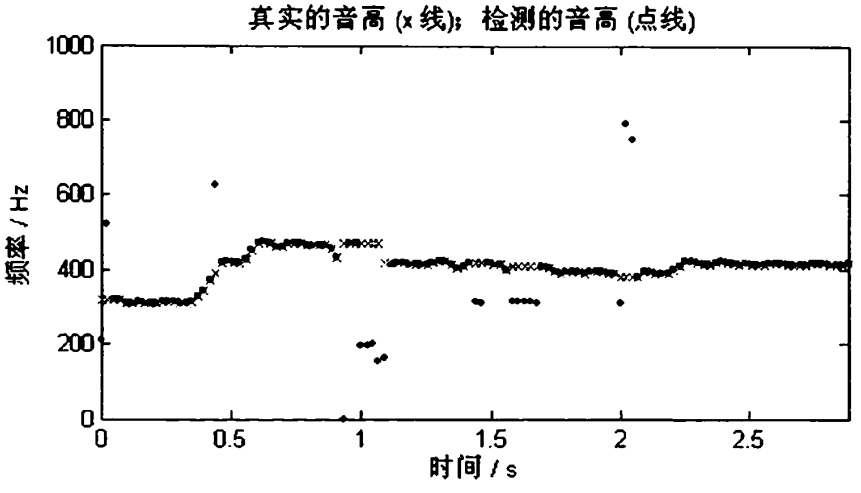

图 3-6 真实音高与检测音高的比较

第四章 歌声分离系统

我们的歌声分离系统由四个部分组成，分别是T-F分解、主音高检测、歌声T-F信息的提取以及歌声的合成。如图4-1所示，这是歌声分离系统的基本框图。根据T-F分解过程的不同，我们提出了两种分离方法，基于STFT的歌声分离和基于Gammatone滤波器的歌声分离。

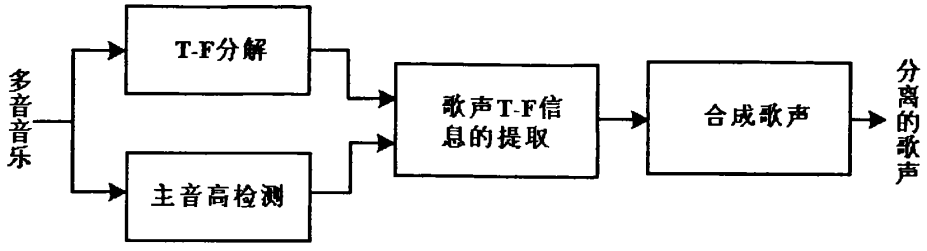

图 4-1 歌声分离系统的基本框图

4.1. 基于 STFT 的歌声分离

STFT 是研究非平稳信号最广泛使用的方法。利用 STFT 对乐音进行分析，其 T-F 结构很明显。而且 STFT 其实是加窗的傅里叶变换，可以很容易进行反变换。基于 STFT 的歌声分离系统是根据检测到的主音高，将每一帧信号的频谱中的各次谐波结构提取出来。对提取的歌声的 STFT 进行反变换，并对变换后的时域波形进行处理，就得到要分离的歌声。这是一种比较传统的方法，但是多数情况下用于含噪语音的分离。将这种方法用于歌声的分离，其优点是乐器的基频及其谐波结构是固定的，不像其它噪声存在于整个频域。

4.1.1. T-F 分解

由于可以认为音乐信号是局部平稳的，所以可以对某一帧音乐信号进行傅里叶变换，即STFT。在2.1节我们已经介绍了连续信号的STFT，当我们要在计算机上实现一个信号的STFT时，该信号必须是离散的。离散STFT的定义为

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m} \qquad (4-1)$$

式（4-1）中 $w(n-m)$是窗口函数序列。同样，不同的窗口函数序列，将得到不同的傅里叶变换结果，我们选择的是汉宁窗。由该式知，STFT有两个自变量：n和ω，所以它既是关于时间n的离散函数，又是关于角频率ω的连续函数。那么，就可以将式（4-1）写为离散STFT的形式

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]e^{-j\omega m} \qquad (4-2)$$

由（4-2）式可以明显的看出，STFT是用一个时间宽度足够窄的固定的窗函数乘时间信号，使取出的信号可以被看成平稳的，然后对取出的这一段信号进行傅立叶变换，便可以反映出该时间宽度中的频谱变化规律，如果让这个固定的窗函数沿着时间轴移动，那就可以得到信号频谱随时间变化的规律了。

在语音的发音过程中，声道通常都处于运动状态，这个运动状态的时变过程同振动过程相比要缓慢得多，因此一般假设语音信号是一种短时平稳信号，在一个很短的时间内（10ms~30ms）是相对平稳的，但在长时间的周期中语音信号的特性会发生变化，这种变化的不同决定了产生语音的不同。音乐信号与语音有类似的特性，一个音符的持续时间要比语音中的一个音要长。一般而言，音乐中的歌声是由声带的准周期振动，经声道共鸣调制，由口鼻辐射出来。不同音符的音色反映在不同的频谱结构中。我们看图 4-3（a），这是一段歌声与钢琴的混合声音的 STFT 图。从图中我们可以清楚的看到，歌声与钢琴随时间的变化，以及在每个时间段内的谐波结构。其中选择的窗长为46.4ms，帧与帧之间的重叠部分是 23.2ms。

4.1.2. 歌声短时谱的提取

无论人声、歌声，还是乐器的声音，它们都不是一个单音，而是一个复合音。也就是由声音的基音和一系列的泛音所构成。这些泛音都是基音频率的倍数。它对音色的特性有非常重要的影响。这些泛音的数量和泛音幅度的不同构成音色的频率特性曲线。这条曲线就体现

了音色的表现力。泛音在频谱上表现为局部峰值会在基频的整数倍反复出现，并随着倍数的增加而衰减。对于一个理想的和谐泛音，频率的泛音部分峰值总是出现在基频的整数倍处。然而，如果经过仔细测量，我们会发现现实世界中某些声音的和谐泛音分布与理想的和谐泛音存在一定误差[34]。例如，对于弦乐乐器，其泛音局部峰值会出现在

$$f_h = hF_0\sqrt{1+\beta(h^2-1)} \tag{4-3}$$

其中，F_0为基频，h为泛音倍数，而β为非和谐因子。当h值较小（≤ 15）时，和谐泛音与理想和谐泛音局部峰值出现位置差别可以忽略。我们所知的大多数乐器以及歌声都满足或近似为泛音局部峰值理想分布。

短时谱提取的作用是对信号进行筛选，保留只属于歌声的特定频段。我们对一帧音乐信号进行傅里叶分析，从图4-2（a）中可以看出，音乐信号具有明显的谐波结构。当我们用1024个点，采样频率为11025Hz，进行傅里叶变换时，其各次谐波的所占的频率宽度大概在100Hz左右。如果在图4-2（a）中，我们只选择谐波上的能量，其余值设为0，如图4-2（c）所示。对其进行傅里叶反变换，也可以得到一个时域信号。图4-2（b）是对原始频谱进行反变换得到的时域波形。图4-2（d）是对经过挑选后的频谱进行反变换得到的时域波形。从图中可以看出两者几乎一样，没有太大的差别。所以，我们将各次谐波的频率定义为

$$f_h = hF_0 \tag{4-4}$$

在第三章我们已经得到了一段混合音乐的音高，也就是每一帧信号的基频确定了。然后根据公式（4-4），提取图4-3（a）中每帧信号的各次谐波，就得到图4-3（b）所示的T-F结构。从图中可以看出，一些乐器的T-F信息被去除了，但是某些帧由于音高检测的不准确，还会保留一些不属于歌声的信息。不过，我们将该T-F结构认为是歌声的短时谱，用来对歌声进行合成。

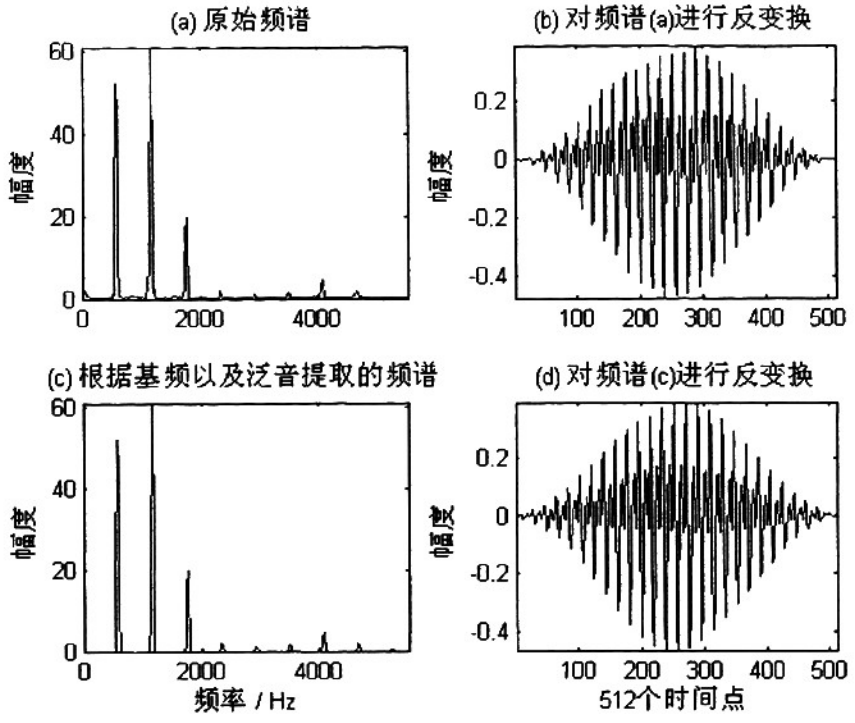

图 4-2 不同的频谱结构进行傅里叶反变换的波形比较

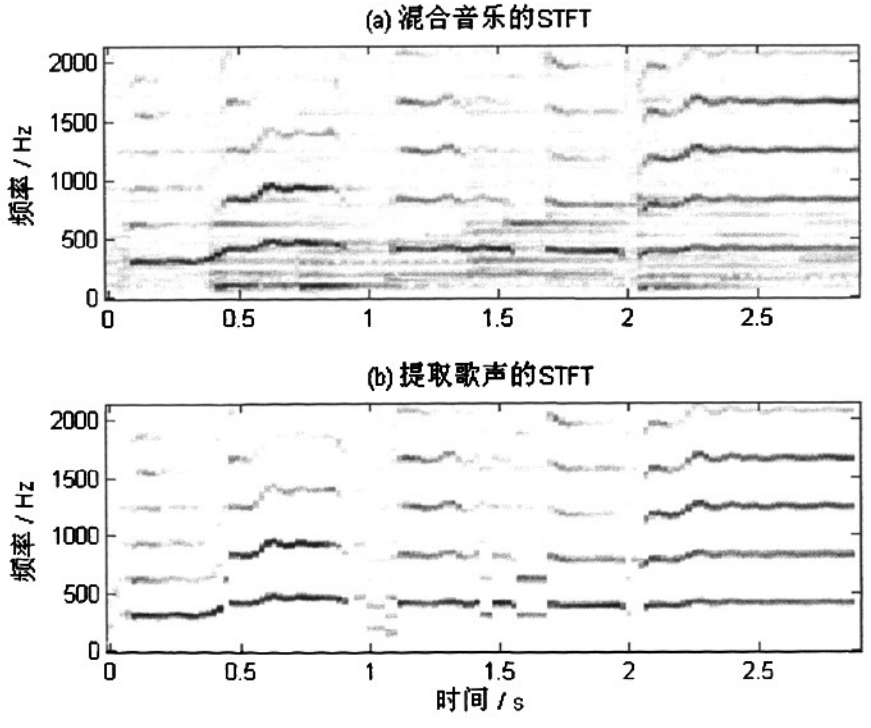

图 4-3 提取歌声的短时谱

4.1.3. 歌声的合成

因为频谱图是 STFT 的绝对平方频率，常常认为有相位损失，因此信号不可能由其恢复。这个论点是不正确的，因为损失的相位是 STFT 的相位，不是信号的相位。事实上，信号的相位和幅度也都出现在 STFT 的相位和幅度之中。因而，有 STFT 幅度可以足够用来恢复信号[35]。由 $X_n(e^{j\omega})$ 恢复 $x(n)$ 通常称为短时综合。两种经典的方法是滤波器组求和法（FBS）和叠加相加法（OLA）。

FBS 是一种传统的短时综合方法，这种方法与 STFT 的滤波器组表示有关，可将 STFT 认为是滤波器组的一列输出。在 FBS 法中，每个滤波器的输出被一个复指数信号所调制，而这些已调制的滤波器输出在每个时刻相加，以得到原始序列相应的时间样点。OLA 方法来源于 STFT 的傅里叶变换观点。实际上，由傅里叶变换观点得到的最简单的方法并非 OLA 方法，而是被称为逆傅里叶变换（IDFT）的方法。在这种方法中，在每个固定时刻，对相应的频率函数做 IDFT，并将所得结果除以分析窗。但是该方法一般不适合实际应用，因为 STFT 的轻微扰动就可能导致合成信号与原始信号大相径庭。在 OLA 法中，对每个固定时刻的离散 STFT 做 IDFT。然而，这里不是将所得到的每个短时段都除以分析窗，而是在短时段之间采用重叠并相加的处理。若分析窗的设计使得重叠相加处理能够从合成序列中有效去除该分析窗，则此方法有效[36]。由上面的分析可知，FBS 法依赖于频域的抽样关系，而 OLA 法依赖于时域上的抽样关系。在 FBS 法中，每个时间样点通过将滤波器输出相加而得到，而 OLA 方法的每个时间样点通过不同短时段相加而得到。

在上一节我们已经得到了要分离的歌声的短时谱，这一节要做的就是如何从短时谱恢复出歌声信号。由于我们已经得到了很好的短时谱，所以源自 STFT 的傅里叶变换观点的 OLA 方法更适合对歌声进行合成。OLA 方法源自序列与其离散时间 STFT 之间的如下关系：

$$x(n)=\frac{1}{2\pi W(0)}\int_{-\pi}^{\pi}\sum_{p=-\infty}^{\infty}X(p,\omega)e^{j\omega n}d\omega \qquad (4-5)$$

其中

$$W(0)=\sum_{-\infty}^{\infty}w(n)$$

给定一离散 STFT $X(n,k)$，OLA 法通过下式合成序列 $y(n)$：

$$y(n)=\frac{1}{W(0)}\sum_{p=-\infty}^{\infty}\left[\frac{1}{N}\sum_{k=0}^{N-1}X(p,k)e^{j\frac{2\pi}{N}kn}\right] \tag{4-6}$$

其中方括号内是 IDFT，且其对每个 p 有：

$$f_p(n)=x(n)w(p-n) \tag{4-7}$$

前提是 DFT 的点数 N 大于窗长 N_w，即 DFT 求逆不存在混叠。所以 $y(n)$ 的表达式变为：

$$y(n)=x(n)\left(\frac{1}{W(0)}\right)\sum_{p=-\infty}^{\infty}w(p-n) \tag{4-8}$$

若离散 STFT 以因子 L 进行了时间抽样，可证明：若分析窗满足

$$\sum_{p=-\infty}^{\infty}w(pL-n)=\frac{W(0)}{L} \tag{4-9}$$

则 $x(n)$ 可由如下关系式综合得到：

$$x(n)=\frac{L}{W(0)}\sum_{p=-\infty}^{\infty}\left[\frac{1}{N}\sum_{k=0}^{N-1}X(pL,k)e^{j\frac{2\pi}{N}kn}\right] \tag{4-10}$$

式（4-9）是 OLA 方法对分析窗的一般性约束条件。

根据 OLA 法，首先对每个固定时刻的离散 STFT 做 IDFT。我们选择其中的一帧信号进行分析。在进行 STFT 处理时，每两帧信号是重叠的，重叠为它们长度的一半。因此，当我们得到连续三帧信号的时域波形后，要恢复出歌声信号采用重叠并相加的处理。也就是前一帧的后半段与后一帧的前半段是重叠部分，我们将这两部分进行相加。图 4-4（a）是第 $n-1$ 帧信号的反变换波形，图 4-4（b）是第 n 帧信号的反变换波形，图 4-4（c）是第 $n+1$ 帧信号的反变换波形。第 $n-1$ 帧的后半段与第 n 帧的前半段叠加，同时第 n 帧的后半段与第 $n+1$ 帧的前半段叠加，将两个叠加结果合并为一帧，就是图 4-4（d）所示的波形，也就是合成的第 n 帧歌声的时域波形。以此类推，对每一帧的 STFT 都做类似的处理，就可以得到分离的歌声信号，如图 4-5 所示。图 4-5（a）为混合的音乐信号，4-5（b）为分离的歌声。

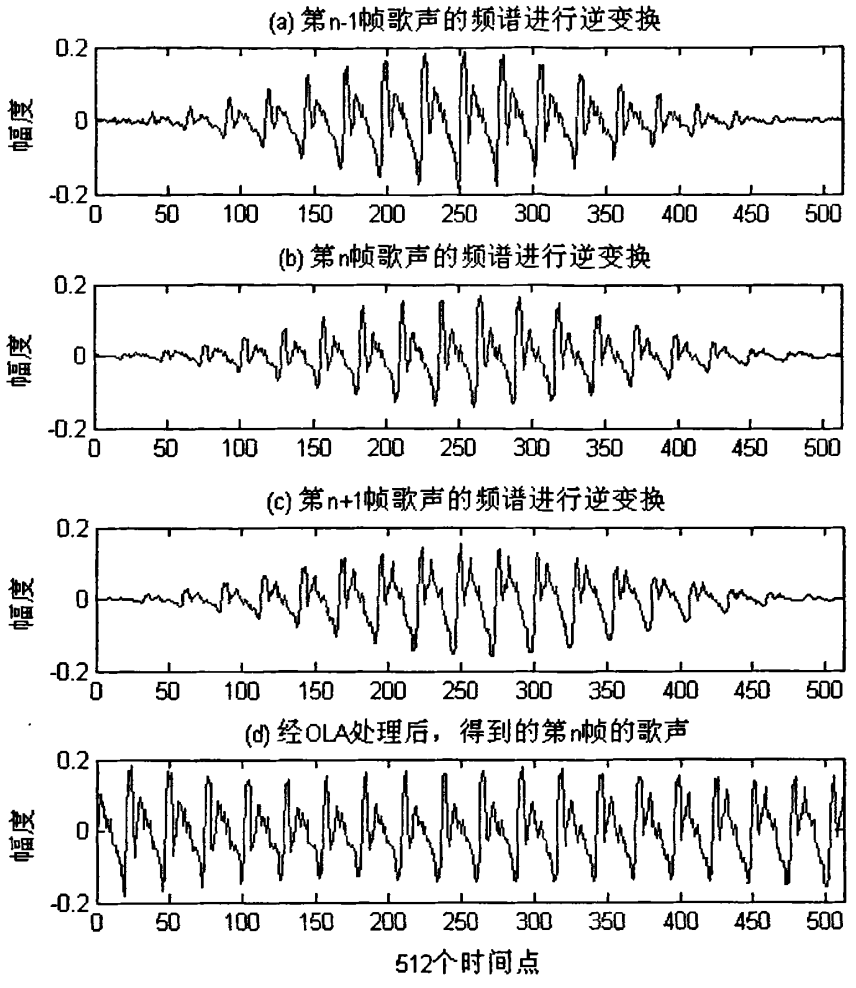

图 4-4 经 OLA 处理得到某帧信号的时域波形

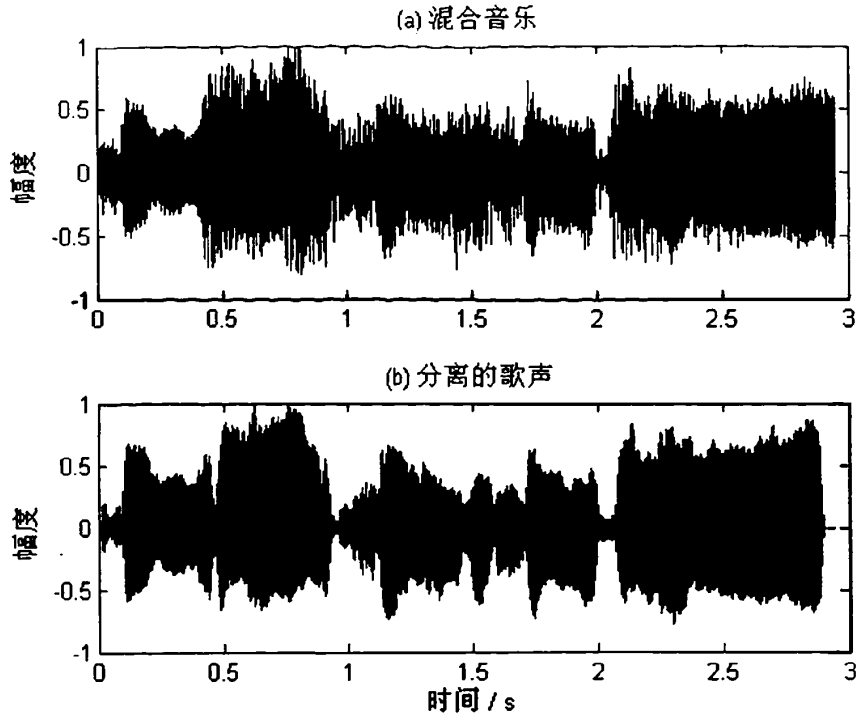

图4-5 混合声音与分离歌声的时域波形比较

4.2. 基于 Gammatone 滤波器组的歌声分离

基于听觉滤波器的歌声分离系统是用一个听觉滤波器将一个输入分解成 T-F 单元，再利用主音高对所有 T-F 单元进行分割和分组，将与主音高一致的那些单元归为歌声，其余的为背景音乐。然后根据端点检测的结果进行重新分组。最后将各通道相加即可得到分离的歌声。

4.2.1. T-F 分解

在 2.4 节我们介绍了 Gammatone 滤波器，在这一节我们就是利用 Gammatone 滤波器对原信号进行 T-F 分解。首先将输入的混合信号通过一个 Gammatone 波器组[37]。N是滤波器的个数，我们取 N=128，即 128 个 Gammatone 滤波器即可较好地模拟人耳的听觉特性。滤波器组所覆盖的频率范围从 80Hz 到 5000Hz，对于乐音而言，这已经足够了。n是滤波器的阶数，我们取 n=4。各个 Gammatone 滤波器的中心频率按照 ERB 的关系，在 80Hz 到 5000Hz 之间呈准对数分布。Gammatone

滤波器的中心频率越高，滤波器的带宽就越宽。当输入信号通过这样一个滤波器组后，就可以得到 128 通道的 T-F 信号。

然后对每个滤波器通道的输出进行分帧，时间帧的帧长为 46.4ms，帧移 23.2ms。经过带通滤波和加窗后，输入信号就被分解为一个二维的 T-F 表示，或称为一个 T-F 单元集合。图 4-10（a）展示一段音乐信号通过 Gammatone 滤波器组后的 T-F 图。

4.2.2. 听觉特征提取

当我们得到了 T-F 分布图时，还需要注意一个重要的信息，同一通道的相邻帧在时域特性上是相似的，而相邻通道的同一帧在频域特性是相似的。根据这一特点，接下来提取下列特征：相关图谱、通道互相关，以及进行端点检测。

（1）相关图谱

相关图谱是基音信息从低频和高频谐波区域相结合得到的一种基音模型。相关图谱在听觉外围和乐音分离之间提供了一个有效的中级听觉。该听觉表示为[21]

$$A_H(c,m,\tau)=\frac{1}{N_c}\sum_{n=0}^{N_c-1}h(c,mT-n)h(c,mT-n-\tau) \tag{4-11}$$

$h(c,n)$表示通道 c在时间步长为 n时的滤波器输出。此处，N_c=512 对应 46.4ms 的帧长，T=256 对应 23.2ms 的帧移。图 4-6 就是某一帧信号的相关图。

（2）通道互相关

由于相邻滤波通道的互相关可以表示该滤波通道是否属于同一声源，可以为后面的分割提供一种有效的特征。因此，对同一 T-F 单元计算各相邻通道间的互相关，以此衡量通道之间的相关程度。通道 c和 c+1之间的通道互相关定义如下：

$$C_H(c,m)=\sum_{\tau=0}^{L-1}\hat{A}_H(c,m,\tau)\hat{A}_H(c+1,m,\tau) \tag{4-12}$$

其中，$\hat{A}_H(c,m,\tau)$表示 $A_H(c,m,\tau)$进行归一化得到的零均值和单位方差的自相关函数。图 4-7 是一段音乐信号的通道互相关图。

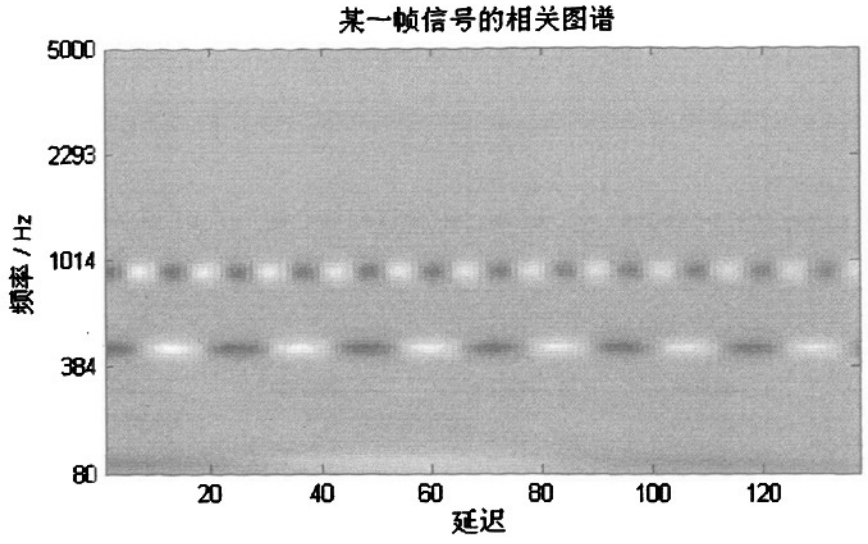

图 4-6 相关图

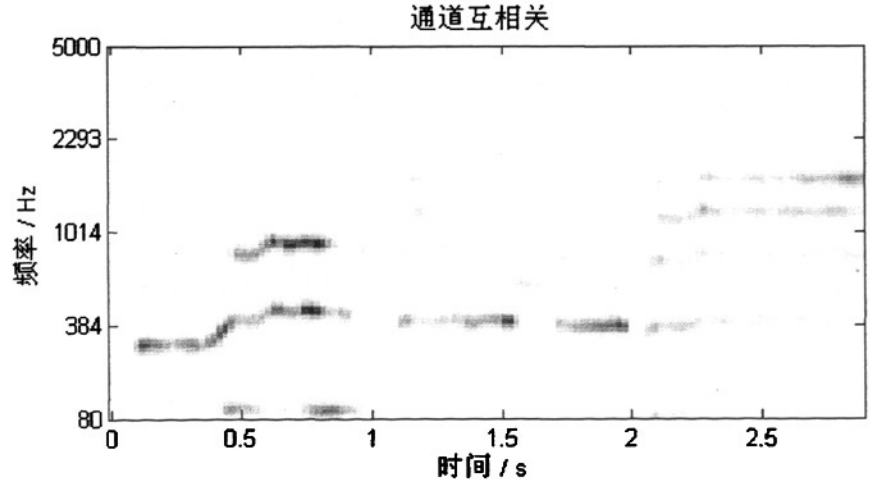

图 4-7 通道互相关

（3）端点检测

乐音信号具有明显的时间连续性，各个音符之间的时域波形在节奏点上幅度大、能量相对集中，而在起始以及结束部分能量变小。根据乐音信号这些特性，我们要为听觉滤波后的 128 个通道输出进行端点检测。相比较对原始信号进行检测，对各个通道进行端点检测会更准确。因为听觉滤波后，各个通道基本对应的是只有一个频率的准正弦信号，每一段波形都属于某一特定声源。我们的端点检测是基于音符的平均能量。首先计算时域信号的平方值，得到时域内大于或等于 0

的能量，然后寻找平方波形的包络。包络确定后，为了检测的准确性，使包络中的峰值点更突出，还要计算一个修正的包络。然后根据音符的平均能量得到端点检测信息。如图 4-8 所示，展示了某一通道的端点检测结果。

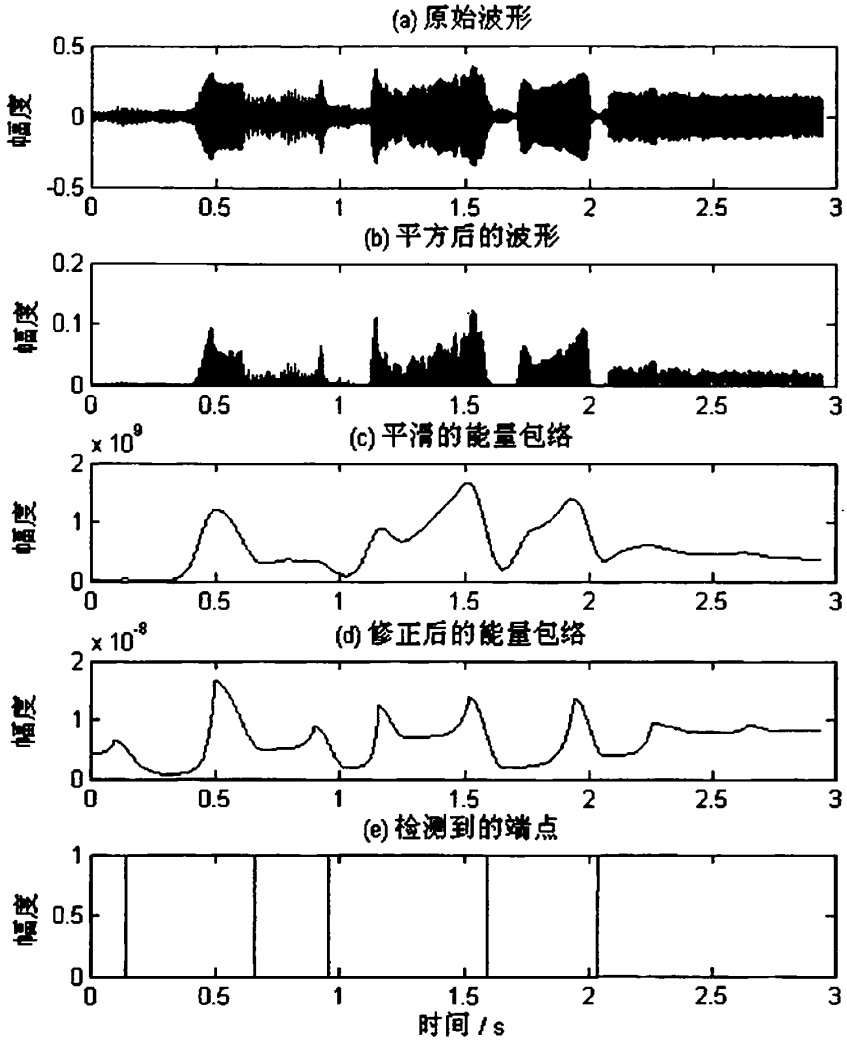

图 4-8 端点检测的过程

4.2.3. 歌声听觉 T-F 信息的提取

该方法分离歌声由三个过程组成：分割、分组和重新分组。分割是对通道互相关特征进行处理，将具有高互相关的 T-F 单元进行合并

形成片段。分组则是在分割完成的基础上，根据主音高将片段归为歌声还是乐器。重新分组是利用检测到的端点信息，对已经分组的各个片段进行修正。

（1）分割

由于相邻通道的频带有很大的重叠，一个确定的谐波通常影响相邻通道，就会产生通道之间的高互相关性。因此，我们根据通道互相关来合并 T-F 单元形成片段。图 4-9（a）是对某一帧信号提取的通道互相关值。从图中可以看出，整个波形有极大值点，也有极小值点，这体现了某些通道之间的相关性。也就是说两个相邻极小值点之间的这几个通道具有高的相关性，代表同一声源。根据这一特点，我们对通道互相关进行分割。图 4-9（b）展示了分割的结果。

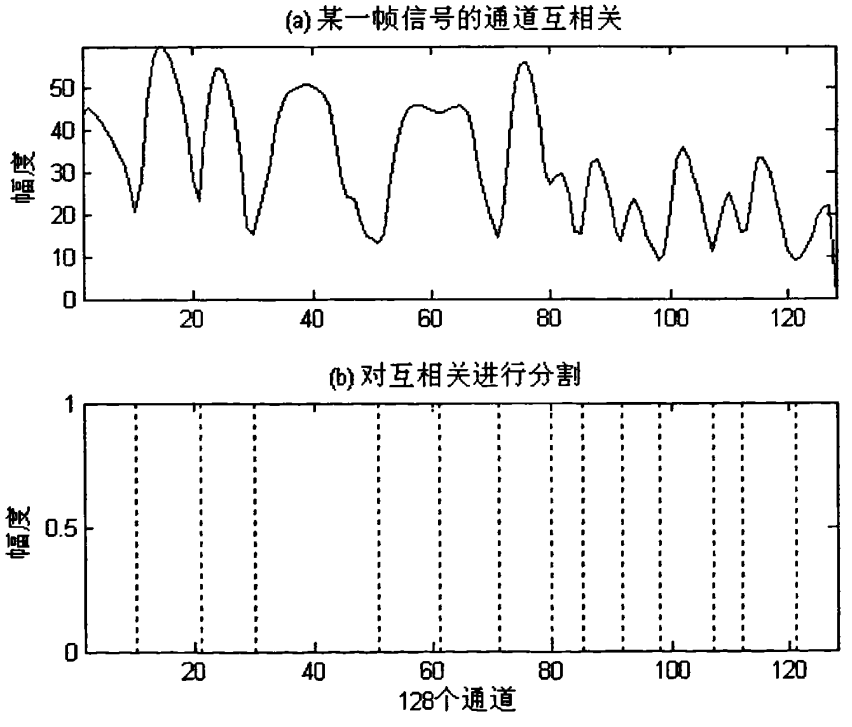

图 4-9 对互相关进行分割的结果

（2）分组

上一节完成了对通道互相关的分割，将 128 个通道分成了几个不同的片段。接下来要利用第三章得到的主音高将这些片段分组，形成一个初始的前景流和一个初始的背景流，两者分别粗略地对应歌声和背景音乐。分组就是对每个 T-F 单元的周期性与主音高进行比较。我

们利用相关图谱这一特征得到每个 T-F 单元的周期，也就是将每个单元自相关的最大值所对应的延迟作为该单元的基音周期。如果这个周期值与相应的主音高周期一致，或者这个周期值与主音高的倍数一致，那么我们就认为该单元与主音高是一致的。对于任何片段，如果该片段中的 T-F 单元有超过一半与主音高或主音高的倍数一致，我们就认为该片段属于歌声，否则就是属于背景音乐。因此就初步得到了属于歌声的所有片段，如图 4-10（b）所示。

（3）重新分组

初步得到了歌声的听觉 T-F 信息，但是还需要进一步的处理。因为我们检测到的主音高存在误差，那么在分组阶段可能会漏掉一些属于歌声的单元，反之会保留一些属于背景音乐的单元。为了进一步获得更准确的听觉 T-F 信息，我们还要进行重新分组。

因为乐音信号的时间连续性，一个音符的持续时间比较长。重新分组是根据检测到的端点信息，对所有的 T-F 单元进行修正。首先对每一通道两两端点之间的 T-F 单元进行检查，也就是检查每个单元的分组情况。如果两个端点之间的单元有超过三分之一被标注为歌声，那么两个端点之间的单元都归为歌声。否则，如果两个端点之间的单元为歌声的片段小于三分之一，将已经标注为歌声的单元去除。对每个通道的两两端点之间进行重复处理，最后就得到了我们需要的歌声的听觉 T-F 信息。图 4-10（c）是我们最终得到的属于歌声的听觉 T-F 信息。与（b）图相比较，恢复了在时间上某些不连续的片段，得到了更准确的歌声 T-F 信息。与（a）图比较，去除了一部分不属于歌声的信息，得到了很好的分离效果。

经过三步处理，我们已经得到了歌声的 T-F 信息。然后将图 4-10（c）各个通道进行叠加，即可得到分离的歌声。图 4-11（a）是输入的混合音乐信号，图 4-11（b）就是我们分离出的歌声的时域波形。

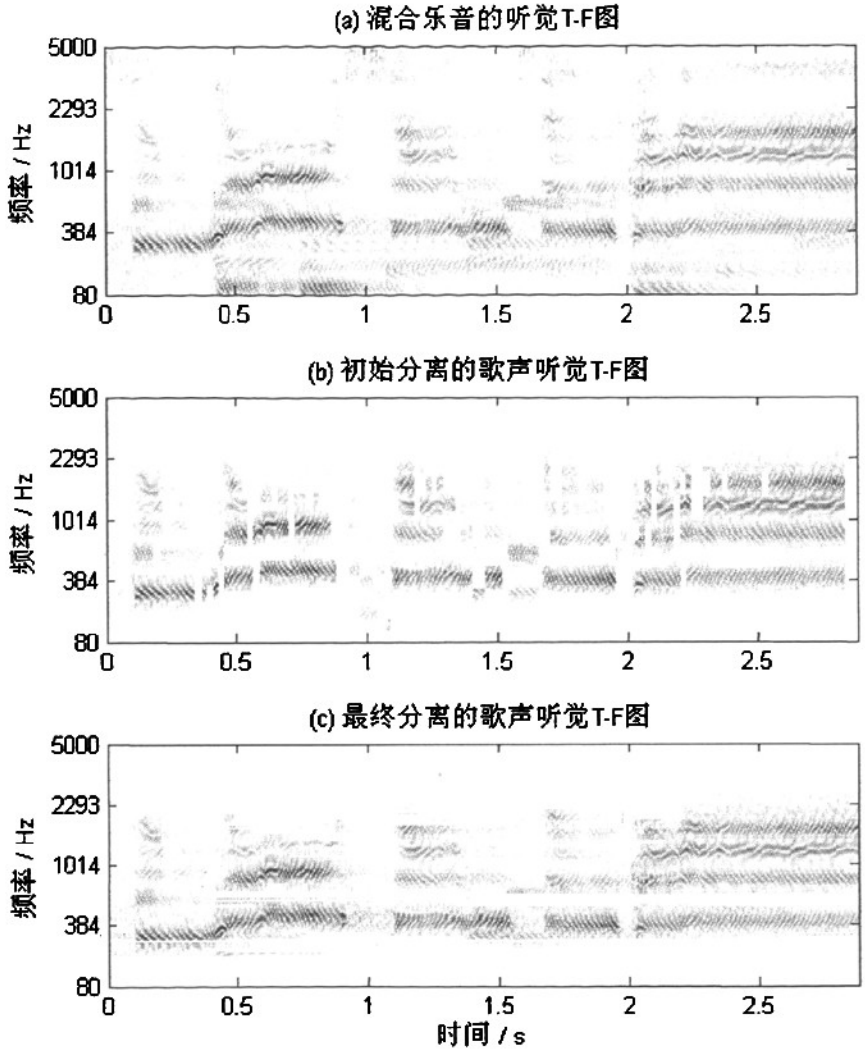

图 4-10 提取歌声的听觉 T-F 信息

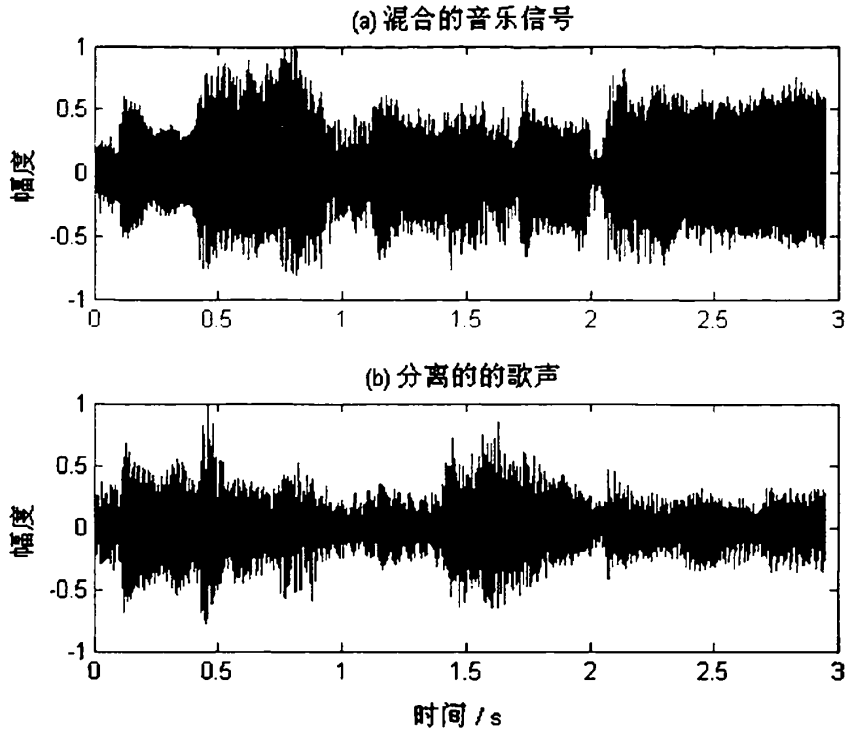

图 4-11 时域波形比较

4.3. 两种分离方法的比较

本章介绍了两种分离方法，基于 STFT 的歌声分离和基于 Gammatone 滤波器的歌声分离。下面分别分析这两种方法的优缺点。

第一种方法是基于 STFT 的歌声分离。这种分离方法的缺点是主要依赖于我们检测到的音高，若是某一帧的音高检测不准确，那恢复出来的时域波形也不正确。但这又是一种相对简单的分离方法。如果音高检测结果比较好，那该系统分离的歌声效果也会很好。

第二种方法是基于 Gammatone 滤波器的歌声分离。与前一种方法相比，该方法还使用了除音高之外的多个特征，可以更准确地得到歌声的 T-F 信息。但是这种方法也存在一个问题。基于 STFT 的分离方法要得到时域信号是利用傅里叶反变换，这是一种比较成熟的技术，能基本恢复原来的波形。而根据听觉滤波器方法得到的 T-F 信息，已经改变了原有的特性，很难恢复出很好的时域波形。这一点是该方法亟待解决的问题。

第五章 总结

本文对乐音信号的分离问题进行了研究，建立了基于T-F分析的歌声分离系统，可以实现乐音中歌声的有效分离。本文从音乐信号的T-F分析、主音高检测、基于STFT的歌声分离和基于Gammatone滤波器的歌声分离几个方面，介绍了歌声分离系统的实现过程。

T-F分析的基本任务是建立一个函数，要求这个函数不仅能够同时用时间和频率描述信号的能量密度，还能够以同样的方式用来计算任何密度。T-F分析的特点就是，既能表现信号随时间的变换，同时还能表现信号在某一时刻频率分布情况。所以我们将T-F分析作为歌声分离系统的基础。详细介绍了几种常见的T-F分析方法，主要有STFT、小波变换，WVD和Gammatone滤波器。通过两点衡量每种T-F分析方法是否适合于建立歌声分离系统。一是该T-F方法将时间信号映射到T-F域以后，还能否转换到时域。另一个是，该T-F方法能否很清楚的根据某一个或多个特征，提取要分离的歌声的T-F信息。经过分析，我们最终选择了STFT和Gammatone滤波器。

主音高检测在本课题中占了很重要的位置。因为检测结果的好坏直接影响到后面分离的效果。到目前为止，对单个基音检测的研究已经很成熟，但是对于多音音乐的研究还有很大的局限性。我们提出的基于NDFT的主音高检测方法，比较有效的检测到了多音音乐中歌声的音高。这主要是因为，在整个音高检测过程中，我们充分利用了歌声与乐器谐波结构的不同，提高了检测结果的正确率。不过，从分析过程中可以看出，对一些乐器声音较大的情况，或者是一个音符的结尾部分能量减小的情况，还不能非常准确地检测到歌声的音高。为了提高歌声分离的效果，还需进一步完善。

在基于 STFT 的歌声分离系统中，我们用 STFT 对信号进行 T-F 分解。从 STFT 的 T-F 图上，可以明显地看到属于歌声的那一部分 T-F 信息。也就是基于这个原理，我们利用音高作为分离歌声的唯一依据，得到歌声的 STFT。因为可以很方便的进行反变换，所以用 OLA 方法，

就可以由 STFT 重构出歌声信号。该方法简单有效，但是过于依赖于主音高特征。基于 Gammatone 滤波器的歌声分离系统，是用 Gammatone 滤波器作为 T-F 分解的依据，得到信号的 T-F 表示。由于滤波后的信号在时间上仍然是连续的，所以可以为每一帧信号提取多个特征，如相关图、通道互相关和端点。然后，将这些特征再结合主音高提取歌声的 T-F 表示。虽然用该方法合成的歌声不太理想，但是它结合了多个特征来提取歌声的 T-F 信息，是一种非常有潜力的研究方法。

实验结果表明，我们提出的基于 T-F 分析的歌声分离系统能有效地分离歌声。要想得到更好的分离效果，还需要对系统的各个环节进一步地完善。在选择 T-F 分析技术时，我们使用了 STFT 和 Gammatone 滤波器。虽然这两种方法有其优势，但是其它方法也不是完全不可行的，尤其是可以考虑其它的听觉滤波器。我们的分离过程主要依赖于主音高特征，很多情况下歌声与乐器的基音以及泛音之间会存在重叠现象，单独利用主音高是很难完全得到歌声的 T-F 信息。人类是通过音色辨别声音的，所以在听歌的过程中，我们能辨认出哪个声音是歌声，哪个是乐器声。也就是说，如果依靠音色作为分离的特征，也许会取得更好的效果。但是，现在对音色的研究还不够深入，仅有的一些音色特征也只能用于声音的识别。不过，我们还是期待有关音色的研究能取得更大的进展。还可以对整个分离过程的细节进行改进。比如基于 STFT 方法中的歌声合成，我们合成的歌声和原始声音是有差别的，主要是因为 OLA 方法也不能完全恢复歌声，需要设计更加合理的短时综合方法。而在基于 Gammatone 的方法中，最需要改进的也是合成歌声的过程，各通道直接相加得不到很好的时域波形，但目前还没有更好的方法。从整体上看，基于 T-F 分析的分离系统在很多方面都可以进一步完善，这也说明该方法拥有很大的发展空间。

参考文献

[1] Y. P Li, D. L. Wang. Separation of Singing Voice from Music Accompaniment for Monaural Recording [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1475-1487.

[2] C. K. Wang, R. Y. Lyu, Y. C. Chiang. An Automatic Singing Transcription System with Multilingual Singing Lyric Recognizer and Robust Melody Tracker [J]. Proceedings of 8th European Conference on Speech Communication and Technology, 2003: 1197-1200.

[3] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, A. Shenoy. Lyrically: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals [J]. IEEE Transactions on Audio, Speech and Language Processing, 2008, 16(2): 338-349.

[4] Shankar Vembu, Stephan Baumann. Separation of Vocals from Polyphonic Audio Recordings [J]. 6th International Conference on Music Information Retrieval, 2005: 337-344.

[5] 胡航. 语音信号处理[M]. 哈尔滨：哈尔滨工业大学出版社，2005.

[6] 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.

[7] 葛哲学，陈仲生. Matlab 时频分析技术及其应用[M]. 北京：人民邮电出版社，2006.

[8] 王文延，曾庆宁，李琴. 时频分析理论及其在非平稳信号处理中的应用[J]. 中国科技论文在线，http://www.paper.edu.cn.

[9] 陈世雄，宫琴. 常见的听觉滤波器[J]. 北京：北京生物医学工程，2008，27(1):94-99.

[10] Y. L. Li, D. L. Wang. Musical Sound Separation Using Pitch-Based Labeling and Binary Time-Frequency Masking [J]. IEEE ICASSP 2008: 173-176.

[11] Tuomas Virtanen. Sound Source Separation in Monaural Music Signals [D]. Ph.D. dissertation, Tampere University of Technology, 2006.

[12] M. R. Every, J. E. Szymansk. Separation of Synchronous Pitched Notes by Spectral Filtering of Harmonics [J]. IEEE Transactions on Audio, Speech, and Languages Processing, 2006, 14: 1845-1856.

[13] S. A. Abdallah. Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic [D]. Ph.D. dissertation, King's College London, Department of Electronic Engineering, 2002.

[14] D. D. Lee, H. S. Seung. Learning the Parts of Objects by Non-negative Matrix Factorization [J]. Nature, 1999, 401: 788-791.

[15] Paris Smaragdis. Non-negative Matrix Factor Deconvolution: Extraction of Multiple Sound Sources from Monophonic Inputs [J]. In International Conference on Independent Component Analysis and Blind Signal Separation, 2004: 494-499.

[16] 邱伟，黄秀轩．计算听觉场景分析介绍[J]．高技术通信，2002:106-110.

[17] David K. Mellinger. Event Formation and Separation in Musical Sound [D]. Ph.D. dissertation, Stanford University, 1991.

[18] D. Godsmark, G. J. Brown. A Blackboard Architecture for Computational Auditory Scene Analysis [J]. Speech Commun., 1999, 27(4): 351-366.

[19] Masataka Goto. A Real-time Music-scene-description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals [J]. Speech Commun., 2004, 43(4): 311-329.

[20] Y. Meron, K. Hirose. Separation of Singing and Piano Sounds [J]. In Proc. 5th Int. Conf. Spoken Lang. Process., 1998.

[21] G. N. Hu, D. L. Wang. Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation [J]. IEEE Transactions

on Neural Networks, 2004, 15(5): 1135-1150.

[22] 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.

[23] 胡广书．现代信号处理教程[M]．北京：清华大学出版社，2004.

[24] 于拾全，景新幸，刘志国．乐器音高检测方法的比较和精度分析[J]．电声技术，2006，7:100-104.

[25] Masataka Goto. PreFEst: A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals [J]. Proceedings of the 18th International Congress on Acoustics, 2004: 1085-1088.

[26] A. P. Klapuri. Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness [J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(6): 804-816.

[27] Y. P. Li, D. L. Wang. Detecting Pitch of Singing Voice in Polyphonic Audio [J]. IEEE Acoustics, Speech, and Signal Processing, 2005, 3: 18-23.

[28] M. Y. Wu, D. L. Wang. A Multipitch Tracking Algorithm for Noisy Speech [J]. IEEE Transaction on Speech and Audio Processing, 2003, 11(3): 229-241.

[29] 赵军，潘永湘．一种基于谐波能量的混跌语音基音提取算法[C]．中国控制与决策学术年会论文集，2005:1439-1442.

[30] 章熙春．翘曲离散傅里叶变换算法与语音处理新方法研究[D]．华南理工大学博士学位论文，2005.

[31] L. R. Rabiner, R. W. Schafer, C. M. Rader. The Chirp z-transform Algorithm [J]. IEEE Transactions Audio and Electroacoustics, 1969, 17(2): 86-92.

[32] A. V. Oppenheim, D. Johanson. Computation of Spectra with Unequal Resolution Using the Fast Fourier Transform [J]. Proceedings of the IEEE, 1971, 59: 293-301.

[33] H. Malik, A. Khokhar, R. Ansari, B. Cappe de Baillon. Predominant pitch Contour Extraction from Audio Signals [J]. Multimedia and Expo. 2002. Proceedings, 2002, 2: 257-260.

[34] 张俊杰．基于和谐泛音检测的主旋律提取技术[D]．上海交通大学工程硕士学位论文，2007.

[35] L.科恩．时-频分析：理论与应用[M]．西安：西安交通大学出版社，1998.

[36] Thomas F. Quatieri．离散时间语音信号处理-原理与应用[M]．北京：电子工业出版社，2004.

[37] Malcolm Slaney. Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work [Z]. 1998.

致谢

论文完成之际，衷心感谢对我的课题及论文进行指导和帮助的所有老师和同学。

首先要感谢我的导师刘若伦老师，感谢他三年来对我的期望、培养和指导。本论文的选题、研究、实验以及成稿的过程中，刘老师一直给予我悉心的指导和深切的关怀，并提供了良好的学术氛围，使我从中不断学习新知识并提高自己。他严格的要求和严谨的学术作风使我受益匪浅，他忘我的工作精神激励我更加努力地研究和工作。他对工作的态度会成为我今后努力的目标。

还要感谢师弟师妹。在每周一次的汇报中，我可以和他们相互交流，使我能够不断获得灵感，不断完善实验设计。希望他们在今后的学习和生活中一切顺利。

感谢山东大学威海分校信息工程学院所有领导和老师的关心与培养，在此向他们表达深深的谢意。

攻读学位期间发表的学术论文

[1] 谢秀琴，刘若论．音乐信号的时频分析声学技术．2008，27（4）:543-546.